



HAL
open science

Voice Attractiveness

Benjamin Weiss, Jürgen Trouvain, Melissa Barkat-Defradas, John J. Ohala

► **To cite this version:**

Benjamin Weiss, Jürgen Trouvain, Melissa Barkat-Defradas, John J. Ohala (Dir.). Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers. In press, 10.1007/978-981-15-6627-1 . hal-02965919

HAL Id: hal-02965919

<https://hal.science/hal-02965919>

Submitted on 13 Oct 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Metadata of the book that will be visualized in SpringerLink

Publisher Name	Springer Singapore	
Publisher Location	Singapore	
Series ID	11951	
SeriesTitle	Prosody, Phonology and Phonetics	
Book ID	470006_1_En	
Book Title	Voice Attractiveness	
Book DOI	10.1007/978-981-15-6627-1	
Copyright Holder Name	Springer Nature Singapore Pte Ltd.	
Copyright Year	2021	
Corresponding Editor	Family Name	Weiss
	Particle	
	Given Name	Benjamin
	Suffix	
	Division	
	Organization	Technische Universität Berlin
	Address	Berlin, Berlin, Germany
	Email	benjamin.de.weiss@outlook.de
Editor	Family Name	Trouvain
	Particle	
	Given Name	Jürgen
	Suffix	
	Division	
	Organization	Saarland University
	Address	Saarbrücken, Saarland, Germany
	Email	trouvain@coli.uni-saarland.de
Editor	Family Name	Barkat-Defradas
	Particle	
	Given Name	Melissa
	Suffix	
	Division	
	Organization	ISEM
	Address	MONTPELLIER, France
	Email	melissa.barkat-defradas@umontpellier.fr
Editor	Family Name	Ohalá
	Particle	
	Given Name	John J.
	Suffix	
	Division	
	Organization	International Computer Science Institute
	Address	Berkeley, CA, USA

Email

ohala@berkeley.edu



Prosody, Phonology and Phonetics

Series Editors

Daniel J. Hirst, CNRS Laboratoire Parole et Langage, Aix-en-Provence, France

Hongwei Ding, School of Foreign Languages, Shanghai Jiao Tong University, Shanghai, China

Qiuwu Ma, School of Foreign Languages, Tongji University, Shanghai, China

UNCORRECTED PROOF

8 The series will publish studies in the general area of Speech Prosody with a
9 particular (but non-exclusive) focus on the importance of phonetics and phonology
10 in this field. The topic of speech prosody is today a far larger area of research than is
11 often realised. The number of papers on the topic presented at large international
12 conferences such as Interspeech and ICPhS is considerable and regularly
13 increasing. The proposed book series would be the natural place to publish
14 extended versions of papers presented at the Speech Prosody Conferences, in
15 particular, the papers presented in Special Sessions at the conference. This could
16 potentially involve the publication of 3 or 4 volumes every two years ensuring a
17 stable future for the book series. If such publications are produced fairly rapidly,
18 they will in turn provide a strong incentive for the organisation of other special
19 sessions at future Speech Prosody conferences.

20 More information about this series at <http://www.springer.com/series/11951>
21



22 Benjamin Weiss · Jürgen Trouvain ·
23 Melissa Barkat-Defradas ·
24 John J. Ohala
25 Editors

26 Voice Attractiveness

27 Studies on Sexy, Likable, and Charismatic
28 Speakers

30
31





36 *Editors*

37 Benjamin Weiss
38 Technische Universität Berlin
39 Berlin, Germany

Jürgen Trouvain
Saarland University
Saarbrücken, Saarland, Germany

40 Melissa Barkat-Defradas
41 ISEM
42 Montpellier, France

John J. Ohala
International Computer Science Institute
Berkeley, CA, USA

50
52

53

54 ISSN 2197-8700 ISSN 2197-8719 (electronic)
55 Prosody, Phonology and Phonetics
56 ISBN 978-981-15-6626-4 ISBN 978-981-15-6627-1 (eBook)
57 <https://doi.org/10.1007/978-981-15-6627-1>

60
61 © Springer Nature Singapore Pte Ltd. 2020

62 This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part
63 of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations,
64 recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission
65 or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar
66 methodology now known or hereafter developed.

67 The use of general descriptive names, registered names, trademarks, service marks, etc. in this
68 publication does not imply, even in the absence of a specific statement, that such names are exempt from
69 the relevant protective laws and regulations and therefore free for general use.

70 The publisher, the authors and the editors are safe to assume that the advice and information in this
71 book are believed to be true and accurate at the date of publication. Neither the publisher nor the
72 authors or the editors give a warranty, expressed or implied, with respect to the material contained
73 herein or for any errors or omissions that may have been made. The publisher remains neutral with regard
74 to jurisdictional claims in published maps and institutional affiliations.

75 This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
76 The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721,
77 Singapore

Editor Proof

Preface

79 At the Interspeech conference 2015, in Dresden, John (Ohala) asked Jürgen
80 (Trouvain) what he thinks about organizing a special session on attractive voices,
81 maybe for the next conference in this series. A former visiting researcher in
82 Berkeley, Melissa (Barkat-Defradas), had already expressed some ideas on such an
83 event on this topic. John has a long-standing interest in evolutionary aspects of
84 speech and voice, Melissa works in an interdisciplinary research team on all kinds
85 of aspects of evolution, and Jürgen has some background in paralinguistic char-
86 acteristics of speech. At the same conference in Dresden, Jürgen introduced
87 Benjamin (Weiss) to John with Benjamin as the optimal complement to this team
88 since he has published several papers on social likeability of voices.

89 It was then at Interspeech in Stockholm 2017, that we were able to organize the
90 planned special session on voice attractiveness. We considered this event as the
91 perfect setting for presenting research dealing with many aspects: perceived vocal
92 preferences of men, women, and synthesized voices in well-defined social situa-
93 tions, acoustic correlates of voice attractiveness/pleasantness/charisma, interrela-
94 tions between vocal features and individual physical and physiological
95 characteristics, consequences for sexual selection, predictive value of voice for
96 personality and for other psychological traits, experimental definition of esthetic
97 standards for the vocal signal, cultural variation of voice attractiveness/pleasantness
98 and standards, and also the link between vocal pathology and vocal characteristics.
99 In Stockholm we agreed on a follow-up publication where the authors have more
100 space than in a conference paper with its strict limitations. Moreover, also those
101 colleagues could be reached that were not participants of this conference.

102 The special session was a success in our view. In total, we had nine accepted
103 contributions. Authors from six papers of this session are also aboard in this vol-
104 ume. In addition to these, there are ten further contributions for this publication,
105 having a total of seventeen papers when we add the introductory chapter. It is our
106 belief that both collections, the nine conference papers, and the seventeen articles in
107 this volume, can provide a useful overview on the state-of-the-art research on voice
108 attractiveness, voice likeability, and vocal charisma. We also hope that these studies



109 represent a fruitful fundament for further thoughts and investigations of an exciting
110 field of speech and voice research.

111 As many book projects of this size, the editing process took longer than
112 expected. This delay is mainly but note entirely due to health reasons of some of the
113 editors. We would like to thank all authors for their patience and the publishing
114 house for the provided support.

116 Berlin, Germany
119 Saarbrücken, Germany
120 Montpellier, France
123 Berkeley, USA
125 April 2020

Benjamin Weiss
Jürgen Trouvain
Melissa Barkat-Defradas
John J. Ohala

UNCORRECTED PROOF



126 Contents

128	Part I General Considerations	
130	1 Voice Attractiveness: Concepts, Methods, and Data	3
131	Jürgen Trouvain, Benjamin Weiss, and Melissa Barkat-Defradas	
133	2 Prosodic Aspects of the Attractive Voice	17
134	Andrew Rosenberg and Julia Hirschberg	
136	3 The Vocal Attractiveness of Charismatic Leaders	41
137	Rosario Signorello	
139	4 Vocal Preferences in Humans: A Systematic Review	55
140	Melissa Barkat-Defradas, Michel Raymond, and Alexandre Suire	
142	Part II Voice	
144	5 What Does It Mean for a Voice to Sound “Normal”?	89
145	Jody Kreiman, Anita Auszmann, and Bruce R. Gerratt	
146	6 The Role of Voice Evaluation in Voice Recall	107
148	Molly Babel, Grant McGuire, and Chloe Willis	
150	7 Voice, Sexual Selection, and Reproductive Success	131
151	Alexandre Suire, Michel Raymond, and Melissa Barkat-Defradas	
153	8 On Voice Averaging and Attractiveness	145
154	Pascal Belin	
155	Part III Prosody	
158	9 Attractiveness of Male Speakers: Effects of Pitch and Tempo	159
159	Hugo Quené, Geke Boomsma, and Romée van Erning	
160	10 The Contribution of Amplitude Modulations in Speech	
162	to Perceived Charisma	171
163	Hans Rutger Bosker	

Editor Proof



164	11 Dress to Impress? On the Interaction of Attire with Prosody	
165	and Gender in the Perception of Speaker Charisma	189
167	Alexander Brem and Oliver Niebuhr	
168	12 Birds of a Feather Flock Together But Opposites Attract!	
169	On the Interaction of F0 Entrainment, Perceived Attractiveness,	
170	and Conversational Quality in Dating Conversations	221
172	Jan Michalsky and Heike Schoormann	
173	Part IV Databases	
176	13 Acoustic Correlates of Likable Speakers in the NSC Database	251
177	Benjamin Weiss, Jürgen Trouvain, and Felix Burkhardt	
178	14 Ranking and Comparing Speakers Based on Crowdsourced	
180	Pairwise Listener Ratings	269
181	Timo Baumann	
182	15 Multidimensional Mapping of Voice Attractiveness	
183	and Listener’s Preference: Optimization and Estimation	
184	from Audio Signal.	287
186	Yasunari Obuchi	
188	Part V Technological Applications	
189	16 Trust in Vocal Human–Robot Interaction: Implications for Robot	
190	Voice Design	305
192	Ilaria Torre and Laurence White	
193	17 Exploring Verbal Uncanny Valley Effects with Vague Language	
194	in Computer Speech	323
196	L. Clark, A. Ofemile, and B. R. Cowan	





Editors and Contributors

197
198

About the Editors

199

200 **Benjamin Weiss** received his Ph.D. in 2008, in phonetics from Humboldt-
201 University, Berlin. Since then he has extensively studied acoustic correlates of
202 pleasant and likable voices, taking into account also speaking styles and conver-
203 sational behavior in order to build quantitative models. He was visiting fellow at the
204 University of Western Sydney and the University of Technology Sydney. In 2019,
205 he completed his habilitation on human dialog and speech-based (multimodal) HCI.
206 Since September 2020, he is an Associate Professor at the School of Intelligence,
207 Hanyang University, Seoul.

209
210 **Jürgen Trouvain** received his Ph.D. in Phonetics in 2004, from Saarland
211 University (Germany), where he works as a Senior Researcher and Lecturer at the
212 Department of Language Science and Technology. His research fields include
213 nonverbal vocalizations such as breathing and laughing, as well as non-native
214 speech and phonetic learner corpora. He has acted as an organizer for several
215 international conferences and workshops.

217
218 **Melissa Barkat-Defradas** obtained her Ph.D. in Forensic Linguistics at the
219 University of Lyon, in 2000, and received the Young Researcher Award for her
220 work in Automatic Language Identification. After a research fellowship at UC
221 Berkeley, she joined the French National Centre for Scientific Research. She is now
222 a full-time Researcher at The Institute of Evolutionary Sciences of Montpellier
223 (France), where she actively contributes to developing interdisciplinary research by
224 bridging the gap between experimental phonetics and evolutionary biology. She is
225 particularly interested in the selective forces that may explain the emergence of
226 articulated language in humans.

228
229
230

Editor Proof



231 **John J. Ohala** is an Emeritus Professor of Linguistics at the University of
232 California, Berkeley, and a Research Scientist at the International Computer
233 Science Institute, Berkeley. He has had a major impact on the field of speech
234 communication. His research interests focus on experimental phonology and pho-
235 netics and ethological aspects of communication, including speech perception,
236 sound change, phonetic and phonological universals, psycholinguistic studies in
237 phonology, and sound symbolism. He proposed an innovative ethological
238 hypothesis, which unifies—via “the frequency code”—such diverse behavioral
239 phenomena as the cross-language use of voice pitch for questions and statements,
240 the systematic use of consonants, vowels, and tones in sound symbolical vocabu-
241 lary, the “smile,” and sexual dimorphism of the vocal anatomy in adult humans.
242

243
244
245 **Contributors**

248 **Anita Auszmann** Department of Head and Neck Surgery and Linguistics,
249 University of California, Los Angeles, CA, USA

250
251
252 **Molly Babel** Department of Linguistics, University of British Columbia, BC,
253 Canada

254
255 **Melissa Barkat-Defradas** Institut des Sciences de l’Evolution de Montpellier,
256 University of Montpellier, Centre National de la Recherche Scientifique, Institut
257 pour la Recherche et le Développement, Ecole Pratique des Hautes Etudes – Place
258 Eugène Bataillon, Montpellier, France

259
260 **Timo Baumann** Universität Hamburg, Language Technology Group, Hamburg,
261 Germany

262
263 **Pascal Belin** Institut de Neurosciences de La Timone, CNRS et Aix-Marseille
264 Université Département de Psychologie, Université de Montréal, Montreal, Canada

265
266 **Geke Boomsma** Utrecht institute of Linguistics, Utrecht University, Utrecht, The
267 Netherlands

268
269 **Hans Rutger Bosker** Max Planck Institute for Psycholinguistics, Nijmegen,
270 The Netherlands;
271 Psychology of Language Department, Donders Institute for Brain Cognition and
272 Behaviour, Radboud University, Nijmegen, The Netherlands

273
274 **Alexander Brem** Innovation and Technology Management, Friedrich-Alexander-
275 Universität Erlangen-Nürnberg, Erlangen, Germany

276
277 **Felix Burkhardt** audeERING GmbH, Berlin, Germany
278



279 **L. Clark** School of Information, & Communication Studies, University College
280 Dublin, Dublin, Ireland;
281 Computational Foundry, Swansea University, Swansea, UK
282

283 **B. R. Cowan** School of Information, & Communication Studies, University
284 College Dublin, Dublin, Ireland
285

286 **Bruce R. Gerratt** Department of Head and Neck Surgery, University of
287 California, Los Angeles, CA, USA
288

289 **Julia Hirschberg** Columbia University, NYC, New York, NY, USA
290

291 **Jody Kreiman** Department of Head and Neck Surgery and Linguistics, University
292 of California, Los Angeles, CA, USA
293

294 **Grant McGuire** Department of Linguistics, University of California Santa Cruz,
295 Santa Cruz, CA, USA
296

297 **Jan Michalsky** University of Oldenburg, Oldenburg, Germany
298

299 **Oliver Niebuhr** Mads Clausen Institute, Centre for Electrical Engineering,
300 University of Southern, Odense, Denmark
301

302 **Yasunari Obuchi** School of Media Science, Tokyo University of Technology,
303 Hachioji, Tokyo, Japan
304

305 **A. Ofemile** English Department, FCT College of Education, Zuba, Abuja, Nigeria
306

307 **Hugo Quené** Utrecht institute of Linguistics, Utrecht University, Utrecht,
308 The Netherlands
309

310 **Michel Raymond** Institut des Sciences de l'Evolution de Montpellier, University
311 of Montpellier, Centre National de la Recherche Scientifique, Institut pour la
312 Recherche et le Développement, Ecole Pratique des Hautes Etudes – Place Eugène
313 Bataillon, Montpellier, France
314

315 **Andrew Rosenberg** Google LLC, NYC, New York, NY, USA
316

317 **Heike Schoormann** University of Oldenburg, Oldenburg, Germany
318

319 **Rosario Signorello** Laboratoire de Phonétique et Phonologie, CNRS & Sorbonne
320 Nouvelle, Paris, France
321

322 **Alexandre Suire** Institut des Sciences de l'Evolution de Montpellier, University
323 of Montpellier, Centre National de la Recherche Scientifique, Institut pour la
324 Recherche et le Développement, Ecole Pratique des Hautes Etudes – Place Eugène
325 Bataillon, Montpellier, France
326

327 **Ilaria Torre** Department of Electronic and Electrical Engineering, Trinity College
328 Dublin, Dublin, Ireland
329

330 **Jürgen Trouvain** Saarland University, Saarbrücken, Germany
331



332 **Romée van Erning** Utrecht institute of Linguistics, Utrecht University, Utrecht,
333 The Netherlands

335 **Benjamin Weiss** Technische Universität Berlin, Berlin, Germany
336

337 **Laurence White** School of Education, Communication and Language Sciences,
338 Newcastle University, Newcastle, UK
339

340 **Chloe Willis** Department of Linguistics, University of California Santa Barbara,
341 Santa Barbara, CA, USA
342

UNCORRECTED PROOF

1
2

Part I

General Considerations

UNCORRECTED PROOF

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Voice Attractiveness	
Series Title		
Chapter Title	Voice Attractiveness: Concepts, Methods, and Data	
Copyright Year	2020	
Copyright HolderName	Springer Nature Singapore Pte Ltd.	
Author	Family Name	Trouvain
	Particle	
	Given Name	Jürgen
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Saarland University
	Address	Campus C7.2, 66123, Saarbrücken, Germany
	Email	trouvain@coli.uni-saarland.de
Corresponding Author	Family Name	Weiss
	Particle	
	Given Name	Benjamin
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Technische Universität Berlin
	Address	Ernst-Reuter-Platz 7, 10405, Berlin, Germany
	Email	benjamin.weiss@tu-berlin.de
Author	Family Name	Barkat-Defradas
	Particle	
	Given Name	Melissa
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	University of Montpellier
	Address	Place Eugène Bataillon cc065, 34090, Montpellier cedex 05, France
	Email	melissa.barkat-defradas@umontpellier.fr
Abstract	This book comprises contributions on vocal aspects of attractiveness, social likability, and charisma. Despite some apparent distinct characteristics of these three concepts, there are not only similarities, but even interdependencies to be considered. This chapter introduces and regards the concepts studied, methods applied, and material selected in the contributions. Based on this structured summary, we argue to increase interdisciplinary and even holistic efforts in order to better understand the concepts for voice and speech in humans and machines.	

Keywords

Attractiveness - Charisma - Likability - Sexual selection - Interdisciplinary - Holistic view -
Structured summary - Speech production - Speech perception

Chapter 1

Voice Attractiveness: Concepts, Methods, and Data



Jürgen Trouvain, Benjamin Weiss, and Melissa Barkat-Defradas

Abstract This book comprises contributions on vocal aspects of attractiveness, social likability, and charisma. Despite some apparent distinct characteristics of these three concepts, there are not only similarities, but even interdependencies to be considered. This chapter introduces and regards the concepts studied, methods applied, and material selected in the contributions. Based on this structured summary, we argue to increase interdisciplinary and even holistic efforts in order to better understand the concepts for voice and speech in humans and machines.

Keywords Attractiveness · Charisma · Likability · Sexual selection · Interdisciplinary · Holistic view · Structured summary · Speech production · Speech perception

1.1 Introduction

Probably, everybody has an idea of the meaning or meanings of *attractive* and *attractiveness* on the one side, and of voice and speaker on the other. It is also likely that everybody has their own ideas, which voices sound attractive—either in general or in specific contexts. But these ideas show by no means homogeneous structures and similar definitions.

A book on voice attractiveness attracts researchers, be it as authors and/or readers, who look at this topic from different angles as the subtitle of this book indicates. A *sexy* speaker is not the same as a *likable* speaker, and a *charismatic* speaker is different

J. Trouvain
Saarland University, Campus C7.2, 66123 Saarbrücken, Germany
e-mail: trouvain@coli.uni-saarland.de

B. Weiss (✉)
Technische Universität Berlin, Ernst-Reuter-Platz 7, 10405 Berlin, Germany
e-mail: benjamin.weiss@tu-berlin.de

M. Barkat-Defradas
University of Montpellier, Place Eugène Bataillon cc065, 34090 Montpellier cedex 05, France
e-mail: melissa.barkat-defradas@umontpellier.fr

© Springer Nature Singapore Pte Ltd. 2020

B. Weiss et al. (eds.), *Voice Attractiveness*, Prosody, Phonology and Phonetics,
https://doi.org/10.1007/978-981-15-6627-1_1

again. These differences of how attractiveness is considered are also reflected in the chapters of this book. Likewise, the definition of speaker and voice is heterogeneously used, too. For this reason, we first attempt to shed some light onto the diversity of concepts we face in the upcoming chapters.

There is a broad range of different methods used in the studies of this volume. Many perform experimental research to investigate aspects of production, acoustics, and perception of attractive speech. There are some studies with a focus on modeling of data with respect to attractiveness, whereas other studies review how speech technology can be applied taking the (missing) attractiveness of voices into account. The data types that were used in the studies of this volume also show a large span. They range from manipulations of monosyllabic stimuli over single words and sentences in controlled settings up to many minutes of spontaneous conversational speech. The recap of the diversity of methods and data in this collection is followed by some concluding remarks on the emerging field of voice attractiveness, a research field that attracts researcher from many disciplines.

1.2 Concepts

1.2.1 Voice, Speaker, and Speech

The contributions of this collection consider the *voice* and *voice attractiveness* in different ways. Voice is not only seen in a narrow sense where it refers only to glottal activity. Voice in a wider sense additionally includes supra-glottal activities such as tongue raising, pharyngeal constriction, nasality or lip spreading (Laver, 1980), so that for instance formants as acoustic correlates of supra-laryngeal resonances are taken into account. For several studies, prosody plays an important role, reflected by fundamental frequency (F0), intensity, pauses and duration from a suprasegmental point of view. Further, timing parameters refer to entrainment in dialogs.

Naively, one would not assume that a voice that is considered as “normal”, “stereotypical” or “average” would correlate to attractiveness. Nevertheless, three papers of this volume look more closely to the acoustic parameters of the “mean” voice and its perception of attractiveness—partially with somewhat surprising results.

Kreiman et al. (this volume) show that listeners differ regarding the question of what it means for a voice to sound “normal”. There seem to be individual, rather consistent, strategies to label how normal or not normal a voice sounds. In their study, listeners assessed a wide range of one second samples of female speakers. From several acoustic parameters, the most relevant for explaining some amount of variance in the labels are fundamental frequency and its variation, as well as the first two formants, but not others that are typically associated with voice quality. However, the authors could not find a simple or generally valid answer, the situation is rather complex because several factors like the listener, the context, the purpose of the judgment, and of course the individual voice have to take into account.

59 The topic of recalling a voice from memory, an everyday task for everybody of
60 us, is analyzed in Babel et al. (this volume). They show in a set of experiments with
61 monosyllabic words as stimulus material that subjective stereotypicality and attrac-
62 tiveness affect the performance to remember a voice. Overall, they found support
63 for the statement that less stereotypical voices and less attractive voices were better
64 memorized.

65 Belin (this volume) reports of findings of experiments where identical short syl-
66 lables of multiple voices of the same sex were averaged. The more voices were
67 averaged the 'speakers' of the averaged voice samples were perceived as more and
68 more attractive. (similar to a visual effect concerning face attractiveness). Obviously,
69 the main responsible factors for this effect are the reduced "distance-to-mean" for
70 differences between F0 and the first formant, and an increased "texture smoothness"
71 reflected by a raised harmonics-to-noise ratio.

72 There are also studies with stimuli to be rated that are longer than just one syllable
73 or just one second. These studies concentrate more on speech prosody. Quené et
74 al. (this volume), for instance, control for tempo and F0 in stimuli sentences, and
75 Bosker (this volume) analyzed amplitude modulation in authentic speech samples.
76 The review of charismatic speech of Rosenberg and Hirschberg (this volume) centers
77 at prosody in all possible aspects, whereas, for instance, Weiss et al. (this volume)
78 investigate acoustic parameters that reflect prosody (F0, intensity, rate), segmental
79 properties (formants, spectral features) but also the voice in a narrow sense (shimmer,
80 jitter, harmonics-to-noise ratio). These examples show that the vocal part in voice
81 attractiveness can be referred to very different aspects of voice and speech when
82 performing research in this field.

83 *1.2.2 Sexual Selection and Voice Attractiveness*

84 A sexy speaker can be seen as somebody who underlines her or his perceived sexual
85 attractiveness—often unconsciously—with her or his voice and speech behavior.
86 Though the voice is the privileged medium for interpersonal communication, it is
87 not solely useful for conveying semantic information to other people. As a matter of
88 fact, voice should also be regarded as a powerful social object, whose role is crucial in
89 the context of human relationships. Indeed, by using oral communication, speakers
90 are not only able to share their ideas and emotions, but they are also able to signal
91 some reliable sociobiological features to their interlocutors such as sex, age, health,
92 and social status, among others. There is a large body of scientific literature, for
93 instance Scherer (1978), which describe the links between voice characteristics and
94 personality traits, or the works by Laver and Trudgill (1979) and Bezooijen (1995),
95 who studied voice as a social and cultural marker, or either still, Banse and Scherer
96 (1996) whose work investigate how voice is used to express one's emotional state.

97 All of these authors, to name a few, have demonstrated that voice goes far beyond
98 its primary linguistic function. Yet, interestingly, researches in Humanities mostly

99 tackled the topic of vocal function independently of any evolutionary considerations.
 100 However, as early as 1890, Darwin addressed the issue within the frame of sexual
 101 selection by drawing intriguing parallels between animal vocalizations and the human
 102 voice:

103 The sexes of many animals incessantly call for each other during the breeding-season; and
 104 in not a few cases, the male endeavors thus to charm or excite the female. This, indeed,
 105 seems to have been the primeval use and means of development of the voice [...]. When
 106 male animals utter sounds in order to please the females, they would naturally employ those
 107 which are sweet to the ears of the species; and it appears that the same sounds are often
 108 pleasing to widely different animals, owing to the similarity of their nervous systems, as
 109 we ourselves perceive in the singing of birds and even in the chirping of certain tree-frogs
 110 giving us pleasure. (Darwin, 1890, pp. 90–96).

111 Darwin's original idea according to which vocalizations allow the transmitter to
 112 attract females' attention and express his reproductive intentions make it legitimate
 113 to address the issue of human voice attractiveness in the specific context of human
 114 mating. As a matter of fact, as it is developed in the first contribution of Suire,
 115 Raymond, and Barkat–Defradas (this volume), it is reasonable to think that sexual
 116 selection—the mechanism which promotes biological and social traits that confer a
 117 reproductive benefit—has also intervened in the shaping of human vocal dimorphism;
 118 the attractiveness of a voice being a proxy, or a reinforcing signal, for other physical
 119 characteristics. By providing an overview of the research that lies at the crossroad
 120 of the human voice and evolutionary biology, the authors aim at demonstrating that
 121 sexual selection provides an interesting theoretical framework to understand the
 122 functional role of the human voice from an evolutionary perspective. Indeed, several
 123 studies have demonstrated the existence of a vocal attractiveness stereotype, which
 124 suggests that voice is an honest signal¹ of phenotypic quality in the same way as
 125 other physical features like, for example, the waist-to-hip ratio.²

126 Such an assumption raises the question of what makes a voice attractive? In
 127 their survey of the literature, Rosenberg and Hirschberg (this volume) examine the
 128 concept of vocal attractiveness itself. The authors consider the concept as highly
 129 context-dependent and discriminate between several types of attraction (i.e., political
 130 charisma, business leadership, nonsexual attraction and, last but not least, romantic
 131 desirability) each one of them being associated with specific articulatory, acoustic,
 132 and prosodic traits. They also show that though voice attractiveness is a complicated
 133 and exceptionally subjective phenomenon, evidence suggests some shared cross-
 134 cultural patterns that must have been shaped in the course of evolution by the selective
 135 pressure induced by the preferences of one sex for the vocal attributes of the other.
 136 The topic of vocal preferences has given rise to a large body of literature on the
 137 evolution of vocal preferences, which generally speaking, reveals that low-pitched

¹Signals are traits that have evolved specifically because they change the behavior of receivers in ways that benefit the signaler. For example, peacock resplendent tail feathers are honest since they truly signal reproductive fitness of their bearer to the receiver.

²The waist-to-hip ratio (WHR) is the dimensionless ratio of the circumference of the waist to that of the hip. WHR correlates with health and fertility (with different optimal values in males and females).

138 masculine voices are universally preferred by women, such voices being perceived
139 as related to a high quality phenotype. Conversely, men tend to prefer high-pitched
140 feminine voices that are perceptually associated with youth and fertility at least
141 in English. For more details of evolutionary mechanisms of attractive voices like
142 mate choice see the systematic review of vocal preferences in humans by Barkat–
143 Defradas, Raymond and Suire (this volume). Quené et al. (this volume) also confirm
144 the expected pattern that men with lower-pitched voices tend to be rated as more
145 attractive by (heterosexual) female listeners. They also reveal the importance of fast
146 tempo in voice attractiveness evaluation. Indeed, their results based on manipulated
147 speech show that the female raters judged masculine voices as less attractive if the
148 F0 was artificially raised and the tempo decreased.

149 In their speed dating study, Michalsky and Schoormann (this volume) investigated
150 the effects of perceived attractiveness and conversational quality on entrainment. In
151 analyzing speed dating dialogs, prosodic disentrainment, in terms of pitch differ-
152 ences, is related to facial attractiveness for interlocutors of opposing sex. However,
153 this result is inhibited by high conversational quality for females, and low conversa-
154 tional quality for males.

155 1.2.3 *Likability and Social Attractiveness*

156 A likable speaker is seen as somebody who underlines her or his perceived social
157 attractiveness or pleasantness with her or his voice and speech behavior. There are
158 several potential aspects that may constitute likability. For example, from the two
159 of the most stable interpersonal concepts for unacquainted persons, benevolence (or
160 warmth, communion) and competence (or agency, capability) (Abele, Cuddy, Judd,
161 & Yzerbyt, 2008; Schaller, 2008; Fiske, Cuddy, & Glick, 2006), the first dimension
162 (benevolence) is often assumed to resemble likability (DePaulo, Kenny, Hoover,
163 Webb, & Oliver, 1987; Fiske et al., 2006; Argyle, 1988). However, liking-aversion
164 may conceptually comprise the second dimension of competence as well (McCroskey
165 & McCain, 1974), even in speech (Putnam & Street, 1984). Actually, there is much
166 evidence from questionnaire analysis in a speech during dimension reduction that
167 evaluative questionnaire items, such as “likable”, can be apparent in both dimen-
168 sions, benevolence and competence, or neither (Cuddy, Fiske, & Glick, 2008; Brown,
169 Strong, & Rencher, 1973, 1985; Hart & Brown, 1974; Street & Brady, 1982; Weirich,
170 2010; Weiss & Möller, 2011). Given these empirical results, it can be argued that
171 the so-called benevolence is just one possible but a very likely attribution to a person,
172 which affects a speaker’s social attractiveness, especially in a first impression.

173 Concerning voice acoustics, there are only few correlates of likability that show at
174 least some robustness to changes in material, most notably increased pitch variability
175 and tempo, while the results of average pitch reveal to be more complex, at least in
176 German (Weiss et al., this volume).

177 While such results aim at correlates of averaged ratings on a scale, paired
 178 comparisons allow for a much finer measure of preference in likability. This method is,
 179 unfortunately, much more effort. Therefore, a crowd-based procedure is presented to
 180 collect such data efficiently, and it was used to train a model for predicting preferences
 181 of pairs of stimuli (Baumann, this volume).

182 In order to better take into regard the individual aspects of attractiveness, a method
 183 is presented that extracts overall voice attractiveness and listeners' preferences from
 184 paired comparisons, so that voices' likability can be estimated by the inner product
 185 of the two vectors of attractiveness and preferences (Obuchi, this volume).

186 **1.2.4 Charisma and Leadership**

187 A charismatic speaker is seen as somebody who underlines her or his perceived lead-
 188 ership, persuasive power, enthusiasm, and passion with her or his voice and speech
 189 behavior. Charisma is, just like likability, a social evaluation. However, likability
 190 typically refers to a dialogic situation, or in passive listening test, to the anticipation
 191 of a dialog—without any predefined difference in social status. In contrast to this,
 192 charisma is typically about an individual affecting a group of people, and thus implies
 193 some kind of social superiority. Charismatic people stand out, formally by social sta-
 194 tus or rank, or situationally by other's acknowledgment of their specialty. Therefore,
 195 the typical domains to study charisma in voice are speeches or talks of famous people,
 196 such as politicians and managers. A passionate and motivating speech by such people
 197 represents an often used, and sometimes even requested and anticipated, method of
 198 leadership. A discursive overview of what a charismatic voice actually is, can be
 199 found in Signorello (this volume).

200 The focus on public speeches and talks when dealing with charisma, complicates,
 201 on the one hand, differentiating between effects of a speech's presentation from
 202 those that originate in the fame, attributions, and social status. On the other hand,
 203 instead of relying on ratings in the laboratory, there a plenty of potentially valid
 204 indicators of charisma of those famous people including type of applause, (social)
 205 media reaction, and election results. For example, during a party conference of the
 206 German social democrats in 1995, the chairman was replaced by his vice-chairman—
 207 atypically early at this specific date—after an inspiring and enthusiastic speech of that
 208 vice-chairman. Given rather similar contents, sometimes even identical formulations,
 209 this outcome of the election was analyzed not regarding rhetorics, but speaking
 210 style instead (Paeschke & Sendlmeier, 1997). Such occurrences not only show that
 211 charisma is blended with power and leadership, but also exemplify the relevance of
 212 voice and speech for charisma. In this volume, the relevance of prosody and attire
 213 is studied for speeches of leading senior managers (Brem & Niebuhr, this volume).
 214 And in Bosker (this volume), a closer look on the modulation spectrum, which is
 215 related to speech rhythm, is taken for speeches from the US presidential campaign
 216 candidates Hillary Clinton and Donald Trump.

217 1.3 Methods

218 From a methodological perspective, we can divide studies on voice attractiveness in
219 three fields. Investigations of the possible effects of different kinds of attractiveness
220 and their vocal correlates are covered by *experimental research*. In addition to this
221 research direction, *modeling* of processes how individual voices in audio samples
222 attract listeners represents a further field of study. Finally, *technological applications*
223 should be viewed as an own field of research in voice attractiveness.

224 1.3.1 Experimental Research

225 Human attractiveness is typically considered as a subjective concept. Therefore,
226 experimental research is dominated by collecting explicit and implicit human rat-
227 ings and decisions. The simplest methodological approach is to present stimuli and
228 explicitly ask for ratings; on a scale if sequentially presented, or as a preference in
229 the case of comparing stimuli. Such listening and ratings are, for example, conducted
230 by Babel et al. (this volume). They collected a variety of subjective characteristics,
231 among them perceptual similarity, applying a comparison of pairs of stimuli on a
232 single scale, and perceptual attractiveness, collecting ratings in a sequential proce-
233 dure for each stimulus individually. The latter method is also frequently used in the
234 studies evaluated by Belin (this volume). Quené et al. (this volume) explicitly argue
235 in favor of the sequential approach with absolute ratings instead of a forced prefer-
236 ence choice of a direct comparison, as they want to avoid drawing attention to the
237 signal manipulations they have conducted. There are various variants applied, often
238 taken advantage of graphical computer interfaces, for example, to sort and assign
239 short stimuli of a set to labels (Kreiman et al., this volume).

240 Instead of explicitly asking for measures of attractiveness, implicit measures can
241 be attempted to collect, in order to avoid a social bias of the subjects. Such approaches
242 comprise observations of social decisions, for example, counting the number of
243 direct interactions in gaming or game-like tasks (Krause, Back, Egloff, & Schmukle,
244 2014). Other observations refer to the number of friends, or offspring (or explicitly
245 asking to disclose the number of sexual partners). Such long-term or retrospective
246 observations and surveys are, however, difficult to relate to specific traits, such as
247 vocal characteristics.

248 1.3.2 Modeling

249 Quantitative modeling of subjective human ratings, such a sexual or social attrac-
250 tiveness, serves in principle two purposes. One is to describe the relations, e.g., cor-
251 relations, found with parameters of interest in a given data set. Such a model could
252 be a starting point for a prediction model, but does not provide explanatory power as

253 in a scientific theory. For the case of voice attractiveness, typical model parameters
254 are acoustic or articulatory measures. Another purpose is to actually explain inter-
255 dependencies between parameters and ratings in a quantitative way. However, in the
256 latter case, the parameters chosen and the kind of relationship have to be confirmed
257 by methodological means ensuring a causal relationship. Synthesizing or resynthe-
258 sizing speech represents the most popular approach to control for the variables in
259 question. It also aims at providing proof for a causal relationship. As the knowledge
260 base is enhanced by empirical studies incrementally, each study might fulfill both
261 purposes to some degree. For example, the linear models of social attractiveness of
262 Weiss et al. (this volume) build on hypotheses drawn from several scientific methods
263 in order to add evidence for acoustic-perceptual relations, but its main result is a
264 simple data description.

265 Baumann (this volume), present a methodological approach, that does comprises
266 not only the acoustic modeling part, but also a method to efficiently collect preference
267 ratings for stimulus pairs. Such pairwise preferences for German spoken Wikipedia
268 articles were acoustically correlated directly, and modeled as relative preferences by
269 means of a recurrent neural network.

270 In a related approach, Obushi (this volume) collected pairwise preferences for a
271 Japanese greeting phrase. The ratings are multidimensionally analyzed, taking into
272 account the listeners' differences as well, and modeled by multiple acoustics features
273 applying machine learning.

274 *1.3.3 Technological Applications*

275 Voice attractiveness can play an essential role in human-machine interaction (HMI)
276 as two contributions in this volume show. There is a tendency that "people tend to
277 attribute personality traits to computers and robots as if they were human agents"
278 (Nass, Moon, Fogg, Reeves, & Dryer, 1995). That means that the human-sounding
279 voices of talking and conversational computers can also be considered as personalized
280 machines. In addition, machines can act for humans, for instance, when a speech
281 synthesizer is used as a speech prosthesis for people who cannot clearly and fluently
282 articulate anymore. From a view of listening to talking machines, we all know that it is
283 most of the time rather boring and less interesting when faced with an artificial voice
284 and synthesized speech, be it when street names are announced in car navigation
285 or when interacting with a dialog system. For conversational agents, e.g., intelligent
286 personal assistants, it is a particular challenge to show skills that are required for
287 smooth dialogs that span aspects of timing up to common grounding. Thus, voice
288 selection and voice modeling should be an integral part of the design in HMI tools.
289 The paper collected in this volume are not empirical studies with existent systems
290 but are reviews in which important thoughts are developed before experiments that
291 test the usability of certain aspects of voice attractiveness are performed.

292 Torre and White (this volume) focus on the characteristics of a robot's voice
293 in human-robot interaction. They are particularly interested in how vocal elements

294 can contribute to the impression of trustworthiness. They review studies in which a
295 robot's voice was analyzed or manipulated, always with a particular view on trust-
296 worthiness. Naturalness and "machine-likeness", cognitive load, incongruity with
297 the robot's behavior in general and the robot's appearance such as its size, gender,
298 accent, and interaction context. Furthermore, they argue that the design of robot
299 voices should come with an unambiguous appearance and function, because unreal-
300 istic expectations of robot performance in human users should be avoided.

301 The human evaluation in regard to different kinds of attractiveness represent
302 immanent social and cognitive processes. Such evaluations are, however, not limited
303 to other living persons. Instead, interactive systems, especially those using speech,
304 are known to evoke similar processes (Reeves & Nass, 1996; Nass & Brave, 2005).
305 And with the emergence of speech interaction with computers in the form of personal
306 smartphone assistants, smart home devices, virtual persons, and human-like (social)
307 robots, the users' appraisal of the verbal and nonverbal behavior of such interactive
308 computers are receiving much attention.

309 One observation specific to anthropomorphic computers is the so-called "uncanny
310 valley" effect. It describes an overall increase in familiarity (or attractiveness or lik-
311 ability) with increasing human-likeness (or level of details) of the systems features
312 and movements that is disrupted by a sudden decrease in familiarity close to perfect
313 human-likeness (Mori, 2012). This awkward or eerie feeling for a close to human,
314 but obviously not natural synthesis is typically explained by a shift in reference
315 from artificial to human and can be circumvented by reducing the level of human-
316 likeliness or choosing an artificial metaphor (e.g., a puppet or cartoon) instead of
317 a human. This effect is mostly studied for visual perceptions of the body and face
318 of a robot or virtual person and their animated movements. However, in Clark (this
319 volume), results for the evaluation of three linguistic strategies, politeness, relational
320 work, and vague language are discussed in their usage for speech interfaces and their
321 potential mismatch with the expectations in human users, and thus their potential to
322 cause an uncanny valley effect.

323 One important sub-concept of social attractiveness is trust (McAleer, Todorov, &
324 Berlin, 2014; Weiss, Wechsung, Kühnel, & Möller, 2015). In Torre and White (this
325 volume) the effects of robot voices' gender, naturalness, prosody, and accent on trust
326 perception in users are presented and systematized. Overall, there are effects, but
327 they depend on the context and user group. For example, a regional accent showed
328 an increased credibility to a standard accent when being knowledgeable, but the
329 opposite in the case of being unknowledgeable.

330 1.4 Data

331 The material used in studies on voice attractiveness varies widely, from monosyllabic
332 stimuli recorded in the lab to large extracts of authentic speech material that was not
333 produced for research. This stylistic diversity is also reflected in the contributions

334 for this volume. Thus, it seems fair to separate three kinds of sources, controlled
 335 experimental data, naturalistic lab data, and natural field data “from the wild”.

336 *1.4.1 Controlled Experimental Data*

337 One major source of the material stems from lab experiments, where new recordings
 338 are conducted for a specific purpose with already defined acoustic and perceptual
 339 analytic methods to be applied on. Such recordings are usually very short, for example
 340 (sustained) vowels, syllables or words. They can also not be considered as socially
 341 authentic, i.e., they do not aim to resemble real-life social communication situations.
 342 Due to its short duration, such material lacks major prosodic aspects, e.g., intonation
 343 contour or emphasis variation, as well as any natural situational grounding, affecting,
 344 e.g., speaking rate. Controlling for such aspects, however, allows to focus on topics
 345 like voice quality and person identification/similarity, while explicitly controlling for
 346 the just mentioned effects.

347 Examples of experimental data are Belin (this volume), who uses averaged short
 348 syllables of multiple voices, for which attractiveness ratings are collected. Kreiman
 349 et al., (this volume) analyzes steady state vowels (one second duration) regarding
 350 “normal” voice quality, whereas Babel et al., and Obuchi (both this volume) used
 351 single (monosyllabic, respectively multisyllabic) words for perception tests.

352 On some occasions, full sentences, or even a paragraph, are read by speakers
 353 in a lab with similar aims. The practical implications include potential laborious
 354 manual work to extract specific segments for analysis, and to take into account richer
 355 linguistic context, while the read speech style in a controlled environment allows to
 356 analyze not only segmental and micro-prosodic, but also macro-prosodic parameters.
 357 Therefore, it is not a coincidence to find a mixture of material types from experimental
 358 data in the cited literature for our topics that refer to social attributions and traits
 359 from speech (Suire et al.; Rosenberg & Hirschberg, both this volume). While some
 360 decisions on the material duration are made because of the costs inflicted by the
 361 prospective methods (see Sect. 1.3), other reasons to select material originate in the
 362 aspects under research.

363 The syllables used by Belin (this volume) were recorded in the lab, and subse-
 364 quently post-processed to study the effect of acoustic averaging over speakers. Such
 365 a manipulation of speech recordings is another kind of experimental data. Manipu-
 366 lations comprise post-processing of the acoustic speech signal, as well as outright
 367 synthesis. Manipulated audio files can be in principle of any duration, but are con-
 368 sidered here still as experimental data due to its similarity in careful and specific
 369 creation in a laboratory, but also due to the aim of controlling influencing factors—
 370 this time by means of inducing a controlled number of manipulations. There are
 371 different reasons for such manipulations, most importantly to verify analysis results
 372 with even more controlled material, producing stimuli for experiments which are
 373 hard or impossible to record, or to obtain speech signal qualities for the domain of
 374 computer speech.

375 The papers in the part on technological applications are good examples, as they
376 all refer to studies in which manipulated or synthesized material, typically shorter
377 utterances in a dialog, are used, or they argue to conduct those (Torre & White; Clark
378 et al., both this volume).

379 **1.4.2 *Naturalistic Data Recorded in the Lab***

380 While strictly controlled speech material from the laboratory is a foundation of
381 basic research, there is always the aim to use naturalistic data in order to estimate
382 the strength of effects for real-life situations and to study situational and dialogic
383 aspects that cannot be simulated with—what we call—experimental data. Typically,
384 this means to elicit naturalistic situations and thus also spontaneous material in the
385 lab, often with the help of some supporting material. In contrast to the aforementioned
386 controlled experiments, the lab recordings of naturalistic data are not controlled to
387 the same degree. Here, experimenters aim to control a good acoustic quality, to
388 initiate conversations, and possibly to instruct conversational tasks. That means that
389 the linguistic and phonetic content is not (strictly) controlled for. However, very
390 specific instructions and support material is often provided to support the subjects
391 to elicit the situation, e.g., a game or task, but databases have been created with far
392 less information provided (Schweitzer, Lewandowski, Duran, & Dogil, 2015).

393 For obtaining attractiveness ratings, Quené et al., (this volume) used sentences
394 from spontaneous interview speech as stimuli that were manipulated. They also used
395 visual data. The situation of speed dating was applied by Michalsky and Schoormann
396 (this volume) to allow for studying the effects of prosodic entrainment in dialog.
397 Simulated telephone conversations on pizza ordering from the Nautilus database,
398 but post-edit to exclude the callee were used by Weiss et al. (this volume).

399 **1.4.3 *Data from the Wild***

400 The last category of the material refers to recordings from real situations. Obtaining
401 such data seems to be the easiest one on the first glance. However, it is often practically
402 impossible to ensure sufficient quality and sufficient amount of material given the
403 available resources, especially if there are requirements on the linguistic conditions
404 to be included. In addition, there is often more information on the speakers required,
405 which might be difficult to collect while or after recording, for example, additional
406 physiological measures. Finally, there might be ethical reasons to avoid taking data
407 from the wild.

408 In this collection, this kind of data was selected to solely study charismatic speak-
409 ers. Bosker (this volume) selected speech fragments of c. 25 s from mass media
410 recordings of US presidential debates. Brem and Niebuhr (this volume) used audio-
411 visual data (video clips of charismatic management leaders). For natural data, this

412 kind of material is the least uncontrolled, as the speakers are not only professional,
 413 but also very aware of the fact of being recorded. Therefore, such field data might
 414 not always be considered as truly “wild”, but of course, it is as natural as it can be
 415 when studying speeches of charismatic leaders.

416 Sometimes, it is not easy to assign data to one of the categories. For example, read
 417 Wikipedia articles used by Baumann (this volume) is comparable on the surface with
 418 other naturalistic speech paragraphs read in the lab, except for the varying recording
 419 quality. But still, the origin of this material is natural, as the speakers truly recorded
 420 themselves with the intention to be listened to by people interested in the Wikipedia
 421 articles.

422 1.5 Conclusions

423 The word “attractiveness” stems from Latin “ad trahere” and means “dragging or
 424 pulling to something”. For our topic, people are dragged or pulled to the voice and
 425 vocal behavior of somebody else. This relationship unfolds in various dimensions:
 426 from sexuality and biology over social likability up to charisma and leadership. It is
 427 this diversity of voice attractiveness that we intended to cover in this book. It is our
 428 hope to raise awareness with this book for this diversity and the broad range of the
 429 various scientific fields involved.

430 What we see in the contributions to this volume is on the one hand a clear and
 431 intended separation of the above-mentioned concepts on the sexual, the likable, and
 432 the charismatic speaker. On the other hand, we recognize the interdependencies
 433 between the three concepts. The classical example is that a person perceived as
 434 beautiful is also regarded as a socially more attractive (Zuckermann & Driver, 1989).

435 In our view, we deal here with a contrast between simultaneous distinctive con-
 436 cepts that have not only mutual influences and mutual conditionality. We see a need
 437 for a unifying theory with respect to the concepts, but also the different methods
 438 and data used in the various scientific disciplines. Several contributions in this book
 439 provide useful suggestions for such a theory, which can be viewed as a starting point
 440 for a more systematic foundation to overcome the current limitations of knowledge.

441 As an example can serve the frequency code by Ohala (1984): Similarities between
 442 languages, cultures, and even species in the use and effect of F0 was argued to orig-
 443 inate in biologically grounded separation between “smaller” and “larger” (vocal)
 444 individuals. This does not only reflect the sexual dimorphism in terms of sexual
 445 selection, but also social aspects of signaling and estimating relational power, sub-
 446 missiveness, even helplessness, and thus supports social roles and interaction. The
 447 universal systematic in F0 observed by Ohala concerns charisma, attractiveness, and
 448 likability alike. Following this road to connect biological and articulatory bases for
 449 acoustic and perceptual effects can be seen as one of the most important elements of
 450 a unifying theory.

451 Interestingly, we observe that *trust* occurs in many contributions and it seems
 452 to have an overarching character. Trust, obviously, represents a link between the

453 concepts of the sexual, the social, and the charismatic attractiveness, as it repre-
 454 sents a positive attitude towards another. Trust may be considered as an immediate
 455 result of attractiveness, whatever the kind of attractiveness and social relation might
 456 be. Therefore, it is an important characteristic of human relationships, but also an
 457 important feature for Human-Computer Interaction.

458 References

- 459 Abele, A. E., Cuddy, A. J. C., Judd, C. M., & Yzerbyt, V. Y. (2008). Fundamental dimensions of
 460 social judgment. Editorial to the Special Issue. *European Journal of Social Psychology*, 38(7),
 461 1063–1065.
- 462 Argyle, M. (1988). *Bodily Communication*. New York: Methuen.
- 463 Banse, R., & Scherer, K. (1996). Acoustic profiles in vocal emotion expression. *Journal of Person-*
 464 *ality and Social Psychology*, 70(3), 614–636.
- 465 Bezooijen, R. V. (1995). Sociocultural aspects of pitch differences between Japanese and Dutch
 466 women. *Language and Speech*, 38, 253–265.
- 467 Brown, B. L., Strong, W. J., & Rencher, A. C. (1973). Perceptions of personality from speech:
 468 effects of manipulations of acoustical parameter. *Journal of the Acoustical Society of America*,
 469 54(1), 29–35.
- 470 Brown, B. L., Giles, H., & Thakerar, J. N. (1985). Speaker evaluation as a function of speech rate,
 471 accent, and context. *Language and Communication*, 5(3), 207–220.
- 472 Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of
 473 social perception: The stereotype content model and the BIAS map. *Advances in Experimental*
 474 *Social Psychology*, 40, 62–149.
- 475 Darwin, C. (1890). *The Expression of the Emotions in Man and Animals*. London: John Murray.
- 476 DePaulo, B. M., Kenny, D. A., Hoover, C. W., Webb, W., & Oliver, P. V. (1987). Accuracy of person
 477 perception: Do people know what kinds of impressions they convey? *Journal of Personality and*
 478 *Social Psychology*, 52(2), 303–315.
- 479 Fiske, S. T., Cuddy, A. J., & Glick, P. (2006). Universal dimensions of social cognition: Warmth
 480 and competence. *Trends in Cognitive Sciences*, 11(2), 77–83.
- 481 Hart, R. J., & Brown, B. L. (1974). Personality information contained in the verbal qualities and in
 482 content aspects of speech. *Speech Monographs*, 41, 271–380.
- 483 Krause, S., Back, M. D., Egloff, B., & Schmukle, S. C. (2014). Implicit interpersonal attraction in
 484 small groups automatically activated evaluations predict actual behavior toward social partners.
 485 *Social Psychological and Personality Science*, 20, 671–679.
- 486 Laver, J., & Trudgill, P. (1979). Phonetic and linguistic markers in speech. In K. R. Scherer & H.
 487 Giles (Eds.), *Social Markers in Speech* (pp. 1–32). Cambridge: Cambridge University Press.
- 488 Laver, J. (1980). *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University
 489 Press.
- 490 McAleer, P., Todorov, A. & Berlin, P. (2014). How do you say 'Hello'? Personality impressions
 491 from brief novel voices. *PLOS ONE* 9(3).
- 492 McCroskey, J., & McCain, T. (1974). *The Measurement of Interpersonal Attraction*. *Speech Mono-*
 493 *graphs*, 41, 261–266.
- 494 Mori, M. (2012). The uncanny valley. *IEEE Robotics and Automation* 19(2). Originally 1970,
 495 Translated by MacDorman, K.F. & Kageki, N. (pp. 98–100).
- 496 Nass, C., & Brave, S. (2005). *Wired for Speech. How Voice Activates and Advances the Human-*
 497 *Computer Relationship*. MIT Press.
- 498 Nass, C., Moon, Y., Fogg, B., Reeves, B., & Dryer, D. (1995). Can computer personalities be human
 499 personalities? *International Journal of Human-Computer Studies*, 43, 223–239.

- 500 Ohala, J. (1984). An ethological perspective on common cross-language utilization of F0 of voice.
 501 *Phonetica*, 41, 1–16.
- 502 Paeschke, A., & Sendlmeier, W. F. (1997). Die Reden von Rudolf Scharping und Oskar Lafontaine
 503 auf dem Parteitag der SPD im November 1995 in Mannheim -Ein sprechwissenschaftlicher und
 504 phonetischer Vergleich von Vortragsstilen. *Zeitschrift für Angewandte Linguistik*, 27, 5–39.
- 505 Putnam, W. B., & Street, R. L. J. (1984). The conception and perception of noncontent speech per-
 506 formance: Implications for speech-accommodation theory. *International Journal of the Sociology
 507 of Language*, 46, 97–114.
- 508 Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and
 509 New Media Like Real People and Places*. Cambridge: Cambridge University Press.
- 510 Schaller, M. (2008). Evolutionary basis of first impressions. In N. Ambady & J. J. Skowronski
 511 (Eds.), *First Impressions* (pp. 15–34). New York: Guilford Press.
- 512 Scherer, K. R. (1978). Personality inference from voice quality: The loud voice of extroversion.
 513 *European Journal of Social Psychology*, 8(4), 467–487.
- 514 Schweitzer, A., Lewandowski, N., Duran, D., & Dogil, G. (2015). Attention, please!—Expanding
 515 the GECO database. In *Proceedings of the 18th International Congress of Phonetic Sciences*,
 516 Glasgow, paper 620.
- 517 Street, Jr. R. L., & Brady, R. M. (1982). Speech rate acceptance ranges as a function of evaluative
 518 domain, listener speech rate and communication context. *Communication Monographs* 49(4),
 519 290–308.
- 520 Weirich, M. (2010). *Die attraktive Stimme: Vocal Stereotypes. Eine phonetische analyse anhand
 521 akustischer und auditiver Parameter*. Saarbrücken: Verlag Dr. Müller.
- 522 Weiss, B., Wechsung, I., Kühnel, C., & Möller, S. (2015). Evaluating embodied conversational
 523 agents in multimodal interfaces. *Computational Cognitive Science*, 1(6), 1–21.
- 524 Weiss, B., & Möller, S. (2011). Wahrnehmungsdimensionen von Stimme und Sprechweise. 22.
 525 Konferenz Elektronische Sprachsignalverarbeitung, Aachen (pp. 261–268).
- 526 Zuckermann, M., & Driver, R. E. (1989). What sounds beautiful is good: The vocal attractiveness
 527 stereotype. *Journal of Nonverbal Behaviour*, 13, 67–82.

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Voice Attractiveness	
Series Title		
Chapter Title	Prosodic Aspects of the Attractive Voice	
Copyright Year	2020	
Copyright HolderName	Springer Nature Singapore Pte Ltd.	
Corresponding Author	Family Name	Rosenberg
	Particle	
	Given Name	Andrew
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Google LLC, NYC
	Address	New York, NY, USA
	Email	rosenberg@google.com
Author	Family Name	Hirschberg
	Particle	
	Given Name	Julia
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Columbia University, NYC
	Address	New York, NY, USA
	Email	julia@cs.columbia.edu
Abstract	<p>A speaker's voice impacts listeners' perceptions of its owner, leading to inference of gender, age, personality, and even height and weight. In this chapter, we describe research into the qualities of speech that are deemed "attractive" by a listener. There are a number of ways that a person can be found attractive. We will review the research into what makes speakers attractive in the political and business domains, and what vocal properties lead to perceptions of trust. We then turn our attention to research into "likeability" and romantic attraction. While the lexical content of a speaker's speech is important to their attractiveness, we focus this survey on prosodic qualities, those acoustic properties that describe "how" the words are said rather than "what" the words are. Of course, attractiveness is subjective; what is attractive to one listener may not be to another. Properties of the listener and other contextual qualities can have a significant impact on the voices which are found to be attractive. The most comprehensive research in this topic includes analyses of both the speaker and the listener, since attraction is frequently a mutual phenomenon; when people are attracted to someone, they want to be found attractive in return. We will also summarize work that has investigated attraction dynamics in two-party conversations.</p>	
Keywords	<p>Likeability - Charisma - Political attractiveness - Business attractiveness - Romantic attraction - Speech prosody - Vocal attractiveness</p>	

Chapter 2

Prosodic Aspects of the Attractive Voice



Andrew Rosenberg and Julia Hirschberg

Abstract A speaker's voice impacts listeners' perceptions of its owner, leading to inference of gender, age, personality, and even height and weight. In this chapter, we describe research into the qualities of speech that are deemed "attractive" by a listener. There are a number of ways that a person can be found attractive. We will review the research into what makes speakers attractive in the political and business domains, and what vocal properties lead to perceptions of trust. We then turn our attention to research into "likeability" and romantic attraction. While the lexical content of a speaker's speech is important to their attractiveness, we focus this survey on prosodic qualities, those acoustic properties that describe "how" the words are said rather than "what" the words are. Of course, attractiveness is subjective; what is attractive to one listener may not be to another. Properties of the listener and other contextual qualities can have a significant impact on the voices which are found to be attractive. The most comprehensive research in this topic includes analyses of both the speaker and the listener, since attraction is frequently a mutual phenomenon; when people are attracted to someone, they want to be found attractive in return. We will also summarize work that has investigated attraction dynamics in two-party conversations.

Keywords Likeability · Charisma · Political attractiveness · Business attractiveness · Romantic attraction · Speech prosody · Vocal attractiveness

A. Rosenberg (✉)
Google LLC, NYC, New York, NY, USA
e-mail: rosenberg@google.com

J. Hirschberg
Columbia University, NYC, New York, NY, USA
e-mail: julia@cs.columbia.edu

© Springer Nature Singapore Pte Ltd. 2020
B. Weiss et al. (eds.), *Voice Attractiveness*, Prosody, Phonology and Phonetics,
https://doi.org/10.1007/978-981-15-6627-1_2

17

2.1 Understanding Vocal Attractiveness

Attraction is central to human social bonding. It is an expression of whom we choose to be close to and whom we choose to avoid. There are as many types of attraction as there are types of interaction. In this chapter we will survey the prosodic qualities of different types of attractive voices.

A person's speech communicates a wide variety of information about the speaker. Not only information that they are trying to communicate, but information about the speaker themselves is important in this regard. This information enables listeners to assess the gender and age of a speaker, their emotional state, and aspects of both their personality, and physicality, all while listening to a person speak. These qualities may be more or less attractive to a listener based on their inherent preferences and other situational factors. For example, in the case of political attractiveness, there are times when anger in a speaker can resonate with a listener and will be perceived positively, while in other contexts anger is deemed inappropriate and, therefore, unattractive.

We divide this survey into five sections, based upon different types of attractiveness. In Sect. 2.2 we discuss political attractiveness. Political figures attract and retain followers through their speeches, interviews, and other public performance. Understanding what allows a speaker to gain political authority has been a source of investigation in political science and sociology for many years. Of late, more computational approaches have been brought to bear in assessing what kind of speech is perceived as charismatic. Also related to this is the kind of charisma that is found in business leaders (cf. Sect. 2.3). The business community takes communication and leadership very seriously. A significant amount of work has examined the speech of entrepreneurs and established (and sometimes beloved) executives in hopes of understanding what draws investors and employees to a business leader. Central to both of these types of attractiveness is trust. In Sect. 2.4 we will survey research that strives to identify what makes a voice sound trustworthy. Researchers also tend to distinguish two more social types of attraction: likeability (Sect. 2.5) and romantic attraction (Sect. 2.6). These types of attractiveness are not identical, but neither are they orthogonal. Types of attraction may overlap with one another. Leaders who are politically attractive may also be perceived as likeable. In addition, physical attraction can impact the degree to which people are trusted. The types of voices that signal qualities of business success may be attractive to some people as friends or romantic partners, but may be unattractive to others.

In all of these analyses of vocal attractiveness, spoken communication is an important avenue to establishing the central social bond. People appear to have relatively consistent preferences regarding vocal attractiveness. Many of these vocal qualities are associated with other speaker properties that are considered attractive; for example, male body size in the case of romantic attractiveness, or enthusiasm and dynamism in the case of political and business leaders, are correlated with attractiveness.

Of course, attractiveness is not an objective phenomenon. Qualities of the listener also contribute to their perceptions of attraction. These can include sexual preference

63 in romantic attraction or political bias in assessing political attractiveness. Similarly,
 64 some voices and messages resonate more or less with a listener on the basis of any
 65 number of factors—memories, contextual relevance, broader business or political
 66 context, or other idiosyncrasies.

67 Another quality that adds a layer of complexity to understanding the attractive
 68 voice is the interplay between inherent and performance qualities of the voice. In
 69 general, studies are looking to assess what makes a voice inherently attractive, but
 70 the same voice may be used in ways that are more or less attractive. Most studies
 71 avoid direct assessment of this distinction. Some will look at the same speaker in dif-
 72 ferent venues or types of speech (cf. Sect. 2.2.1). Other work, particularly in studying
 73 romantic attractiveness (cf. Sect. 2.6), will contextualize speech in two-party conver-
 74 sations and consider qualities and assessments of the two speakers. Distinguishing
 75 the influence of the voice itself and the way it is used in a particular stimulus remains
 76 an open question in these studies. Overall assessments of attractiveness in each of
 77 these domains is a combination of both inherent qualities of the voice and how it is
 78 being used in the specific utterance that is being assessed.

79 Moreover, attraction is often a dynamic process in which conversational partners
 80 are simultaneously being attracted (or repelled) by an interlocutor while demonstrat-
 81 ing their own preference for their partner to be attracted to (or repelled by) them.
 82 This contemporaneous perception and performance can make analysis challenging.
 83 For example, male voices which are spoken lower in the speakers' pitch range and
 84 with a relatively large formant dispersion tend to be found attractive by heterosexual
 85 women. But men who are attracted and are signaling their attraction to a conversa-
 86 tional partner demonstrate the same qualities. So should we consider this voice to be
 87 attractive or flirtatious?

88 While there are relatively few clear, consistent, and universal answers to what
 89 makes speech attractive even in a specific context, to a specific group, there are some
 90 broad conclusions in the literature centered around identifying prosodic properties of
 91 an attractive voice. This chapter is an attempt to summarize the current understanding,
 92 highlight gaps and inconsistencies, and provide some directions for future inquiry.

93 2.2 Political Attractiveness and Charisma

94 Charisma is defined as the ability to persuade and command authority by virtue of
 95 personal qualities rather than through formal institutional (political, organizational,
 96 or military) structures (Weber, 1947). Viewed from this perspective, charisma is a
 97 challenge for institutional stability because it represents a path to leadership that
 98 eschews standard institutional pathways to power. Alternately, charisma is an impor-
 99 tant driver of revolutionary change specifically because it does not require specific
 100 structures to grant power; rather, it is a quality attributed to a person by her or his
 101 followers.

102 There is a wealth of political science and sociology research on charismatic leaders
 103 and movements, including importantly (Weber, 1947; Boss, 1976; Marcus, 1961). In

104 this section, we will survey research that has used empirical techniques to investigate
 105 charismatic political speech. In Sect. 2.2.1, we will survey studies that have looked
 106 at spoken correlates of charismatic speech. We will summarize work that has sought
 107 to define charisma empirically in Sect. 2.2.2.

108 2.2.1 *Vocal Correlates of Charisma*

109 Rosenberg and Hirschberg (2005, 2009) describe the first set of studies that attempt
 110 to measure the vocal and lexical correlates of charisma in American English. This
 111 study presented 45 speech segments to eight subjects. Materials were chosen to
 112 balance speakers, topics, and genres. A small set of speakers were chosen from those
 113 whose public speech covered a similar set of topics, and for whom speech tokens
 114 could be found in a wide variety of genres, or speaking styles. Since the experiment
 115 was designed during the winter and spring of 2004, there was abundant speech
 116 material available for the nine candidates running at that time for the Democratic
 117 Party's nomination for President. Speakers were limited to Democrats in this study
 118 to confine the range of opinions presented in the tokens, as it has been suggested
 119 in the literature (Boss, 1976; Dowis, 2000; Weber, 1947) that a listener's agreement
 120 with a speaker bears upon their judgment of that speaker's charisma. Segments were
 121 selected from a variety of topics in order to test the influence of topic on subject
 122 judgments of charisma. Five speech tokens were chosen from each speaker, one
 123 on each of the following topics: health care, postwar Iraq, Pres. Bush's tax plan,
 124 the candidate's reason for running, and a content-neutral topic (e.g., greetings). For
 125 these five tokens, genre was also varied among the following types: interview, debate,
 126 stump speech, campaign ad.

127 Subjects were presented with each of the stimuli twice, with a 2 s silence between
 128 presentations. They were asked to respond to 26 statements about the speaker includ-
 129 ing "The speaker is charismatic." The order of presentation of stimuli and statements
 130 was randomized for each subject.

131 Using the subject responses, a mean score measuring the degree to which the
 132 speech in each token was calculated in order to examine the extent to which the
 133 subject believed that the speaker was charismatic. Colloquially this was referred to
 134 this as "how charismatic" the utterance was—despite charisma being a quality of
 135 the speaker rather than the speech itself. With this mean charisma score for each
 136 token, it was possible to analyze acoustic–prosodic qualities of the speech to iden-
 137 tify correlates with charisma. These qualities were identified by measuring pitch,
 138 intensity, speaking rate, and duration features of the tokens in the experiment and
 139 then measuring the degree of correlation between these features and subject ratings
 140 of the charismatic statement. Results of these analyses showed significant positive
 141 correlations between charisma ratings and the duration of the speech, whether mea-
 142 sured in words, seconds, or number of phrases. These results also showed positive
 143 correlations between enthusiastic and passionate ratings and mean and maximum F0,
 144 intensity, and speaking rate. More colloquially this means, higher pitched, louder,

145 and faster speech is considered to be more passionate and more enthusiastic (with
 146 caveats that the perceptual properties of pitch and loudness are not identical to the
 147 acoustic measurements of mean and maximum F0 and intensity). Additionally, a
 148 positive correlation between standard deviation of F0 and ratings of enthusiastic and
 149 passionate speech was observed in male speakers.

150 In a later study, Rosenberg and Hirschberg (2009) extended this analysis to include
 151 ToBI labeling (Beckman & Hirschberg, 2005) of the segments. In this study, phrase
 152 boundary prosody was classified into three types: rising pitch (L-H%; H-H%), falling
 153 pitch (L-L%; L-), and plateau or flat pitch (H-L%; H-). Results showed that the rate
 154 of rising tokens negatively correlates with charisma. Rising intonation is used in
 155 questions, and can be associated with uncertainty. Neither of these qualities is con-
 156 sistent with “persuasiveness,” a component of charisma. Consistent with this, the
 157 L*+H pitch accent type, also associated with uncertainty, had a negative correlation
 158 with charisma. The L*+H pitch accent is realized with low pitch on a prominent
 159 syllable nucleus which rises, typically reaching a peak after the nucleus boundary.
 160 In addition, prosody associated with “new” information (H* pitch accents) was posi-
 161 tively correlated with charisma, while prosody associated with “given” information
 162 (downstepped contours: H* !H* L-L%) was negatively correlated. H* pitch accents
 163 are high tone pitch peaks that are more or less time-synchronized with intensity peaks
 164 occurring within syllable nuclei. Downstepped high pitch accents, !H*, are H* pitch
 165 accents that occur after a previous high tone, and have a lower pitch height during
 166 their high tone. The “downstepped” contour is a shorthand to describe a high tone,
 167 followed by one or more downstepped high tones with a L-L% phrase ending.

168 Other notable efforts in measuring vocal correlates to charisma have investi-
 169 gated political speech in other languages and countries. From this work we can look
 170 for evidence of linguistic and or cultural biases in the perception or production of
 171 charisma. Disentangling these factors (linguistic vs. cultural; perception vs. produc-
 172 tion) is virtually impossible given the size of these studies and additional confounds
 173 (speaker/listener demographics and other biases, political, social, and temporal con-
 174 text to name a few) that all analyses in this space are subject.

175 Cullen and Harte (2018) analyzed a relatively large set (945 utterances) of longi-
 176 tudinal speech material from a single speaker, over seven years (2007–2012). This
 177 material, compiled as the Irish Political Speech Database, has a number of useful qual-
 178 ities. By focusing on a single speaker, many political biasing elements are controlled
 179 for. By including many recording contexts (talk shows, parliamentary speeches) dif-
 180 ferences in genre can be accounted for. The longitudinal aspect also allows polling
 181 data to be associated with the politician’s speech, facilitating investigation of how
 182 popularity or standing impact communication. This work also included automatic
 183 classification of charisma based on acoustic–prosodic features. The authors found
 184 that prosodic features, based on pitch, intensity, and duration, outperformed spectral
 185 features. The specific performance of this classifier is somewhat immaterial—the
 186 broad applicability of a single speaker model for a paralinguistic task is *extremely*
 187 limited. But the relative value of the acoustic signal is revealing—charisma is found
 188 here to be a function of suprasegmental qualities more than voice quality (as captured
 189 by spectral features).

190 Biadsy, Rosenberg, Carlson, Hirschberg, and Strangert (2008) significantly
 191 extended the studies described in Rosenberg and Hirschberg (2005, 2009). The orig-
 192 inal American English stimuli were additionally rated by native Swedish and Pales-
 193 tinian Arabic speakers, and a subsequent study presenting Palestinian Arabic speech
 194 to speakers of the American English and Palestinian Arabic was conducted. Compar-
 195 ative analysis of the original study with these four new studies allowed the identifica-
 196 tion of some vocal correlates of charisma that appear to be robust to differences in the
 197 language of the speaker or listener. Others appeared to be sensitive to the language
 198 of the listener, regardless of the language of the speaker, and still others are specific
 199 to the speaker/listening configuration. For example, across all experiments, mean
 200 pitch, pitch range, mean and standard deviation of intensity, and stimulus duration
 201 all positively correlated with charisma ratings regardless of the language spoken and
 202 the native language of the rater. Conversely, the presence of disfluencies negatively
 203 correlated with charisma in all experiments, though this correlation was weakest for
 204 Swedish judgments of American English.

205 The studies also found that raters tended to pattern similarly in response to many
 206 aspects of the stimuli regardless of their native language. For instance, when assessing
 207 English stimuli, minimum F0 was positively correlated with charisma. However,
 208 when assessing Palestinian Arabic utterances, this feature was negatively correlated
 209 for Palestinian subjects, and not significant for American subjects. Also both groups
 210 judging Arabic data rated speech to be more charismatic that exhibits larger standard
 211 deviations in F0 but none of the groups judging English showed the same effect.

212 Finally, a third group of correlates appeared to be specific to the language of both
 213 speaker and listener. For example, the speaking rate was positively correlated with
 214 charisma judgments only for American and Swedish ratings of English: the faster the
 215 speech, the more charismatic the speaker was deemed to be. However, when Pales-
 216 tinians judged Arabic speakers, speaking rate approached a negative correlation with
 217 charisma, with no correlation between speaking rate and charisma when Palestinians
 218 judged American English or Americans judged Palestinian Arabic.

219 This is not the only work that has looked at cross-cultural biases in perceptions and
 220 production of charisma. Though not every investigation found clear differences on the
 221 basis of culture or nationality. For example, Cullen et al. (2014) also found that native
 222 Irish raters and Amazon Mechanical Turk workers, who are largely American, were
 223 quite consistent in their assessment of Irish Political speech with respect to charisma.

224 Pejčić (2014) investigated persuasiveness in Serbian and British political speech,
 225 which appears clearly related to charisma. This study presented five samples of Ser-
 226 bian political speeches and five samples of British speeches to 113 Serbian subjects
 227 asking them to respond to a subset of the 26 statements used in Rosenberg and
 228 Hirschberg (2009) on a 5-point Likert scale. Acoustic analysis was performed on the
 229 tokens from both languages considered as a common population, and also on each
 230 language in isolation. When pooling both languages, relatively few statistically sig-
 231 nificant correlates with persuasiveness were observed. These were the standard devia-
 232 tions for F0 peaks in narrow-focused rising nuclear tones, their percentage in Tone
 233 Units' F0 range and the maximum F0 of their Tone Units. Anecdotal observations

234 suggest roughly that larger F0 excursions were positively associated with persuasion
 235 in Serbian speech, but negatively associated with British speech, at least when rated
 236 by Serbian speakers.

237 In addition to these studies, there are a number of descriptive investigations of the
 238 speaking style of politicians, particularly concerning the recognition of charisma.
 239 Pèrez (2016) contrasted the speech of the Venezuelan politicians, Hugo Chávez and
 240 José Luis Rodríguez Zapatero, characterizing Chávez as using a “revolutionary”
 241 style, consistent with charismatic authority, whereas Zapatero uses a more “tra-
 242 ditional” style, consistent with institutional authority. Ryant and Liberman (2016)
 243 proposed a number of visualization techniques to investigate and compare prosodic
 244 qualities of speech, using U.S. Presidents Barack Obama and George W. Bush as
 245 examples.

246 2.2.2 Defining Charisma

247 Careful reading will reveal that the studies described in Sect. 2.2.1 side-step any
 248 definition of “charisma.” Specifically, subjects in Rosenberg and Hirschberg (2005)
 249 were simply asked to respond to the statement “The speaker is charismatic,” which
 250 does very little to identify the personal or vocal qualities that lead to this perception.

251 Researchers in other fields have posited a number of factors that contribute to
 252 perceptions of charisma. Boss (1976) sees charismatic leaders emerging from an
 253 *important crisis* met by an *inspiring message* delivered by a messenger with a *gift*
 254 *of grace*. Marcus takes a more specific view identifying charisma as a product of the
 255 faith of a potential leader’s *listener-followers* (Marcus, 1961). While these are useful
 256 perspectives on political attractiveness and authority, they provide little direction
 257 when we try to empirically quantify charisma and charismatic speech.

258 In Rosenberg and Hirschberg (2005), subjects were asked to respond to the state-
 259 ment “the speaker is charismatic.” But the subjects also responded to 25 other state-
 260 ments about the speaker and his or her speech. Most of these were of the form
 261 “The speaker is X,” where X was one of the following: *charismatic, angry, spon-*
 262 *taneous, passionate, desperate, confident, accusatory, boring, threatening, informa-*
 263 *tive, intense, enthusiastic, persuasive, charming, powerful, ordinary, tough, friendly,*
 264 *knowledgeable, trustworthy, intelligent, believable, convincing, reasonable*. These
 265 attributes represent a subset of those often associated in the literature with charisma.
 266 “The speaker’s message is clear” and “I agree with the speaker” were also included
 267 as statements to be rated. Using these ratings, along with the ratings of charisma, it was
 268 possible to determine which *other* qualities were highly correlated with charisma, to
 269 help in developing a “functional” definition of this term. Rather than offering a for-
 270 mal definition of charisma as a sociopolitical concept or a vocal characteristic, these
 271 results indicate how the subjects themselves understood charisma and how they were
 272 using the term. Specific results can be found in Table 2.1. These results confirmed
 273 some of the conventional wisdom of what we mean when we say charismatic—
 274 specifically, a charismatic speaker is **charming**—and what we believe charisma to

Table 2.1 Statements showing the most consistent subject responses with the statement “The speaker is charismatic”

Statement	κ
The speaker is enthusiastic	0.606
The speaker is charming	0.602
The speaker is persuasive	0.561
The speaker is boring	-0.513
The speaker is passionate	0.512
The speaker is convincing	0.503

275 be used for—a charismatic speaker is **convincing** and **persuasive**. However, they
 276 also provide support for claims found in Dowis (2000) and Boss (1976) that charis-
 277 matic speakers should be passionate and enthusiastic and, by extension, not boring.
 278 It was also interesting to see that responses to the *desperate*, *threatening*, *accusatory*,
 279 and *angry* qualities showed no positive or negative ($|\kappa| < 0.15$) agreement with
 280 the charismatic statement. Apparently, a charismatic speaker *can* demonstrate these
 281 qualities, but, at least among the subjects in this study, they neither promote nor
 282 inhibit perceptions of charisma.

283 A similar approach to defining charisma was undertaken in Signorello, D’Errico,
 284 Poggi, and Demolin (2012). This study administered a free-form web survey, asking
 285 58 French participants to provide adjectives that are consistent or inconsistent with the
 286 term “charisma” as they understood it. Retaining only adjectives that were reported by
 287 more than one subject, the authors identified 40 terms that were positively associated
 288 with charisma and 27 that were negatively associated. To facilitate understanding,
 289 the authors grouped these into five categories (1) Pathos, (2) Ethos Benevolence,
 290 (3) Ethos Competence, (4) Ethos Dominance, and (5) Emotional Induction Effects.
 291 Table 2.2 is reproduced from Signorello et al. (2012). Note that *charming*, *persuasive*,
 292 *enthusiastic*, and *‘boring’* appear in both Signorello et al. (2012) and Rosenberg
 293 and Hirschberg (2009) despite the studies using French and American participants,
 294 respectively.

295 One divergent finding did appear however: while Rosenberg and Hirschberg
 296 (2009) found no correlation between *threatening* and *anger*, Signorello et al. (2012)
 297 identified through factor analysis an *Authoritarian-Threatening* factor which in their
 298 study is a factor, including the terms *determined*, *authoritarian*, *leader*, *confident* as
 299 well as the more aggressive terms *Who Scares*, *cold*, *dishonest* and *menacing*.

300 While not directly related to defining charisma, but related to political speech,
 301 an interesting idea presented in Cullen and Harte (2018) addresses vocal attractive-
 302 ness more broadly. The Irish Political Speech Database is labeled for six attributes:
 303 *charisma*, *boring*, *enthusiastic*, *inspiring*, *likeable*. From these six, Cullen and Harte
 304 (2018), define Overall Speaker Appeal (OSA) as the average of these six ratings
 305 (including negative boredom ratings). The correlation of these attributes may limit

Table 2.2 The 67 positive and negative adjectives related to charisma. Reproduced from

Dimension	Positive adjectives	Negative adjectives
Pathos	Passionate, empathetic, enthusiastic, reassuring	Cold, indifferent
Ethos benevolence	Extroverted, positive, spontaneous, trustworthy, honest, fair, friendly, easygoing, makes the others feel important	Untrustworthy, dishonest, egocentric, individualistic, introverted
Ethos competence	Visionary, organized, smart, sagacious, creative, competent, wise, enterprising, determined, resolute, who propose, seductive, exuberant, sincere, clear, communicative	Inefficient, inadequate, uncertain, faithless, unclear, menacing
Ethos dominance	Dynamic, calm, active, courageous, confident, vigorous, strong, leader, authoritarian, captivating, who persuade, who convince	Apathetic, timorous, weak, conformist, unimportant, who scare
Emotional induction effects	Charming, attractive, pleasant, sexy, bewitching, eloquent, influential	Boring

306 the efficiency of this measure, but the attempt to summarize these signals into a single
 307 measure is potentially valuable, even if the specific formulation might benefit from
 308 modification.

309 **2.3 Business Attractiveness**

310 Business organizations are an area in which leadership and authority have clear
 311 impacts. There are many organizational structures that are used in business activi-
 312 ties, but all instill participants with distinct, decision-making authority. Within these
 313 structures, charismatic authority can be manifested the way (Weber, 1947) formulated
 314 it—as an alternative to established, institutional authority. This would be revealed
 315 by a situation where employees look to a co-worker who is not in a management or
 316 reporting structure for direction rather than their direct manager. A more common
 317 way to think about charismatic leadership in a business context is when charismatic
 318 authority is aligned with institutional authority. This allows us to think about “how
 319 charismatic” is one manager, one CEO, or one founder over another.

320 While there is always an element of “trust” in a leader–follower relationship, this
 321 is somewhat more quantifiable in business relationships. Investors are entrusting their
 322 capital to the efforts of a founder when they invest in a business. While the specific
 323 leadership of a founder may be more essential to a start-up, opinions about the CEO
 324 can have an impact on institutional investing in well-established corporations.

Editor Proof

325 We previously noted some of the complications in defining charisma. The use of
 326 a limited number of speakers who have a cultural consensus of being charismatic
 327 is one way to get around a broader definition. One thread of work undertaken by
 328 Oliver Niebuhr and colleagues has been to study Steve Jobs, former CEO and co-
 329 founder of Apple Inc., as an exemplar of a charismatic business leader. Niebuhr,
 330 Brem, Novák-Tót, and Voße (2016b) posit a profile of charismatic speech based on
 331 a reading of previous political studies (cf. Sect. 2.2). This is summarized as having
 332 high and varied pitch, high and varied intensity, a fast speaking rate, few disfluencies,
 333 a large number of emphatic words, but with varied realizations and high rhythmic
 334 variation. By automatic analysis of two landmark speeches (launching the iPhone 4
 335 and iPad 2) they find that Steve Jobs does in fact fit this profile.

336 This research direction is continued in a number of works via a contrastive anal-
 337 ysis of Steve Jobs and Mark Zuckerberg, founder and CEO of Facebook (Mixdorff,
 338 Niebuhr, & Hönemann, 2018; Niebuhr, Voße, & Brem, 2016a, 2018b). The approach
 339 here is based on the common perceptions of Steve Jobs as a charismatic speaker and
 340 Mark Zuckerberg as a less charismatic speaker, though both were CEOs of major
 341 corporations at the time their speech was collected for analysis.

342 Niebuhr et al. (2016a) find that Jobs has shorter phrases, fewer and shorter hesi-
 343 tations, and a more dynamic use of pitch and rhythm than Zuckerberg. While Jobs
 344 speaks quickly (compared to “normal” speech), Zuckerberg’s speaking rate is even
 345 higher. This contributes to strong phonetic reductions in his speech which may neg-
 346 atively impact perceptions of charisma. Applying the Fujisaki model of intonation
 347 Fujisaki and Hirose (1984), Mixdorff et al. (2018) enable a more specific analysis
 348 of how the two CEOs manipulate pitch in their speeches. In general, this analysis
 349 brings insight into the earlier (and overly simplistic) findings that high pitch leads
 350 to perceptions of charisma. These two speakers differ more in how they reset their
 351 pitch ranges across phrases and the strength of their excursions. This work is then
 352 expanded upon in Niebuhr et al. (2018b) where the timing and shape of pitch accents
 353 are examined. Moreover, the authors find that a large vowel space, limited place of
 354 assimilation, and a clear differentiation between voiced and unvoiced stops all dif-
 355 ferentiate Jobs from Zuckerberg. These factors all contribute to fast, dynamic speech
 356 that is clearly pronounced.

357 While analysis of specific business leaders enables clear contrastive discussion,
 358 there is more work that looks at business speech in entrepreneurship more gener-
 359 ally. Weninger, Krajewski, Batliner, and Schuller (2012) extracted speeches from
 360 143 male business leaders that were shared on YouTube. They collected ratings of
 361 charisma and attempted to automatically predict the human ratings with acoustic
 362 and linguistic features. The raters were 10 psychology Ph.D. students, 5 male and
 363 5 female.¹ This work investigated a large number (1,582) of acoustic–prosodic fea-
 364 tures, in addition to lexical features derived from automatic speech recognition tran-
 365 scripts of the speeches. This work finds that charisma can be automatically detected
 366 with 61.9% accuracy, significantly over chance level, based on acoustic-prosodic and
 367 lexical features.

¹No statistically significant gender effects in the ratings of charisma were discovered.

368 While the previous studies looked at established business leaders (Niebuhr, Brem,
369 & Tegtmeier, 2017) investigated start-up state entrepreneurs, since “a decisive part
370 of their strategy and daily work is to persuade others.” Leaders of these early stage
371 businesses need to convince both investors, suppliers, and customers of the legitimacy
372 of their nascent technology, developing products and services, and of the likely market
373 demand. In this study, 45 participants gave the same elevator pitch, 15 practiced with
374 no feedback, 15 received visual feedback, and 15 received feedback based on the
375 Steve-Jobs-as-charismatic-exemplar acoustic model described above. They found
376 that speakers who received acoustic feedback about their speech were rated 41%
377 more charismatic following training, significantly more than those who received no
378 feedback (24% more charismatic) or those who received visual feedback (12% more
379 charismatic).

380 Extending this investigation of entrepreneurial speech into spectral qualities con-
381 tributing to voice quality, Niebuhr et al. (2018a) found that a fuller and less breathy
382 voice also led to higher speaker charisma ratings. This may be consistent with find-
383 ings that suggest that clear or easily understood speech is an important element to
384 charisma.

385 Much of the study of business attractiveness has been focused on analysis of
386 speech spoken by men. On one hand, this can limit variability to facilitate analysis.
387 On the other, it perpetuates patriarchal norms, implicitly treating charisma—and
388 specifically business leadership—as a quality only associated with male speech.
389 This thus limits our ability to understand charisma in female speakers. Novák-Tót,
390 Niebuhr, & Chen, 2017) investigated the bias in the perception of speeches delivered
391 by American female executives Oprah Winfrey and Ginni Rometti and male executive
392 Steve Jobs. No information as to the gender of the raters was provided. They found
393 that female speech that is judged to be as charismatic as male speech demonstrates
394 more and stronger acoustic cues to charisma. This suggests that this gender bias may
395 be compensated for by making a greater effort by the female speakers. Significantly
396 more work is necessary with regard to the charisma of female leaders both in business
397 and politics alike.

398 2.4 Vocal Correlates of Trust

399 Trust and attractiveness are closely related. Some studies have found that people
400 trust romantically attractive strangers more than unattractive ones, e.g., Wilson and
401 Eckel (2006). While others have found that the relationship is not so simple. Sofer,
402 Dotsch, Wigboldus and Todorov (2015) found that more “typical” faces elicited
403 more trust, rather than the most attractive faces. In this work, “typical” faces were
404 constructed as an averaged composite of 92 faces, while the “attractive” face was
405 an averaged composite of the 12 most attractive in the used data set. However, in
406 an investigation of responses to dating profiles, McGloin and Denes (2018) found
407 that attractive men were considered trustworthy, but attractive women were not.
408 It is worth noting that in both of these studies, the presented face was exhibiting

409 a “neutral” expression. Smiling or grimacing would likely impact impressions of
410 attractiveness, pleasantness, trustworthiness, and likeability in unanticipated ways.
411 When we think about attractiveness more broadly, as we have done in this chapter,
412 trust is a necessary component to political, business, and nonsexual attractiveness.
413 In the political and business roles, attractiveness can endow abilities to the person.
414 They can obtain political power via elections or they can obtain commercial power
415 through investment. Trusting the person is necessary when granting these abilities
416 and responsibilities to the person.

417 In an analysis of deceptive and truthful, trusted, and mistrusted speech in
418 the Columbia Cross-Cultural Deception (CXD) corpus, Levitan, Mareida, and
419 Hirschberg, Levitan et al. (2018) found significant differences in trusted and mis-
420 trusted speech. The CXD corpus is a study of deceptive versus nondeceptive speech
421 from native speakers of Standard American English (SAE) and Mandarin Chinese
422 (MC), all speaking in English. The participants were balanced between male and
423 female speakers and native speakers of English and Chinese. It contains interviews
424 between 340 subjects in 122h of speech. A variation of a fake resume paradigm
425 was used to collect the data. All subjects were previously unacquainted, and pairs
426 of subjects played a “lying game” with each other. Each subject filled out a 24-item
427 biographical questionnaire and was instructed to create false answers for a random
428 half of the questions. They also reported demographic information including gender
429 and native language, and completed the NEO-FFI personality inventory. The speech
430 was recorded in a double-walled sound booth, where the two subjects were sepa-
431 rated by a curtain to ensure no visual contact. For the first half of the game, one
432 subject assumed the role of the interviewer, while the other answered the biographi-
433 cal questions, lying for half and telling the truth for the other; questions chosen in
434 each category were balanced across the corpus. For the second half of the game, the
435 subjects’ roles were reversed, and the interviewer became the interviewee. During
436 the experiment, the interviewer was encouraged to ask follow-up questions to aid
437 them in determining the truth of the interviewee’s answers. Interviewers recorded
438 their judgments for each of the 24 questions, providing information about human
439 perception of deception. Subjects were incentivized monetarily: for every response
440 to the 24 questions that the interviewer judged correctly, the interviewer received
441 an extra \$1, while every incorrect judgment cost them \$1. Every false answer the
442 interviewee persuaded the interviewer was true gained the interviewee \$1, while
443 every false answer the interviewer judged false lost the interviewee \$1. The interview-
444 ees annotated each of their statements during the interview by pressing a “truth” or
445 “false” key on a computer keyboard. We aligned these annotations with transcriptions
446 of the interviews obtained by speech recognition with crowdsourced corrections and
447 automatically aligned the transcripts with the speech recordings.

448 Overall, the researchers found that the mistrusted speech in their corpus (interview-
449 ee responses that were not believed by interviewers) was significantly more intense
450 (louder) and spoken in a higher pitch range, while the speech that interviewers tend to
451 trust was spoken more rapidly. However, they also found differences between male
452 and female and English and Mandarin Chinese native speakers in these features.
453 While male speakers did tend not be trusted when they spoke in a high pitch range,

454 this was not true of female speakers (note that all features were z-score normalized,
 455 so these findings were not influenced by a speaker’s “normal” range or loudness
 456 or speaking rate). Both genders were trusted more when they spoke more rapidly.
 457 Female speakers, however, were trusted more when their voice quality exhibited
 458 more jitter and shimmer—instabilities in their pitch and intensity associated with per-
 459 ceived “roughness” or “breathiness.” There were also differences in trustworthiness
 460 in speakers’ native language backgrounds, although all speakers spoke in English. In
 461 general, native speakers of Standard American English were more trusted when they
 462 exhibited high jitter and shimmer while this was not a significant factor for native
 463 speakers of Mandarin Chinese, who were more trusted when they spoke more rapidly.
 464 These Chinese speakers were less likely to be trusted when they spoke in a high pitch
 465 range and when their overall mean pitch was high; they were also mistrusted when
 466 their maximum intensity was high and when their Harmonics-to-Noise (HNR) ratio
 467 (another measure of voice quality disorders) was high.

468 The researchers also examined the gender and the native language of the inter-
 469 viewers that correlated with their judgments whether interviewers are lying or telling
 470 the truth. Overall, all interviewers mistrusted speech with a high pitch range and a
 471 high maximum intensity and trusted speech spoken rapidly. However, there were
 472 major differences between genders. Male interviewers distrusted speech with high
 473 mean pitch and maximum intensity and trusted fast speaking rate while females only
 474 mistrusted high jitter and shimmer. Comparing native English speakers to native
 475 Mandarin speakers, the researchers found fewer differences: both mistrusted high-
 476 intensity speech and trusted faster speaking rate, but only native English speakers
 477 mistrusted high pitch range.

478 2.5 Likeability or Nonsexual Social Attractiveness

479 The distinction between finding a voice “pleasant” to listen to, and finding the speaker
 480 to be socially attractive as in “I like this person” is difficult to distinguish in research
 481 protocols. These two facets may overlap, they may even be identical for some lis-
 482 teners, but there may be differences that are elided in the research in this space.

483 There are several factors that have been found to contribute to likeability in speech.
 484 Strangert and Gustafson (2008) found that the speaker should be proficient. That
 485 is, the speech should include limited disfluencies and a reasonably high speaking
 486 rate. For clear speech, Weiss and Burkhardt (2010) found that warm/relaxed speech
 487 correlated significantly with likeability.² This included less pressed, more breathy
 488 voice quality and lower spectral center of gravity.

489 Weiss and Burkhardt (2012) performed a focused analysis of 30 speakers rated
 490 as highly likeable and 30 that were highly not-likeable, drawn from the material
 491 used in the 2012 Interspeech paralinguistics challenge (which is discussed in detail

²Note the difference in likeability correlating with *relaxed* speech, while charismatic speech (cf. Sect. 2.2.1 correlates with passion and enthusiasm.

below). The presence of positive factors of likeability was found in all speakers. These included minimal disfluencies and no discernible accent. However, unlikable speakers show higher pitch, lower articulation rate, and lower pronunciation precision. This suggests that these factors can make a speaker “unlikeable,” although perhaps the mere absence of negative attributes is sufficient for an unknown speaker of a relatively short amount of speech to be viewed as “likeable.”

Regarding the “no discernible accent” finding of Weiss and Burkhardt (2012), there appears to be a more nuanced relationship between social factors like likeability and trust and a speaker’s accent. For example, Tavernier (2007) examined perceptions of Flemish speaker’s responses to English speech. They found the highest social attractiveness and trust ratings to come from RP (Native British) speech, with the lowest ratings coming from Flemish-accented English, despite the raters being Flemish speakers themselves. Looking at American English, Preston (1999) found broad differences in social assessments on the basis of the internal regional accent of American speakers, including a finding that northern speakers are considered to be less friendly than southern speakers by students in Michigan.

Baumann (2017) collected pairwise likability ratings from more than 220 speakers and over 160 raters. This work found very limited acoustic correlations with rater preferences. Only measures related to the acoustic fidelity of the recording showed significant correlations, while prosodic qualities showed trends that did not reach statistical significance. However, the authors did find an interesting relationship between gender and likeability. Both male and female raters responded to male speech similarly. However, female speech was rated as much more likeable by female raters than by male raters.

As in the study of charisma, qualities of the *listener* do not receive as much research attention as qualities of the *speaker*. This is particularly true in the case of likeability. Social attractiveness necessarily involves two parties and is a subjective quality. We do not all want to be friends with the same people. The attitude and behaviors of the listener can impact the speaker and reveal the dynamics of establishing, maintaining, or undermining social attractiveness.

Schweitzer, Lewandowski, and Duran (2017) directly addressed this facet of likeability. This work examined dialogs between pairs of German female speakers who both rated their dialog partners following their conversation. This work treats likeability as social and participatory. By investigating only dialogs between two female participants, this study avoids the biasing on the part of speaker or listener based on gender. While it was not explicitly measured, there is an assumption in this work that the participants were all heterosexual, therefore, the potential for overlap between likeability (social attractiveness) and sexual attractiveness is diminished. It is worth mentioning that in work that investigates social and sexual attractiveness, the sexual preferences of the participants are particularly relevant. As such, it is necessary to collect or verify information about the sexual preferences of subject participants.

The experiment consisted of 46 two-party dialogs between 13 participants. Dialogs were collected in situations where the speakers could see each other, and where they were visually separated. Each dialog was spontaneous and unconstrained,

536 and lasted approximately 25 min. After the conversation both participants responded
537 to a questionnaire about how likeable, competent, friendly, and self-confident they
538 found their conversational partner.

539 The authors found limited confirmation of pitch and voice quality correlates to
540 likeability in this study. Specifically, they found no effect of absolute pitch or pitch
541 range. Neither were effects of shimmer, jitter, or HNR observed. However, they
542 did find a number of entrainment or “convergence” based effects. These relate to
543 how the acoustic–prosodic and lexical qualities of two (or more) speakers either
544 become more or less similar over the course of a dialog. The authors found that
545 lexical entrainment, when interlocutors use the same words, is a reliable predictor of
546 likeability. In multimodal conversations, where the participants could see each other,
547 they found convergence of peak F0 height made a speaker appear *less* likeable.

548 The Interspeech Paralinguistics Challenge is an annual shared task with results
549 presented at the Interspeech Conference each fall. The organizers distribute speech
550 data sets labeled for some paralinguistic quality which are partitioned into train,
551 development, and evaluations sets. Previous tasks have included classification of
552 emotion, sleepiness, and intoxication among many others. The 2012 challenge
553 included a task to classify the likeability of a speaker on the basis of a short utterance.
554 Sentences were drawn from the aGender corpus (Burkhardt, Eckert, Johannsen, &
555 Stegmann, (2010), and originally collected for the prediction of age and gender. The
556 longest utterance for each speaker was selected. This resulted in 800 speakers bal-
557 anced between male and female and divided into three age ranges (young: 15–24;
558 middle: 25–54; senior: 55–85). These were rated on a 7-point Likert scale of like-
559 ability by 32 participants (17M; 15F) aged 20–42 years. Ratings were adjusted based
560 on evaluator reliability and discretized into Likeable and Not-Likeable classes for
561 classification. The organizers of the challenge found no impact of the rater’s age
562 or gender on ratings, but the age and gender of the *speaker* did have a significant
563 impact. These challenges have served as a venue for the broader research community
564 to test the limits of automatic analysis of paralinguistics. In many situations, in part
565 because of the short time frame, and limited meta data available for the challenge
566 data sets, a good number of submissions associated with these challenges tend to be
567 applications of feature selection, e.g., Pohjalainen, Kadioglu, and Räsänen (2012),
568 Wu (2012) and classification approaches, e.g., Cummins, Epps, and Kua (2012), Lu
569 and Sha (2012), Brueckner and Schuller (2012), Sanchez, Lawson, Vergyri, and Bratt
570 (2012). Some of these are quite novel to these tasks yet include only limited analyses
571 of the underlying phenomena. One exception can be found when participants develop
572 novel acoustic features for analysis. This was undertaken by Buisman and Postma
573 (2012) in this likability challenge. They found that spectral information extracted
574 via log-gabor-filter-based features were able to predict likeability with higher accu-
575 racy than a much larger set of features included in the OpenSmile baseline (Eyben,
576 Wöllmer, & Schuller, 2010).

577 Additionally, Montacié and Caraty (2012) developed specific pitch and intona-
578 tion feature sets based on MOMEL (Hirst, 1987) and INTSINT (Louw and Barnard,
579 2004). MOMEL is a stylization technique which smooths out microprosody from
580 a pitch contour, while INTSINT discretizes the contour into “key ranges” describ-

581 ing the speaker’s pitch range, and “contextual labels” describing the relationship
 582 between the pitch at a given target to the previous target. A set of features based on
 583 the MOMEL and INTSINT processes were developed to help predict likability and
 584 also personality traits (another task of the 2012 Interspeech Paralinguistics Chal-
 585 lenge). While the specific correlations between likeability and these novel features
 586 are not presented, the use of intonational features was useful for the prediction of
 587 likeability where they were not useful for predicting personality traits. This suggests
 588 that these features may be particularly well suited to likeability, rather than being
 589 generally valuable features for paralinguistic analysis. There are conflicting results
 590 about correlations between pitch and likeability. These seem to suggest that either the
 591 specific formulation of intonational features is critical, or the relationship is nuanced
 592 and significantly influenced by other factors.

593 2.6 Romantic Attractiveness

594 Romantic attraction is a complicated phenomenon that involves the synthesis of a
 595 wide array of signals to determine romantic interest. The current understanding of
 596 this topic involves an interplay of influences too complicated to summarize here.
 597 Here we will focus only on the work that has investigated qualities of the voice that
 598 lead a listener to find a speaker romantically attractive, or not.

599 While romantic attractiveness is exceptionally subjective, research has been
 600 undertaken to identify voices that are typically found to be more (or less) attrac-
 601 tive. In this work, compared to much of the work surveyed elsewhere in the paper,
 602 characteristics of the listener are measured, and generally controlled for. However, a
 603 significant number of studies in this area conflate the influence of gender and sexual
 604 orientation in considering the qualities of the listener. Some studies investigate how
 605 males react to female voices or faces and others will study how females respond to
 606 male voices. In doing this, there is an assumption that all of the participants are, in
 607 fact, attracted romantically or sexually to members of the opposite sex. When these
 608 studies do not report the sexual orientation of the subjects, it stands to reason that
 609 the question was not asked of the participants. This is a significant methodological
 610 problem with this body of work. Through this section we will highlight whether a
 611 study has in fact reported the sexual orientation of the subjects or not, and suggest
 612 that future studies take this into consideration. We would also suggest that gender
 613 questions in recruitment for these studies be broadened to gain an understanding of
 614 how transgender, nonbinary, and intersex people assess attractiveness by the voice.
 615 None of the surveyed papers address these populations.

616 In an example of this, Collins and Missing (2003) investigated subject ratings of
 617 attractiveness of female voices, and female faces. To account for sexual preference,
 618 they used only male raters. However, they do not report whether all participants
 619 were heterosexual. In this work, they found strong agreement as to what was an
 620 attractive voice, and what was an attractive face, and moreover, attractive voices
 621 belonged to attractive faces. They found that voices of younger women are typically

622 higher pitched, as are voices of smaller women, while taller women demonstrate
 623 a narrower formant dispersion. The authors' findings suggest that both the visual
 624 and auditory signals are communicating complementary information regarding age
 625 and body shape. The finding that men find high-pitched women's voices attractive
 626 has been identified elsewhere as well, including by Feinberg, DeBruine, Jones, and
 627 Perrett (2008b).

628 On the other hand, Feinberg, Jones, Little, Burt, and Perrett (2005), Collins and
 629 Missing (2003), and Hodges-Simeon, Gaulin and Puts (2010) all found that women
 630 find men with lower pitched voices to be more attractive. Feinberg, DeBruine, Jones,
 631 and Little (2008a) found that both male and female subjects consistently rated the
 632 masculinity of male faces and voices and demonstrated preferences for more mascu-
 633 line voices. The claim here is that testosterone information is similarly communicated
 634 via the voice and the face. This supports a finding by Saxton et al. (2006) that men
 635 with attractive voices also have attractive faces. Interestingly, this result was found
 636 in adolescent and adult women, but not in female children. Of these, only Hodges-
 637 Simeon et al. (2010) reported the sexual orientation of the participants reported.

638 Many of these findings are predicated on the idea that attractiveness of a voice is
 639 being used as a proxy or a reinforcing signal for other physical characteristics. While
 640 there are plausible evolutionary justifications (cf. Puts, Doll, & Hill, 2014) for why
 641 some secondary sexual traits are attractive, the value of an attractive voice is less
 642 obvious. There is, however, some evidence that attractive voices are correlated with
 643 other physical traits that are themselves attractive. For example, Bruckert, Liénard,
 644 Lacroix, Kreutzer, and Leboucher (2006) found that male speech with low-frequency
 645 formants correlate with age, height, and weight. However, female listeners were only
 646 able to reliably estimate the age and weight of a male speaker based on enunciation
 647 of vowels. González (2006) found that the pitch of human speech reveals very little
 648 about body size when age and gender are controlled for. However, formant dispersion
 649 does carry this information. Despite the fact that it is a poor signal, listeners do rely
 650 on pitch information to estimate body size. Babel, King, McGuire, Miller, & Babel
 651 (2011) investigated the vocal correlates of attractiveness particularly as it relates to
 652 body size in the perception of opposite-sex voices by both male and female listeners.
 653 They found that the ratings of both genders were highly correlated, though males
 654 generally rated other males as less attractive than females did. They also found
 655 that attractive female voices had high second formants in high vowels, breathy voice
 656 quality, reduced pitch variance, and longer durations. However, attractive male voices
 657 had shorter durations (consistent with faster speaking rate), higher vowels, lower first
 658 formants overall, and higher second formant in /u/s. While this work was motivated
 659 by a search for body size correlates, the authors found a much more complicated
 660 relationship than expected.

661 In addition to pitch qualities, speaking rate also matters. Quené, Boomsma, and
 662 van Erning (2016) investigated the attractiveness of male voices by heterosexual
 663 female listeners as a function of both pitch and speaking rate. They found that faster
 664 and lower pitched speech was more attractive. However, tempo only matters if the
 665 pitch component is present. Fast but relatively high-pitched speech was not consis-
 666 tently rated as attractive.

667 In general, there are relatively few published findings about the relationship
 668 between voice quality and attractiveness. Babel et al. (2011) found breathy voice
 669 to be an indicator of attractiveness in female voices. Barkat-Defradas et al. (2015)
 670 found that male voices that are slightly rough (R1 on the GRBAS scale, a measure
 671 of dysphonia) are rated as the most attractive by women. The sexual orientation of
 672 subjects was not reported in either study.

673 Given these findings that there are vocal correlates to attractiveness, Fraccaro
 674 et al. (2013) investigated whether subjects could intentionally sound more or less
 675 attractive. They asked male and females to intentionally raise and lower the pitch
 676 of their voice. They found that when male speakers lowered their pitch and female
 677 speakers raised theirs, these manipulations did not necessarily lead to increased
 678 attractiveness. Additionally, when the male speakers raised their pitch and women
 679 lowered theirs, their attractiveness was lowered. This suggests that it is difficult to
 680 “fake” an attractive voice. Although we will return to the idea of intention when we
 681 discuss entrainment and communication of interest (i.e., flirting).

682 These trends, that lower pitched (and therefore more masculine) men are consid-
 683 ered more attractive, are not independent of other qualities of the subject. Valentová,
 684 Roberts, and Havlíček (2013) investigated ratings of attractiveness and masculin-
 685 ity of male voices and faces by homosexual men and heterosexual women. These
 686 authors also collected information about the relationship status and sexual restrictiv-
 687 ness. Homosexual male subjects also self-rated themselves on a masculine–feminine
 688 scale. (Heterosexual female subjects were not asked to perform this self-rating.) They
 689 found no consistent preference for masculine faces by either homosexual men or het-
 690 erosexual women. Moreover, a preference for masculine voices was only found in
 691 coupled heterosexual women and single homosexual men, While a preference for
 692 less masculine faces was observed in coupled homosexual men. Homosexual men
 693 who considered themselves to be more masculine tended to prefer more masculine
 694 voices, but more feminine faces. These findings highlight the complexity of iden-
 695 tifying romantically attractive voices. Perceptions of attractiveness are conditioned
 696 not only on gender, but also sexual preference, and the gender expression of both
 697 the listener and speaker, in addition to other subjective idiosyncrasies. While this
 698 (and other) work by Valentova et al. goes further than most in acknowledging and
 699 investigating these factors, there remains a wide range of unstudied questions and
 700 interactions in this space.

701 The studies that we have surveyed so far have studied the perceptions of listeners
 702 who are not also conversational participants. While there are, of course, situations
 703 where this occurs, listening to the radio, an audiobook, a lecture, or other presentation,
 704 romantic attraction is more commonly established in two-party conversations. Here
 705 attraction is both assessed and performed and the voice is used to both express
 706 attraction and promote attractiveness. While this is a more complicated process, a
 707 number of efforts have been made to understand how romantic attractiveness works
 708 in a conversational setting.

709 Leongómez et al. (2014) investigated this by examining how adult heterosex-
 710 ual participants spoke when addressing attractive and unattractive potential partners
 711 (opposite-sex conversational partners) and potential competitors (same-sex conversa-

712 tional partners). The scenario followed a design similar to video dating and was con-
713 ducted in both Czech and English. Subjects watched a stimulus video and recorded
714 a response video introducing themselves. In the case of opposite-sex stimuli, the
715 response video was to be played to the person who recorded the initial video. In the
716 case of same-sex stimuli, the response video would be played along with the stimulus
717 video to all opposite-sex subjects. Participants were instructed to explain whether
718 and why they would like to date the potential partner in opposite-sex stimuli, and to
719 explain why they should be chosen over the subject in same-sex stimuli. The stimuli
720 videos were rated for attractiveness by an independent set of raters and comprised the
721 three most and least attractive men and women drawn from a set of 40 participants
722 (20 male and 20 female). They found that male F0 varied most in speech toward
723 attractive women, but female F0 varied more in response to attractive competitors.
724 Also, male minimum pitch was lowered when addressing attractive women. In a
725 follow-up study, the experimenters also found that speech directed *toward* attractive
726 participants was itself considered to be more attractive.

727 Dating scenarios are especially useful for investigating romantic attractiveness.
728 The previous study used a video-dating paradigm. Another body of work looks at
729 speed dating. In speed dating, participants engage in short (approximately 5 min)
730 face-to-face conversations with potential partners and then fill out a questionnaire
731 about their partner including an opportunity to indicate whether they would like to
732 see the person again. In a speed-dating session, each participant may repeat this
733 experience 10 or more times. In this work, all participants have self-selected to be
734 interested in opposite-sex romantic partners. McFarland, Jurafsky, and Rawlings
735 (2013) recorded speed-dating participants, and analyzed their speech, the content
736 of their conversations, and their responses toward each other. While their analyses
737 are quite comprehensive, we focus on vocal qualities here. Both genders described
738 increased “connection” when they expressed excitement toward their partner. Male
739 participants expressed this excitement through laughter, varied loudness, and reduced
740 pitch variance. Female participants, however, raised and varied their pitch, spoke
741 softer, varied loudness, and took shorter turns. They also found that women felt they
742 “clicked” more with male partners who interrupted them. While this is somewhat
743 unexpected—conventional understanding of interruption is that it is rude—closer
744 inspection of these interruptions suggest that the overlapping speech that leads to a
745 sense of connection was used to demonstrate understanding, through backchanneling
746 and agreement. This is not to say that all interruption is “constructive” or used
747 to demonstrate connection. Interruption can also be rude or dismissive. However,
748 distinguishing the pragmatic effect of interruption can be challenging especially via a
749 reliable automated technique. The study also found that entrainment, the convergence
750 or divergence of vocal qualities between partners, is associated with attractiveness.
751 Specifically, they found that partners who described a connection mimicked each
752 others rate of speech, use of function words, and use of laughter.

753 Michalsky and Schoormann (2017) also looked at the role of entrainment in
754 attractiveness, again investigated in a speed-dating setting. They focused on measures
755 of pitch convergence. They found that speakers become more similar over time in
756 both register and range, but that this degree of convergence was influenced by how

757 attractive subjects found their conversational partner. In a later study, Michalsky and
 758 Schoormann (2018) found that listener reactions of attraction were sensitive to pitch
 759 height relative to the speaker’s natural pitch range rather than an absolute measure.
 760 That is, attractive male voices are not simply low, but they are low in the speaker’s
 761 pitch range. Conversely, female voices that are considered attractive are high in the
 762 woman’s pitch range, not just naturally high pitched.

763 Examining vocal qualities in conversations forces experimenters to attempt to
 764 disentangle those aspects that are perceptible (being attractive) from those which
 765 are performative (expressing attractiveness). Puts et al. (2011) found that increased
 766 pitch and increased formant dispersion in women is found to be attractive and to be
 767 perceived as flirtations by other women. Jurafsky, Ranganath, and McFarland (2009)
 768 found that women who are labeled as “flirting” by men on speed dates spoke faster
 769 and with higher pitch and laugh more. These prosodic qualities overlap completely
 770 men who are labeled as “flirting,” but men also speak more quietly. When women
 771 labeled their male partner as flirting (whether or not they actually were), they laughed
 772 more and lowered their intensity. But when men labeled their female partner as
 773 flirting, they raised their pitch. These analyses were developed and systematized in
 774 Ranganath, R., Jurafsky, and McFarland Ranganath et al. (2009). This work attempted
 775 to automatically detect flirting in speed-date speech. The most interesting qualities of
 776 this work come from identifying which features are used in the perception of flirting
 777 but are *not* used in the expression of flirting. For example, men are perceived to
 778 flirt when they overlap less and use fewer appreciations, but this is not significant in
 779 men who indicated that they were flirting. Similar faster speaking rate has a stronger
 780 influence on the perception of flirting than the performance of flirting. For women,
 781 laughing, taking fewer longer turns, and asking repair questions are strong indicators
 782 of a woman intending to flirt, but are not perceived by their partners as flirtatious.

783 2.7 Conclusions

784 In this chapter, we have surveyed the literature on four types of attraction and trust
 785 as it relates to a person’s speech. We have used the term “charismatic” to describe a
 786 speaker who is politically attractive. In general, charismatic speakers are dynamic,
 787 passionate, and enthusiastic. These assessments are consistent across a range of
 788 listeners. American, Irish, Swedish, and Palestinian subjects have come to similar
 789 conclusions. However, the vocal realizations of this passion and dynamism vary by
 790 speaker. In general, charismatic political speakers vary their use of pitch, intensity,
 791 and speaking rate. Some research suggests that clear comprehensible pronunciation
 792 with relatively few disfluencies is also important.

793 Considering attraction in the business domain, business leaders considered charis-
 794 matic often demonstrate the same qualities as political leaders. They pronounce words
 795 clearly, are rarely disfluent, and demonstrate more varied speech.

796 In the cases of business and political attractiveness, male and female subjects
 797 tended to assess speakers similarly. However, across research in both of these

798 domains, far greater attention has been given to charisma in male speakers. One
799 area that needs further study is what qualities of the female voices lead listeners to
800 find them to be charismatic.

801 Regarding trust in a speaker, evidence suggests that listeners trust people who
802 speak quickly. Male voices spoken with high pitch led to mistrust and female voices
803 with more breathiness were more trusted. It is worth noting that these qualities are
804 strongly linked to measures of political or business-based charisma.

805 Considering likeability, listeners tend to prefer voices that clearly enunciate—
806 they have a higher pronunciation precision, but also a higher speaking rate. Other
807 prosodic properties have less of an impact on assessment.

808 Romantic attraction as it relates to the voice has received quite a bit of research
809 attention. The broad and most consistent finding here suggests that men with low
810 voices and greater formant dispersion are attractive as are women with higher voices
811 and more breathiness. The dynamics of romantic attraction in two-party conversa-
812 tions create an interesting area for research. The voice is involved both as an object
813 of attraction and also a mechanism to demonstrate attraction. When heterosexual
814 male speakers flirt, they lower their pitch, while flirting heterosexual women raise
815 their pitch. Also, when participants are mutually attracted they tend to entrain on
816 a number of prosodic dimensions including speaking rate, the use of laughter, and
817 intensity.

818 One important caveat in the assessment of romantic attraction is that in many
819 cases the gender of a listener is assumed to be a proxy for sexual preference. This is
820 a methodological problem that can be found in a number of the reviewed studies.

821 While we have presented these types of attraction as related to each other, they have
822 their own idiosyncrasies both in terms of how they operate socially and in how they are
823 communicated via the voice. These forms of attraction may interact in unpredictable
824 ways. The current research does not consider ways in which the qualities that make a
825 voice attractive in one context may make it more or less attractive in another context
826 or for a distinct social assessment. For example, are voices that are socially likeable
827 more or less like voices that are attractive in business leaders?

828 In all, our understanding of what makes a voice attractive is fairly limited. There
829 are a number of broad findings, but none of these in isolation is sufficient to either reli-
830 ably predict attractiveness, or to provide overwhelmingly useful feedback to speakers.
831 This ambiguity of findings can be found in individual studies but is even more clear
832 through this survey. It is possible that it results from the fact that there is more inter-
833 listener variability in both what is attractive and what signals are being relied on to
834 make this decision.

835 While there is clearly more work to be done on this subject, major areas for further
836 study include (1) investigation of business and political charisma in female speakers,
837 (2) likeability and romantic attraction in nonheterosexual participants, and (3) more
838 thorough consideration of qualities of the listener in identifying not just what is
839 attractive in the speaker's voice, but what particular types of listeners find attractive.

840 **References**

- 841 Babel, M., King, J., McGuire, G., Miller, T., & Babel, M. (2011). Acoustic determiners of vocal
 842 attractiveness go well beyond apparent talker size. In *Laboratory Report: University of British*
 843 *Columbia and University of California, Santa Cruz*.
- 844 Barkat-Defradas, M., Fauth, C., Didirkova, I., de La Bretèque, B. A., Hirsch, F., Dodane, C.,
 845 & Sauvage, J. (2015). "Dysphonia is beautiful: A perceptual and acoustic analysis of vocal
 846 roughness. In *18th International Congress of Phonetic Sciences (ICPhS-18)*.
- 847 Baumann, T. (2017). Large-scale speaker ranking from crowdsourced pairwise listener ratings. In
 848 *Proceedings of INTERSPEECH*.
- 849 Beckman, M., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and the
 850 evolution of the ToBI framework. In (pp. 9–54).
- 851 Biadys, F., Rosenberg, A., Carlson, R., Hirschberg, J., & Strangert, E. (2008). A cross-cultural com-
 852 parison of American, Palestinian, and Swedish perception of charismatic speech. In *Proceedings*
 853 *of Speech Prosody* (Vol. 37).
- 854 Boss, G. P. (1976). Essential attributes of the concept of charisma. *Southern Journal of Communi-*
 855 *cation*, 41(3), 300–313.
- 856 Bruckert, L., Liénard, J.-S., Lacroix, A., Kreutzer, M., & Leboucher, G. (2006). Women use voice
 857 parameters to assess men's characteristics. *Proceedings of the Royal Society of London B: Bio-*
 858 *logical Sciences*, 273(1582), 83–89.
- 859 Brueckner, R., & Schuller, B. (2012). Likability classification—a not so deep neural network approach.
 860 In *Proceedings of INTERSPEECH*.
- 861 Buisman, H., & Postma, E. (2012). The log-Gabor method: Speech classification using spectrogram
 862 image analysis. In *Proceedings of INTERSPEECH*.
- 863 Burkhardt, F., Eckert, M., Johannsen, W., & Stegmann, J. (2010). A database of age and gender
 864 annotated telephone speech. In *Proceedings of LREC*.
- 865 Collins, S. A., & Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal*
 866 *Behaviour*, 65(5), 997–1004.
- 867 Cullen, A., & Harte, N. (2018). A longitudinal database of Irish political speech with annotations
 868 of speaker ability. *Language Resources and Evaluation*, 52(2), 401–432.
- 869 Cullen, A., Hines, A., & Harte, N. (2014). Building a database of political speech: Does culture
 870 matter in charisma annotations? In *Proceedings of the 4th International Workshop on Audio/Visual*
 871 *Emotion Challenge* (pp. 27–31). ACM.
- 872 Cummins, N., Epps, J., & Kua, J. M. K. (2012). A comparison of classification paradigms for
 873 speaker likeability determination. In *Proceedings of INTERSPEECH*.
- 874 Dowis, R. (2000). *The lost art of the great speech: how to write it, how to deliver it*. Amacom Books.
- 875 Eyben, F., Wöllmer, M., & Schuller, B. (2010). OpenSmile: The Munich versatile and fast open-
 876 source audio feature extractor. In *Proceedings of the 18th ACM International Conference on*
 877 *Multimedia* (pp. 1459–1462). ACM.
- 878 Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M., & Perrett, D. I. (2005). Manipulations of
 879 fundamental and formant frequencies influence the attractiveness of human male voices. *Animal*
 880 *Behaviour*, 69(3), 561–568.
- 881 Feinberg, D. R., DeBruine, L. M., Jones, B. C., & Little, A. C. (2008a). Correlated preferences for
 882 men's facial and vocal masculinity. *Evolution and Human Behavior*, 29(4), 233–241.
- 883 Feinberg, D. R., DeBruine, L. M., Jones, B. C., & Perrett, D. I. (2008b). The role of femininity and
 884 averageness of voice pitch in aesthetic judgments of women's voices. *Perception*, 37(4), 615–623.
- 885 Fraccaro, P. J., O'Connor, J. J., Re, D. E., Jones, B. C., DeBruine, L. M., & Feinberg, D. R. (2013).
 886 Faking it: Deliberately altered voice pitch and vocal attractiveness. *Animal Behaviour*, 85(1),
 887 127–136.
- 888 Fujisaki, H., & Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative
 889 sentences of Japanese. *Journal of the Acoustical Society of Japan (E)*, 5(4), 233–242.
- 890 González, J. (2006). Research in acoustics of human speech sounds: Correlates and perception of
 891 speaker body size. *Recent Research Developments in Applied Physics*, 9, 1–15.

- 892 Hirst, D. (1987). La description linguistique des systèmes prosodiques: une approche cognitive.
893 Ph.D. thesis. Thèse de doctorat d'Etat, Université de Provence.
- 894 Hodges-Simeon, C. R., Gaulin, S. J., & Puts, D. A. (2010). Different vocal parameters predict
895 perceptions of dominance and attractiveness. *Human Nature*, 21(4), 406–427.
- 896 Jurafsky, D., Ranganath, R., & McFarland, D. (2009). Extracting social meaning: Identifying inter-
897 actional style in spoken conversation. In *Proceedings of HLT/NAACL* (pp. 638–646). Association
898 for Computational Linguistics.
- 899 Leongómez, J. D., Binter, J., Kubicová, L., Stolařová, P., Klapilová, K., Havlíček, J., et al. (2014).
900 Vocal modulation during courtship increases perceptivity even in naive listeners. *Evolution and
901 Human Behavior*, 35(6), 489–496.
- 902 Levitan, S. I., Maredia, A., & Hirschberg, J. (2018). Acoustic-prosodic indicators of deception and
903 trust in interview dialogues. In *Proceedings of INTERSPEECH* (pp. 416–420).
- 904 Louw, J., & Barnard, E. (2004). Automatic intonation modeling with INTSINT. In *Proceedings of
905 the Pattern Recognition Association of South Africa* (pp. 107–111).
- 906 Lu, D., & Sha, F. (2012). Predicting likability of speakers with Gaussian processes. In *Proceedings
907 of INTERSPEECH*.
- 908 Marcus, J. T. (1961). Transcendence and charisma. *The Western Political Quarterly*, 14(1), 236–241.
- 909 McFarland, D. A., Jurafsky, D., & Rawlings, C. (2013). Making the connection: Social bonding in
910 courtship situations. *American Journal of Sociology*, 118(6), 1596–1649.
- 911 McGloin, R., & Denes, A. (2018). Too hot to trust: Examining the relationship between attractive-
912 ness, trustworthiness, and desire to date in online dating. *New Media & Society*, 20(3), 919–936.
- 913 Michalsky, J. & Schoormann, H. (2017). Pitch convergence as an effect of perceived attractiveness
914 and likability. In *Proceedings of INTERSPEECH* (pp. 2253–2256).
- 915 Michalsky, J., & Schoormann, H. (2018). Opposites attract! Pitch divergence at turn breaks as cause
916 and effect of perceived attractiveness. In *Proceedings of Speech Prosody* (pp. 265–268).
- 917 Mixdorff, H., Niebuhr, O., & Hönemann, A. (2018). Model-based prosodic analysis of charismatic
918 speech. In *Proceedings of Speech Prosody* (pp. 1–5).
- 919 Montaciè, C., & Caraty, M. -J. (2012). Pitch and intonation contribution to speakers' traits classi-
920 fication. In *Proceedings of INTERSPEECH*.
- 921 Niebuhr, O.; Brem, A., Novák-Tót, E., & Voße, J. (2016a). Charisma in business speeches-A contra-
922 stative acoustic-prosodic analysis of Steve Jobs and Mark Zuckerberg. In *Proceedings of Speech
923 Prosody*.
- 924 Niebuhr, O., Voße, J., & Brem, A. (2016b). What makes a charismatic speaker? A computer-based
925 acoustic-prosodic analysis of Steve Jobs tone of voice. *Computers in Human Behavior*, 64, 366–
926 382.
- 927 Niebuhr, O., Brem, A., & Tegtmeier, S. (2017). Advancing research and practice in entrepreneurship
928 through speech analysis—From descriptive rhetorical terms to phonetically informed acoustic
929 charisma profiles. *Journal of Speech Sciences*, 6(1), 3–26.
- 930 Niebuhr, O., Thumm, J., & Michalsky, J. (2018a). Shapes and timing in charismatic speech-Evidence
931 from sounds and melodies. In *Proceedings of Speech Prosody*.
- 932 Niebuhr, O.; Skarnitzl, R., & Tylečková, L. (2018b). The acoustic fingerprint of a charismatic
933 voice-Initial evidence from correlations between long-term spectral features and listener ratings.
934 In *Proceedings of Speech Prosody* (pp. 359–363).
- 935 Novák-Tót, E., Niebuhr, O., & Chen, A. (2017). A gender bias in the acoustic-melodic features of
936 charismatic speech? In *Proceedings of INTERSPEECH* (pp. 2248–2252).
- 937 Pejčić, A. (2014). Intonational characteristics of persuasiveness in Serbian and English Political
938 debates. *Nouveaux cahiers de linguistique française*, 31, 141–151.
- 939 Pèrez, C. P. (2016). A study of the phono-styles used by two different Spanish-speaking political
940 leaders: Hugo Chávez and José L. R. Zapatero. In *Proceedings of Speech Prosody* (pp. 410–414).
- 941 Pohjalainen, J., Kadioglu, S., & Räsänen, O. (2012). Feature selection for speaker traits. In *Pro-
942 ceedings of INTERSPEECH*.
- 943 Preston, D. R. (1999). A language attitude analysis of regional US Speech: Is northern US English
944 not friendly enough? *Cuadernos de filología inglesa*, 8, 129–146.

- 945 Puts, D. A., Barndt, J. L., Welling, L. L., Dawood, K., & Burriss, R. P. (2011). Intrasexual compe-
 946 tition among women: Vocal femininity affects perceptions of attractiveness and flirtatiousness.
 947 *Personality and Individual Differences*, 50(1), 111–115.
- 948 Puts, D. A., Doll, L. M., & Hill, A. K. (2014). Sexual selection on human voices. In *Evolutionary*
 949 *perspectives on human sexual psychology and behavior* (pp. 69–86). Springer.
- 950 Quenè, H., Boomsma, G., & van Erming, R. (2016). Attractiveness of male speakers: Effects of
 951 voice pitch and of speech tempo. In *Proceedings of Speech Prosody* (Vol. 8, pp. 1086–1089).
- 952 Ranganath, R., Jurafsky, D., & McFarland, D. (2009). It's not you, it's me: Detecting flirting and its
 953 misperception in speed-dates. In *Proceedings of EMNLP* (pp. 334–342). Association for Com-
 954 putational Linguistics.
- 955 Rosenberg, A., & Hirschberg, J. (2005). Acoustic/prosodic and lexical correlates of charismatic
 956 speech. In *EUROSPEECH*.
- 957 Rosenberg, A., & Hirschberg, J. (2009). Charisma perception from text and speech. *Speech Com-*
 958 *munication*, 51(7), 640–655.
- 959 Ryant, N., & Liberman, M. (2016). Automatic analysis of phonetic speech style dimensions. In
 960 *Proceedings of INTERSPEECH* (pp. 77–81).
- 961 Sanchez, M. H., Lawson, A., Vergyri, D., & Bratt, H. (2012). Multi-system fusion of extended
 962 context prosodic and cepstral features for paralinguistic speaker trait classification. In *Proceedings*
 963 *of INTERSPEECH*.
- 964 Saxton, T. K., Caryl, P. G., & Craig Roberts, S. (2006). Vocal and facial attractiveness judgments of
 965 children, adolescents and adults: The ontogeny of mate choice. *Ethology*, 112(12), 1179–1185.
- 966 Schweitzer, A., Lewandowski, N., & Duran, D. (2017). Social attractiveness in dialogs. In *Proceed-*
 967 *ings of Interspeech* (pp. 2243–2247).
- 968 Signorello, R., D'Errico, F., Poggi, I., & Demolin, D. (2012). How charisma is perceived from
 969 speech: A multidimensional approach. In *2012 ASE/IEEE International Conference on Social*
 970 *Computing*.
- 971 Sofer, C., Dotsch, R., Wigboldus, D. H., & Todorov, A. (2015). What is typical is good: The influence
 972 of face typicality on perceived trustworthiness. *Psychological Science*, 26(1), 39–47.
- 973 Strangert, E., & Gustafson, J. (2008). What makes a good speaker? Subject ratings, acoustic mea-
 974 surements and perceptual evaluations. In *Proceedings of INTERSPEECH*.
- 975 Tavernier, J. (2007). Attitudes towards native and non-native accents of English. Ph.D. thesis. Ghent
 976 University.
- 977 Valentová, J., Roberts, S. C., & Havlíček, J. (2013). Preferences for facial and vocal masculinity
 978 in homosexual men: The role of relationship status, sexual restrictiveness, and self-perceived
 979 masculinity. *Perception*, 42(2), 187–197.
- 980 Weber, M. (1947). *The theory of social and economic organization*. New York: Oxford University
 981 Press.
- 982 Weiss, B., & Burkhardt, F. (2010). Voice attributes affecting likability perception. In *Proceedings*
 983 *of INTERSPEECH*.
- 984 Weiss, B., & Burkhardt, F. (2012). Is "not bad" good enough? Aspects of unknown voices' likability.
 985 In *Proceedings of INTERSPEECH*.
- 986 Weninger, F., Krajewski, J., Batliner, A., & Schuller, B. (2012). The voice of leadership: Models
 987 and performances of automatic analysis in online speeches. *IEEE Transactions on Affective*
 988 *Computing*.
- 989 Wilson, R. K., & Eckel, C. C. (2006). Judging a book by its cover: Beauty and expectations in the
 990 trust game. *Political Research Quarterly*, 59(2), 189–202.
- 991 Wu, D. (2012). Genetic algorithm based feature selection for speaker trait classification. In *Pro-*
 992 *ceedings of INTERSPEECH*.

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Voice Attractiveness
Series Title	
Chapter Title	The Vocal Attractiveness of Charismatic Leaders
Copyright Year	2020
Copyright HolderName	Springer Nature Singapore Pte Ltd.
Corresponding Author	Family Name Signorello Particle Given Name Rosario Prefix Suffix Role Division Organization Laboratoire de Phonétique et Phonologie, CNRS & Sorbonne Nouvelle Address Paris, France Email rosario.signorello@gmail.com
Abstract	Social attractiveness in human leaders is defined as charisma, the set of leadership characteristics such as vision, emotions, and dominance used by leaders to share beliefs, persuade listeners, and achieve goals. Charisma is expressed through voice quality manipulations reflecting physiologically-based qualities and culturally-acquired habits to display leadership. These manipulations are adapted by the speakers to the social environment where they intend to be perceived as charismatic. Charisma in political speech is observed here to unveil the biological abilities versus the culturally-mediated strategies in leaders' speech according to different social contexts in which political communication takes place. Manipulations of vocal pitch, loudness, and phonation types are shown to cause both cross-cultural and culture-specific social attractiveness and consequently, are key factors for charisma effectiveness. Charismatic voice is then intentionally and unintentionally controlled by the human leaders to carry the perlocutionary salience of persuasive speech and influence listeners' choice of leadership.
Keywords	Vocal charisma - Political speech - Attractiveness of leadership - Biological abilities in vocal persuasion - Cultural descriptors of charisma - Perceived charisma from speech

Chapter 3

The Vocal Attractiveness of Charismatic Leaders



Rosario Signorello

Abstract Social attractiveness in human leaders is defined as charisma, the set of leadership characteristics such as vision, emotions, and dominance used by leaders to share beliefs, persuade listeners, and achieve goals. Charisma is expressed through voice quality manipulations reflecting physiologically-based qualities and culturally-acquired habits to display leadership. These manipulations are adapted by the speakers to the social environment where they intend to be perceived as charismatic. Charisma in political speech is observed here to unveil the biological abilities versus the culturally-mediated strategies in leaders' speech according to different social contexts in which political communication takes place. Manipulations of vocal pitch, loudness, and phonation types are shown to cause both cross-cultural and culture-specific social attractiveness and consequently, are key factors for charisma effectiveness. Charismatic voice is then intentionally and unintentionally controlled by the human leaders to carry the perlocutionary salience of persuasive speech and influence listeners' choice of leadership.

Keywords Vocal charisma · Political speech · Attractiveness of leadership · Biological abilities in vocal persuasion · Cultural descriptors of charisma · Perceived charisma from speech

3.1 Introduction

3.1.1 *Charisma Defined as the Social Attractiveness of Group Leaders*

In modern literature, the term “charisma” was first popularized by sociologist Max Weber (1920). According to Weber, “charismatic” leaders generally emerge in times of great crisis for a nation, responding to the necessity of strong leadership to over-

R. Signorello (✉)

Laboratoire de Phonétique et Phonologie, CNRS & Sorbonne Nouvelle, Paris, France
e-mail: rosario.signorello@gmail.com

© Springer Nature Singapore Pte Ltd. 2020

B. Weiss et al. (eds.), *Voice Attractiveness*, Prosody, Phonology and Phonetics,
https://doi.org/10.1007/978-981-15-6627-1_3

41

24 come the crisis. This author defines charisma as an “extraordinary quality” of a
 25 person who is believed to be endowed with superhuman properties, in such a way as
 26 to induce people to acknowledge him as a leader, to the point of making a cult of him.
 27 Weber calls this quality “charisma” (from Greek *charis*, grace), thus considering it a
 28 grace, a divine gift that only some enlightened people may possess. Weber does not
 29 describe this gift at length, and even considers it beyond human comprehension; yet,
 30 the very notion of charisma has been alternatively redefined and challenged.

31 Some first sketches of charisma may be retrieved from ancient philosophy.
 32 According to Heraclitus, only a few individuals are endowed with particular physical
 33 and mental skills and virtues, that include, in accordance with Socrates, fast learning
 34 capacities, memory, open mind, and vision. These virtues are innate, according to
 35 Plato, and make a chief the object of trust, faith, and veneration by other people, which
 36 results in the cult of the leader (Cavalli, 1995). Such idea of the charismatic leader
 37 was personified in the great dictators of the twentieth century: Hitler, Mussolini, and
 38 Stalin.

39 Previously, research on charisma was mainly conducted in social psychology
 40 within the general framework of leadership studies. Some authors consider leadership
 41 as an internal trait of individuals (House & Howell 1992). For example, transforma-
 42 tional leaders, which Burns (1978) and Bass (1985) consider to be charismatic, show
 43 high values in four of the Big Five factors: extraversion, openness, agreeableness,
 44 and conscientiousness (Bono & Judge 2004).

45 An opposing view—the contingency perspective, which also includes the contex-
 46 tual approach, contends that leadership and charisma are strongly determined by the
 47 context: contextual factors trigger or inhibit particular leadership behaviors, and lead-
 48 ership is interactively constructed by the relationship between leader and followers
 49 (Haslam et al., 2011). This contextualist view further develops into the transactional
 50 leadership perspective, in which the strength and effectiveness of leadership is deter-
 51 mined by a cost-benefit computation, where followers agree to comply with the
 52 leader’s will to the extent they feel this is functional to their goals. Their behavior
 53 is stimulated by rewards and punishments more than trust and identification. This
 54 is not the case, however, for transformational leadership, which, introduced by the
 55 so-called neo-charismatic school, views a true leader as an authentically charismatic
 56 person (Lowe et al., 1996), endowed with vision and capacity for inspiring followers,
 57 who works in their interest and aims at their growth (Burns, 1978; Bass, 1985). Neo-
 58 charismatic scholars stress the ethical impact of transformational leadership, and
 59 warn of the “dark side” of charisma and the inauthentic or pseudo-transformational
 60 leaders, who with self-serving aims act in bad faith, consciously or unconsciously.
 61 Actually, the charismatic/transformational view integrates sociological and psycho-
 62 logical aspects since it sees charisma as a “social process” in which the perception
 63 of followers becomes a very central aspect (Shamir, 2000).

64 The discussion among these diverse perspectives, based on personality or context,
 65 transaction or transformation, makes the definition of a charismatic leader and the
 66 singling out of charismatic attributes particularly complex. In fact, charisma is a
 67 multidimensional construct: it is certainly affected (and constructed) by the values,
 68 needs, motivations, and discourses of potential followers, but it also, indubitably

69 depends on the leader's skills, choices, and characteristics. External displays are the
 70 perceivable expression of the internal features, and we can distinguish two kinds:
 71 one which we call the "charisma of the body" and the other, "charisma of the mind"
 72 (Signorello, 2014). Actually, the external features may stem either from the mind or
 73 from the body of the leader. Aspects of the charisma of the mind, such as creative and
 74 charming ideas or feelings, are displayed by a person's words or actions, while the
 75 charisma of the body is displayed by specific aspects of their visual and/or acoustic
 76 appearance, determined by their body's multimodal physical traits and behaviors
 77 (Shamir, Zakay, Breinin, & Popper, 1993; Bull, 1986; Atkinson, 1984; Rosenberg &
 78 Hirschberg 2009).

79 The athletic and proud gait of Barack Obama is a way of moving that conveys
 80 dignity. But, take Mahatma Gandhi, who was a short, thin shy man, without a loud
 81 voice, and who even sometimes stuttered: the features of his charisma did not emanate
 82 from his voice or gait, but from the strength of his message, and what revolutionary
 83 ideas came from his words and his political action. The first example is a case of the
 84 charisma of the body, while the latter is an example of the charisma of the mind: the
 85 meaning of a discourse by Gandhi (Bligh & Robinson, 2010). It is through words that
 86 his charismatic qualities shine forth. These two forms of expression of charisma—
 87 body and mind—may sometimes appear in combination, for example, Barack Obama
 88 may be seen as charismatic both for the concepts he proposes and the way he exposes
 89 them: he has charisma both of the body and of the mind (Bono & Judge 2004).

90 In sum, charismatic persons may have different kinds of charisma which depend
 91 on the type of internal charismatic features they possess, the external features that
 92 express them, and on their combinations. The aim of the present work is then to
 93 highlight the multidimensionality of charisma, and to explore in detail a specific
 94 display of political leaders' attractiveness: their voice. The hypothesis of this study is
 95 that the charisma of a person can be disentangled into a set of "charismatic features",
 96 and that in different persons, particular combinations of these features cluster into
 97 peculiar kinds of charisma. So what are the internal features of charisma, and how
 98 can we find them out?

99 **3.1.2 Charisma and Voice Behavior: The Charismatic Voice**

100 Group leaders use their voices to communicate their charisma, the set of leadership
 101 characteristics, such as vision, emotions, and dominance used by leaders to share
 102 beliefs, persuade, and achieve goals. Voice quality reflects leaders' physiologically-
 103 based vocal characteristics and culturally-acquired habits and strategies used to
 104 shape those characteristics qualitatively. Political speech is studied in order to unveil
 105 the biological abilities versus the culturally-mediated strategies of group leaders'
 106 charismatic voices. Through voice acoustic analyses and perceptual studies, a cross-
 107 culturally similar use of vocal pitch, loudness levels, and ranges in political speech
 108 and a culture-specific perceptual effect of overall vocal characteristics like phona-
 109 tion types, prosodic factors, and temporal characteristics were found. Charismatic

110 voices reflect individuals' (a) biological needs to have easy access to resources and
 111 (b) cultural needs to show skills that reflect high social status and power.

112 Voice quality results from speakers' biologically-derived differences in vocal
 113 apparatus combined with learned linguistic and cultural habits used to convey their
 114 personal identity (Garvin & Ladefoged, 1963; Kreiman & Sidtis, 2011). Voice quality
 115 conveys individuals' physical (e.g., size, Ohala, 1994; Pisanski et al., 2014, attrac-
 116 tiveness, Zuckerman & Driver, 1989; Collins, 2000), psychological (e.g., personality
 117 traits, Scherer, 1972; emotional status, Patel, Scherer, Björkner, & Sundberg, 2011)
 118 and social characteristics (e.g., leadership; Surawski & Ossoff, 2006; Tigue et al.
 119 2012; Klofstad, Anderson, & Peters, 2012; dominance, Ohala, 1984). These studies
 120 raise the question about whether particular features characterizing political speak-
 121 ers' voices are biologically versus culturally determined, and which type of feature
 122 is primary in distinguishing individuals chosen as group leaders from non-leaders.

123 Besides theoretical discussions on the nature of charisma, some studies investi-
 124 gated how charisma is perceived from voice. Tackling the relationship between the
 125 acoustic-prosodic characteristics of a political leader's speech and the perception of
 126 his/her charisma, Touati (1993) investigated the prosodic features of rhetoric utter-
 127 ances in French political speech in pre and post-elections discourses. Strangert and
 128 Gustafson (2008) examined the relationship between prosodic features and the per-
 129 ception of a speaker as a "good communicator", while Rosenberg and Hirschberg
 130 (2009), studied the correlation between acoustic, prosodic, and lexico-syntactic char-
 131 acteristics of political speech and the perception of charisma.

132 The overview above, introduced our conceptual definition of charisma focused
 133 on its psychological multidimensionality that affects social attractiveness, as well as
 134 a few theoretical insights, on the use of voice and speech as nonverbal behaviors to
 135 convey vocal attractiveness in political speech. This chapter reports investigations on
 136 the perceptual features that characterize vocal attractiveness in charismatic political
 137 discourse. This work highlights the features of charisma conveyed by the speakers
 138 and its social attractiveness on listeners speaking several languages. In the following
 139 sections, I first present a tool developed to measure the differences between vocal
 140 qualities of speaking individual political leaders. I later introduce studies that aimed
 141 to distinguish various kinds of charisma while singling out the features of voice that
 142 are responsible for their perception.

143 3.2 Charismatic Voices

144 3.2.1 Cultural- and Language-Based Descriptors of 145 Charisma

146 In contemporary literature about the perception of charisma from voice, scholars ask
 147 participants to rate voices in terms of adjectives that in previous studies had been
 148 connected to charisma (e.g., Rosenberg & Hirschberg, 2009). In our research, stud-

ies testing how people describe the charisma of group leaders in different languages and cultures were carried out in order to make a scale for the rating of charisma (Signorello et al., 2012a, 2012b). Through an empirical and non-biased approach, positive and negative traits of charisma in several languages (American English, French, Italian, and Brazilian Portuguese) were collected to develop the “Multi-dimensional Adjective-based Scale of others’ Charisma Perception” (MASCharP) (Signorello, 2014), a psychometric tool to be used in research on the perception of charismatic traits from individuals’ perceivable behaviors, such as voice. This approach entailed three experimental phases.

The first phase involved the collection of lexical and semantic descriptions of charismatic traits communicated through an individual’s perceivable behaviors from subjects of the languages being studied. This part entailed the gathering of adjectives that describe charismatic, as well as noncharismatic prototypes of leadership. It is fundamental to understand that the language in question is inseparable from its culture. These two factors act as filters in the attribution of an individual’s traits.

The second phase involved dimensions of theoretical classification of the adjectives gathered. As in Di Blas and Forzi (1998), the adjectives were selected by their frequency of usage. Only the most frequently used terms that are representative and descriptive of charismatic traits in the participants’ language were retained. In the first stage of data sorting, adjectives with a frequency higher than 1 were retained, indicating a cognitive commonality between at least two individuals who agree on a semantic-representational connection that designates the adjective as a trait of charisma. The adjectives used most frequently to describe charisma were then categorized in dimensions that were deduced from aspects of the persuasive process illustrated in the Sect. 3.2 of this chapter. The data were then organized according to semantic closeness, as in the cases of Saucier (2009) and Di Blas and Forzi (1998), corresponding to the dimensions of Poggi’s theory of persuasion (Poggi 2005). An example of the definitive selection of adjectives and dimensional classification constitutes the MASCharP as represented in Table 3.1 (American English).

The third phase involved the creation of a psychometric tool to perform the perceptual tests and measure the perception of charisma from voice. Each adjective from MASCharP could be evaluated through a Likert scale (Likert, 1932). An interface based on the server-side software Limesurvey® (The LimeSurvey project team, (2011)) was developed to collect the data. This software is written in PHP and uses a MySQL database to store data. The interface features the combination of the MASCharP with the 7-point Likert scale. The use of this tool has already been validated in several studies to measure the traits and types of charismatic leadership conveyed by voice (Signorello et al. 2012a, 2012b, 2014b D’Errico et al., 2012, 2013).

Table 3.1 Positive and negative interpersonal traits of perceived other's charisma in American English. Classification according to Signorello (2014)

Positive Charisma Traits	Negative Charisma Traits
Caring, Passionate, Kind, Enthusiastic, Understanding	Rude, Mean, Cold, Unkind, Egotistical
Extroverted, Optimistic, Trustworthy, Outspoken, Friendly, Genuine, Sociable	Introverted, Pessimistic, Dishonest, Selfish, Hostile, Aloof
Intelligent, Witty, Humble, Brave, Determined, Bold, Respectful, Assertive, Well-spoken	Ignorant, Stubborn, Closed-minded, Arrogant, Reserved
Dynamic, Confident, Energetic, Strong, Leader, Engaging, Persuasive	Aggressive, Angry, Apathetic, Shy, Weak, Overbearing, Dull, Obnoxious, Intimidating
Charming, Funny, Attractive, Humorous, Interesting, Relatable, Personable	Boring, Annoying, Uninteresting, Depressing

3.2.2 Charisma Perception in Cross-Language Settings

The following study was conducted to understand what in the voice perceptual domain could be considered as universal versus language and culture-based. The perception of charismatic speaker identity from voice might be influenced unpredictably by one vocal characteristic or by a whole complex pattern resulting from source and filter characteristics, mode of vocal fold vibration, temporal characteristics, articulatory settings and characteristics, degree of nasality, prosodic line, and syllable structure (Kreiman & Sidtis, 2011).

To do so, this study first assessed how listeners use the vocal pitch as a biological cue to detect speakers' charismatic traits from voice and how they use this cue to assess leadership fitness and choose their leader. In several studies vocal pitch has emerged as a feature that serves as an important biological cue that signals social and physical dominance (e.g., Ohala, 1982, 1983, 1984, 1994, 1996; Puts et al. 2007), conveys leadership (Klofstad et al., 2012; Anderson & Klofstad, 2012, and that influences the choice of a leader (Darwin, 1871; Tigue et al., 2012). In an experiment, 40 French listeners evaluated the dominance conveyed by different voice quality patterns in the voice of an Italian speaker and political leader (Umberto Bossi, former leader of the Lega Nord party from 1980 to 2012). The results showed significant negative correlations between the perceived dominant type of charismatic leadership and average F0 ($r = -0.19$, $p < 0.05$, linear regression), wide F0 range ($r = -0.18$, $p < 0.05$), and maximum F0 ($r = -0.18$, $p < 0.05$). Meanwhile, higher F0 mean ($r = 0.52$, $p < 0.01$), minimum F0 ($r = 0.49$, $p < 0.01$), maximum F0 ($r = 0.55$, $p < 0.01$), and the F0 range ($r = 0.53$, $p < 0.01$) are significantly and positively correlated with a nondominant type of charismatic leadership.

To confirm and extend these results, the investigations were repeated with the manipulation of F0 for vocal stimuli from two different leaders (Luigi de Magistris, an Italian leader; François Hollande, a French leader). Forty-eight Italians were then asked to rate vocal stimuli from the French leader and 48 French listeners were

216 asked to rate vocal stimuli from the Italian leader. Results show that French and Italian
217 listeners perceive leaders as having a less dominant charisma when they use a high F0
218 (average of 200 Hz for the French speaker; 212 Hz for the Italian speaker) and a wide
219 F0 range (16 semitones for French listeners; 12 semitones for Italian listeners). This
220 experiment studied the way in which listeners assess leadership fitness from voice. A
221 voice sounding more dominant (low frequencies of F0 and a narrow F0 range) would
222 be perceived as more effective by Italian listeners ($r = 0.61$, $p < 0.0001$; simple
223 linear regression), whereas French participants perceive effective leadership from
224 higher pitched voices ($r = 0.41$, $p = 0.004$). Results from the two experiments imply
225 that low frequencies of F0 and a narrow F0 range convey a dominant charismatic
226 leadership and that higher F0 average and wider F0 range, cause the perception
227 of a nondominant charismatic leader. These different types of leadership would be
228 perceived as more or less effective in different cultures.

229 Finally, the perception of specific charismatic traits from overall vocal charac-
230 teristics was studied taking into account the role of the language and the culture
231 of listeners. The study first assessed the way in which different patterns of voice
232 quality convey the different charismatic traits of leaders. Forty French participants
233 assessed the charisma of the Italian leader Umberto Bossi from natural voice sam-
234 ples. Detailed profiles based on the correlation between voice acoustics, perception
235 of charismatic traits, emotional states aroused, and choice of leader were created.
236 A profile with a voice pattern characterized by a medium pitch range (13 semi-
237 tones), moderate falling pitch contour movements, modal phonation, phrase-final
238 harsh-high (middle-range) vowels and long inter-word pauses (~ 1 s) communicate an
239 Authoritarian-Threatening type of charisma where the leader is perceived as individ-
240 ualistic, untrustworthy, influential, confident, organized, resolute, egocentric, deter-
241 mined, authoritarian, menacing, scary, and cold (see Table 3.2), and moreover arouses
242 negative emotional states in the listeners like anxiety. A second profile shows that a
243 voice pattern characterized by a wide pitch range (16 semitones) from very low to very
244 high frequencies, abrupt pitch contour movements, harsh or modal phonation, and
245 sentence-final vowels in creaky phonation communicate a Proactive-Attractive type
246 of charisma. Listeners who perceived the Proactive-Attractive type of charismatic
247 leadership described the leadership of the speakers as vigorous, active, dynamic,
248 charming, and attractive (see Table 3.2), arousing positive emotions like amusement,
249 admiration, enthusiasm, reassertion, and calmness. French listeners would be most
250 likely to choose a leader perceived as Proactive-Attractive. The third profile shows
251 a voice pattern characterized by a narrow pitch range from low to medium-high fre-
252 quencies (9–13 semitones), but not as high as the two vocal patterns above, smooth
253 pitch contour movements, harsh-low, harsh-mid, or modal phonation types, and an
254 increasing duration of the vocalization (from ~ 1 s to 6.5 s). This pattern commu-
255 nicates the Competent-Benevolent type of charismatic leadership, characterized by
256 participant-selected adjectives such as wise, prudent, calm, trustworthy, fair, intelli-
257 gent, easygoing, honest, sagacious, and sincere (see Table 3.2), arousing amusement
but not calmness emotions. This type of leadership communicates the image of a

Table 3.2 Charisma types and interpersonal traits. Speaker: Umberto Bossi. Assessed perceptually through the MASCharP tool. Exploratory Factor Analysis: Varimax Rotation that extracted three factors which explained 45% of the variance; significant Bartlett's test of sphericity ($p = 0.000$); Kaiser–Mayer Olkin (KMO) measure of Sampling Adequacy (0.83); high level of reliability (Proactive-Attractive: $\alpha = 0.92$, i.i. = 0.52; Calm-Benevolent: $\alpha = 0.87$, i.i. = 0.44; Authoritarian-Threatening: $\alpha = 0.90$, i.i. = 0.41)

Authoritarian-Threatening		Proactive-Attractive		Calm-Benevolent	
Determined	0.508	Vigorous	0.837	Wise	0.825
Menacing	0.775	Active	0.767	Prudent	0.737
Who scares	0.767	Dynamic	0.766	Calm	0.731
Dishonest	0.762	Charming	0.738	Trustworthy	0.689
Cold	0.679	Attractive	0.709	Fair	0.645
Individualistic	0.642	Courageous	0.701	Intelligent	0.605
Authoritarian	0.585	Convincing	0.687	Easygoing	0.585
Leader	0.578	Captivating	0.676	Honest	0.576
Untrustworthy	0.563	Seductive	0.642	Sagacious	0.527
Influent	0.552	Bewitching	0.604	Sincere	0.514
Confident	0.523	Sexy	0.592		
Organized	0.509	Eloquent	0.553		
Resolute	0.506	Determined	0.54		
Egocentric	0.485	Who propose	0.54		
		Visionary	0.472		
Variance	22.52%		12.6%		10.83%

258 leader competent enough to access vital resources and benevolent enough to share
 259 those resources with other individuals. French listeners in the sample studied would
 260 not choose this type of leadership.

261 3.3 Conclusions

262 3.3.1 Leaders' Social Attractiveness

263 Since Weber (1920), first launched the notion of charisma, the definition has gone
 264 through various changes. The notion itself may have seemed too difficult to opera-
 265 tionalize, while the literature has fluctuated from serious investigation to skeptical
 266 consideration. This may be partly due to the very nature of charisma, which lives
 267 at the crossroad of various psychosocial dimensions and takes very different forms
 268 (Shamir, 2000). This work has defined charisma as a set of internal and physical
 269 qualities of a person that make him or her capable of influencing other people by
 270 wakening their most positive emotions, and hence inducing them to do what she/he

271 wants very willingly and exploiting their internal motivation. These qualities are
 272 related to various perceived aspects of the group leaders persona (moral, intellec-
 273 tual, affective), of power management, as well as esthetic and even erotic aspects.
 274 Charisma is a multidimensional psychosocial notion: the studies presented in this
 275 chapter tried to discover and disentangle its dimensions from participants' descrip-
 276 tion of charismatic and noncharismatic persons using a scale of charisma perception.
 277 The present research found out that dimensions may combine to give rise to different
 278 types of charisma. The type of perceived charisma depends on whether the esthetic
 279 and dynamic dimensions prevail, resulting in a Proactive-Attractive charisma, or
 280 whether they are moderated by the intellectual and ethical side, thus enhancing
 281 a calm-benevolent charisma; or finally whether the dimensions of dominance and
 282 deliberate influence cluster in an Authoritarian-Threatening charisma.

283 Besides discovering these internal features and their combinations, this investi-
 284 gation focused on a peculiar property of charismatic political leaders, their vocal
 285 communication, showing that charisma resides in particular types of speech acts, but
 286 also in particular parameters of the leader's voice that, depending on given variations,
 287 may become less charismatic, or take up a different type of charisma. Two issues we
 288 specifically investigated in this connection were the change in charisma caused by
 289 a switch from modal to dysphonic voice, and the different perception of charisma
 290 caused, in the French and the Italian culture, by a change in pitch and pause duration.

291 Results on the former issue—that the modal voice conveys a proactive-attractive,
 292 or even an authoritarian-threatening charisma, whereas the disordered one bears a
 293 calm-benevolent one—may be accounted for by an evolutionary perspective that
 294 views a dynamic leader as more functional to the effectiveness of the group.

295 As to the issue of whether charisma perception is universal or cultural, our results
 296 may be interpreted as follows: The single traits attributed to a charismatic leader
 297 tend to be different between cultures and may arise at two levels: first, the single
 298 properties may cluster in different ways for two cultures, in that a type of charisma
 299 may be more salient in one culture and dispersed in single properties in another;
 300 second, as seen in the third phase of study, each specific type of charisma may be
 301 evoked by some vocal parameters in one language or culture and by other parameters
 302 in another.

303 These results may help answer some questions concerning charisma. For instance,
 304 one possible objection to the very existence of such a notion is that a person may
 305 appear as charismatic to some people but not to others. In other words, is it true
 306 that -beauty is in the eyes of the beholder-? In our view, this is not so. Different
 307 perceptions of charisma may well be accounted for by its multidimensionality. In
 308 this sense, interactive accounts that view charisma as determined by the intertangling
 309 between a leader and their followers may be sound. -Charismatic leadership- may
 310 hold per se, but also, followers can contribute their perceptual preferences to its
 311 emergence (Shamir, 2000).

312 In the same vein, the multidimensionality account might answer the question
 313 whether and why the perception of charisma varies across cultures. Since cultures
 314 definitely attribute different importance to different dimensions of life, cognitive
 315 functioning and social interaction, two cultures may well see the same leader as

316 charismatic or not, depending on the dimensions they value the most. Yet, this leads
 317 to another question: aren't there any aspects of charisma that are universal, that is,
 318 any characteristics of a leader (or of a person) that are perceived as charismatic by
 319 people of all cultures?

320 An answer in line with the "emotional culture" approach above (Ekman & Friesen,
 321 1971, Turner, 1976; Gordon, 1989; Matsumoto, 1990; Bagozzi, Verbeke, & Gavino,
 322 2003) would be that leaders are perceived as charismatic to the extent to which they
 323 adapt to the communicative norms of their culture. Yet, we might contend that, on the
 324 contrary, the charismatic leader, does not "adapt to", but rather, "leads" his followers,
 325 imposing new norms and values, and thus also changing the relative preference of
 326 the charismatic dimensions. Therefore, a primary and possibly universal dimension
 327 of charisma might be just the visionary skill that makes a leader point at something
 328 new.

329 A final issue, among others, that is raised by our investigation is how the notion of
 330 charisma proposed here can be applied not only to political leaders but to a broader
 331 domain: not only social leaders can be charismatic, but actors, singers, managers,
 332 and teachers. Our theoretical explanation of charisma could be applied generally to
 333 all charismatic individuals.

334 3.3.2 *The Charismatic Voice*

335 The present research demonstrates how a specific vocal pattern used by leaders can
 336 convey different traits and types of their charisma, and also how several patterns can
 337 influence the perception of the same type of characteristic leadership when perceived
 338 by different individuals or social groups. The acoustics of voice in political speech
 339 is a cue to the perception of charisma in leaders. We used a cross-cultural approach
 340 to assess and distinguish the physiological/anatomical and cultural influence in the
 341 production and perception of voice in charismatic leadership.

342 In the perceptual domain, the research described above, first found evidence that
 343 vocal pitch is a cross-cultural signal to distinguish dominant versus less dominant
 344 charisma. This result is consistent with previous studies on the perception of dominance
 345 versus submission related to vocal pitch (e.g., Collins, 2000; Feinberg et al.,
 346 2006). Higher fundamental frequency and wider range are used by the speaker while
 347 addressing a more diverse audience (in terms of sex, age and social status). Lower
 348 fundamental frequency and narrower range are used by the leader-speaker when
 349 addressing an audience of similar social status (other leaders). Healthy vocal range
 350 is used by leaders in informal contexts of communication (during which no political
 351 topics are addressed and the leadership is not questioned).

352 This work then found that certain vocal quality patterns used by the speaker-leader
 353 fit the listener's expectations about the vocal style that best conveys charisma in a
 354 given language and culture. The same vocal pattern can convey both an Authoritarian-
 355 Threatening and a Proactive-Attractive charisma that are perceptually distinguished

356 in different languages and cultures. Competent-Benevolent charismatic leadership
357 can be conveyed by several vocal quality patterns.

358 These results may help to better distinguish between the biological components
359 on the one hand, and language and cultural components on the other, present in voice
360 behavior that fit listeners' expectations and influences the choice of the social group's
361 leader. Listeners seem capable of accurately distinguishing these vocal features of the
362 charismatic leader and these results might explain why some leaders have been found
363 to be endowed with a cross-language and cultural charisma (e.g., Barack Obama was
364 found to be the most charismatic leader in the general sense in several cultures),
365 and some other leaders not endowed with effective speaking (Bligh & Robinson,
366 2010), are mostly endowed with a circumscribed charisma restricted within social
367 groups and languages (Gandhi is only charismatic if we understand English or if it
368 is translated).

369 References

- 370 Anderson, R. C., & Klofstad, C. A. (2012). Preference for leaders with masculine voices holds in
371 the case of feminine leadership roles. *PLoS ONE*, 7(12), e51216. <https://doi.org/10.1371/journal.pone.0051216>.
- 372 Aristotle (1991). *Rhetoric*, translated by George A. Kennedy. Location: Acheron Press, Kindle ed.
- 373 Atkinson, M. (1984). *Our Masters' voices. The language and body language of politics*. London:
374 Routledge.
- 375 Bagozzi, R. P., Verbeke, W., & Gavino, J. C. Jr. (2003). Culture moderates the self-regulation of
376 shame and its effects on performance: The case of salespersons in The Netherlands and the
377 Philippines. *Journal of Applied Psychology*, 88(2), 219.
- 378 Baken, R., & Orlikoff, R. (2000). *Clinical measurement of speech and voice*, 2nd (rev ed.). San
379 Diego: Singular Publishing Group.
- 380 Bass, B. M. (1985). *Leadership and performance beyond expectations*. New York: Free Press.
- 381 Bass, B. M. (1990). *Bass and Stogdill's handbook of leadership: Theory, research, and managerial
382 applications* (3rd ed.). New York: Free Press.
- 383 Biadys, F., Hirschberg, J., Rosenberg, A., & Dakka, W. (2007). Comparing American and Pales-
384 tinian perceptions of charisma using acoustic-prosodic and lexical analysis. In *Proceedings of
385 Interspeech*
- 386 Biadys, F., Rosenberg, A., Carlson, R., Hirschberg, J., & Strangert, E. (2008). A cross-cultural com-
387 parison of American, Palestinian, and Swedish perception of charismatic speech. In *Proceedings
388 of the 4th Conference on Speech Prosody*, (pp. 579–582). Campinas, Brazil.
- 389 Bligh, M. C. and Kohles, J. C. (2009). The enduring allure of charisma: How barack obama won
390 the historic 2008 presidential election. *The Leadership Quarterly*, 20(3), 483–492.
- 391 Bligh, M. C., & Robinson, J. L. (2010). Was Gandhi 'Charismatic'? Exploring the rhetorical lead-
392 ership of Mahatma Gandhi. *The Leadership Quarterly*, 21(5), 844–55.
- 393 Bono, J. E. and Judge, T. A. (2004). Personality and transformational and trans- actional leadership:
394 A meta-analysis. *Journal of Applied Psychology*, 89(5): 901–910.
- 395 Boss, P. (1976). Essential attributes of charisma. *Southern Speech Communication Journal*, 41(3),
396 300–13.
- 397 Bull, P. (1986). The use of hand gesture in political speeches: Some case studies. *Journal of Language
398 and Social Psychology*, 5(2), 103–118.
- 399 Burns, J. (1978). *Leadership*. New York, NY, USA: Harper & Row.

- 401 Castelfranchi, C. & Falcone, R. (2000). Trust is much more than subjective probability: Mental
 402 components and sources of Trust. In *32nd Hawaii International Conference on System Sciences—*
 403 *Mini Track on Software Agents*. Maui: IEEE Press.
- 404 Cavalli, L. (1995). *Carisma: la qualità straordinaria del leader*. Rome, Italy: Laterza.
- 405 Cicero (1967). *De Oratore*, translated by E. W. Sutton. Cambridge, MA: Harvard University Press.
- 406 Collins, S. A. (2000). Men's voices and Women's choices. *Animal Behavior*, 60(6), 773–80.
- 407 Darwin, C. (1871). *The descent of man, and selection in relation to sex*. Murray, J., London, UK.
- 408 Dastur, Y. (2016). *Charisma perception in the Japanese language*. Department of Linguistics:
 409 University of Southern California.
- 410 Den Hartog, D. N., House, R. J., Hanges, P. J., Ruiz-Quintanilla, S. A., & Dorfman, P. W. (1999).
 411 Culture specific and cross-culturally generalizable implicit leadership theories: Are attributes of
 412 charismatic/transformational leadership universally endorsed? *The Leadership Quarterly*, 10(2),
 413 219–56.
- 414 D'Errico, F., Signorello, R., Demolin, D., & Poggi, I. (2013). The perception of charisma from
 415 voice: A cross-cultural study. In *Proceedings of the 2013 Humaine Association Conference on*
 416 *Affective Computing and Intelligent Interaction* (pp. 552–557). IEEE Computer Society.
- 417 D'Errico, F., Signorello, R., & Poggi, I. (2012). Le Dimensioni del Carisma. In M. Cruciani & F.
 418 Ceconi (Eds.), *IX Convegno Annuale dell'Associazione Italiana di Scienze Cognitive–AISC* (pp.
 419 245–52). Rome: Università di Trento.
- 420 Di Blas, L., & Forzi, M. (1998). The circumplex model for interpersonal trait adjectives in Italian.
 421 *Personality and Individual Differences*, 24(1), 47–57.
- 422 Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of*
 423 *Personality and Social Psychology*, 17(2), 124.
- 424 Emerich, K. A., Titze, I. R., Švec, J. G., Popolo, P. S., & Logan, G. (2005). Vocal range and intensity
 425 in actors: A studio versus stage comparison. *Journal of Voice*, 19(1), 78–83.
- 426 Esling, J. (2006). Voice quality. In K. Brown (Ed.), *The Encyclopedia of Language and Linguistics*
 427 (2nd ed., pp. 470–4). Oxford: Elsevier.
- 428 Feinberg, D. R., Jones, B. C., Law Smith, M. J., Moore, F. R., DeBruine, L. M., Cornwell, R. E.,
 429 et al. (2006). Menstrual cycle, trait estrogen level, and masculinity preferences in the human
 430 voice. *Hormones and Behavior*, 49(2), 215–22.
- 431 Foti, R. J., & Luch, C. H. (1992). The influence of individual differences on the perception and
 432 categorization of leaders. *Leadership Quarterly*, 3, 55–66.
- 433 Garvin, P., & Ladefoged, P. (1963). Speaker identification and message identification in speech
 434 recognition. *Phonetica*, 9, 193–199.
- 435 Gordon, S. L. (1989). *Institutional and impulsive orientations in selectively appropriating emotions*
 436 *to self The sociology of emotions: Original essays and research papers* (pp. 115–35).
- 437 Haslam, S. A., Reicher, S. D., & Platow, M. J. (2011). *The New psychology of leadership*. Hove,
 438 UK: Psychology Press.
- 439 Hofstede, G. (1993). Cultural constraints in management theories. *Academy of Management Exec-*
 440 *utive*, 7(1), 81–94.
- 441 Hollander, E. P., & Julian, J. W. (1970). Studies in leader legitimacy, influence, and innovation.
 442 *Advances in Experimental Social Psychology*, 5, 33–69.
- 443 House, R. J. and Howell, J. M. (1992). Personality and charismatic leadership. *The Leadership*
 444 *Quarterly*, 3(2), 81–108.
- 445 Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, 39, 31–6.
- 446 Klofstad, C. A., Anderson, R. C., & Peters, S. (2012). Sounds like a winner: Voice pitch influences
 447 perception of leadership capacity in both men and women. *Proceedings of the Royal Society B:*
 448 *Biological Sciences*, 279(1738), 2698–2704.
- 449 Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to*
 450 *voice production and perception*. Oxford, UK: Wiley-Blackwell.
- 451 Kruskal, W. H., & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *Journal of*
 452 *the American Statistical Association*, 47(260), 583–621.

- 453 Lamarche, A., Ternström, S., & Hertegård, S. (2009). Not just sound: Supplementing the voice
454 range profile with the singer's own perceptions of vocal challenges. *Logopedics, Phoniatrics,*
455 *Vocology*, 34(1), 3–10.
- 456 Lewin, K. (1952). *Field theory in social science: Selected theoretical papers*. London: Tavistock.
- 457 Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology*, 140, 1–55.
- 458 Lowe, K. B., Kroeck, K. G., & Sivasubramaniam, N. (1996). Effectiveness correlates of transforma-
459 tional and transactional leadership: A meta-analytic review of the Mlq Literature. *The Leadership*
460 *Quarterly*, 7(3), 385–425.
- 461 Matsumoto, D. (1990). Cultural similarities and differences in display rules. *Motivation and Emo-*
462 *tion*, 14(3), 195–214.
- 463 Offermann, L. R., Kennedy, J. K., & Wirtz, P. W. (1994). Implicit leadership theories: Content
464 structure, and generalizability. *Leadership Quarterly*, 5(1), 43–55.
- 465 Ohala, J. (1996). Ethological theory and the expression of emotion in the voice. In *Proceedings of*
466 *the 4th International Conference on Spoken Language Processing (ICSLP 96)* (vol. 3, pp. 1812–
467 1815). Philadelphia, PA, USA.
- 468 Ohala, J. J. (1982). The voice of dominance. *Journal of the Acoustical Society of America*, 72(S1),
469 S66–S66.
- 470 Ohala, J. J. (1983). Cross-language use of pitch: An ethological view. *Phonetica*, 40(1), 1–18.
- 471 Ohala, J. J. (1984). An ethological perspective on common cross-language utilization of F0 of voice.
472 *Phonetica*, 41(1), 1–16.
- 473 Ohala, J. J. (1994). The frequency codes underlies the sound symbolic use of voice pitch. *Sound*
474 *Symbolism* (pp. 325–347). Cambridge, MA, USA: Cambridge University Press.
- 475 Patel, S., Scherer, K. R., Björkner, E., & Sundberg, J. (2011). Mapping emotions into acoustic space:
476 The role of voice production. *Biological Psychology*, 87(1), 93–98.
- 477 Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J. M., Röder, S., Andrews, P. W., et al.
478 (2014). Vocal indicators of body size in men and women: A meta-analysis. *Animal Behaviour*,
479 95, 89–99.
- 480 Poggi, I., D'Errico, F., & Vincze, L. (2011). Discrediting moves in political debate. In P. Ricci-Bitti
481 (Ed.), *Proceedings of UMMS* (pp. 84–99). Heidelberg: Springer.
- 482 Poggi, I. (2005). The goals of persuasion. *Pragmatics & Cognition*, 13(2), 297–336.
- 483 Puts, D. A., Hodges, C. R., Cárdenas, R. A., & Gaulin, S. J. C. (2007). Men's voices as dominance
484 signals: Vocal fundamental and formant frequencies influence dominance attributions among
485 men. *Evolution and Human Behavior*, 28(5), 340–344.
- 486 Raymond, M. (2008). *Cro-magnon toi-même! Petit Guide Darwinien de la vie Quotidienne*. Paris:
487 Seuil.
- 488 Reboul, O. (1998). *Introduction à la Rhétorique* (3rd ed.). Paris: Presses Universitaires de France.
- 489 Reicher, S., Haslam, S. A., & Hopkins, N. (2005). Social identity and the dynamics of leader-
490 ship: Leaders and followers as collaborative agents in the transformation of social reality. *The*
491 *Leadership Quarterly*, 16(4), 547–68.
- 492 Rhee, N., & Signorello, R. (2016). The acoustics of charismatic voices in Korean political speech:
493 A cross-gender study. *Journal of the Acoustical Society of America*, 139(4), 2123–3.
- 494 Riggio, R. E., Chaleff, I., & Lipman-Blumen, J. (2008). *The art of followership: How great followers*
495 *create great leaders and organizations*. San Francisco: Jossey-Bass.
- 496 Rosenberg, A., & Hirschberg, J. (2009). Charisma perception from text and speech. *Speech Com-*
497 *munication*, 51(7), 640–55.
- 498 Rudolph, S. H., & Rudolph, L. I. (1983). *Gandhi: The traditional roots of charisma*. Chicago:
499 University of Chicago Press.
- 500 Saucier, G. (2009). Semantic and linguistic aspects of personality. In P. J. Corr & G. Matthews
501 (Eds.), *The Cambridge handbook of personality psychology* (pp. 379–99). Cambridge: Cambridge
502 University Press.
- 503 Scherer, K. R. (2010). Voice appeal and its role in political persuasion. In *International Workshop*
504 *on Political Speech, Rome*.

- 505 Scherer, K. R. (1972). Judging personality from voice: A cross-cultural approach to an old issue in
 506 interpersonal perception. *Journal of Personality*, 40, 191–210.
- 507 Shamir, B. (2000). Taming charisma for better understanding and greater usefulness: A response to
 508 Beyer. *The Leadership Quarterly*, 10(4), 555–562.
- 509 Shamir, B., House, R. J., & Arthur, M. B. (1993). The motivational effects of charismatic leadership:
 510 A self-concept based theory. *Organization Science, A Journal of the Institute of Management*
 511 *Sciences*, 4(4), 577.
- 512 Shamir, B., Zakay, E., Breinin, E., & Popper, M. (1998). Correlates of charismatic leader behavior in
 513 military units: Subordinates' attitudes, unit characteristics, and Superiors' appraisals of leader
 514 performance. *Academy of Management Journal*, 41(4), 387–409.
- 515 Signorello, R. (2014). *La Voix Charismatique: Aspects Psychologiques et Caractéristiques Acous-*
 516 *tiques*. Ph.D diss. Université de Grenoble, France and Università degli Studi Roma Tre, Italy.
- 517 Signorello, R. (2014b). The biological function of fundamental frequency in leaders' charis-matic
 518 voices. *The Journal of the Acoustical Society of America*, 136(4), 2295–2295.
- 519 Signorello, R., & Demolin, D. (2013). The physiological use of the charismatic voice in political
 520 speech. In *Proceedings of the 14th Annual Conference of the International Speech Communication*
 521 *Association (Interspeech)* (pp. 987–991).
- 522 Signorello, R., D'Errico, F., Poggi, I., & Demolin, D. (2012b). How charisma is perceived from
 523 speech: a multidimensional approach. *ASE/IEEE International Conference on Social Computing*
 524 (pp. 435–440).
- 525 Signorello, R.; D'Errico, F.; Poggi, I.; Demolin, D. & Mairano, P. (2012a). "Charisma Perception
 526 in Political Speech: A Case Study." In *Proceedings of the VIIth GSCP International Conference:*
 527 *Speech and Corpora*, edited by H. Mello, M. Pettorino, and T. Raso, 343–8. Firenze: Firenze
 528 University Press.
- 529 Signorello, R., & Rhee, N. (2016). The voice acoustics of the 2016 United States presidential
 530 election candidates: A cross-gender study. *Journal of the Acoustical Society of America*, 139(4),
 531 2123–3.
- 532 Signorello, R., Demolin, I., Poggi, D., & D'Errico, F. (2011). *Il Carisma del Corpo: Caratteristiche*
 533 *Acustiche della Voce Carismatica*. In X Giornate della Ricerca: Università degli Studi Roma Tre.
- 534 Strangert, E. & Gustafson, J. (2008). "What Makes a Good Speaker? Subject Ratings, Acoustic
 535 Measurements and Perceptual Evaluations." In *Proceedings of the 9th Annual Conference of the*
 536 *International Speech Communication Association (Interspeech 2008)*: 1688–91.
- 537 Surawski, M. K., & Ossoff, E. P. (2006). The effects of physical and vocal attractiveness on impres-
 538 sion formation of politicians. *Current Psychology*, 25(1), 15–27.
- 539 The LimeSurvey project team (2011). LimeSurvey. Web-based com-puter program. Retrieved June
 540 26, 2011, from <http://www.LimeSurvey.org/>.
- 541 Tigue, C. C., Borak, D. J., O'Connor, J. J. M., Schandl, C., & Feinberg, D. R. (2012). Voice Pitch
 542 Influences Voting Behavior. *Evolution and Human Behavior*, 33(3), 210–216.
- 543 Touati, P. (1993). Prosodic Aspects of Political Rhetoric. *ESCA Workshop on Prosody*, 168–71.
- 544 Tuppen, C. (1974). Dimensions of Communicator Credibility: an Oblique Solution. *Speech Mono-*
 545 *graphs*, 41(3), 253–60.
- 546 Turner, R. H. (1976). The Real Self: From institution to impulse. *American Journal of Sociology*,
 547 989–1016.
- 548 Weber, M. (1920). The theory of social and economic organization. Oxford University Press, New
 549 York, USA.
- 550 Zuckerman, M., & Driver, R. (1989). What sounds beautiful is good: The vocal attractiveness
 551 stereotype. *Journal of Nonverbal Behavior*, 13(2), 67–82.

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Voice Attractiveness
Series Title	
Chapter Title	Vocal Preferences in Humans: A Systematic Review
Copyright Year	2020
Copyright HolderName	Springer Nature Singapore Pte Ltd.
Corresponding Author	Family Name Barkat-Defradas
	Particle
	Given Name Melissa
	Prefix
	Suffix
	Role
	Division
	Organization Institut des Sciences de l'Evolution de Montpellier, University of Montpellier, Centre National de la Recherche Scientifique, Institut pour la Recherche et le Développement, Ecole Pratique des Hautes Etudes – Place Eugène Bataillon
	Address 34095, Montpellier, France
	Email melissa.barkat-defradas@umontpellier.fr
Author	Family Name Raymond
	Particle
	Given Name Michel
	Prefix
	Suffix
	Role
	Division
	Organization Institut des Sciences de l'Evolution de Montpellier, University of Montpellier, Centre National de la Recherche Scientifique, Institut pour la Recherche et le Développement, Ecole Pratique des Hautes Etudes – Place Eugène Bataillon
	Address 34095, Montpellier, France
	Email michel.raymond@umontpellier.fr
Author	Family Name Suire
	Particle
	Given Name Alexandre
	Prefix
	Suffix
	Role
	Division
	Organization Institut des Sciences de l'Evolution de Montpellier, University of Montpellier, Centre National de la Recherche Scientifique, Institut pour la Recherche et le Développement, Ecole Pratique des Hautes Etudes – Place Eugène Bataillon
	Address 34095, Montpellier, France

Email

alexandre.suire@umontpellier.fr

Abstract

Surprisingly, the study of human voice evolution has long been conducted without any reference to its biological function. Yet, following Darwin's original concept, John Ohala was the first linguist to assume the functional role of sexual selection to explain vocal dimorphism in humans. Nevertheless, it is only at the very beginning of the millennial that the study of voice attractiveness developed, revealing that beyond its linguistic role, voice also conveys important psycho-socio-biological information that have a significant effect on the speaker's mating and reproductive success. In this review article, our aim is to synthesize 20 years of research dedicated to the study of vocal preferences and to present the evolutionary benefits associated with such preferences.

Keywords

Vocal preferences - Perception - Language evolution - Sexual selection - Evolutionary biology - Acoustics - Voice - Fundamental frequency - Formant dispersion - Voice attractiveness

Chapter 4

Vocal Preferences in Humans: A Systematic Review



Melissa Barkat-Defradas, Michel Raymond, and Alexandre Suire

Abstract Surprisingly, the study of human voice evolution has long been conducted without any reference to its biological function. Yet, following Darwin's original concept, John Ohala was the first linguist to assume the functional role of sexual selection to explain vocal dimorphism in humans. Nevertheless, it is only at the very beginning of the millennial that the study of voice attractiveness developed, revealing that beyond its linguistic role, voice also conveys important psycho-socio-biological information that have a significant effect on the speaker's mating and reproductive success. In this review article, our aim is to synthesize 20 years of research dedicated to the study of vocal preferences and to present the evolutionary benefits associated with such preferences.

Keywords Vocal preferences · Perception · Language evolution · Sexual selection · Evolutionary biology · Acoustics · Voice · Fundamental frequency · Formant dispersion · Voice attractiveness

4.1 Introduction

Darwin thought of mate choice as a purely aesthetic experience, a selection of beauty for its own sake (Darwin, 1871). However, his view has not been embraced by modern evolutionary biology, for which mate choice results from human adaptive preferences, a mechanism that has evolved because of dimorphic physical features or sexual ornaments (such as the female waist-to-hip ratio, the male shoulder-to-

M. Barkat-Defradas (✉) · M. Raymond · A. Suire
Institut des Sciences de l'Évolution de Montpellier, University of Montpellier, Centre National de la Recherche Scientifique, Institut pour la Recherche et le Développement, Ecole Pratique des Hautes Etudes – Place Eugène Bataillon, 34095 Montpellier, France
e-mail: melissa.barkat-defradas@umontpellier.fr

M. Raymond
e-mail: michel.raymond@umontpellier.fr

A. Suire
e-mail: alexandre.suire@umontpellier.fr

hip ratio, facial traits, breast size, voice, and so on) that are assumed to be reliable indicators of mate quality (Arak & Enquist, 1993). Indeed, the mere sound of a person's voice contains important, embedded biological information. Consequently, a large amount of research has been dedicated to identifying men's preferences for women's secondary sexual characteristics and vice versa, as well as the evolutionary benefits associated with such preferences.

Preferences partly proceed from an unconscious mechanism: an individual may be aware of the factors that have led him to choose one sexual partner instead of another, but it does not necessarily mean s/he is conscious of the link existing between his or her preference and the property conveyed by the cue itself. A good example to illustrate this statement rests on women's preference for masculine low-pitched voices. Though female subjects are often conscious of their attraction for this type of vocal attribute in males, they are hardly aware that it indicates men's phenotypic quality as well as part of their heritable genotypic value as potential mates (Apicella, Feinberg, & Marlowe, 2007). In human species, mate's selective value includes several phenotypic qualities among which: state of health, fertility, age, intelligence, social status, and so on ... (Buss, 1989; Geary, Vigil, & Byrd-Craven, 2004; Sugiyama, 2015). All these qualities are displayed through the face, the body, and the voice. For example, health is indicated by skin complexion, the body shape is a proxy of nutritional status, and the vocal height is determined by testosterone level. Therefore, it is reasonable to assume that female typical preference for men exhibiting deep voices has been shaped by evolution as an honest signal of masculinity related to an increased level of androgens, a high physical strength, a good immune system, etc., all of these features favoring men's—and thus women's—fitness. However, masculine versus feminine preferences for the ornaments exhibited by the other sex are not the same since some of the traits that are associated to desirable qualities in men may differ from those linked to desirable phenotypic qualities in women. Consequently, men and women do not grant the same importance to the different socio-biological cues driving mate choice. Generally speaking, and at least in Western industrialized societies, men tend to attach a great importance to women's beauty, and as early as Ancient Greece, the concept of beauty has been closely associated with physical attractiveness, especially feminine physical attractiveness (for a detailed review of the evolution of feminine beauty see Bovet, 2018). But when choosing a mate, men and women also use non-physical features, such as smell, movements, behaviors, and voice. Although these traits are not all equally weighted in mating decisions, they all likely contribute to the general evaluation of a potential partner.

Our aim here is not to explore the diverse effects of physical attractiveness but rather to examine the role of voice in the mating context by showing which vocal features are considered attractive by men and/or women and why. Previous research on vocal attractiveness (i.e., the perceived attractiveness of voices when isolated from other cues, such as visual or olfactory cues) has suggested that vocal attractiveness plays a role in mate choice in humans (e.g., Apicella et al., 2007; Hill et al., 2013; Leongomez et al., 2014). For example, individuals possessing vocal characteristics that are correlated with attractiveness report greater reproductive potential (as indexed by reported number of sexual partners, Kordsmeyer, Hunt, Puts, Ostner, & Penke,

2018; Hill et al., 2013) and, at least in hunter-gatherers, have greater reproductive fitness (Apicella et al., 2007). People also alter their vocal attractiveness in mating contexts, such as when interacting with an attractive potential mate (Leongomez et al., 2014; Pisanski, Bhardwaj, & Reby, 2018; Suire, Raymond, & Barkat-Defradas, 2018). In accordance to the runaway selection mechanism,¹ we assume preferences may contribute to the shaping of attractiveness in human voices. Our goal therefore is to show that preferences for some vocal attributes are likely the result of sexual selection. Although the acoustic features associated with vocal attractiveness are not exhaustively studied here (i.e., the prosodic dimension, in particular, could be further developed), we propose an exhaustive review of the different studies (n = 37, over a period of 40 years covering the years 1979–2020) that tackled the issue of vocal preferences for men and women (see Table 4.1). Subsequently, we will focus on the evolutionary mechanisms driving our preferences. Before fully entering our topic, it should be noted that only the studies that have clearly identified the acoustic correlates behind vocal preferences were considered.

Overall, a first remarkable point appears to be the importance ascribed to the study of F0 and the formant position. Secondly, one will immediately notice that English speakers are overrepresented in comparison with speakers of other languages. From a methodological point of view, it appears that the number and the nature of vocal stimuli used in the perceptual experiments are quite variable (i.e., spontaneous speech, isolated words or vowels, reading versus oral speech ...). Likewise the number of auditory judges is extremely heterogeneous from one study to another. As for the acoustic analyses themselves, we distinguish between two types of approaches: on the one hand, there are correlational studies, which basically aim at relating acoustic characteristics and vocal attractiveness from auditory judge's scores on Likert's scales and on the other hand, there are experimental studies that try to establish causal relations between acoustic features. All these studies help us pinpoint some general trends about human vocal preferences.

A brief overview in Table 4.1 reveals that among the different measures that were investigated for qualifying vocal attractiveness across studies, it is undoubtedly vocal height (i.e., F0) that has most often aroused the authors' interest. Nevertheless some other articulatory and acoustic features have lead to interesting results suggesting vocal attractiveness is not confined to the realm of fundamental frequency but also extend to other aspects, which effects on perceived vocal attractiveness are also reviewed in the next sections.

¹Runaway selection is a mechanism whereby a secondary sexual trait expressed in one sex is correlated with a preference for the trait in the other sex. The genetic coupling of the trait and the preference leads to self-reinforcing loops of coevolution between the trait and preference for the trait (Travers, 2017).

Table 4.1 Studies are characterized by the language under study, the number and nature of tested the stimuli, the number and gender of auditory judges, the methodology (Likert's scale versus forced choice), results by gender and the direction of observed correlations. For Likert's scales, the lowest score (i.e., 1) corresponds to the less attractive stimuli, the highest to the most attractive voice; when forced choice is used the judge has to choose between two stimuli the one he perceives the most attractive. By "manipulated speech" we mean that the subjects were recorded after they were asked to modify their voices following the experimenters' instructions. Note that for studies based on modified stimuli (whether naturally or not) forced choice is often used since it allows judges to select the most attractive stimuli between two versions of the same voice (i.e., natural versus modified). NB: CA stands for Canadian, AU for Australian, U.S. for American, and U.K. for British variants of English

References	Language(s) under study	No. of stimuli	No. of auditory judges	Type of stimuli	Method and no. of evaluated stimuli	Acoustic features under study and direction of the observed correlations
Tuomi and Fisher (1979)	English (CA)	10♀ 5♂	10♀ 10♂	Spontaneous sentences	Likert's scale n = 15 (i.e., all)	- Low F0 +attractive ♀ and ♂ (n.s.)
Zuckerman and Miyake (1993)	English (U.S.)	62♀ 48♂	8♀ 9♂	Speech reading	Likert's scale n = 110 (i.e., all)	- Low F0 +attractive ♂ - Low F0min +attractive ♂ - Low energy +attractive ♂ - Less pausing time +attractive ♂ - n.s. ♀
Oguchi and Kikuchi (1997)	(i) Japanese - Experience 1	4♂	25♀	Read sentences	Likert's scale n = 4 (i.e., all)	- Low F0 +attractive ♀ and ♂ - Low F0-SD +attractive ♀ and ♂
	(ii) Japanese -Experience 1	8♀ 8♂	42♀ 20♂	Read sentences	Likert's scale n = 16 (i.e., all)	- Low F0 +attractive ♀ and ♂ - Low F0-SD +attractive ♀ and ♂

(continued)

Table 4.1 (continued)

References	Language(s) under study	No. of stimuli	No. of auditory judges	Type of stimuli	Method and no. of evaluated stimuli	Acoustic features under study and direction of the observed correlations
Collins (2000)	Dutch	34 σ	54 ϕ	Isolated vowels (natural speech)	Likert's scale n = between 10 & 14	- Spectral distribution in low frequencies +attractive σ - Low formant spacing +attractive σ
Collins and Missing (2003)	English (U.K.)	30 ϕ	30 σ	Isolated vowels (natural speech)	Likert's scale n = 10	- Spectral distribution in high frequencies +attractive σ - High formants +attractive σ - High formant spacing +attractive σ
Feinberg et al. (2005)	English (CA)	10 σ	68 ϕ	Isolated vowels (manipulated speech)	Likert's scale n = 10 (i.e., all)	- Low F0 +attractive σ - Lower formant spacing + attractive σ
Bruckert et al. (2006)	French	26 σ	102 ϕ	Isolated vowels (natural speech)	Likert's scale n = 6	- Low F0 +attractive σ - High F0-SD +attractive σ

(continued)

Table 4.1 (continued)

References	Language(s) under study	No. of stimuli	No. of auditory judges	Type of stimuli	Method and no. of evaluated stimuli	Acoustic features under study and direction of the observed correlations
Riding et al. (2006)	English (U.S.)	9 σ	54 φ	Spontaneous then manipulated speech	Likert's scale n = 11	- High F0 -attractive σ - F0-SD n.s. σ
Saxton et al. (2006)	English (U.K.)	12 σ	40 φ 7-10 y.o 40 φ 12-15 y.o 40 φ 20-34 y.o	Number counting (natural speech)	Forced choice n = 6 or 12	- Low F0 +attractive for 12-15 and 20-34 y.o.
Feinberg, DeBruine, Jones, and Little (2008a, 2008b)	-Experience 1 English (CA)	123 φ	10 σ	Isolated vowels (manipulated speech)	Likert's scale n = 61 or 62	- High F0 +attractive σ
Hughes et al. (2008)	-Experience 2 English (CA)	15 σ	263 φ 342 σ	Number counting (from 1 to 10)	Forced choice n = 15 pairs	- High F0 +attractive σ
	English (U.S.)	31 φ 40 σ	50 φ 51 σ	Numbers recitation (from 1 to 10)	Likert's scale n = 71 (each voice being evaluated by 13 or 15 judges)	- Low F0min +attractive σ - F0, F0max, F0 range, median F0, Intensity, Duration, Jitter, Shimmer, HNR n.s. σ - n.s. φ
Leaderbrand et al. (2008)	English (U.S.)	1 φ 1 σ	39 φ 9 σ	Sentences (manipulated speech)	Likert's scale n = 4	- Low F0min +attractive σ and φ

(continued)

Table 4.1 (continued)

References	Language(s) under study	No. of stimuli	No. of auditory judges	Type of stimuli	Method and no. of evaluated stimuli	Acoustic features under study and direction of the observed correlations
Vukovic et al. (2008)	English (U.K.)	36 σ	58 φ (+contraceptive) 65 φ (-contraceptive)	Sentences (manipulated speech)	– Forced choice n = 16 pairs – Likert's scale for each preferred voice	– Low F0 +attractive σ – No effect of contraception on vocal preferences
Saxton et al. (2009)	English (U.K.)	6 φ 6 σ 11–13 y.o. 6 φ 6 σ 13–15y.o. 6 σ 13–15 y.o	148 φ 177 σ (same category of age)	Isolated vowels (manipulated speech)	– Forced choice n = 6 pairs – Likert's scale for each preferred voice	– High F0 +attractive φ (for 11–13 y.o. σ only) – Low F0 +attractive σ (for 13–15 y.o. φ only)
Jones et al. (2010)	English (U.K.)	4 φ	30 φ 30 σ	Spontaneous sentences (natural speech)	Forced choice n = 16 pairs	– High F0 +attractive φ
Fraccaro et al. (2011)	English (CA)	6 φ	178 σ	Isolated vowels (manipulated speech)	Forced choice n = 6 pairs	– High F0 +attractive φ (long-term relationship condition only)

(continued)

Table 4.1 (continued)

References	Language(s) under study	No. of stimuli	No. of auditory judges	Type of stimuli	Method and no. of evaluated stimuli	Acoustic features under study and direction of the observed correlations
Hodges-Simeon et al. (2010)	English (U.S.)	111 σ	142 φ	Spontaneous speech	Likert's scale n = 30 or 31	<ul style="list-style-type: none"> - Low F0 +attractive σ (no effect of short vs. long relationship condition, no effect of φ menstrual cycle phase) - Low F0-SD +attractive σ (in long term + fertile context and in short term + unfertile context) - Spectral distribution in the low frequencies +attractive σ (in short/long term + fertile conditions)
Hughes, Farley and Rhodes (2010)	English (U.S.)	25 φ 20 σ	27 φ 12 σ	Truncated phone calls + speech manipulation	Forced choice n = 45 (i.e., all)	<ul style="list-style-type: none"> - Low F0 +attractive σ and φ
Jones et al. (2018)	English (U.K.)	6 φ 6 σ	100 φ 100 σ	Isolated vowels (manipulated speech)	Forced choice n = 6 pairs	<ul style="list-style-type: none"> - High F0 +attractive φ - Low F0 +attractive σ
Borkowska and Pawlowski (2011)	Polish	58 φ	144 σ	Isolated vowels (manipulated speech)	Likert's scale n = 13 voices	<ul style="list-style-type: none"> - High F0 +attractive φ (non linear relation)

(continued)

Table 4.1 (continued)

References	Language(s) under study	No. of stimuli	No. of auditory judges	Type of stimuli	Method and no. of evaluated stimuli	Acoustic features under study and direction of the observed correlations
Pisanski and Rendall (2011)	English (CA)	2♀ 6♂	30♀ 31♂	Words list (natural then manipulated speech)	Likert's scale n = 40 voices	– Low F0 and formants +attractive ♂ (same trend observed for natural and manipulated speech) – Low F0 and formants-attractive ♀ (same trend observed for natural and manipulated speech)
Puts et al. (2011)	English (U.S.)	72♀	63♂	Text reading (manipulated speech)	Likert's scale n = 18 pairs	– High F0 +attractive ♀ – Spectral distribution in the high frequencies +attractive ♀
Liu and Xu (2011)	English (U.K.)	1♀	10♂	3 repetitions of 1 single emotion-free sentence (natural then manipulated speech)	Likert's scale n = 81 (i.e., all)	– High F0 +attractive ♀ – Small vocal length tract +attractive ♀
Simmons et al. (2011)	English (AU)	54♂	15♀	Isolated vowels (natural speech)	Likert's scale n = 54 (i.e., all)	– Low F0 +attractive ♂

(continued)

Table 4.1 (continued)

References	Language(s) under study	No. of stimuli	No. of auditory judges	Type of stimuli	Method and no. of evaluated stimuli	Acoustic features under study and direction of the observed correlations
Barkat-Defradas et al. (2012)	French	62 σ	92 φ	Text reading + Isolated vowel (natural speech)	Likert's scale n = 34	- F0 n.s. - Mild roughness degree +attractive σ - Low breathiness +attractive σ
Re et al. (2012)	English (CA)	1 φ 1 σ	9 φ 10 σ	Isolated vowels (manipulated speech)	Forced choice n = 50 pairs + supplementary pairs 6 σ and 42 φ	- High F0 +attractive φ - Low F0 +attractive σ
Fraccaro et al. (2013)	English (CA)	4 φ 4 σ	104 φ 110 σ	Isolated vowels (manipulated speech)	Forced choice n = 16 pairs	- High F0 +attractive φ - Low F0 +attractive σ
O'Connor et al. (2013)	English (CA)	4 φ 4 σ	128 σ	Words (manipulated speech)	Likert's scale n = 40	- High F0 +attractive φ - Low F0 +attractive σ

(continued)

Table 4.1 (continued)

References	Language(s) under study	No. of stimuli	No. of auditory judges	Type of stimuli	Method and no. of evaluated stimuli	Acoustic features under study and direction of the observed correlations
Xu et al. (2013)	English (U.K.)	– Experience 1 1♀ – Experiences 2–5 1♀ 1♂	10♂ 16♀ 16♂	Sentences (manipulated speech) Sentences (synthesized speech)	Likert's scale n = 81 (i.e., all) Likert's scale n = 81 (i.e., all)	– Low F0 +attractive ♂ – Formants: n.s. – High F0 +attractive ♀ – Low F0 +attractive ♂ – High breathiness +attractive ♀ ♂ – Low formants +attractive ♂ – Formants n.s. ♀
Babel et al. (2014)	English (U.S.)	30♀ 30♂	15♀ 15 ♂	Words (natural speech)	Likert's scale n = 15 (one single voice for each trial)	– Spectral distribution in high frequencies +attractive ♀ – Low F0 +attractive ♀ (n.s.) – Breathily voices +attractive ♀ – Spectral distribution in low frequencies +attractive ♂ – Shorter duration +attractive ♂

(continued)

Table 4.1 (continued)

References	Language(s) under study	No. of stimuli	No. of auditory judges	Type of stimuli	Method and no. of evaluated stimuli	Acoustic features under study and direction of the observed correlations
Hughes et al. (2014)	English (U.S.)	20♀ 20♂	20♀ 20♂	Number recitation (from 1 to 10) (manipulated speech)	Likert's scale n = 40 voices	<ul style="list-style-type: none"> - High hoarseness +attractive ♀ n.s. ♂ - Longer duration +attractive ♂ ♀ - Low F0 +attractive ♂ n.s. ♂ - Loudness n.s. ♂ ♀
Skrinda et al. (2014)	Latvian	60♂	29♂	Isolated vowels (natural speech)	Likert's scale n = unspecified	<ul style="list-style-type: none"> - Low F0 +attractive ♂ - Low F2 values +attractive ♂ - Other formants n.s.
Tsantani et al. (2016)	English (U.K.)	10 ♀ 9♂	183♀ 57♂	"Hello" (manipulated speech)	Forced choice n = 40 pairs	<ul style="list-style-type: none"> - Low F0 +attractive ♂ n.s. ♀
Sebesta et al. (2017)	Cross-linguistic	45♂ (Cameroonians) 48♂ (Namibians)	62 ♀ Czechs	Sentence (natural speech)	Likert's scale n = 45 – 48 pairs	<ul style="list-style-type: none"> - Low F0 +attractive Cameroonians ♂ - Low formant position + attractive Namibians ♂ - High breathiness +attractive Namibians ♂

(continued)

Table 4.1 (continued)

References	Language(s) under study	No. of stimuli	No. of auditory judges	Type of stimuli	Method and no. of evaluated stimuli	Acoustic features under study and direction of the observed correlations
Shirazi et al. (2018)	Cross-linguistic	6 σ (EN U.S)	20 breastfeeding + 20 nulliparous Filipinos φ	Sentences (manipulated speech)	Likert's scale n = 12 stimuli	- High F0 +attractive σ - n.s. between the 2 groups of φ
Suire et al. (2018)	French	58 σ	137 φ	Sentence	Forced choice n = 11 pairs	- Low F0 +attractive σ - High F0-SD +attractive σ - Other acoustic measures n.s.
Suire et al. (2019)	French	13 φ	135 σ 2 conditions: short- versus long-term relationship	Read sentences (natural speech)	Forced choice n = 13 pairs	- High speaking rate +attractive φ - Low F0 +attractive φ - Spectral distribution in high frequencies +attractive φ - High roughness (high Jitter values) +attractive φ - Low breathiness (high HNR values) +attractive φ

(continued)

Table 4.1 (continued)

References	Language(s) under study	No. of stimuli	No. of auditory judges	Type of stimuli	Method and no. of evaluated stimuli	Acoustic features under study and direction of the observed correlations
Zheng et al. (2020)	Chinese	80♀, 35♂	88♀, 79♂	Isolated vowels (natural speech)	3 conditions: – habitual voice – raised in pitch (+20 Hz) – lowered in pitch (–20 Hz)	– High F0 +attractive ♂ both for male and female raters – High F0 +attractive ♂ for male raters – Low F0 +attractive ♂ for female raters (but very low pitches are perceived less attractive > cue for laryngeal damage or below the intelligibility threshold)

4.2 Preferences for Vocal Height

Most of the previous studies, whether they are correlational or experimental, have revealed a negative correlation between vocal height and attractiveness of men. Such a regular trend shows that women, whatever their linguistic environments and/or cultural backgrounds, are predominantly attracted to men exhibiting deep low voices (Bruckert, Lienard, Lacroix, Kreutzer, & Leboucher, 2006; Feinberg et al., 2005; Hodges-Simeon, Gaulin, & Puts, 2010; Hughes, Farley, & Rhodes, 2010; Jones, Feinberg, DeBruine, Little, & Vukovic, 2010; Pisanski & Rendall, 2011; Vukovic et al., 2008; Xu, Lee, Wu, Liu, & Birkholz, 2013; Suire, Raymond, & Barkat-Defradas, 2019). Still, a few exceptions are to be considered. As a matter of fact, Babel, McGuire, and King (2014) and Hughes, Mogilski, and Harrison (2014) reported no significant correlation between vocal height and attractiveness in American men. Likewise, Barkat-Defradas et al. (2012) demonstrated F0 does not seem to be the most salient perceptual feature to assess masculine voice attractiveness as compared to roughness at least in clinical context when patients range into a comparable vocal height category (i.e., ± 125 Hz) irrespective of their global dysphonic grade. Lastly, Shirazi, Puts, and Escasa-Dorne (2018) obtained an unexpected opposite result with Filipino women judging male vocal samples produced in English by American speakers. As for women vocal attractiveness, the vast majority of studies reach the same results with men being consistently attracted by high-pitched feminine voices (Borkowska & Pawlowski, 2011; Collins & Missing, 2003; Feinberg et al., 2008a, 2008b; Jones et al., 2010; Puts, Barndt, Welling, Dawood, & Burriss, 2011; Re et al., 2012). But here again, the results obtained by Leaderbrand, Dekam, Morey, and Tuma (2008), Oguchi and Kikuchi (1997) go in the opposite direction when those by Hughes et al. (2010, 2014) reveal interesting trends. In Hughes et al. (2010), the authors show that women tend to lower their voices when interacting with men they consider as particularly attractive while they significantly raise their pitch when facing men they are not attracted to. The same kind of unexpected result is observed for men who judge those low-pitched women as sexier. More recently, Pisanski et al. (2018) replicated the same results. In a second study, in which female subjects were asked to modify their voice so as they might be perceived as more attractive by male auditory judges, it has been shown that in such an evoked seductive context, women are also inclined to deepen their voices, and interestingly the subsequent perceptual study revealed that the female voices attesting the lower pitch values are also those that were perceived as the most attractive by the group of male auditory judges (Hughes et al., 2014). The results launched by Zheng, Compton, Heyman, and Jiang (2020) in what must be to our knowledge the most recent available study tackling the subject aimed at determining more precisely the effect of raised versus lowered pitch on voice perceived attractiveness. In order to answer this question, the authors used a method based on voluntarily pitch-shifted voices. Their findings suggest that indeed pitch shifts do affect voice attractiveness in the sense that female voices are perceived—both for male and female raters—as more attractive when vocal pitch is raised (+20 Hz from a digitally computed average



143 pitch at 237Hz). As for male voices, they typically show that lowered pitch lead to
 144 better evaluations by female raters (up to certain limits beneath which low voices are
 145 perceived either as pathological or unintelligible). But surprisingly, they also come
 146 to the result that their male raters consider high-pitched masculine voices as more
 147 attractive. According to the authors this may be explained by the fact that in real-life
 148 conditions, men are more often placed into the position of evaluating sex-opposite
 149 attributes using morphological signals, like waist-to-hip ratio,² but also vocal cues so
 150 as to find information of phenotypical compatibility, which makes their perceptual
 151 evaluation biased either by a lack of experience or by the unconscious usage of a
 152 perceptual grid of evaluation that is structured around feminine vocal references and
 153 which is consequently quiet unsuitable for the evaluation of male voices.

154 4.3 Preferences for Vocal Modulation

155 If studies dealing with the effect of mean F0 on vocal attractiveness are relatively
 156 numerous, those based on the measure of F0-SD (i.e., the increased versus reduced
 157 mean fundamental frequency variations, which the listener perceives, respectively, as
 158 rather flat versus highly modulated speech) are rather scarce. Yet, Hodges-Simeon et
 159 al. (2010) have shown that male speakers producing speech with very little variations
 160 in F0 are perceived as more masculine and attractive by female raters. Given that
 161 the extent of F0-excursions is affected by attitudinal and emotional factors (Traun-
 162 möller & Eriksson, 1995), such a trend appears to be kind of difficult to explain at
 163 first glance. Indeed, as it is well admitted the non-verbal characteristics of voices
 164 can play a significant role in signaling emotional as well as health state, like for the
 165 latter, major depression that is regularly reflected through reduced vocal modulation,
 166 female preferences for small melodic variations in male voices may be explained both
 167 by vocal dimorphism (since it has been regularly shown lively speech is related with
 168 feminine talking style (Polce-Lynch, Myers, Kilmartin, Forssmann-Falck, & Kliewer
 169 1998; Hall, 1978) and social factors (as the extensive vocal expression of emotions
 170 is more often associated with female behavior (Fischer & Manstead, 2000). There-
 171 fore, assuming pitch variations are perceived along a continuum (from monotonic
 172 to highly expressive speech), the receivers may have assigned monotonous voices to
 173 masculinity and, reversely, dynamic speech to femininity. Besides, Suire et al. (2020)
 174 have shown males' sexual orientation can be inferred more accurately from F0-SD
 175 than mean F0, suggesting vocal modulation is a more reliable acoustic cue for gays'
 176 vocal feminization than vocal height. Moreover, though previous studies assessed

²The WHR has been used as an indicator of health and the risk of developing serious health conditions. WHR correlates with fertility (with different optimal values in males and females). The concept and significance of WHR as an indicator of attractiveness has been theorized by Singh (1993) who argued the WHR is a consistent estrogen marker, and thus a reliable proxy of fertility. Women with a 0.7 WHR are usually rated as more attractive by men from Indo-European cultures (Singh & Young 2001), but preferences may vary according to the culture under study (Fisher & Voracek, 2006).

177 that reduced fundamental frequency variations are rather linked to vocal masculin-
 178 ity, two other studies lead to unexpected opposite results. According to Bruckert
 179 et al. (2006), monotonous voices are judged as significantly less attractive for men
 180 while Leongómez et al. (2014) found modulated voices are rated as more attractive
 181 for both sexes. Further researches are thus needed to disentangle these inconsistent
 182 results. But yet for now, it is interesting to notice that the same criterion may lead
 183 to different auditory impressions, which valences are somehow contradictory. For
 184 example, although perceived as more attractive, those masculine speakers exhibit-
 185 ing monotonous, low-pitched voices are also perceived as being less cooperative
 186 (Tognetti et al., 2019), more threatening, and their likelihood to have extramarital
 187 affairs is considered as higher. This claim does not result from unfounded subjective
 188 impressions since there is also evidence that suggest men with masculine voices
 189 report a higher number of extra-pair sex partners and are more often chosen by
 190 women as extra-pair partners (Hughes et al., 2004).

191 The above suggests that men with relatively more masculine voices—that are
 192 negatively correlated with testosterone levels (Evans, Neave, Wakelin, & Hamil-
 193 ton, 2008)—may present a greater infidelity risk to their partners, though it is still
 194 unclear whether observers assess infidelity risk via vocal cues to underlying testos-
 195 terone levels. Likewise, women with relatively high-pitched, modulated voices—that
 196 are linked both with youth, higher fertility, and increased perceived attractiveness—are
 197 also seen as more conspicuous and more likely to commit adultery (O’Connor, Re,
 198 & Feinberg, 2011). But, while there is substantial evidence for a positive relationship
 199 between testosterone, deep voice, and “unbridled” sexuality among men, the rela-
 200 tionship between women’s sexuality and feminine vocal features is more complex
 201 (for a review, see Bancroft, 2005). We should therefore be cautious and presume that
 202 women with attractive voices may be more likely to be unfaithful due to a greater
 203 opportunity for extra-pair sex given their desirability as a mate as their attractive
 204 voices are more often chosen by paired men as extra-pair partners (Hughes, Dis-
 205 penza, & Gallup, 2004).

206 4.4 Preferences for Timbre

207 Sounding vocalizations are the product of multiple acoustic parameters, including
 208 formant position and formant dispersion. Formant dispersion is a measure of the
 209 average spacing between the formants (Fitch, 1997). It is a function of the length and
 210 shape of the vocal tract and corresponds to the space through which sound waves must
 211 travel from the vocal folds to the oral cavity. Until sexual maturity, vocal tract length
 212 grows without any sexual dimorphism between boys and girls (Vorperian et al., 2005),
 213 but at puberty, under the influence of androgens, males’ larynges descend farther than
 214 females’ (Fitch & Giedd, 1999). Indeed, working through hormone receptors in the
 215 epithelial cells of the laryngeal tissue, testosterone enlarges the larynx on the one
 216 hand and lengthens and thickens the vocal folds on the other. The consequence of
 217 these remarkable anatomic modifications is a longer vocal tract and the acoustic

218 result is a lower vocal height and a deeper and more resonating voice in adult males.
219 On average, the vocal tract is about 15% longer in men than women (Fant, 1960) and
220 this results in perceptible sex differences in formant dispersion, with males exhibiting
221 formants of lower frequency (measured through formant position) as well as lower
222 formant dispersion (Hanson, 1997).

223 Studies trying to correlate vocal resonances and perceived attractiveness have
224 lead to controversial results. For instance, Hodges-Simeon et al. (2010), Pisanski
225 and Rendall (2011) showed that the lower the formant dispersion, the more attractive
226 the masculine voices. The same tendency was observed by Sebesta et al. (2017) for
227 whom the formant position was the acoustic variable of interest. Conversely, Skrinda
228 et al. (2014) and Xu et al. (2013) found no correlation between low resonances and
229 male voice attractiveness. Interestingly, two other studies led to original results.
230 Using formant dispersion, Babel et al. (2014) showed that only tall women tend to
231 prefer low resonances in males' voices. Likewise, Feinberg et al. (2005) observed
232 the same preferences but only for the two high vowels /i/ and /u/, which are perceived
233 more attractive when the spacing between F1 and F2 is reduced. Such a result may
234 be explained by basic acoustic principles. Indeed, Holmberg et al. (1995) showed
235 that the relative amplitude of the harmonics is closely related with the adduction
236 of the vocal folds, with the higher the adduction, the lower the harmonics at the
237 glottal exit. Moreover, using fiberoptics to characterize vocal closure as function
238 of speakers' gender, Södersten, Lindstad, and Hammarberg (1991) showed female
239 speakers' higher degree of incomplete closure is correlated with increased harmonics.
240 Therefore, the results of Feinberg et al. are in line with theoretical analysis and
241 observations in experimental acoustics, since sounds with greater low-frequency
242 and weaker high-frequency components are recognized to result from more adducted
243 glottal considerations that are, themselves, more typical of male speakers (Hanson,
244 1996).

245 Collins and Missing (2003) investigated the relationship between male human
246 vocal characteristics and female judgments about the speaker and showed that, in
247 general, women found men's voices with harmonics that are closer together and
248 lower in frequency more attractive. This corroborates the findings of earlier studies
249 where less masculine sounding speakers were described as having higher formant
250 frequencies (Avery & Liss, 1996). In their study aiming in testing listeners' weighting
251 of F0 and/or formant frequency for the rating of vocal attractiveness, Pisanski and
252 Rendall (2011) reached the same conclusion, that is, voices with relatively low F0
253 and/or low formant frequencies rated as more attractive if male and less attractive if
254 female. Interestingly, the authors also showed that, in assessing attractiveness, listeners
255 appeared to weigh formant frequency cues more heavily than F0, an unpredicted
256 result which suggests female listeners might interpret lower frequency cues as indi-
257 cating greater masculinity and thus greater attractiveness in male voices. Finally, the
258 results obtained by Xu et al. (2013) also showed male voices sounded more attractive
259 when they are low pitched and with densely distributed formants associating such
260 characteristics with the large body size projected.

4.5 Preferences for Voice Quality

Among the various complex acoustic features that give a voice its quality, the variations of the glottal source waveform hold a special place. The values of the parameters that describe the glottal waveform can vary depending on the glottal configuration and/or the quality of the vocal fold vibrations, and it is expected that these variations may lead to different voice qualities. Some voice qualities are usually associated with disordered voice, such as harshness (also referred to as vocal roughness or hoarseness), but since our main concern here is vocal attractiveness, we will focus on those that may occur for voices that are not perceived to be pathological. Voice qualities that occur frequently in normal speech are described to be “modal,” that is, smooth and acoustically brilliant voices (Laver, 1980; Titze, 1994), but there are also some voice qualities that are commonly related to dysphonia but may also occur in normal (i.e., non-pathological) conversational speech and still be perceived attractive (Barkat-Defradas et al. 2012). It is typically the case for both moderately breathy and rough voices. According to Fairbanks (1960: 179), “breathy quality” (also called murmured voice or whispery voice) is described as an inefficient laryngeal vibration: “(...) *In the coordination of normal voice quality the vibrating vocal folds approximate in the midline once per cycle, closing the glottis and interrupting the airflow. In breathy quality the vocal folds vibrate, but the intermittent closure fails and the airflow is continuous.*” Interestingly, the author also underlines breathy voice lowers voice pitch and is almost invariably accompanied by limited vocal intensity. As for vocal roughness, or “harsh quality,” it is defined as an “*irregular, aperiodic noise in the vocal fold spectrum caused by an excessive laryngeal tension*” (Fairbanks, 1960: 179; Laver, 1980: 133, 1994: 477). Though the indication of psychological attributes conveyed through voice quality has aroused researchers’ attention since ancient times (Laver, 2009: 38), this belief has long found rather eccentric and impressionistic assertions. For example, a breathy quality was supposed to show that men were “aesthetic” and women “pretty and callow”; flat that men are “distant” and women “hard and lethargic”; nasal that men are “unattractive and self-effacing” and women the same; tense that men are “cantankerous” and women “high-strung”; throaty that men are “stable” and women “oafish”; orotund (or loud) that men are “suave” and women “aggressive”; and so on. The idea that personality characteristics are correlated with voice quality has recently been tested more scientifically, and although some controversy remains, it must be admitted some correlations do exist. Among the few studies that have tackled the topic of vocal breath and/or vocal roughness and their effects on perceived voice attractiveness, it has been shown that harsh voices are regularly correlated with more aggressive, dominant, and authoritative personalities while breathy ones are more frequently associated with self-effacing, submissive, and weak temperaments. A way to quantify breathiness—which is caused by glottal air leakage—is to measure harmonics-to-noise ratio (henceforth HNR), a measure that quantifies the relative amount of additive noise.³ As for vocal roughness, it

³At the physiological level, low HNR values are believed to be related to insufficient vocal fold adduction during the so-called “closed” interval of the phonatory cycle. Insufficient closure would

302 results from irregular vocal fold vibrations. These vibratory perturbations have come
 303 to be more commonly referred to as vocal jitter. As a matter of fact, a number of
 304 investigators have demonstrated a significant correlation between increased levels
 305 of jitter and perceived roughness (Lieberman, 1963; Moore & Thomson, 1965). For
 306 example, Babel et al. (2014) and van Borsel et al. (2009) found female voices were
 307 perceived more attractive when breathy. Unexpectedly, Sebesta et al. (2017) and
 308 Xu et al. (2013) showed significant relations between vocal breath and attractiveness
 309 for both sexes. A plausible explanation for male vocal attractiveness unexpectedly
 310 enhanced by breathiness in this particular study lies in the fact that this predomi-
 311 nantly feminine vocal feature may presumably soften the aggressiveness regularly
 312 associated with low deep voices.

313 Though some other phonetic characteristics could be addressed so as to charac-
 314 terize vocal attractiveness (e.g., preferences for speech tempo), the above overview
 315 offers an exhaustive assessment of the state of the art regarding the topic and under-
 316 lines the necessity to question both understudied acoustic parameters that may be
 317 relevant for vocal pleasantness and the effect of language/culture on perceived attrac-
 318 tiveness.

319 4.6 Sources of Variations in Vocal Preferences

320 Though some general tendencies emerge from studies dealing with vocal preferences,
 321 some sources of variations should be mentioned. These are mainly of two different
 322 natures. Some sources of variation seem to be due to physiological matters (i.e.,
 323 variations in hormonal levels) while some others are more concerned with cultural
 324 arguments (i.e., social representations).

325 4.6.1 *The Effect of Menstrual Cycle on Females' Vocal* 326 *Preferences*

327 It has been suggested that women's preferences maybe affected both by menstrual
 328 cycle (i.e., whether they are in their ovulatory versus follicular and/or luteal phase)
 329 and the context of mating they are looking for (i.e., short- versus long-term rela-
 330 tionships). Feinberg et al. (2006), Pisanski et al. (2014), and Puts (2005) have put
 331 forward the hypothesis of "*good genes ovulatory shift*" which suggests that women
 332 in ovulatory phase tend to prefer more masculine men (higher masculinity being
 333 associated with a better genotypic quality according to the theory of immunocompe-

allow excessive airflow through the glottis, giving rise to a turbulence noise component in the quasi-periodic source signal. This friction noise would result in a higher noise level in the spectrum, especially in the higher frequencies.

334 tence handicap⁴) more particularly in the context of short-term relationships (Jünger
 335 et al., 2018). Conversely, in the context of long-term relationships, women in their
 336 follicular and/or luteal phases tend to prefer men exhibiting less masculine traits,
 337 indicating they are more likely to invest themselves in parental care. Such variability
 338 in females' preferences would account for an adaptive strategy allowing women to
 339 optimize their fitness (i.e., reproductive success) in function of their menstrual cycle.

340 As for vocal preferences specifically, Puts (2005) noted that for the same vocal
 341 stimulus, women in their ovulatory phase judge low-pitched masculine voices (i.e.,
 342 low F0) more attractive when looking for a short-lived relationship. Likewise, Fein-
 343 berg et al. (2006) and Pisanski et al. (2014) observed this choice is even more marked
 344 for women in their fertility window. Hodges-Simeon et al. (2010) also investigated
 345 the effect of vocal resonance (i.e., formant dispersion) on females' vocal preferences
 346 and, though they could not find any effect of the type of relationship (i.e., short
 347 or long) specifically linked to this feature, they showed women are more likely to
 348 judge attractive masculine voices that exhibit a low dispersion of formants (i.e., deep
 349 voices). They also notice a shift in women's preferences as function of both menstrual
 350 cycle and duration commitment: monotonous masculine voices (low F0-SD) being
 351 judged as more attractive by infertile women in the context of short-term liaisons
 352 while the same vocal stimuli are perceived as more attractive for fertile women who
 353 are engaged in a long-term relationship. Those somehow inconsistent results lead
 354 some authors to question the validity of menstrual cycle as a reliable explanatory
 355 factor for women's variations in their attractiveness preferences. For example, Jones
 356 et al. (2018) and Marcinkowska, Galbarczyk, and Jasienska (2018) found no effect
 357 of female's menstrual cycle on body and face attractiveness evaluations of men.
 358 Likewise, Jünger et al. (2018)—using a robust methodology—could not confirm
 359 any effect neither of cycle phases nor of steroids to explain females' variations in
 360 their choices. As for feminine voices, since laryngeal epithelial cells are known to
 361 be highly sensitive to hormonal variations (Haselton, Mortezaie, Pillsworth, Bleske-
 362 Rechek, & Frederick, 2007; Miller et al., 2007; Higgins & Saxman, 1989; Abitbol
 363 et al., 1999; Amir & Biron-Shental, 2003; Bryant & Haselton, 2009; Fischer et al.,
 364 2011), women's voices undergo perceivable variation in their quality. As a matter
 365 of fact, Pipitone and Gallup (2008) have shown that feminine voices—which are
 366 higher pitched when women approach their fertile period—are perceived as more
 367 attractive by men whereas they sound lower pitched outside the ovulatory phase and
 368 are, consequently, judged less appealing (Bryant & Haselton, 2009; Fischer et al.,
 369 2011). These variations in females' vocal quality are essentially due to changes in
 370 estrogens and progesterone levels across the menstrual cycle, which lead to physio-

⁴The theory of immunocompetence handicap (Zahavi, 1975) suggests that androgen-mediated traits accurately signal condition due to the immunosuppressive effects of androgens. This immunosuppression may be either because testosterone alters the allocation of limited resources between the development of ornamental traits and the immune system or because heightened immune system activity has a propensity to launch autoimmune attacks against gametes, such that suppression of the immune system enhances fertility. Therefore, only healthy individuals can afford to suppress their immune system by raising their testosterone levels, which also augments secondary sexual traits and displays (among which low deep voices for men).

371 logical modifications in the mass, the tension, and the viscosity of the vocal folds,
 372 which in turn modify their oscillatory properties. It has been suggested these cyclic
 373 vocal quality variations could have been adaptive since they could contribute to the
 374 enhancement of women's attractiveness and facilitate mating when the risk of con-
 375 ception is higher and, therefore, the chance to conceive higher (Fischer et al., 2011;
 376 Pipitone & Gallup, 2008; Puts et al., 2013).

377 **4.6.2 The Effect of Sociocultural Environment on Vocal** 378 **Quality**

379 Though they are remarkably scarce, the few existing studies that have investigated
 380 the effect of sociocultural environment on vocal preferences have shown they are
 381 not universal but language/culture dependent. For example, van Bezooijen (1995)
 382 demonstrated that Japanese women exhibited the highest vocal pitch among a large
 383 sample of natural languages (i.e., 232 Hz) while the mean fundamental frequency
 384 of American women is around 214 Hz and that of Dutchwomen close to 196 Hz.
 385 Vaissière (2015) found French women's voice are even lower pitched with a mean
 386 F0 close to 190 Hz. It has been suggested that these significant differences in female
 387 vocal height could be constrained by specific cultural requirements that are them-
 388 selves shaped by social values and expectations that are linked to the roles allocated
 389 to women versus men and, more generally, to the stereotypes of femininity versus
 390 masculinity defined by the culture in question. Stereotypes of gender therefore vary
 391 among different cultures as well as among different ethnic groups (Landrine, 1985;
 392 Harris, 1994). In this way, the figure of femininity in Japanese culture is traditionally
 393 related to modesty, innocence, gentleness, subordination, physical fragility, and psy-
 394 chological submission (Sughira & Katsurada, 1999); these personality traits being
 395 vocally signaled to Japanese men who share the same cultural background through
 396 that famous "*voix de petite fille*" which has been subtly described by Léon (1981).
 397 Conversely, in the Netherlands—a country described as more egalitarian—women
 398 exhibit more masculine (i.e., low pitched) voices since their culture favors psycho-
 399 logical traits that are associated with female independence. In conclusion, it seems
 400 that the acoustic features that are typical of feminine versus masculine voices are
 401 not only due to anatomical and/or physiological criteria (i.e., vocal length tract and
 402 hormonal level) but also to cultural aspects depending on the social values attributed
 403 to sex roles in a given society. Besides, the studies conducted by Sebesta et al. (2017)
 404 and Shirazi et al. (2018) have shown that cultural expectations do not only con-
 405 cern vocal height. For example, in a Namibian population, male attractiveness is
 406 not predicted by F0 but by the degree of vocal breathiness they exhibit. Likewise,
 407 in the Philippines, females tend to prefer men with higher pitched voices. Though
 408 the effect of sociocultural representations on voice has been focused on, there is, to
 409 our knowledge, no study that aimed at identifying the factors of this variation. Yet,
 410 it does not seem to occur randomly in the same way as it has been observed for

411 the evolution of the waist-to-hip ratio (Bovet & Raymond, 2015; Bovet, 2019), the
412 body mass index, or the stature, in which variations have been shown to be partly
413 due to the ecology (see Pisanski & Feinberg, 2013 for a discussion), and that is why
414 cross-cultural surveys are still needed to evaluate the weight of culture on vocal pref-
415 erences. The scope of research dealing with voice attractiveness should also consider
416 the issue of preferences limitations. As a matter of fact, there are very few studies that
417 tackle the topic of superior and/or inferior limits above/below which a voice is no
418 longer perceived as attractive. Among these, Re et al. (2012) have shown women’s
419 preferences do not vary when male vocal pitch is below 96 Hz, but when they have to
420 choose between two stimuli above this value, they regularly prefer the lower voice.
421 As for men, to our knowledge, two studies were interested in determining a vocal
422 height threshold (in the range 160–300 Hz) below/above which female voices would
423 no longer be perceived as attractive (Feinberg et al., 2008a, 2008b; Re et al., 2012).
424 Results show men always consider high-pitched voices as more attractive for women.
425 Moreover, Borkowska and Pawlowski (2011) reported a non-linear relation between
426 vocal height and attractivity, the latter starting to decrease when F0 is close to 260 Hz.
427 According to the authors, this may be due to the fact that high-pitched voices are
428 commonly associated to sexually immature females. Though works dealing with the
429 determination of perceptible thresholds from which vocal attractiveness is affected are
430 still in the pipeline, several studies have shown that straight after a voice is perceived
431 as too distant from the norm, it is often categorized as pathological and associated
432 with negative personality traits (Barkat-Defradas et al., 2015; Revis, 2017).

433 Conversely, vocal attractiveness has a profound influence on listeners—a bias
434 known as the “*what sounds beautiful is good*” vocal attractiveness stereotype—with
435 tangible impact on a voice owner’s success at mating, job applications, and/or elec-
436 tions (Zuckerman & Driver, 1989). This led some authors, like Bruckert et al. (2010),
437 to test the effect of averaging voices via auditory morphing on perceived attractiv-
438 ity. Overall, their results reveal that the larger the number of voices averaged, the
439 more attractive the result. This is partly because composite voices have a smoother,
440 more regular texture and also because they sound more like the average voice and
441 reflect norm-based encoding of vocal stimuli. Preferences for some voices may also
442 be explained by the principle of sparseness. It has been demonstrated that human
443 perceptual systems (visual, auditory, and olfactory) have been selected so as to code
444 the information efficiently that is to say quickly and as parsimoniously as possible
445 to be in line with the principle of least effort (Renoult, Bovet, & Raymond, 2016).
446 Such a cognitive process relies on the elimination of the redundant components of a
447 signal, by which processing is consequently more accurate and less costly while the
448 storage and the retrieval of relevant information is more efficient. Nevertheless, the
449 neuropsychological mechanisms driving the coding of acoustic signals in relation
450 with vocal attractivity has received little scientific attention and, to our knowledge,
451 there is no study investigating these aspects specifically. Yet, since clear evidence for
452 interference between facial and vocal information has been observed (Aben, Pflügera,
453 Koppensteiner, Coquerellee, & Grammer, 2015), it seems reasonable to claim that
454 vocal and facial cues convey redundant information about a speaker’s mate value
455 and thus may serve as a backup signal for human mate choice decisions.

4.7 How Evolution Shaped Human Voice via Opposite Sex's Preferences

Though it is easy to understand how morpho-anatomical, physiological as well as behavioral differences between species result from natural selection and environmental adaptations, in some famous cases, those well-known mechanisms fail to explain the existence of certain remarkable features (Darwin, 1871). The iconic example that is traditionally invoked to illustrate this point is the male peacock's tail (*Pavo cristatus*), which is adorned with iridescent feathers. Darwin himself recognized this extravagant ornament contradicted his theory of natural selection. As a matter of fact, no doubt the male peacock's tail represents a critical bulk for his flight, and its outstanding colors has the disadvantage to attract his predators' attention. Besides, noting their absence in females and juveniles, the author concludes such an ornament cannot serve the animal's survival. Indeed, if peacocks' tail feathers were useful against predators then females and juveniles would exhibit the same. Therefore, he suggests the presence of some morphological characteristics cannot be explained solely by the advantages they provide to their bearers in terms of survival (which refers to "natural selection" itself) but also in terms of mating and fitness (which refers to a complementary concept, he defines as "sexual selection"). According to Darwin, sexual selection is restricted to secondary sex characteristics⁵—among which body size—and explains why many species exhibit sexual dimorphism at sexual maturity through the spectacular feathers of the birds-of-paradise, the impressive antlers of the male members of the deer family and, last but not least, vocal dimorphism in humans, among other dimorphic traits. The theory of Ohala's frequency code (1984)—inspired by Morton (1977)⁶—indicates that despite the development of highly complex language capable of conveying fine subtleties in meaning, humans still use an encoding strategy similar to the one widely used by nonhuman animals, namely, (i) by using relatively low-frequency sounds to indicate they are likely to attack versus (ii) more high-frequency sounds to indicate they are submissive, appealing, or fearful. Here pattern (i) is to project a large body size so as to threaten the receiver, because a larger animal has a better chance at winning a physical confrontation. Pattern (ii) is to project a small body size to attract the receiver, because a smaller animal is less likely to be a threat (Morton, 1977). Following this reasoning, Ohala (1984) argues the longer vocal folds of human males may have evolved under

⁵Secondary sex characteristics are features that appear during puberty in humans, and at sexual maturity in other animals. Secondary sex characteristics include, for example, the manes of male lions, the bright facial and rump coloration of male mandrills, and horns in many goats and/or antelopes. In humans, visible secondary sex characteristics include pubic hair, enlarged breasts and widened hips of females, facial hair, Adam's apples on males, etc.

⁶In a famous article dealing with vocal communication in animals, Morton (1977) introduces his « motivation-structural rules » theory, which suggests physical proprieties of acoustic signals (sounds of high versus low frequencies) are motivated since they reflect the vocalizer's body size and inform about his/her intentions and/or emotional state. He argues a large number of birds and mammals use low-frequency sounds to express hostility, threat, and aggression whereas high-frequency sounds are rather used to express fear, submission, and "amicability."

489 a selection pressure to compete with other males in achieving dominance for the
 490 sake of gaining access to female mates (i.e., intra-sexual selection). Likewise, the
 491 longer vocal tract of males may have evolved under the same pressure, as it may also
 492 reflect a larger body size and attract females (i.e., inter-sexual selection, see Puts
 493 et al., 2006 for an exhaustive presentation of the role of intra-selection in males).
 494 Extending the mechanism further, the shorter vocal folds and vocal tract of females
 495 may have developed under a pressure in the opposite direction, i.e., to project a small
 496 body size in order to attract male mates. To sum it up, by making an analogy between,
 497 on the one hand, the appearance of antlers in male deers, which develop when they
 498 attain sexual maturity and, on the other hand, voice change in pubescent boys, Ohala
 499 was a pioneer in assessing the functional role of sexual selection for the emergence
 500 of vocal dimorphism in humans.

501 I think the enlargement of the vocal apparatus also occurs to enhance aggressive
 502 displays. Males, by their role in the family unit and the fact that they compete for the
 503 favors of the female—i.e, they are subject to what Darwin called sexual selection—
 504 would be the ones to develop such deviations from the ‘norm’. However, they would
 505 only need these aggressive decorations when they are ready to compete and retain
 506 the favors of a female, that is, at the time of sexual maturity (Ohala, 1984: 14).

507 4.8 Conclusion

508 This contribution aimed at showing the mechanism of sexual selection formalized by
 509 Darwin as early as 1871 constitutes a crucial force in the evolution of voice, which
 510 directly intervenes in reproductive strategies. Though such an argument has been
 511 considered as obvious for many species, it is only at the very beginning of the 2000s
 512 that the phenomenon of vocal dimorphism has been tackled in relationship with Dar-
 513 win’s theory. As a matter of fact, it is surprising that the study of language activity has
 514 long been conducted without any reference to its biological function. Traditionally,
 515 humanities (anthropology, linguistics ...) used to consider language as a pure cultural
 516 product, which had been created by humans in the same ways as writing or art (Levi-
 517 Strauss, in Charbonnier, 1959: 48; Noble and Davidson, 1996: 214; Tomassello,
 518 1999: 94), and which developed irrespective of any selective pressure (Chomsky,
 519 1975: 75). In this purely cultural conception, the study of ultimate (or distal) causes
 520 explaining the existence of vocal dimorphism in terms of evolutionary forces has
 521 been left aside for the benefit of extensive analyses of proximal mechanisms, which
 522 explain its biological function in terms of immediate physiological or environmental
 523 factors. Yet, a transdisciplinary approach—at the crossroad of linguistics and evolu-
 524 tionary biology—is of a great interest to better understand the whys and wherefores
 525 of the evolution of articulated language in the human lineage. Indeed, beyond its
 526 evidenced social function (Dunbar, Duncan, & Nettle, 1995), vocal behavior should
 527 undoubtedly be regarded as a reliable way to display one’s phenotypic value (Puts,
 528 2010). Moreover, the existence of a low laryngeal configuration—an indispensable
 529 condition for language—in many non-speaking species undermines the hypothesis

of a specific adaptation to language in humans (Fitch & Reby, 2001). Reversely, considering such a disposition is present in several animals of different species clearly indicates it has evolved during phylogenesis to respond to other functions.

References

- Aben, P., Pflügera, L., Koppensteiner, M., Coquerellee, M., & Grammer, K. (2015). Sound of female shape: A redundant signal of vocal and facial attractiveness. *Evolution and Human Behavior*, 36(3), 174–181.
- Abitbol, J., Abitbol, P., & Abitbol, B. (1999). Sex hormones and the female voice. *Journal of Voice*, 13, 424–446.
- Amir, O., & Biron-Shental, T. (2003). The impact of hormonal fluctuations on female vocal folds. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 12, 180–184.
- Apicella, C. L., Feinberg, D. R., & Marlowe, F. W. (2007). Voice pitch predicts reproductive success in male hunter-gatherers. *Biology Letters*, 3(6), 682–684.
- Arak, A., & Enquist, M. (1993). Hidden preferences and the evolution of signals. *Philosophical Transactions of the Royal Society of London B*, 340, 207–213. <https://doi.org/10.1098/rstb.1993.0059>.
- Avery, J. D., & Liss, J. M. (1996). Acoustic characteristics of less masculine-sounding male speech. *The Journal of the Acoustical Society of America*, 99, 3738–3748.
- Babel, M., McGuire, G., & King, J. (2014). Towards a more nuanced view of vocal attractiveness. *PLoS One*, 9(2), e88616.
- Bancroft, J. (2005). The endocrinology of sexual arousal. *Journal of Endocrinology*, 186, 411–427.
- Barkat-Defradas, M., Busseuil, C., Chauvy, O., Hirsch, F., Fauth, C., Revis, J., & Amy de la Bretéque, B. (2012). Dimension esthétique des voix normales et dysphoniques: Approches perceptives et acoustique. *TIPA* 28.
- Barkat-Defradas, M., Fauth, C., Didirkova, F., Amy de la Breteque, B., Hirsch, F., Dodane, C., & Sauvage, J. (2015). Dysphonia is beautiful: A perceptual and acoustic analysis of vocal roughness. *International Congress of Phonetic Sciences*, 18th ICPhS, Glasgow 10–14 August 2015, Scotland, UK.
- Borkowska, B., & Pawlowski, B. (2011). Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour*, 82(1), 55–59.
- Bovet, J. (2018). The evolution of feminine beauty. In Z. Kapoula et al. (Eds.), *Exploring Transdisciplinarity in Art and Sciences* (pp. 327–348). Springer International Publishing AG, Part of Springer Nature.
- Bovet, J. (2019). Evolutionary theories and men's preferences for women's waist-to-hip ratio: Which hypotheses remain? *A Systematic Review. Frontiers in Psychology*, 10, <https://doi.org/10.3389/fpsyg.2019.01221>.
- Bovet, J., & Raymond, M. (2015). Preferred women's waist-to-hip ratio variation over the last 2, 500 years. *PLoS One*, 10, e0123284.
- Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., et al. (2010). Vocal attractiveness increases by averaging. *Current Biology*, 20(2), 116–120.
- Bruckert, L., Lienard, J.-S., Lacroix, A., Kreutzer, M., & Leboucher, G. (2006). Women use voice parameters to assess men's characteristics. *Proceedings of the Royal Society B: Biological Sciences*, 273(1582), 83–89.
- Bryant, Gregory A., & Haselton, Martie G. (2009). Vocal cues of ovulation in human females. *Biology Letters*, 5, 12–15.
- Buss, D. M. (1989). Sex differences in human mate preferences: Evolutionary hypotheses tested in 37 cultures. *Behavioral and Brain Sciences*, 12(1), 1–14.

- 577 Charbonnier, G. (1959). Entretien avec Claude Levi-Strauss. <https://www.jpbu.com/philos/.../Levi->
578 [Strauss_Charbonnier_Culture-langage.rtf](https://www.jpbu.com/philos/.../Levi-Strauss_Charbonnier_Culture-langage.rtf).
- 579 Chomsky, A. N. (1975). *Reflections on Language*. New York: Pantheon Books.
- 580 Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour*, *60*(6), 773–780.
- 581 Collins, S. A., & Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal*
582 *Behaviour*, *65*(5), 997–1004.
- 583 Darwin, C. (1871). *The descent of man, and selection in relation to sex*. London, UK: John Murray.
- 584 Dunbar, R., Duncan, N., & Nettle, D. (1995). Size and structure of freely forming conversational
585 groups. *Human nature*, *6*(1), 67–78.
- 586 Evans, S., Neave, N., Wakelin, D., & Hamilton, C. (2008). The relationship between testosterone
587 and vocal frequencies in human males. *Physiology and Behavior*, *93*, 783–788.
- 588 Fairbanks, G. (1960). *Voice and articulation*. J. Cotler Books, 2nd revised edition.
- 589 Fant, G. (1960). *Acoustic Theory of Speech Production*. Berlin: De Gruyter.
- 590 Feinberg, D. R., DeBruine, L. M., Jones, B. C., & Little, A. C. (2008a). Correlated preferences for
591 men's facial and vocal masculinity. *Evolution and Human Behavior*, *29*(4), 233–241.
- 592 Feinberg, D. R., DeBruine, L. M., Jones, B. C., & Perrett, D. I. (2008b). The role of femininity and
593 averageness of voice pitch in aesthetic judgments of women's voices. *Perception*, *37*(4), 615–623.
- 594 Feinberg, D. R., Jones, B. C., DeBruine, L. M., Moore, F. R., Law Smith, M. J., Cornwell, R. E.,
595 et al. (2005). The voice and face of woman: One ornament that signals quality? *Evolution and*
596 *Human Behavior*, *26*(5), 398–408.
- 597 Feinberg, D. R., Jones, B. C., Law Smith, M. J., Moore, F. R., DeBruine, L. M., Cornwell, R. E.,
598 et al. (2006). Menstrual cycle, trait estrogen level, and masculinity preferences in the human
599 voice. *Hormones and Behavior*, *49*(2), 215–222.
- 600 Fischer, J., Semple, S., Fickenscher, G., Jürgens, R., Kruse, E., Heistermann, M., et al. (2011). Do
601 women's voices provide cues of the likelihood of ovulation? The importance of sampling regime.
602 *PLoS One*, *6*(9).
- 603 Fischer, A. H., & Manstead, A. S. (2000). The relation between gender and emotions in different
604 cultures. *Gender and emotion: Social psychological perspectives*, *1*, 71–94.
- 605 Fisher, M. L., & Voracek, M. (2006). The shape of beauty: determinants of female physical attrac-
606 tiveness. *Journal of Cosmetic Dermatology*, *5*(2), 190–4.
- 607 Fitch, W. T., & Reby, D. (2001). The descended larynx is not uniquely human. *Proceedings of the*
608 *Royal Society of London. Series B: Biological Sciences*, *268*(1477), 1669–1675.
- 609 Fitch, W. T. (1997). Vocal tract length and format frequency dispersion correlated with body size
610 in rhesus macaques. *The Journal of the American Society of America*, *102*, 1213–1222.
- 611 Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study
612 using magnetic resonance imaging. *The Journal of the American Society of America*, *106*, 1511–
613 1522.
- 614 Fraccaro, P. J., Jones, B. C., Vukovic, J., Smith, F. G., Watkins, C. D., Feinberg, D. R., et al. (2011).
615 Experimental evidence that women speak in a higher voice pitch to men they find attractive.
616 *Journal of Evolutionary Psychology*, *9*(1), 57–67.
- 617 Fraccaro, P. J., O'Connor, J. J. M., Re, D. E., Jones, B. C., DeBruine, L. M., & Feinberg, D. R.
618 (2013). Faking it: Deliberately altered voice pitch and vocal attractiveness. *Animal Behaviour*,
619 *85*(1), 127–136.
- 620 Geary, D. C., Vigil, J., & Byrd-Craven, J. (2004). Evolution of human mate choice. *Journal of Sex*
621 *Research*, *41*(1), 27–42.
- 622 Hall, J. A. (1978). Gender effects in decoding nonverbal cues. *Psychological Bulletin*, *85*, 845–857.
- 623 Handson, H. (1997). Glottal characteristics of female speakers: Acoustic correlates. *The Journal of*
624 *the American Society of America*, *101*(1), 466–81.
- 625 Hanson, H. M. (1997). Glottal characteristics of female speakers: Acoustic correlates. *The Journal*
626 *of the American Society of America*, *101*(1), 466–481.
- 627 Harris, A. C. (1994). Ethnicity as a determinant of sex role identity: A replication study of item
628 selection for the Bem Sex Role Inventory. *Sex Roles*, *31*, 241–273.

- 629 Haselton, M. G., Mortezaie, M., Pillsworth, E. G., Bleske-Rechek, A., & Frederick, D. A. (2007).
 630 Ovulatory shifts in human female ornamentation: Near ovulation, women dress to impress. *Hor-*
 631 *mones and Behavior*, *51*, 40–45.
- 632 Higgins, M. B., & Saxman, J. H. (1989). Variations in vocal frequency perturbation across the
 633 menstrual cycle. *Journal of Voice*, *3*, 233–243.
- 634 Hill, A. K., Hunt, J., Welling, L. L. M., Cárdenas, R. A., Rotella, M. A., Wheatley, J. R., et al.
 635 (2013). Quantifying the strength and form of sexual selection on men's traits. *Evolution and*
 636 *Human Behavior*, *34*(5), 334–341.
- 637 Hodges-Simeon, C. R., Gaulin, S. J. C., & Puts, D. A. (2010). Different vocal parameters predict
 638 perceptions of dominance and attractiveness. *Human Nature*, *21*(4), 406–427.
- 639 Holmberg, E. B., Hillman, R. E., Perkell, J. S., Guiod, P. C., & Goldman, S. L. (1995). Compar-
 640 isons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice.
 641 *Journal of Speech, Language, and Hearing Research*, *38*, 1212–1223.
- 642 Hughes, S. M., Dispenza, F., & Gallup, G. G. (2004). Ratings of voice attractiveness predict sexual
 643 behavior and body configuration. *Evolution and Human Behavior*, *25*, 295–304.
- 644 Hughes, S. M., Farley, S. D., & Rhodes, B. C. (2010). Vocal and physiological changes in response
 645 to the physical attractiveness of conversational partners. *Journal of Nonverbal Behavior*, *34*(3),
 646 155–167.
- 647 Hughes, S. M., Mogilski, J. K., & Harrison, M. A. (2014). The perception and parameters of
 648 intentional voice manipulation. *Journal of Nonverbal Behavior*, *38*(1), 107–127.
- 649 Hughes, S. M., Pastizzo, M. J., & Gallup, G. G. (2008). The sound of symmetry revisited: Subjective
 650 and objective analyses of voice. *Journal of Nonverbal Behavior*, *32*(2), 95–108.
- 651 Jones, B. C., Feinberg, D. R., DeBruine, L. M., Little, A. C., & Vukovic, J. (2010). A domain-
 652 specific opposite-sex bias in human preferences for manipulated voice pitch. *Animal Behaviour*,
 653 *79*(1), 57–62.
- 654 Jones, B. C., Hahn, A. C., Fisher, C. I., Wang, H., Kandrik, M., Han, C., et al. (2018). No compelling
 655 evidence that preferences for facial masculinity track changes in women's hormonal status. *Psy-*
 656 *chological Science*, *29*(6), 10.
- 657 Jünger, J., Motta-Mena, N. V., Cardenas, R., Bailey, D., Rosenfield, K. A., Schild, C., et al. (2018).
 658 Do women's preferences for masculine voices shift across the ovulatory cycle? *Hormones and*
 659 *Behavior*, *106*, 122–134.
- 660 Kordsmeyer, T. L., Hunt, J., Puts, D. A., Ostner, J., & Penke, L. (2018). The relative importance of
 661 intra- and intersexual selection on human male sexually dimorphic traits. *Evolution and Human*
 662 *Behavior*, *39*(4), 424–436.
- 663 Landrine, H. (1985). Race x class stereotypes of women. *Sex Roles*, *13*, 65–75.
- 664 Laver, J. (1980). *The phonetic description of voice quality*, (vol. 2009, 1st edn). Cambridge Uni-
 665 versity Press.
- 666 Laver, J. (1980). *The phonetic description of voice quality*. Cambridge University Press.
- 667 Laver, J. (1994). *Principles of phonetics*. Cambridge University Press.
- 668 Leaderbrand, K., Dekam, J., Morey, A., & Tuma, L. (2008). The effects of voice pitch on perceptions
 669 of attractiveness: Do you sound hot or not? *Winona State University Psychology Student Journal*,
 670 *6*.
- 671 Léon, P. (1981). BB ou la voix charmeuse, petite fille et coquette. *Studia Phonetica*, *18*, 159–171.
- 672 Leongomez, J. D., Binter, J., Kubicova, L., Stolařova, P., Klapilova, K., Havlíček, J., et al. (2014).
 673 Vocal modulation during courtship increases perceptivity even in naive listeners. *Evolution and*
 674 *Human Behavior*, *35*(6), 489–496.
- 675 Lieberman, P. (1963). Some acoustic measures of the fundamental periodicity of normal and patho-
 676 logic larynges. *The Journal of the American Society of America*, *35*(3), 344–353.
- 677 Liu, X., & Xu, Y. (2011). What makes a female voice attractive? In *Proceedings of ICPhS. Hong-*
 678 *Kong* (pp. 1274–1277).
- 679 Marcinkowska, U. M., Galbarczyk, A., & Jasienska, G. (2018). La donna è mobile? Lack of cyclical
 680 shifts in facial symmetry, and facial and body masculinity preferences—A hormone based study.
 681 *Psychoneuroendocrinology*, *88*, 47–53.

- 682 Miller, G. F., Tybur, J., & Jordan, B. (2007). Ovulatory cycle effects on tip earnings by lap-dancers:
 683 Economic evidence for human estrus? *Evolution and Human Behavior*, 6, 375–381.
- 684 Moore, P., & Thomson, C. L. (1965). Comments on physiology of hoarseness. *Archives of Oto-*
 685 *laryngology Head and Neck Surgery*, 81(1), 97–102. [https://doi.org/10.1001/archotol.1965.](https://doi.org/10.1001/archotol.1965.00750050102022)
 686 [00750050102022](https://doi.org/10.1001/archotol.1965.00750050102022).
- 687 Morton, E. S. (1977). On the occurrence and significance of motivation-structural rules in some
 688 bird and mammal sounds. *The American Naturalist*, 111(981), 855–869.
- 689 Noble, W., & Davidson, I. (1996). *Human evolution, language and mind*. Cambridge: Cambridge
 690 University Press.
- 691 O'Connor, J. J. M., Fraccaro, P. J., Pisanski, K., Tigue, C. C., & Feinberg, D. R. (2013). Men's
 692 preferences for women's femininity in dynamic cross-modal stimuli. *PLoS One*, 8(7), e69531.
- 693 O'Connor, J. J., Re, D. E., & Feinberg, D. R. (2011). Voice pitch influences perceptions of sexual
 694 infidelity. *Evolutionary Psychology*, 9(1), 147470491100900109.
- 695 Oguchi, T., & Kikuchi, H. (1997). Voice and interpersonal attraction. *Japanese Psychological*
 696 *Research*, 39(1), 56–61.
- 697 Ohala, J. J. (1984). An ethological perspective on common cross- language utilisation of F0 of
 698 voice. *Phonetica*, 41(1), 1–16.
- 699 Pipitone, R. N., & Gallup, G. G. (2008). Women's voice attractiveness varies across the menstrual
 700 cycle. *Evolution and Human Behavior*, 29(4), 268–274.
- 701 Pisanski, K., Bhardwaj, K., & Reby, D. (2018). Women's voice pitch lowers after pregnancy. *Evo-*
 702 *lution and Human Behavior*, 39(4), 457–463.
- 703 Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J. M., Röder, S., Andrews, P. W., et al.
 704 (2014). Vocal indicators of body size in men and women: A meta-analysis. *Animal Behaviour*,
 705 95, 89–99.
- 706 Pisanski, K., & Rendall, D. (2011). The prioritization of voice fundamental frequency or formants in
 707 listeners' assessments of speaker size, masculinity and attractiveness. *The Journal of the American*
 708 *Society of America*, 129(4), 2201–2212.
- 709 Polce-Lynch, M., Myers, B. J., Kilmartin, C. T., Forssmann-Falck, R., & Kliewer, W. (1998). Gender
 710 and age patterns in emotional expression, body image, and self-esteem: A qualitative analysis.
 711 *Sex Roles*, 38(11–12), 1025–1048.
- 712 Puts, D. A. (2005). Mating context and menstrual phase affect women's preferences for male voice
 713 pitch. *Evolution and Human Behavior*, 26(5), 388–397.
- 714 Puts, D. A. (2010). Beauty and the beast: Mechanisms of sexual selection in humans. *Evolution*
 715 *and Human Behavior*, 31(3), 157–175.
- 716 Puts, D. A., Bailey, D. H., Cárdenas, R. A., Burriss, R. P., Welling, L. L. M., Wheatley, J. R., et al.
 717 (2013). Women's attractiveness changes with estradiol and progesterone across the ovulatory
 718 cycle. *Hormones and Behavior*, 63(1), 13–19.
- 719 Puts, D. A., Barndt, J. L., Welling, L. L. M., Dawood, K., & Burriss, R. P. (2011). Intrasexual com-
 720 petition among women: Vocal femininity affects perceptions of attractiveness and flirtatiousness.
 721 *Personality and Individual Differences*, 50(1), 111–115.
- 722 Puts, D. A., Gaulin, S. J. C., & Verdolini, K. (2006). Dominance and the evolution of sexual
 723 dimorphism in human voice pitch. *Evolution and Human Behavior*, 27(4), 283–296.
- 724 Renoult, J. P., Bovet, J., & Raymond, M. (2016). Beauty is in the efficient coding of the beholder.
 725 *Royal Society Open Science*, 3(3), 160027.
- 726 Re, D. E., O'Connor, J. J. M., Bennett, P. J., & Feinberg, D. R. (2012). Preferences for very low
 727 and very high voice pitch in humans. *PLoS One*, 7(3), e32719.
- 728 Revis, J. (2017). *La voix et soi: Ce que notre voix dit de nous*. France: Solal.
- 729 Riding, D., Lonsdale, D., & Brown, B. (2006). The effects of average fundamental frequency and
 730 variance of fundamental frequency on male vocal attractiveness to women. *Journal of Nonverbal*
 731 *Behavior*, 30(2), 55–61.
- 732 Saxton, T. K., Caryl, P. G., & Craig R. S. (2006). Vocal and facial attractiveness judgments of
 733 children, adolescents and adults: The ontogeny of mate choice. *Ethology*, 112(12), 1179–1185.

- 734 Saxton, T. K., Debruine, L. M., Jones, B. C., Little, A. C., & Roberts, S. C. (2009). Face and voice
735 attractiveness judgments change during adolescence. *Evolution and Human Behavior*, 30(6),
736 398–408.
- 737 Sebesta, P., Kleisner, K., Turecek, P., Kočnar, T., Akoko, R. M., Trebicky, V., et al. (2017). Voices of
738 Africa: Acoustic predictors of human male vocal attractiveness. *Animal Behaviour*, 127, 205–211.
- 739 Shirazi, T. N., Puts, D. A., & Escasa-Dorne, M. J. (2018). Filipino women's preferences for male
740 voice pitch: Intra-individual, life history, and hormonal predictors. *Adaptive Human Behavior
741 and Physiology*, 4(2), 188–206.
- 742 Simmons, L. W., Peters, M., & Rhodes, G. (2011). Low pitched voices are perceived as masculine
743 and attractive but do they predict semen quality in men? *PLoS One*, 6(12).
- 744 Singh, D. (1993). Adaptive significance of female physical attractiveness: Role of waist-to-hip ratio.
745 *Journal of Personality and Social Psychology*, 65(2), 293–307.
- 746 Singh, D., & Young, R. K. (2001). Body weight, waist-to-hip ratio, breasts, and hips: Role in
747 judgments of female attractiveness and desirability for relationships. *Ethology and Sociobiology*,
748 16(6), 483–507.
- 749 Skrinda, I., Krama, T., Kecko, S., Moore, F. R., Kaasik, A., Meija, L., et al. (2014). Body height,
750 immunity, facial and vocal attractiveness in young men. *Naturwissenschaften*, 101(12), 1017–
751 1025.
- 752 Södersten, M., Lindestad, P. A., & Hammarberg, B. (1991). Vocal fold closure, perceived breathi-
753 ness, and acoustic characteristics in normal adult speakers. In J. Gauffin & B. Hammarberg (Eds.),
754 *Vocal fold physiology: Acoustic, perceptual, and physiological aspects of voice mechanisms* (pp.
755 217–224).
- 756 Sughira, Y., & Katsurada, E. (1999). Masculinity and femininity in Japanese culture. *Sex Roles*,
757 40(7/18), 635–646.
- 758 Sugiyama, L. S. (2015). Physical attractiveness in adaptationist perspective. In D. M. Buss (Ed.),
759 *The handbook of evolutionary psychology* (pp. 292–343). Wiley.
- 760 Suire, A., Raymond, M., & Barkat-Defradas, M. (2019). Male vocal quality and its
761 relation to females' preferences. *Evolutionary Psychology*, 1–12, [https://doi.org/10.1177/
762 1474704919874675](https://doi.org/10.1177/1474704919874675).
- 763 Suire, A., Tognetti, A., Durand, V., Raymond, M., & Barkat-Defradas, M. (2020). The influence
764 of sexual orientation and circulating testosterone levels on speech acoustic features. *Archives of
765 Sexual Behavior*, 10, 1–9. <https://doi.org/10.1007/s10508-020-01665-3>.
- 766 Suire, A., Raymond, M., & Barkat-Defradas, M. (2018). Vocal behavior within competitive and
767 courtship contexts and its relation to mating success in humans. *Evolution and Human Behavior*,
768 39, 684–691.
- 769 Titze, I. R. (1994). *Principles of voice production*. Englewood Cliffs, N.J.: Prentice Hall.
- 770 Tognetti, A., Durand, V., Barkat-Defradas, M., & Hopfensitz, A. (2019). Does he sound cooperative?
771 Acoustic correlates of cooperativeness. *British Journal of Psychology*, 1–17. [https://doi.org/10.
772 1111/bjop.12437](https://doi.org/10.1111/bjop.12437).
- 773 Tomassello, M. (1999). *The cultural origins of human cognition*. Harvard University Press.
- 774 Traunmüller, H., & Eriksson, A. (1995). *The frequency range of the voice fundamental in the speech
775 of male and female adults*. Unpublished manuscript.
- 776 Travers, L. M. (2017). Runaway selection. In J. Vonk & T. Shackelford (Eds.), *Encyclopedia of
777 animal cognition and behavior*. Springer.
- 778 Tsantani, M. S., Belin, P., Paterson, H. M., & McAleer, P. (2016). Low vocal pitch preference drives
779 first impressions irrespective of context in male voices but not in female voices. *Perception*, 45(8),
780 946–963.
- 781 Tuomi, S. K., & Fisher, J. E. (1979). Characteristics of simulated sexy voice. *Folia Phoniatrica and
782 Logopaedica*, 31(4), 242–249.
- 783 Vaissière, J. (2015). *La phonétique*. Paris: Presses Universitaires de France.
- 784 van Bezooijen, R. (1995). Sociocultural aspects of pitch differences between Japanese and Dutch
785 women. *Language and Speech*, 38(3), 253–265.

- 786 van Borsel, J., Janssens, J., & De Bodt, M. (2009). Breathiness as a feminine voice characteristic:
787 A perceptual approach. *Journal of Voice*, 23(3), 291–294.
- 788 Vorperian, H. K., Kent, R. D., Lindstrom, M. J., Kalina, C. M., Gentry, L. R., & Yandell, B. S. (2005).
789 Development of vocal tract length during early childhood: A magnetic resonance imaging study.
790 *The Journal of the American Society of America*, 117, 338–350.
- 791 Vukovic, J., Feinberg, D. R., Jones, B. C., DeBruine, L. M., Welling, L. L. M., Little, A. C.,
792 et al. (2008). Self-rated attractiveness predicts individual differences in women’s preferences for
793 masculine men’s voices. *Personality and Individual Differences*, 45(6), 451–456.
- 794 Xu, Y., Lee, A., Wu, W.-L., Liu, X., & Birkholz, P. (2013). Human vocal attractiveness as signaled
795 by body size projection. *PLoS One*, 8(4), e62397.
- 796 Zahavi, A. (1975). Mate selection—A selection for handicap. *Journal of Theoretical Biology*, 53,
797 205–214.
- 798 Zheng, Y., Compton, B. J., Heyman, G. D., & Jiang, Z. (2020). Vocal attractiveness and voluntarily
799 pitch-shifted voices. *Evolution and Human Behavior*.
- 800 Zuckerman, M., Driver, R. E. (1989). What sounds beautiful is good: The vocal attractiveness
801 stereotype. *Journal of nonverbal behavior*, 13, 67–82, <https://doi.org/10.1007/BF00990791>
- 802 Zuckerman, M., & Miyake, K. (1993). The attractive voice: What makes it so? *Journal of Nonverbal*
803 *Behavior*, 17(2), 119–135.

1
2

Part II
Voice

UNCORRECTED PROOF

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Voice Attractiveness	
Series Title		
Chapter Title	What Does It Mean for a Voice to Sound “Normal”?	
Copyright Year	2020	
Copyright HolderName	Springer Nature Singapore Pte Ltd.	
Corresponding Author	Family Name	Kreiman
	Particle	
	Given Name	Jody
	Prefix	
	Suffix	
	Role	
	Division	Department of Head and Neck Surgery and Linguistics
	Organization	University of California
	Address	Los Angeles, CA, USA
	Email	jkreiman@ucla.edu
Author	Family Name	Ausmann
	Particle	
	Given Name	Anita
	Prefix	
	Suffix	
	Role	
	Division	Department of Head and Neck Surgery and Linguistics
	Organization	University of California
	Address	Los Angeles, CA, USA
	Email	ausmannanita@gmail.com
Author	Family Name	Gerratt
	Particle	
	Given Name	Bruce R.
	Prefix	
	Suffix	
	Role	
	Division	Department of Head and Neck Surgery
	Organization	University of California
	Address	Los Angeles, CA, USA
	Email	bgerratt@ucla.edu
Abstract	<p>It is rather unclear what is meant by “normal” voice quality, just as it is often unclear what is meant by “voice quality” in general. To shed light on this matter, listeners heard 1-sec sustained vowels produced by 100 female speakers, half of whom were recorded as part of a clinical voice evaluation and half of whom were undergraduate students who reported no vocal disorder. Listeners compared 20 voices at a time in a series of sort-and-rate trials, ordering the samples on a line according to the severity of perceived pathology. Any voices perceived as normal were placed in a box at one end of the line. Judgments of</p>	

“normal” versus “not-normal” status were at chance. Listeners were relatively self-consistent, but disagreed with one another, especially about what counts as normal. Agreement was better, but still limited, about what counts as “not normal.” Strategies for separating “normal” from “not normal” differed widely across individual listeners, as did strategies for determining how much a given voice deviated from normal. However, acoustic modeling of listeners’ responses showed that several acoustic measures—F0, F1 and F2, and F0 coefficient of variation—appeared more often than others as significant predictors of both categorical judgments and of scalar normalness ratings. These variables did not account for most of the variance in these analyses, and did not appear together in the perceptual models for even half of the listeners, but they did appear individually in most analyses, suggesting that in practice the concept of “normal” may have some small core of meaning based on F0 and vowel quality. Thus, the answer to our initial question of what it means for a voice to sound normal is a complex one that depends on the listener, the context, the purpose of the judgment, and other factors as well as on the voice.

Keywords

Voice quality - Normal voice - Dysphonia - Voice perception - Voice disorders - Listener - Agreement

Chapter 5

What Does It Mean for a Voice to Sound “Normal”?



Jody Kreiman, Anita Auszmann, and Bruce R. Gerratt

Abstract It is rather unclear what is meant by “normal” voice quality, just as it is often unclear what is meant by “voice quality” in general. To shed light on this matter, listeners heard 1-sec sustained vowels produced by 100 female speakers, half of whom were recorded as part of a clinical voice evaluation and half of whom were undergraduate students who reported no vocal disorder. Listeners compared 20 voices at a time in a series of sort-and-rate trials, ordering the samples on a line according to the severity of perceived pathology. Any voices perceived as normal were placed in a box at one end of the line. Judgments of “normal” versus “not-normal” status were at chance. Listeners were relatively self-consistent, but disagreed with one another, especially about what counts as normal. Agreement was better, but still limited, about what counts as “not normal.” Strategies for separating “normal” from “not normal” differed widely across individual listeners, as did strategies for determining how much a given voice deviated from normal. However, acoustic modeling of listeners’ responses showed that several acoustic measures—F0, F1 and F2, and F0 coefficient of variation—appeared more often than others as significant predictors of both categorical judgments and of scalar normalness ratings. These variables did not account for most of the variance in these analyses, and did not appear together in the perceptual models for even half of the listeners, but they did appear individually in most analyses, suggesting that in practice the concept of “normal” may have some small core of meaning based on F0 and vowel quality. Thus, the answer to our initial question of what it means for a voice to sound normal is a complex one that depends on the listener, the context, the purpose of the judgment, and other factors as well as on the voice.

J. Kreiman (✉) · A. Auszmann
Department of Head and Neck Surgery and Linguistics, University of California,
Los Angeles, CA, USA
e-mail: jkreiman@ucla.edu

A. Auszmann
e-mail: auszmannanita@gmail.com

B. R. Gerratt
Department of Head and Neck Surgery, University of California, Los Angeles, CA, USA
e-mail: bgerratt@ucla.edu

© Springer Nature Singapore Pte Ltd. 2020

B. Weiss et al. (eds.), *Voice Attractiveness*, Prosody, Phonology and Phonetics,
https://doi.org/10.1007/978-981-15-6627-1_5

89

24 **Keywords** Voice quality · Normal voice · Dysphonia · Voice perception · Voice
25 disorders · Listener · Agreement

26 **5.1 Introduction**

27 The voice literature provides surprisingly little insight into what it means for a voice
28 to be “normal,” despite the fact that much depends on the concept of a normal voice.
29 Many studies have shown that a listener’s perception of vocal abnormality may
30 lead to negative assessments of the personality, health, intelligence, or social desir-
31 ability/social attractiveness of the speaker. For example, Amir and Levine-Yundof
32 (2013) found significant differences between speakers with voice disorders and non-
33 dysphonic speakers with respect to listeners’ judgments of attractiveness, agreeable-
34 ness, reliability, potency, aggressiveness, and tenseness. Similarly, Maryn and Debo
35 (2015) found a correlation of $r = 0.85$ between clinicians’ ratings of severity of dys-
36 phonia and naïve listeners’ ratings of healthiness. Similar results have been reported
37 for adult or child listeners, and for expert and naïve judges (Table 5.1). Results also
38 appear to apply to both child and adult speakers, and are robust cross-culturally (e.g.,
39 Altenberg & Ferrand, 2006; Irani et al., 2014). These kinds of effects can cause
40 embarrassment and interfere with job performance; in the worst case, they can lead
41 to reduced career opportunities and social isolation.

42 In clinical settings, a clear understanding of “normal” voice would seem to under-
43 lie the entire diagnosis-and-treatment enterprise. A sense that a voice does not sound
44 normal leads patients to initiate treatment, and “normal” serves as a target for deter-
45 mining when therapy is complete. Studies of treatment efficacy logically depend on
46 defining a normal voice as a target, and the practice of establishing normative values
47 for instrumental measures of voice assumes that “normal” has at least a relatively
48 constant meaning.

49 Despite the importance of “normal” in understanding voice and voice disorders,
50 authors discussing the nature of normal voice have typically emphasized the difficulty
51 of pinning down exactly what it is, echoing Sundberg’s (1988) lament that everyone
52 knows what voice is until they try to be specific. Discussions of normal quality
53 have focused on two main themes. The first and more common one describes a
54 normal voice as one that properly presents the person speaking—their age, sex,
55 emotional state—and that adequately meets the speaker’s occupational and social
56 communication needs (e.g., Behlau & Murry, 2012; Dehqan et al., 2010; Greene &
57 Mathieson, 1992; Johnson et al., 1965; Aronson & Bless, 2009). Such definitions
58 emphasize the functionality of a voice. For example, Greene and Mathieson (1992)
59 wrote:

60 The simplest definition of normal voice is it is ‘ordinary’: it is inconspicuous with nothing
61 out of the ordinary in its sound. To achieve this standard of acceptability, the voice must
62 be loud enough to be heard, and appropriate for the age and sex of the speaker. It must be
63 reasonably pleasing to the ear of the listener, modulated and clear, not droning and flat or
64 hoarse and breathy. It must be appropriate to the context and not too loud or assertive. (p. 43)

Editor Proof

Table 5.1 Representative studies showing perceptual and social sequelae of perceived disordered voice or speech

Speakers	Listeners	Attribute judged	Result	References
Normal and hypernasal children	Children	Social acceptance	Negative responses increased with increasing hypernasality	Blood and Hymen (1977)
Normal and hypernasal children	Children	Social acceptance	Even mild-to-moderate hypernasality decreased social acceptance	Watterson et al., (2013)
Normal and dysphonic female adolescents	Teachers	Personality	Voice disorders increased negative perceptions	Zacharias et al., (2013)
Normal and dysphonic adult females	Adults; monolingual and bilingual, younger and older	Personality	Even mild voice disorders led to negative impressions, for all listener groups	Altenberg and Ferrand (2006)
Normal and dysphonic adult females	Adults	Personality, attractiveness	Nasality and breathy/harsh quality both associated with worse perceptions	Blood et al., (1979)
Normal, dysphonic, and hypernasal females	Students with and without information about voice disorders	Social desirability	Ratings were more negative for speakers with voice disorders	Lallh and Rochet (2000)
Normal and dysphonic adults and children	Adult SLPs; naïve listeners	Healthiness	Even slight dysphonia produced the perception of unhealthiness	Maryn and Debo (2015)
Normal and dysphonic speakers of Hebrew	Young and older adults	Personality	Dysphonia associated with negative perceptions, for women more than for men	Amir and Levine-Yundof (2013)

65 It follows from this definition that standards and judgments will vary across lis-
 66 teners and contexts. For example, Moore (1971) wrote:

67 It is apparent that the voice is abnormal for a particular individual when he or she judges it
 68 to be so regardless of the circumstances. Judgment implies a set of standards that are learned
 69 through experience and that are related to the judge's own aesthetic and cultural criteria.
 70 Judgment also implies that standards are not fixed, that there is opportunity for more than
 71 one conclusion. This flexibility in determining the defectiveness of voices does not alter the
 72 validity of the basic definition of voice disorders, but it does underscore the observation that
 73 vocal standards are culturally based and environmentally determined. (p. 535)

74 However, to our knowledge the nature and extent of this variability have not been
 75 studied, nor have the factors conditioning variability in perceived vocal abnormality.

76 A second definitional approach emphasizes physical normalness, without partic-
 77 ular concern for vocal quality or for use of the voice in communication. For example,
 78 normal voice can be characterized as the acoustic product of a normal vocal tract
 79 that is functioning normally (Mathieson, 2000) or as a voice produced by a speaker
 80 with no current or previous voice complaint and that passes a perceptual evaluation
 81 by a speech-language pathologist (Bonilha & Deliyski, 2008).

82 To our knowledge, no empirical data exist in support of either of these views. In the
 83 face of the importance a perceived voice disorder can have for a speaker, clinicians
 84 and scientists have proceeded as if “normal” unambiguously exists. For example,
 85 numerous studies propose algorithms devised to automatically separate normal from
 86 pathological phonation, arguing that such algorithms bring needed objectivity to
 87 clinical voice evaluation (e.g., Arias-Londoño et al., 2011; Orozco-Arroyave et al.,
 88 2015; Wang et al., 2011; Moro-Velázquez et al., 2016). “Normal” in these studies
 89 remains an unexamined concept, and algorithms typically show good classifica-
 90 tion accuracy (usually >90% correct), suggesting this approach is not unreasonable.
 91 Similarly, many more studies have reported normative values for acoustic (e.g., Goy,
 92 Fernandes et al., 2013; Wuyts et al., 2002), physiological (e.g., Xue & Hao, 2006),
 93 and/or aerodynamic measures of voice (e.g., Lewandowski et al., 2017), again imply-
 94 ing that it is possible to define “normal” as a quality with clear boundaries. The voice
 95 literature thus presents a paradox. Clinical concerns combined with the demonstrated
 96 social and personal importance of sounding normal lead researchers to design studies
 97 that assume a clear boundary between normal and not-normal phonation, while at the
 98 same time arguing that no such boundaries exist in theory, all of this in the absence
 99 of empirical evidence about what sounds normal or not normal to listeners.

100 This study is intended to address this situation. Our goals are to gather listeners’
 101 assessments of the extent to which voices sound normal, and to seek insight into the
 102 factors that determine whether a voice sounds better or worse to a particular listener.

5.2 Methods

5.2.1 *Speakers and Voice Samples*

The voices of 100 female speakers were used in this experiment. Females (as opposed to males) were selected for this preliminary study because of recent research interest in the perception of normal versus abnormal female voice quality, particularly with respect to vocal fry and “creaky voice” (Yuasa, 2010; Anderson et al., 2014; Oliveira et al., 2016). Fifty voice samples were drawn from an existing database of recordings of speakers who had a diagnosis from an otolaryngologist (“not normal”). Voices were unselected with respect to diagnoses, which included functional and neurogenic disorders, mass lesions, reflux, and age-related dysphonia. Samples ranged from extremely mild to very severe vocal pathology. An additional 50 voices were drawn from the UCLA Speaker Variability Database (Keating et al. 2018), which includes multiple voice samples from over 200 male and female UCLA undergraduate students, all of whom reported no history of voice or speech complaints (“normal”). Note that although voices were categorized as \pm normal based on diagnostic status, no assumptions were made about the normal or abnormal quality of the voices, and no attempt was made to select “normal” or “not-normal” voices that sounded more or less normal, beyond informally ensuring that the “not-normal” samples represented a broad range of severity of perceived pathology.

All speakers sustained the vowel /a/ as part of their recording sessions, and all were recorded with a Brüel and Kjær 1/2” microphone. Steady-state vowels were studied rather than continuous speech, to allow listeners to focus on voice quality and not on articulation or native/nonnative status of the speakers. Previous studies (e.g., Gerratt et al., 2016) have shown negligible effects of stimulus type on quality assessment. Samples were directly digitized at a 20kHz (clinical samples) or 22kHz (normal samples) sampling rate, edited to 1 s duration, and then downsampled to 10kHz prior to acoustic analyses and testing.

5.2.2 *Listeners and Listening Task*

Stimuli were assembled into blocks of 20 voices each, which in turn were assembled into five sets of nine trials (each trial comprising one 20-voice block), such that across the five sets of trials, every voice was compared at least once to every other voice and every voice received a total of 90 ratings. Each listener heard 9, 20-voice trials, for a total of 180 judgements/listener: each stimulus voice was judged at least once/listener, with 80 voices repeated in 2 different trials so that test–retest reliability could be assessed. (No voices were repeated within a single trial.)

All experimental procedures were approved by the UCLA Institutional Review Board. Ten UCLA students and staff (aged 18–68; mean age = 21.5; sd = 9.67) heard each set of trials, for a total of 50 listeners. All listeners reported normal

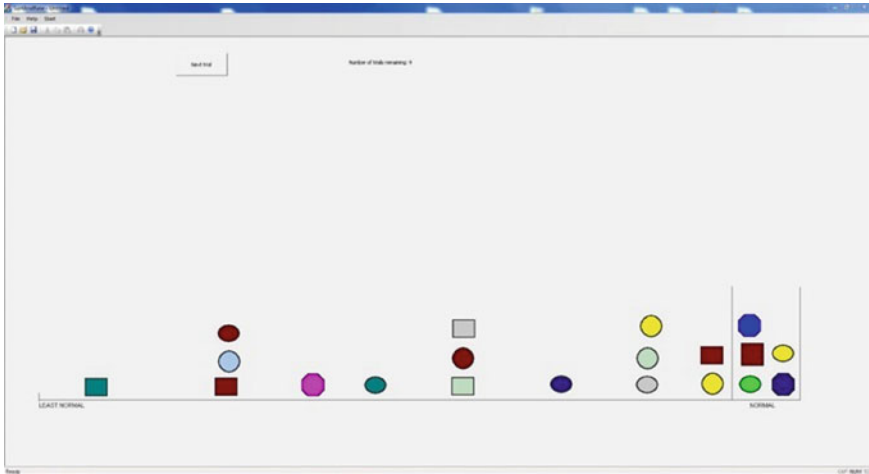


Fig. 5.1 The testing interface for the sort-and-rate task. Listeners played each voice by clicking its icon, and then dragged the icon to indicate (1) whether the voice sounded normal, in which case the icon was placed in the box on the right and (2) if it did not sound normal, how close to normal it sounded. The most abnormal-sounding voices were placed toward the left end of the line; those that approached normal were placed near the box

141 hearing and received course credit in return for their participation. Clinicians were
 142 not targeted separately during subject selection because evidence indicates they do
 143 not differ significantly from naïve listeners when judging the severity of dysphonia
 144 (Eadie et al. 2010).

145 Subjects heard the stimuli over Etymotic insert earphones (model ER-1) at a
 146 comfortable constant listening level. The testing interface is shown in Fig. 5.1. Each
 147 icon in the figure represents a single voice token, randomly assigned to that icon.
 148 Listeners played each voice by clicking its icon, then dragged the icon to a line to
 149 indicate (1) whether the voice sounded normal, in which case the icon was placed in
 150 the box on the right end of the line and (2) if it did not sound normal, how close to
 151 normal it sounded (a visual sort-and-rate task; Granqvist, 2003). The most abnormal-
 152 sounding voices were placed toward the left end of the line; those that approached
 153 normal were placed near the box. Voices judged as equally dysphonic were to be
 154 stacked on the line. Because the box for “normal” voices appeared rather small on
 155 the screen, listeners were explicitly instructed that box size did not mean that there
 156 were only a few normal voices in the set, and that they could place as many or as few
 157 icons as desired in the box. Listeners were encouraged to play the voices as often
 158 as required, in any order, until they were satisfied with their sort, after which testing
 159 advanced to the next trial. The experiment was self-paced and listeners were allowed
 160 to take breaks as needed. They were not told how many total speakers were included
 161 in the experiment. Total testing time was less than 1 h.

Table 5.2 Acoustic variables. Means and coefficients of variation were calculated for all measures using VoiceSauce software

Variable	Definition and reference
H1*-H2*	Relative amplitudes of the first and second harmonics, corrected for the effects of formants on amplitude (Hanson, 1997; Iseli & Alwan, 2004)
H2*-H4*	Relative amplitudes of the second and fourth harmonics, corrected for the effects of formants on amplitude
H4*-H2kHz*	Relative amplitudes of the fourth harmonic and the harmonic nearest 2 kHz, corrected for the effects of formants on amplitude
H2kHz*-H5kHz	Relative amplitudes of the harmonic nearest 2 kHz and that nearest 5 kHz; H2kHz is corrected for the effects of formants on amplitude
Cepstral peak prominence (CPP)	The relative amplitude of the cepstral peak in relation to the expected amplitude as derived via linear regression; a measure of aperiodicity (Hillenbrand et al., 1994)
Energy Root Mean Square (RMS)	Energy, calculated over five pitch pulses.
Subharmonic-to-harmonic ratio (SHR)	The amplitude ratio between subharmonics and harmonics; characterizes speech with alternating pulse cycles (period-doubling; Sun, 2002)
Fundamental frequency (F0)	The frequency of the first harmonic
F1, F2, F3, F4	Center frequencies of the first four formants

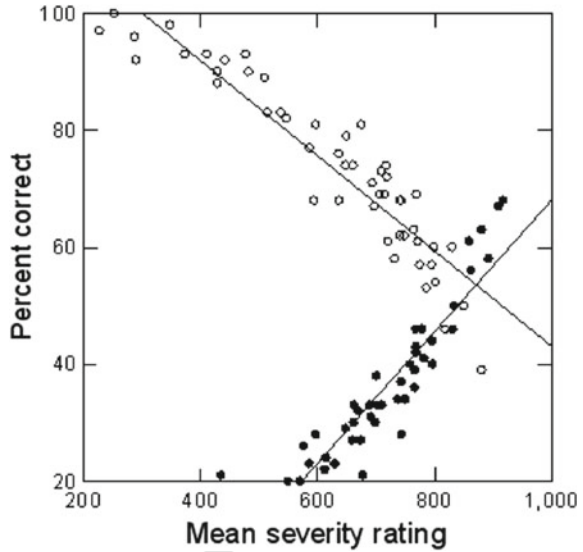
5.2.3 Acoustic Analyses

Acoustic measurements (Table 5.2) were made on all stimuli to facilitate interpretation of listeners’ perceptual strategies. As a set, these measures constitute a psychoacoustic model of voice quality (Kreiman et al., 2014) and were chosen because as a set they are sufficient to model the perceived quality of virtually any sustained phonation. Variables were measured every 5 ms using VoiceSauce software (Shue et al., 2011), and then averaged across the entire utterance. Coefficients of variation were also calculated as estimates of signal variability.

5.3 Results

Analyses fall into two groups, corresponding to the two approaches to defining “normal” discussed in the Introduction. The first analyses treated “normal” (i.e., placed in the box by a listener; Fig. 5.1) and “not-normal” (placed on the line outside the box) responses as straightforwardly categorical, consistent with definitions of nor-

Fig. 5.2 Accuracy of classification judgments as a function of the mean rating (where a larger rating = a more normal voice). Open circles represent “not-normal” voices; filled circles represent “normal” voices.



mal as “lacking a diagnosis.” The second set of analyses treated ratings as forming a continuum from most severe (=0), to normal (=1000), consistent with the idea that perceived normalness varies continuously as a function of listening context (Gerratt et al., 1993), social and/or communicative context, and other such factors. Both sets examined (1) listener agreement about (the degree of) perceived deviation and (2) the acoustic cues that explained listeners’ judgments.

Figure 5.2 shows the relationship between these two measurement approaches in a plot of categorization accuracy as a function of mean normalness ratings. In this figure, a priori “normal” voices are plotted as filled circles and a priori “not-normal” voices are plotted as open circles. Note that accuracy is greater for “not-normal” voices than for “normal” voices: It is apparent from this figure that many voices with diagnoses sound quite normal, and many nominally normal voices sound rather abnormal on average. The majority of “normal” voices were judged normal less than 50% of the time, while only a few “not-normal” voices were incorrectly categorized more than 50% of the time. Also note that the range of severity ratings for “normal” voices completely overlaps that for “not-normal” voices, but not vice versa. This pattern occurs because the normal end of the scale has an absolute ending point—a voice cannot be more normal than normal—but one can always imagine a worse voice, so that the left end of the scale can extend infinitely.

Editor Proof

194 **5.3.1** *Categorical Judgments of “Normal” Versus*
 195 *“Not-Normal” Voice Quality*

196 **5.3.1.1** **Can Listeners Accurately Separate Nominally Normal from**
 197 **Nominally Not-Normal Voices?**

198 If the boundary between normal and not-normal voice quality is ill-defined, as sug-
 199 gested by the papers reviewed in the Introduction, then it should be difficult for
 200 listeners to make categorical decisions regarding the status of a voice sample. This
 201 proved to be the case. For voices deemed normal a priori, listener performance
 202 ranged from 1.1 to 67.8% correct classification, with a mean of 34.1% correct (sd
 203 = 14.64%), where chance is 50%. Performance was somewhat better for a priori
 204 not-normal voices, which were correctly classified an average of 73.6% of the time
 205 (sd = 14.99%), with a range of 45.6–100%. Chi square analyses indicated that lis-
 206 teners heard only 2/50 a priori normal voices as normal at above chance levels, but
 207 agreed at above chance levels that 30/50 normal voices were not normal. For a priori
 208 not-normal voices, 35/50 were significantly often classified as not normal, and none
 209 was incorrectly classified as normal.

210 Finally, d' analysis (e.g., Green & Swets, 1966) was used to assess overall cate-
 211 gorization accuracy across the entire group of listeners. In this context, d' measures
 212 listeners' ability to correctly identify voices as normal or not normal, independent
 213 of response biases in favor of “normal” or “not-normal” responses. Ratings on the
 214 normal/not-normal scale were quantized to range from 1 to 10, where 1 represented
 215 the worst voice quality and 10 meant the voice had been classified as normal. These
 216 rescaled values were then used to calculate d' for each listener and for the group as a
 217 whole (Macmillan & Creelman, 2005). Results for both the pooled listeners and for
 218 individuals indicated that performance was at chance levels. For the pooled listeners,
 219 d' equaled 0.21, while across individual listeners, values averaged 0.24, with a range
 220 of -0.27 – 0.81 (sd = 0.28). We conclude from these data that listeners were unable
 221 to distinguish nominally normal from nominally not-normal voices at above chance
 222 rates, due to misclassifications both of normal voices as dysphonic and of not-normal
 223 voices as normal.

224 **5.3.1.2** **Do Listeners Agree with One Another in Their Categorical**
 225 **Judgments?**

226 Although listeners were inaccurate in their categorical responses, it is possible that
 227 this occurred because some of the clinical voice samples were very mildly deviant,
 228 and some of the nominally normal voices were characterized by high or low F0,
 229 vibrato, vocal fry, and/or breathiness, which could be interpreted as abnormal. This
 230 is especially possible when not normal is defined entirely in terms of physiology,
 231 because abnormal-appearing vocal folds can sometimes occur without any perceptual
 232 consequences. If this is the case, listeners might agree in their normal/not-normal

233 judgments, even though these do not correspond to the clinically defined state of
 234 affairs.

235 To assess this possibility, we examined listener agreement about vocal status,
 236 independent of the existence of a diagnosis. Listeners did not agree unanimously
 237 that any voice was normal; they were unanimous regarding only a single not-normal
 238 voice. Significant agreement was almost as uncommon as unanimous agreement.
 239 Chi square analyses showed that listeners agreed at above chance levels that only
 240 2/100 voices were normal (both of which were in fact normal; $p < 0.05$); they
 241 agreed at above chance levels that 65/100 voices were not normal (30 of which were
 242 nominally normal, as noted above; $p < 0.05$). We conclude that listeners are no more
 243 in agreement than they are accurate when asked to judge whether or not a voice is
 244 normal.

245 5.3.1.3 Are Listeners Self-consistent in Their Judgments?

246 Two possibilities emerge from the findings that listeners are highly inaccurate and
 247 disagree widely when asked to judge whether a voice is or is not normal. First, it is
 248 possible that “normal” is truly meaningless in practice. However, it is also possible
 249 that every listener has his/her own consistent idea of what “normal” is, but that these
 250 ideas differ from listener to listener. To examine these possibilities, we calculated
 251 intrarater agreement in normal/not-normal judgments for the 80 repeated ratings
 252 each listener provided. Average intrarater agreement equaled 75.8%, with a range
 253 from 57.5 to 94.4% ($sd = 9.22\%$; chance = 50%). Three of 50 listeners were self-
 254 consistent at rates below 60%; 30/50 were self-consistent at rates of 75% or above.
 255 These results indicate that most listeners are reasonably reliable when they report
 256 that a voice is or is not normal, but suggest that the basis for these judgments may
 257 vary across listeners, leading to self-consistency but low interrater agreement. We
 258 pursue this possibility in the next section.

259 5.3.2 Can We Predict Listeners' Categorical Responses from 260 Voice Acoustics?

261 Linear discriminant (LD) analysis was used to determine how well listeners' cat-
 262 egorical “normal” versus “not-normal” judgments could be predicted from voice
 263 acoustics (regardless of the existence/non-existence of a diagnosis). All variables
 264 from the psychoacoustic model were entered simultaneously into the analysis. One
 265 eigenfunction accounted for 100% of the variance in the data (canonical correlation
 266 = 0.263; Wilks' lambda = 0.931; chi square = 642.72, $df = 14$, $p < 0.001$). 70%
 267 of stimuli were correctly classified as perceptually normal or not normal. Predictors
 268 with weights ≥ 0.3 (~10% variance accounted for) included F2 (weight = -0.52),
 269 F0 (weight = 0.33), and F0 cv (weight = -0.30). These results suggest that, even

Table 5.3 Patterns of weights on eigenfunctions resulting from LD analyses relating individual listeners’ categorical normal/not-normal judgments to acoustic variables

Primary predictor variable	Additional significant predictors	Number of listeners
Variability		14
Vowel quality		7
Vowel quality	Variability	2
Vowel quality	Noise	5
Vowel quality	F0	5
F0		1
F0	Noise	3
F0	Spectral shape	5
Noise		6
Spectral shape		2

270 when considered as a group, listeners are not responding randomly, but also show
 271 that only a few rather simple variables (vowel quality, pitch, and pitch variability)
 272 are apparently shared across listeners.

273 To examine differences among listeners, we repeated the LD analyses for each
 274 of the 50 individual listeners. Results showed significant classification based on
 275 acoustic measures for all but 1 listener; across individuals, voices were correctly
 276 categorized as “perceived to be normal” or “perceived to be not normal” 81.35% of
 277 the time ($sd = 6.64$; range = 67.8–96.7%). However, listeners differed widely in the
 278 measures that emerged from these analyses. For brevity of presentation, the acoustic
 279 parameters were grouped into five categories: variability (coefficients of variability
 280 for all measures), vowel quality (F1, F2, F3, F4); spectral noise (CPP, energy, SHR),
 281 F0, and source spectral shape (H1*-H2*, H2*-H4*, H4*-H2kHz*, H2kHz*-H5kHz).
 282 Variables that weighted at 0.3 or higher on the eigenvector for each listener are
 283 tallied in Table 5.3. As in the group analyses just described, F0 and vowel quality
 284 were important for explaining individual listeners’ normal/not-normal decisions, but
 285 overall acoustic variability and noise also emerged as important predictors.

286 Finally, context effects are well known to affect ratings of vocal severity. For
 287 example, a given voice will sound rougher in the context of normal voices, and less
 288 rough in the context of voices with severe vocal pathology (Gerratt et al., 1993). To
 289 examine the influence such effects might have had on perceptual strategies in the
 290 present task, we repeated the LD analyses separately for each of the five groups of
 291 listeners. Recall that all listeners heard all the voices, but voices were grouped into
 292 different sets of 20, so the context in which each voice was judged varied from group
 293 to group. Results appear in Table 5.4. Groups did differ somewhat in the acoustic
 294 variables that predict overall categorical response patterns. Notably, spectral shape
 295 parameters appear in the solutions for two groups, and CPP appears in two other
 296 solutions. However, the increased complexity of the sets of predictor variables did

Table 5.4 Discriminant analysis results for the five groups of listeners. All analyses $p < 0.001$; only weights exceeding 0.3 are listed

Listener group	Variables (weights)	% Correct classification
1	CPP (0.46), CPP cv (-0.41), F2 (-0.35), F1 (0.34)	70.3
2	F2 (-0.50), H4*-2kHz* (-37), H2*-H4* (0.31), F0 cv (-0.30)	66.7
3	F2 (-0.49)	70.8
4	F0 (0.52), CPP cv (-0.48), F0 cv (-0.44)	77.8
5	F2 (-0.66), H4*-2kHz* (-0.30)	64.0

297 not result in improved correct classification rates, which generally remained well
 298 below those observed for individual listeners. This suggests that, although context
 299 effects exist, individuals in even small groups ($n = 10$) vary enough in perceptual
 300 strategies that controlling context effects does not improve correct classification to
 301 any measurable extent.

302 To summarize, across all listeners, parameters associated with F0, F0 variability,
 303 and vowel quality appear to be important for separating normal from not-normal
 304 voices for many, but not most, listeners, and thus provide at best moderate prediction
 305 of how a voice will be judged. Listeners' strategies vary with listening context, but
 306 modeling this aspect of variation does not improve overall prediction. However,
 307 LD analyses indicated that individual listeners' strategies can be well predicted from
 308 acoustics, but that listeners differ widely from one another. We conclude that listeners
 309 disagree because they are using rather different perceptual strategies, which are more
 310 idiosyncratic than they are context dependent. We examine this possibility further in
 311 the next section.

312 5.3.3 Do Listeners at Least Sort Voices in Similar Fashions?

313 A final possible explanation for our findings is that listeners rank the voices similarly
 314 on a scale from normal to maximally not normal, but differ in where they place the
 315 dividing line between categories. This could also have occurred if listeners differed
 316 in their interpretation of the size of the "normal" box in the experimental interface. To
 317 investigate these possibilities, we calculated Spearman correlations between scalar
 318 ratings for all pairs of listeners within a group. Rank-order correlations averaged only
 319 0.267 ($sd = 0.107$; range = -0.093 – 0.583), indicating that listeners do not agree
 320 even about the relative normalness/not-normalness of the voices.

5.3.4 *Can We Predict the Extent to Which a Voice Sounds Not Normal? What Parameters Are Associated with Increasing Perceived Vocal Deviance for Individual Listeners?*

Analyses in previous sections have demonstrated that listeners are individually self-consistent but inaccurate and in disagreement when separating normal from not-normal voices. To investigate this further, we modeled each listener’s perceptual strategy with a series of correlation and multiple regression analyses using only the voices categorized as not normal. First, for each listener, we calculated a multiple regression between the scalar not-normal ratings and the complete set of acoustic measures, entered into the equation in five blocks (F1, F2, F3, and F4; the coefficients of variation; F0; CPP, energy, and SHR; and the four spectral shape parameters). Order of entry was determined by the overall importance of the sets of variables in the LD analyses (Table 5.3). Next, for each listener, we calculated Pearson’s correlation between each acoustic measure and the scalar rating on the normal/not-normal scale for that listener, again including only the voices that the listener categorized as not normal. Finally, we calculated additional multiple regressions again relating ratings to acoustic measures for each listener, but this time using only the variables that were significant predictors in the first regression for that listener plus any additional variables that were significantly correlated with that listener’s not-normal ratings.

Results are shown in Table 5.5. All the regressions were statistically significant ($p < 0.01$), but all accounted for rather small amounts of variance in listeners’ judgments (mean $r = 0.477$; $sd = 0.126$; range = 0.227–0.699). As Table 5.5 shows, every variable contributed significantly to predicting ratings for at least one listener, but F0, F1, F2, and F0 cv stand out as more important across listeners than the rest. Recall that these same variables were associated with categorical normal/not-normal judgments for many listeners, as described above. This suggests that, for at least some listeners, deciding whether or not a voice sounds normal and establishing exactly how not normal it sounds depend on the same cues and thus are essentially the same process.

5.4 Discussion and Conclusions

To summarize our findings, judgments of diagnostically “normal” versus “not-normal” status were at chance. Listeners were relatively self-consistent in their judgments, but disagreed with one another, especially about what counts as normal. Agreement was better, but still limited, about what counts as “not normal.” This may have occurred because of differences in the possible ranges of the two labels. As noted above, the range of perceived not-normal quality can extend essentially limitlessly. As a result, there will always be voices that are so far from the boundary between normal and not normal that little or no ambiguity exists with respect to

Table 5.5 The frequency with which each acoustic variable emerged as a significant predictor in multiple regressions relating acoustic variables to the degree of perceived not-normalness. The most important predictors are listed in **bold type**. The maximum possible value is 50 (=the number of listeners)

Variable	# listeners for whom that variable was a significant predictor of perceived not-normalness
H1*-H2*	4
H2*-H4*	7
H4*-H2kHz*	3
H2kHz*-H5kHz	5
CPP	8
Energy	3
SHR	5
F0	19
F1	14
F2	24
F3	3
F4	3
H1*-H2* cv	1
H2*-H4* cv	4
H4*-H2kHz* cv	8
H2kHz*-H5kHz cv	3
CPP cv	9
Energy cv	7
SHR cv	3
F0 cv	26
F1 cv	2
F2 cv	2
F3 cv	10
F4 cv	4

360 their status. In contrast, logically a voice cannot be more normal than “normal,” and
 361 any deviation in quality, however slight, creates ambiguity (and hence disagreement)
 362 about the voice’s status. The surprising aspect of our results was how completely the
 363 category “normal” was compromised by this process.

364 The overall picture that emerges from the present data is one of differences
 365 between listeners, but less so within listeners, in the attributes they pay attention
 366 to when deciding that a voice is or is not normal. Strategies for separating “normal”
 367 from “not normal” differed widely across individual listeners, as did strategies for
 368 determining how much a given voice deviated from normal, and all variables in the
 369 psychoacoustic model played a role in decisions for at least one listener. However,

370 several variables—F0, F1 and F2, and F0 cv—appeared more often than the others as
 371 significant predictors of both categorical judgments and of scalar normalness ratings.
 372 These variables did not account for most of the variance in these analyses, and did not
 373 consistently appear as a set in the perceptual models for even half of the listeners, but
 374 they did appear individually in most analyses, suggesting that in practice the concept
 375 of “normal” has some small core of meaning based on F0 and vowel quality.

376 We note that the “core” variables are also important determinants of individual
 377 voice quality (see Kreiman & Sidtis, 2011, for review), which is judged in terms of
 378 a central category member and idiosyncratic deviations from that “average” voice.
 379 Thus, it is possible that (at least some of the time), listeners assess normalness much
 380 as they assess individual voice quality in general, with respect to a central pattern and
 381 the deviations from that pattern that appear in the particular voice sample at hand.
 382 Thus, the answer to our initial question—What does it mean for a voice to sound
 383 normal?—is a complex one that depends on the listener, the context, the purpose of
 384 the judgment, and other factors as well as on the voice.

385 A few limitations to this research should be noted. First, stimuli were steady-state
 386 vowels rather than connected speech. This means that many details that can char-
 387 acterize disordered speech were not available for consideration, including prosody,
 388 articulation, pausing, and other vocal attributes. However, it seems unlikely that
 389 inclusion of more complex stimuli would improve overall listener agreement, par-
 390 ticularly with respect to which voices sound normal. This study was also restricted
 391 to female speakers. While it is likely that different parameters will emerge from
 392 studies of normal versus not-normal male voices, the fact that listeners’ behavior is
 393 consistent with broader models of voice perception makes it unlikely that the over-
 394 all pattern of results would differ substantially. Studies of male voices are currently
 395 underway in our laboratory. Finally, the relatively small size of the response box
 396 for “normal” voices in the testing interface (Fig. 5.1) may have discouraged some
 397 listeners from categorizing too many voices as normal, despite instructions that any
 398 number of voices could be placed in the box. However, we note that correlation
 399 analyses showed very poor agreement among listeners, suggesting that the effect of
 400 this design issue on the overall pattern of results is minimal.

401 In conclusion, these results have implications for ongoing efforts to identify acous-
 402 tic measures to screen for vocal pathology or the provision of normative values for
 403 single acoustic measure. The finding that listeners are self-consistent but highly indi-
 404 vidual in their perceptual strategies for determining what is and is not normal suggest
 405 that automatic protocols or screening based on normative values may be of limited
 406 clinical or theoretical use. Clear communication between clinicians and patients in
 407 a context of cultural awareness would seem to be the straightest path to satisfactory
 408 treatment outcomes.

409 **Acknowledgments** This work was supported by NIH grant DC01797, and by NSF grants IIS
 410 1704167 and IIS 1450992. A preliminary version was presented at the 175th Meeting of the Acous-
 411 tical Society of America, Minneapolis, MN, May 2018. We thank Norma Antoñanzas for program-
 412 ming support, Meng Yang for help with acoustic analyses, Jordan Shavalian for assistance with
 413 subject testing and data analysis, and Pat Keating and Marc Garellek for helpful comments.

414 **References**

- 415 Altenberg, E. P., & Ferrand, C. T. (2006). Perception of individuals with voice disorders by mono-
 416 lingual English, bilingual Cantonese-English, and bilingual Russian-English women. *Journal of*
 417 *Speech, Language, and Hearing Research*, *49*, 879–887.
- 418 Amir, O., & Levine-Yundof, R. (2013). Listeners' attitude toward people with dysphonia. *Journal*
 419 *of Voice*, *27*, 524.e1–524.e10.
- 420 Anderson, R., Klofstad, C., Mayew, W., & Venkatachalam, M. (2014). Vocal fry may undermine
 421 the success of young women in the labor market. *PLOS ONE*, *9*, e97506.
- 422 Arias-Londoño, J. D., Godino-Llorente, J. I., Markaki, M., & Stylianou, Y. (2011). On combin-
 423 ing information from modulation spectra and mel-frequency cepstral coefficients for automatic
 424 detection of pathological voices. *Logopedics Phoniatrics Vocology*, *36*, 60–69.
- 425 Aronson, A. E., & Bless, D. M. (2009). *Clinical voice disorders*. New York, N.Y.: Thieme.
- 426 Behlau, M., & Murry, T. (2012). International and intercultural aspects of voice and voice disorders.
 427 In D. E. Battle (Ed.), *Communication disorders in multicultural and international populations*
 428 (4th ed., pp. 174–207). St. Louis, MO: Mosby.
- 429 Blood, G. W., & Hyman, M. (1977). Children's perception of nasal resonance. *Journal of Speech*
 430 *and Hearing Disorders*, *42*, 446–448.
- 431 Blood, G. W., Mahan, B. W., & Hyman, M. (1979). Judging personality and appearance from voice
 432 disorders. *Journal of Communication Disorders*, *12*, 63–67.
- 433 Bonilha, H. S., & Deliyski, D. D. (2008). Period and glottal width irregularities in vocally normal
 434 speakers. *Journal of Voice*, *22*, 699–708.
- 435 Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method
 436 of paired comparisons. *Biometrika*, *39*(3/4), 324–345.
- 437 Dehqan, A., Ansari, H., & Bakhtiar, M. (2010). Objective voice analysis of Iranian speakers with
 438 normal voices. *Journal of Voice*, *24*, 161–167.
- 439 Eadie, T. L., Kapsner, M., Rosenzweig, J., Waugh, P., Hillel, A., & Merati, A. (2010). The role of
 440 experience on judgments of dysphonia. *Journal of Voice*, *24*, 564–573.
- 441 Gerratt, B. R., Kreiman, J., & Antoñanzas-Barroso. (1993). Comparing internal and external stan-
 442 dards in voice quality judgments. *Journal of Speech and Hearing Research*, *36*, 14–20.
- 443 Gerratt, B. R., Kreiman, J., & Garellek, M. (2016). Comparing measures of voice quality from
 444 sustained phonation and continuous speech. *Journal of Speech Hearing Research*, *59*, 994–1001.
- 445 Goy, H., Fernandes, D. N., Pichora-Fuller, M. K., & van Lieshout, P. (2013). Normative voice data
 446 for younger and older adults. *Journal of Voice*, *27*, 545–555.
- 447 Granqvist, S. (2003). The visual sort and rate method for perceptual evaluation in listening tests.
 448 *Logopedics Phoniatrics Vocology*, *28*, 109–116.
- 449 Greene, M.C., & Mathieson, L. (1992). *The voice and its disorders*. San Diego, CA: Singular.
- 450 Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Krieger.
- 451 Hanson, H. M. (1997). Glottal characteristics of female speakers: Acoustic correlates. *Journal of*
 452 *the Acoustical Society of America*, *101*, 466–481.
- 453 Hillenbrand, J., Cleveland, R. A., & Erickson, R. L. (1994). Acoustic correlates of breathy vocal
 454 quality. *Journal of Speech, Language, and Hearing Research*, *37*, 769–778.
- 455 Irani, F., Abdalla, F., & Hughes, S. (2014). Perceptions of voice disorders: A survey of Arab adults.
 456 *Logopedics Phoniatrics Vocology*, *39*, 87–97.
- 457 Iseli, M., & Alwan, A. (2004). An improved correction formula for the estimation of harmonic
 458 magnitudes and its application to open quotient estimation. In *Proceedings of ICASSP'04* (pp.
 459 669–672), Montreal, Canada.
- 460 Johnson, W., Brown, S.F., Curtis, J.F., Edney, C.W., & Keaster, J. (1965). *Speech handicapped*
 461 *school children*. New York: Harper & Brothers. Cited in Aronson & Bless (2009).
- 462 Keating, P. A., Kreiman, J., & Alwan, A. (2018). The UCLA speaker variability database. Retrieved
 463 July 26, 2018, from <https://ucla.box.com/s/81ho14uypkmv2nn1s3tvy1gvdqtq70gcf>.
- 464 Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies*. Walden, MA: Wiley-Blackwell.

- 465 Kreiman, J., Gerratt, B. R., Garellek, M., Samlan, R., & Zhang, Z. (2014). Toward a unified theory
466 of voice production and perception. *Loquens*, 1(1), 1–9. [https://doi.org/10.3989/loquens.2014.](https://doi.org/10.3989/loquens.2014.009)
467 009.
- 468 Lallh, A. K., & Rochet, A. P. (2000). The effect of information on listeners’ attitudes toward speakers
469 with voice or resonance disorders. *Journal of Speech, Language, and Hearing Research*, 43, 782–
470 795.
- 471 Lewandowski, A., Gillespie, A. I., Kridgen, S., Jeong, K., Yu, L., & Gartner-Schmidt, J. (2018).
472 Adult normative data for phonatory aerodynamics in connected speech. *The Laryngoscope*, 128,
473 909–914.
- 474 Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user’s guide* (2nd ed.). Mahwah,
475 NJ: Erlbaum.
- 476 Maryn, Y., & Debo, K. (2015). Is perceived dysphonia related to perceived healthiness? *Logopedics*
477 *Phoniatrics Vocology*, 40, 122–128.
- 478 Mathieson, L. (2000). Normal-disordered continuum. In R. D. Kent & M. J. Ball (Eds.), *Voice*
479 *quality measurement* (pp. 3–12). San Diego, CA: Singular.
- 480 Moore, G. P. (1971). Voice disorders organically based. In L. E. Travis (Ed.), *Handbook of speech*
481 *pathology and audiology* (pp. 535–569). Englewood Cliffs, NJ: Prentice-Hall.
- 482 Moro-Velázquez, L., Gómez-García, J., & Godino Llorente, J. (2016). Voice pathology detection
483 using modulation spectrum-optimized metrics. *Frontiers in Bioengineering and Biotechnology*,
484 4. <https://doi.org/10.3389/fbioe.2016.00001>.
- 485 Oliveira, G., Davidson, A., Holczer, R., Kaplan, S., & Paretzky, A. (2016). A comparison of the use
486 of glottal fry in the spontaneous speech of young and middle-aged American women. *Journal of*
487 *Voice*, 30, 684–687.
- 488 Orozco-Arroyave, J. R., Belalcazar-Bolanos, E. A., Arias-Londoño, J. D., Vargas-Bonilla, J. F.,
489 Skodda, S., & Rusz, J., & Nöth, E., (2015). Characterization methods for the detection of multiple
490 voice disorders: Neurological, functional, and laryngeal diseases. *IEEE Journal of Biomedical*
491 *and Health Informatics*, 19, 1820–1828.
- 492 Shue, Y.-L., Keating, P., Vicenik, C., & Yu, K. (2011). VoiceSauce: A program for voice analysis. In
493 *2011 Proceedings of International Congress of Phonetic Sciences (ICPhS) XVII* (pp. 1846–1849),
494 Hong Kong.
- 495 Sun, X. (2002). Pitch determination and voice quality analysis using subharmonic-to-harmonic ratio.
496 In *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*
497 (Vol. 1, pp. 1–333). IEEE. <https://doi.org/10.1109/ICASSP.2002.5743722>.
- 498 Sundberg, J. (1988). *The science of the singing voice*. DeKalb, IL: Northern Illinois University
499 Press.
- 500 Wang, X., Zhang, J., & Yan, Y. (2011). Discrimination between pathological and normal voices
501 using GMM-SVM approach. *Journal of Voice*, 25, 38–43.
- 502 Watterson, T., Mancini, M., Brancamp, T. U., & Lewis, K. E. (2013). Relationship between the per-
503 ception of hypernasality and social judgments in school-aged children. *Cleft Palate Craniofacial*
504 *Journal*, 50, 498–502.
- 505 Wuyts, F. L., Heylen, L., Mertens, F., De Bodt, M., & Van de Heyning, P. H. (2002). Normative
506 voice range profiles of untrained boys and girls. *Journal of Voice*, 16, 460–465.
- 507 Xue, S. A., & Hao, J. G. (2006). Normative standards for vocal tract dimensions by race as measured
508 by acoustic pharyngometry. *Journal of Voice*, 20, 391–400.
- 509 Yuasa, I. P. (2010). Creaky voice: a new feminine voice quality for young urban-oriented upwardly
510 mobile American women? *American Speech*, 85, 315–337.
- 511 Zacharias, S. R., Kelchner, L. N., & Creaghead, N. (2013). Teachers’ perceptions of adolescent
512 females with voice disorders. *Language, Speech, and Hearing Services in Schools*, 44, 174–182.

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Voice Attractiveness
Series Title	
Chapter Title	The Role of Voice Evaluation in Voice Recall
Copyright Year	2020
Copyright HolderName	Springer Nature Singapore Pte Ltd.
Author	Family Name Babel Particle Given Name Molly Prefix Suffix Role Division Department of Linguistics Organization University of British Columbia Address 2613 West Mall Vancouver, BC, V6T 1Z4, Canada Email molly.babel@ubc.ca
Corresponding Author	Family Name McGuire Particle Given Name Grant Prefix Suffix Role Division Department of Linguistics Organization University of California Santa Cruz Address 1156 High St, Santa Cruz, CA, 95060, USA Email gmcguir1@ucsc.edu
Author	Family Name Willis Particle Given Name Chloe Prefix Suffix Role Division Department of Linguistics Organization University of California Santa Barbara Address Santa Barbara, CA, 93106, USA Email chloemwillis@umail.ucsb.edu
Abstract	This chapter examines the relationship among a suite of voice evaluation metrics—vocal attractiveness, voice typicality, gender categorization fluency, intelligibility, acoustic similarity, and perceptual similarity—in a set of 60 American English voices with the goal of understanding how these evaluation metrics predict listeners' abilities to accurately recall voices. This question of what makes a voice memorable has been studied from a range of perspectives, as it raises critical theoretical issues about auditory memory and phonetic encoding, in addition to having applied concerns in the context of eyewitness testimony. We find

that the more subjective voice evaluation measures of stereotypicality and attractiveness predict listeners' ability to recall voices more so than the more objective measures related to voice similarity and processing. These results suggest that listeners' cognitive organization of voices is influenced by social assessments of voices.

Keywords

Voice recall - Talker recognition - Voice evaluation - Voice typicality - PCA - Voice organization

Chapter 6

The Role of Voice Evaluation in Voice Recall



Molly Babel, Grant McGuire, and Chloe Willis

Abstract This chapter examines the relationship among a suite of voice evaluation metrics—vocal attractiveness, voice typicality, gender categorization fluency, intelligibility, acoustic similarity, and perceptual similarity—in a set of 60 American English voices with the goal of understanding how these evaluation metrics predict listeners' abilities to accurately recall voices. This question of what makes a voice memorable has been studied from a range of perspectives, as it raises critical theoretical issues about auditory memory and phonetic encoding, in addition to having applied concerns in the context of earwitness testimony. We find that the more subjective voice evaluation measures of stereotypicality and attractiveness predict listeners' ability to recall voices more so than the more objective measures related to voice similarity and processing. These results suggest that listeners' cognitive organization of voices is influenced by social assessments of voices.

Keywords Voice recall · Talker recognition · Voice evaluation · Voice typicality · PCA · Voice organization

M. Babel
Department of Linguistics, University of British Columbia, 2613 West Mall Vancouver, BC V6T 1Z4, Canada
e-mail: molly.babel@ubc.ca

G. McGuire (✉)
Department of Linguistics, University of California Santa Cruz, 1156 High St, Santa Cruz, CA 95060, USA
e-mail: gmcguir1@ucsc.edu

C. Willis
Department of Linguistics, University of California Santa Barbara, Santa Barbara, CA 93106, USA
e-mail: chloemwillis@umail.ucsb.edu

© Springer Nature Singapore Pte Ltd. 2020

B. Weiss et al. (eds.), *Voice Attractiveness*, Prosody, Phonology and Phonetics,
https://doi.org/10.1007/978-981-15-6627-1_6

107

6.1 Introduction

This chapter examines the relationship between vocal attractiveness, voice typicality, and other related vocal evaluation metrics along with listeners' ability to recall voices from memory. What makes a voice memorable has been studied from a range of perspectives as it raises critical theoretical issues about auditory memory and phonetic encoding, in addition to having applied concerns in the context of earwitness testimony. In this work, we explore some of the qualities of the voices that improve and detract from voice recall performance.

Talker recognition or listeners' ability to recall voices they have been previously exposed to is highly affected by what is referred to as the *language familiarity effect*. Listeners are more accurate at recalling voices that speak the same language as the listener population (Goggin, Thompson, Strube, & Simental, 1991; Perrachione & Wong, 2007; Thompson, 1987; Winters, Levi, & Pisoni, 2008; Perrachione, Del Tufo, & Gabrieli, 2011; Xie & Myers, 2015; Orchard & Yarmey, 1995; Bregman & Creel, 2014) or speak with a familiar accent (Goggin et al., 1991; Stevenage, Clarke, & McNeill, 2012; Senior et al., 2018; Thompson, 1987; Perrachione, Chiao, & Wong, 2010). The mechanism behind these findings is generally considered to be one of listeners' familiarity with the phonetic distribution of sounds in the language or accent. When listeners are familiar with a language or accent, they are better able to determine which acoustic-phonetic features in the speech stream are language-specific and which are attributes of a particular speaker's voice (Winters et al., 2008; Perrachione, in press).

While this literature has established that voices with familiar languages and accents are generally more accurately recalled, voices within a language variety are not equally memorable. Within a language variety, what makes a voice more or less memorable? Several studies have found that subjective listener ratings of distinctiveness, typicality, memorability, among other evaluative qualities can predict which voices have better recall accuracy (Papcun, Kreiman, & Davis, 1989; Kreiman, & Papcun, 1991; Yarmey, 1991; O'Toole et al., 1998).

For example, Papcun et al. (1989) exposed listeners to 10 voices that had been previously rated on a scale from easy- to hard-to-remember and tested voice recall in an open set task with 1-, 2-, and 4-week delays. Subjects were generally better at rejecting novel voices rather than correctly identifying the voices that they had been exposed to. Specifically, the voices did not differ greatly in accuracy of recall, but did differ in false identifications, such that "hard" voices engendered more false positives. Papcun and colleagues invoke a prototype model to explain these results, hypothesizing that listeners characterize and remember voices in terms of a prototype and deviations therefrom. Thus, more prototypical voices are hard-to-remember as they are more similar to other voices and are more likely to be misidentified as a previously heard voice. Papcun and colleagues propose that easy-to-remember voices are less stable in memory because the voice-specific traits that make a voice easy-to-remember fade as a function of time, as the voice coalesces toward the prototype, resulting in more false alarms in the longer test delays. The authors attribute this



58 to “a psychological analog to statistical regression to the mean” and suggest that
59 hard-to-remember (prototypical) voices are more stable in memory than easy-to-
60 remember (atypical) ones (Papcun et al. 1989, p. 923). In a follow-up study, Kreiman
61 and Papcun (1991) examined the discrimination and recognition accuracy of voices
62 from Papcun et al. (1989). Overall, results were similar to the previous experiment:
63 voices that were rated easier to remember were less likely to be confused with other
64 voices while hard-to-remember voices were easily confused. Of special interest in
65 this study is that the accuracy results were compared with various acoustic and sub-
66 jective quality predictors (made by a separate group of listeners) that were assessed
67 via a multidimensional-scaling solution. The authors interpret the most predictive
68 dimensions for the discrimination results to be roughly equivalent to “masculinity,”
69 “creakiness,” “variability,” and “mood” while the recognition results were best pre-
70 dicted by what was interpreted as dimensions relating to “masculinity,” “breathiness,”
71 and “liveliness.” These descriptors and their relationship to voice discrimination and
72 recognition are applicable only to the set of 10 voices used in Kreiman and Papcun’s
73 studies, but the applicability of these dimensions illustrates the features in which
74 listeners cognitively organize this set of voices.

75 Voice typicality was the explicit subject evaluation under consideration in
76 Mullennix et al. (2011). Mullennix and colleagues asked listeners to evaluate 40
77 voices for typicality, using these judgments to prune the larger set for a memory
78 task. The voices with the highest (4 male, 4 female) or lowest (4 male, 4 female)
79 typicality ratings were selected. An independent group of listeners were exposed
80 to the 16 subset voices in a vowel identification task, and were then given a sur-
81 prise memory task. Overall, listeners were more accurate with the voices they had
82 previously trained on, but showed a bias to make recognition errors when typical
83 voices were used as foils, especially listeners exposed to typical voices. A recur-
84 ring theme across these studies is that unique or distinctive voices are more easily
85 remembered. What listeners rate when evaluating voices in terms of distinctiveness
86 or typicality is not clear, but it appears to be a measureable quality that listeners
87 exhibit agreement on. Typicality and distinctiveness may be connected to speech
88 clarity and the predictability of phonetic variation. Voices vary in how clearly they
89 produce linguistic contrasts, and this variation in contrast clarity has implications
90 for how listeners process and recognize the speech stream (Bradlow, Torretta, & Pisoni,
91 1996; Newman, Clouse, & Burnham, 2001). How an individual manifests a phonetic
92 contrast is a talker-specific feature that listeners track and exploit in subsequent pro-
93 cessing, spilling over into perceptual events beyond the moment of comprehension
94 (Theodore, Myers, & Lomibao, 2015). Too much phonetic variation can affect lis-
95 teners’ confidence in their categorization of speech sounds (Clayards, Tanenhaus,
96 Aslin, & Jacobs, 2008). Unexpected or unfamiliar phonetic variation associated with
97 accents or dialects that are different from one’s own makes comprehension and recog-
98 nition more challenging (Clopper & Pisoni, 2004; Bradlow & Bent, 2008), and this
99 is often attributed to lack of exposure and experience. While this may be intuitive
100 when thinking about nonnative speakers, the evidence is mixed as to whether non-
101 native speakers are more variable in their acoustic–phonetic realizations than native
102 speakers (Vaughn et al., 2020; Wade, Jongman, & Sereno, 2007). Talker variability

103 occurs within an accent or speech community as well (Strand, 1999; Bradlow et al.,
 104 1996; Babel & McGuire, 2015), resulting in intelligibility and memory benefits for
 105 familiar speakers (Nygaard & Pisoni, 1998). Accents that may be less familiar, but
 106 are the standard variety, often, however, show similar processing benefits to famil-
 107 iar varieties (Clopper, 2014; Clopper, Tamati, & Pierrehumbert, 2016), suggesting
 108 that the cognitive organization of voices is not exclusively tailored to the quantity
 109 of experience, but may involve some preferential encoding of socially prestigious
 110 exemplars (Babel, 2012; Babel, McGuire, & King, 2014b; Sumner, Kim, King, &
 111 McGowan, 2014).

112 How does the social evaluation of voices affect processing or the cognitive orga-
 113 nization of voices? As is clear from the topic of this book, there is extensive evidence
 114 that listeners assess voices in terms of their attractiveness. The patterns by which
 115 voices are deemed attractive seem to be a combination of culturally acquired (Babel,
 116 McGuire, Walters, & Nicholls, 2014a; Bezooijen, 1995) and more strongly evolution-
 117 arily encoded (Zuckerman & Miyake, 1993; Puts, Gaulin, & Verdolini, 2006; Riding
 118 et al., 2006; Saxton et al., 2006; Feinberg, DeBruine, Jones, & Perrett, 2008; Apicella,
 119 Feinberg, & Marlowe, 2007) preferences that tap into acoustic–phonetic dimensions
 120 that are related to sexually dimorphic traits. Many of the culturally acquired compo-
 121 nents appear to stem from what is typical or standard within a speech community.
 122 While there may be initial appeal in thinking of typicality or standardness in terms
 123 of the pattern that is the most common or at the peak of a community’s acoustic–
 124 phonetic distribution, linguistic standardness is much more of an imposed concept.
 125 Listeners tend to show stronger recognition patterns for pronunciation variants that
 126 are standard, despite a different pronunciation variant being far more frequent in the
 127 input (Sumner & Samuel, 2005) and listeners exhibit more false memories for a less
 128 socially prestigious accent compared to a more prestigious accent, despite equiva-
 129 lence in experience with the two (Sumner & Kataoka, 2013). Media is one means
 130 through which standardness and socially conditioned social preferences appear to
 131 be formed for speech communities (Kinzler & DeJesus, 2013; Lippi-Green, 2012).
 132 Overall, this body of literature makes clear that not all voices are treated equivalently
 133 in terms of processing and that both exposure and social preference play a role in
 134 voice evaluation.

135 To better understand the dimensions on which listeners may organize voices and
 136 how this organization may affect voice recall, we first report on a set of experiments
 137 and analyses intended to quantify the typicality of a set of voices from 60 American
 138 English speakers. These experiments provide two response time-based measures—
 139 Intelligibility and Categorization Fluency—designed to better reflect exposure by
 140 tapping into online frequency effects. Previous research has shown that response
 141 latency to voices is a proxy for familiarity; words are more likely to be recognized
 142 quickly if heard in a familiar voice rather than an unfamiliar voice (Goldinger, 1996).
 143 For the intelligibility task, listeners were asked to shadow voices embedded in noise
 144 and in the Categorization Fluency task, listeners identified voices as male or female
 145 in a speeded fashion. In both cases, faster responses indicate easier processing for a
 146 given voice. Additionally, we provide two subjective assessments, perceived Attract-
 147 iveness and perceived Stereotypicality. For both of these assessments listeners were

asked to subjectively rate the voices on either their attractiveness or typicality. We expect these measures to better tap into social preference. Because previous studies demonstrate that more similar voices are less likely to be remembered and are more likely to be considered a previously heard voice, we also include two measures of similarity, one based on auditory–acoustic measures, Acoustic Similarity, and one based on comparative listener ratings, Perceptual Similarity. After reporting the methods and results of each of these experiments, we examine to what extent these measures tap into similar dimensions in Sect. 6.2.7. Following this, Sect. 6.3 reports on a voice recall experiment, which we analyze with the voice evaluation metrics to assess which voice metrics best predict voice recall performance.

6.2 Voice Evaluation Experiments

6.2.1 Materials for All Experiments

The voice stimuli used in all the experiments reported here were from participants recruited as part of a previous study (Babel, 2012). They consist of 30 female (mean age 24, range 18–57) and 30 male (mean age 24, range 18–47) native speakers of American English reading 50 low-frequency monosyllabic words. For the present study a subset of 15 words which contain /i a u/ as the syllable nucleus were selected for each voice, 5 words per vowel (Table 6.1).

6.2.2 Intelligibility

To quantify the intelligibility of the voices, we used a speeded shadowing task where the response time to the onset of vocalization is taken as a proxy for how easy it was for listeners to understand the utterance.

Table 6.1 Words used in the experiments organized by the vowel category for each item

/i/	/a/	/u/
deed	cot	boot
key	pod	dune
peel	sock	hoop
teal	sod	toot
weave	tot	zoo

170 **6.2.2.1 Participants**

171 Thirty participants (15 male, 15 female) were recruited from the University of Cal-
172 ifornia, Santa Cruz, undergraduate population and were compensated with course
173 credit. All were native speakers of American English from the state of California.
174 Ages ranged from 18 to 23, mean 20.4 years.

175 **6.2.2.2 Materials**

176 The same voices and words used in the gender categorization fluency task were used
177 in this task. Each individual sound file was embedded in pink noise at +6 dB signal
178 to noise ratio (SNR). The noise began at the onset of each word and ended at the
179 offset of each word.

180 **6.2.2.3 Procedure**

181 Participants were seated in a sound-attenuated booth at a computer workstation wear-
182 ing AKG HSC271 model headset with integrated condenser microphone. The stimuli
183 were presented in a randomized order at a comfortable listening volume (approx-
184 imately 70 dB). Subjects were asked to repeat each word, initiating their repetition as
185 quickly as possible without compromising accuracy. Response times were measured
186 from the onset of the stimulus to the onset of the subject's production as registered
187 by a microphone connected to a PST serial response box. The response time for each
188 trial was displayed on the computer monitor to participants to help motivate fast
189 response times. This feedback screen was displayed for 1000 ms, after which a new
190 trial began. Each word production was recorded as a unique .wav file.

191 **6.2.2.4 Results**

192 Response time was automatically calculated for each production, and the accuracy
193 of each shadowed production was determined by manual coding. A custom-written
194 program brought up each individual sound file and provided an orthographic tran-
195 scription of the intended word. Each production was categorized as correct or incor-
196 rect. Productions with disfluencies, missing phones, or the wrong lexical item were
197 considered incorrect.

198 Accuracy of the repeated item is a measure of recognition. Female ($M = 81\%$
199 correct, $SD = 39$) voices achieved higher recognition rates than the male ($M = 76\%$
200 correct, $SD = 42$) voices [$t(51.67) = 2.47, p = 0.02$], indicating that female voices
201 were overall more intelligible than the male voices. Correct responses for reaction
202 times within two standard deviations of the group mean were then aggregated across
203 words for each voice. Using response time to correctly identified items as a proxy for
204 intelligibility, we found no significant differences between male and female voices

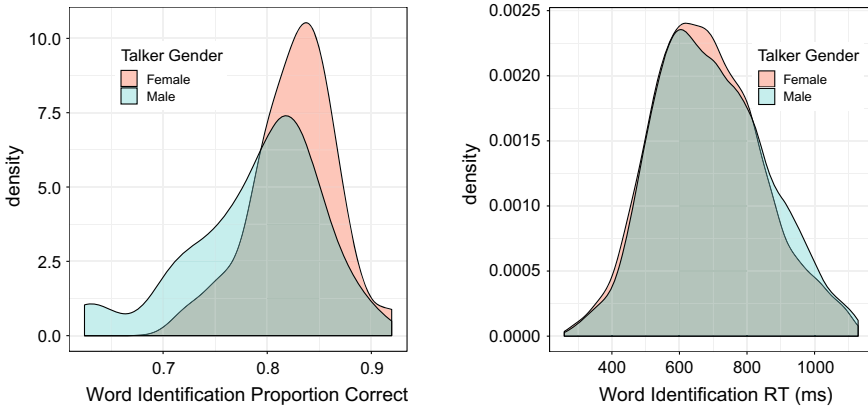


Fig. 6.1 Density plots showing the distribution of accuracy of correctly identifying each item (left panel) in a speeded shadowing task and the distribution of voice intelligibility, as measured by response lag (right panel) in a speeded shadowing task

205 [t (56.04) = 1.68, p = 0.098]. When items were accurately recognized, there was
 206 no difference in the intelligibility of those items for female and male voices. These
 207 aggregate measures mask the talker-specific variability of these measures. Figure 6.1
 208 provides density plots to illustrate the range of recognition scores (left panel) and
 209 intelligibility (right panel).

210 6.2.3 Gender Categorization Fluency

211 In order to have an online estimate of typicality, the voices were assessed using
 212 a gender categorization fluency task. This is a speeded classification task where
 213 subjects heard a single word and quickly decided the gender of the voice. Previous
 214 work has used this for evaluation of typicality for faces (Orena, Theodore, & Polka,
 215 2015) and voices (Strand, 1999).¹

216 6.2.3.1 Participants

217 Thirty participants (15 male, 15 female) were recruited from the University of Cal-
 218 ifornia, Santa Cruz, undergraduate population and were compensated with course
 219 credit. All were native speakers of American English from the state of California.
 220 Ages ranged from 18 to 24 years, with a mean of 21.

¹The data from this experiment were originally reported in Babel and McGuire (2015).

221 6.2.3.2 Materials

222 In order for the task to be feasible for the participants to complete in 45 min, the word
223 list was pruned to nine words for each talker (9 words \times 60 voices = 540 stimuli). The
224 original word list was presented to an independent group of university students ($n =$
225 23) who rated how likely each word was to be used by males or females. The words
226 *teal*, *weave*, *pod*, *sod*, *toot*, and *dune* were identified as the most gender-valenced of
227 the word set and were removed from the list.

228 6.2.3.3 Procedure

229 Listeners were presented with the individual words, one per trial. Words and voices
230 were randomized across all voices, and participants were prompted to respond to
231 each word by selecting whether the voice that said the word was male or female.
232 Reaction time feedback was given after each trial and listeners were asked to respond
233 in less than 500 ms. Each trial timed out after 1500 ms if no response was given.

234 6.2.3.4 Results

235 Response times for correct responses (98% of the data) made within two standard
236 deviations of the mean were then aggregated across words for each voice. The speed
237 at which listeners identified male ($M = 523$ ms, $SD = 17.5$) and female ($M = 525$
238 ms, $SD = 14$) voices differed was nonsignificant [$t(55.93) = 0.56$, $p = 0.58$].

239 6.2.4 Acoustic Similarity

240 To assess the voices in terms of their raw acoustic–auditory similarity, we calcu-
241 lated voice similarity using mel-frequency cepstral coefficients (MFCCs). While
242 MFCCs have no straightforward perceptual interpretations, they provide a global and
243 unbiased acoustic assessment of the speech signal. This type of unbiased acoustic
244 measurement is useful when trying to determine the extent to which listeners’ orga-
245 nization of sound patterns are faithful to acoustic–auditory parameters or whether
246 they are influenced by listeners’ experiences (Cristiá, Mielke, Daland, & Peperkamp,
247 2013; Mielke, 2012). The choice to use MFCCs, as opposed to resonant frequen-
248 cies or other spectral properties more readily connected to listeners’ perception of
249 phoneme categories, allows us to side-step any explicit decision about which aspects
250 of the speech spectrum to explicitly measure.

6.2.4.1 Materials

The set of 15 words produced by the 60 talkers was used in this analysis.

6.2.4.2 Procedure

The MFCC acoustic similarity algorithm implemented in Phonological CorpusTools (PCT; Hall, Allen, Fry, Mackie, & McAuliffe, 2018) was used to quantify acoustic vocal distinctiveness within the voice set. In this analysis, twenty-six mel-scaled triangular filters are applied to a windowed signal, and the resulting spectrum is the log of the power of each filter. The mel-frequency cepstrum is calculated using a discrete cosine transform, resulting in twelve coefficients. MFCCs are then compared using a dynamic time warping algorithm, which ultimately returns the summed distances of the best path through the data matrix. This comparison was done between matched words and each voice in the data set. While dynamic time warping may eliminate durational differences among tokens, and thus one cue to gender, it is a reliable way to directly compare the tokens. We chose this method over correlation-based approaches to quantifying spectral similarity because of precedent in the speech literature (Mielke, 2012) and the challenges of correlating signals of different lengths.

6.2.4.3 Results

To compare the acoustic vocal distinctiveness in the voice set, the similarity values for each voice comparison were averaged and used to create a distance matrix. Distance matrices were created separately for male and female voices as a combined analysis resulted in a first dimension that simply separated male and female voices. For both female and male voice sets, a scree plot of stress suggested an elbow at the fourth dimension, therefore a four-dimensional multidimensional-scaling solution was fit to each matrix using isoMDS() from the MASS package in R (Venables & Ripley, Venables and Ripley (2002)). For the female set, the stress of the four-dimensional solution was 8.28, while the stress of the four-dimensional solution was 6.78 for the male set.² The visualization of the first two dimensions for both the female and male voices sets are presented in the left panel of Fig. 6.2. We have made no attempt to identify the dimensions.

To use the similarity scores alongside the other voice evaluation metrics, we created a distance score for each voice. Given that talker gender was a robust dimension on which the voices were separated in the MDS space, the voice distance score was calculated separately for female and male voices. Following methods of calculating vowel dispersion (e.g., Ménard et al. 2013), acoustic voice similarity was calculated using the four dimensions of the MDS solution for each gender by taking the

²Note that these stress values are not indicative of a particularly strong fit, indicating that more dimensions might ultimately provide a better characterization of the data.

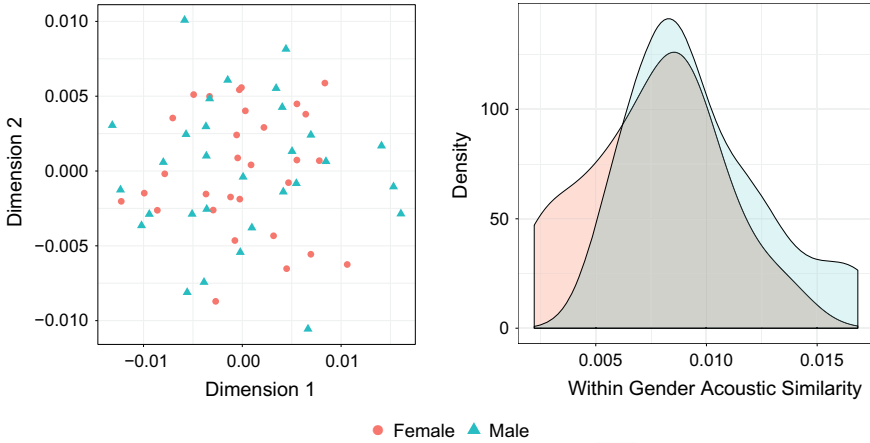


Fig. 6.2 The first two dimensions of the four-dimensional scaling solutions for the MFCC acoustic similarity of the 60 voices (left panel) and a density plot showing the distribution of within-gender acoustic variability for the 60 voices (right panel). Higher values along the x-axis in the density plot indicate more acoustically dissimilar voices. Female data are in red, and male in cyan

286 Euclidean distance of a voice from the average four-dimensional values for all other
 287 voices of that voice's gender. The distribution of these values was relatively normal,
 288 and is shown in the right panel in Fig. 6.2.

289 6.2.5 Perceptual Similarity

290 Even when measures of acoustic similarity use a transformation that models the
 291 human auditory system (like the mel-scale used in Sect. 6.2.4), such analyses may
 292 not adequately weigh or represent the cues that perceivers rely on when assessing
 293 voices. To address this, we conducted a similarity rating experiment using the voice
 294 corpus.

295 6.2.5.1 Participants

296 A research assistant who was a female native speaker of West Coast English (age =
 297 19) completed this task with all 60 voices.³

³While having just a single listener does affect the potential generalizability of our conclusions, we ultimately feel this single data point is better than no data point.

298 **6.2.5.2 Materials**

299 The 15 words spoken by the 60 voices were used as stimuli in this task.

300 **6.2.5.3 Procedure**

301 On a given trial, a random selection of nine words (three from each vowel group)
302 from a voice were presented in randomized order with 500 ms interstimuli interval,
303 followed by 1000 ms break, then a second voice comprising the same nine words.
304 After the presentation of the second voice, the participant rated the similarity of
305 the voices on a scale from 1 (very dissimilar) to 9 (very similar) using a computer
306 keyboard. All possible nonidentical pairs were presented in both orders resulting in
307 3480 trials (60 voices, 602 pairs = 3600, minus $60 \times 2 = 120$ identical pairs). Given
308 the tedious and repetitive nature of this task, it was conducted at the participant's
309 convenience over the course of several months.

310 **6.2.5.4 Results**

311 The ratings matrix was simplified in a similar way to the acoustic similarity data.
312 Again, a combined analysis demonstrated that the first dimension was based on
313 voice gender, so separate within-gender analyses were fit. A scree plot of stress
314 suggested an elbow at four dimensions for both analyses and thus a four-dimensional
315 nonmetric multidimensional-scaling solution was fit to each matrix using isoMDS()
316 from the MASS package in R (Venables and Ripley, 2002). The stress of the four-
317 dimensional solution was 8.36 for the female set and 7.48 for the male set of voices.⁴
318 The visualization of the first two dimensions for both the female and male voices
319 sets are presented in the left panel of Fig. 6.2.

320 For comparison with the other measures, perceptual voice similarity was calcu-
321 lated in an identical way to the similarity data. That is, separate distance scores were
322 created for male and female voices by using the four dimensions from the MDS
323 solutions and finding the mean Euclidean distance for each voice by gender. The
324 distribution of these values is shown in right panel of Fig. 6.3.

325 **6.2.6 Subjective Voice Ratings**

326 To examine how listeners' subjective impressions of a voice's attractiveness and
327 stereotypicality affect voice memory alongside the more objective measures described
328 above, we collected the metrics described below.⁵

⁴Again, these high stress values suggest that more dimensions could provide a better fit to the data.

⁵These subjective voice ratings were previously reported in Babel and McGuire (2015).

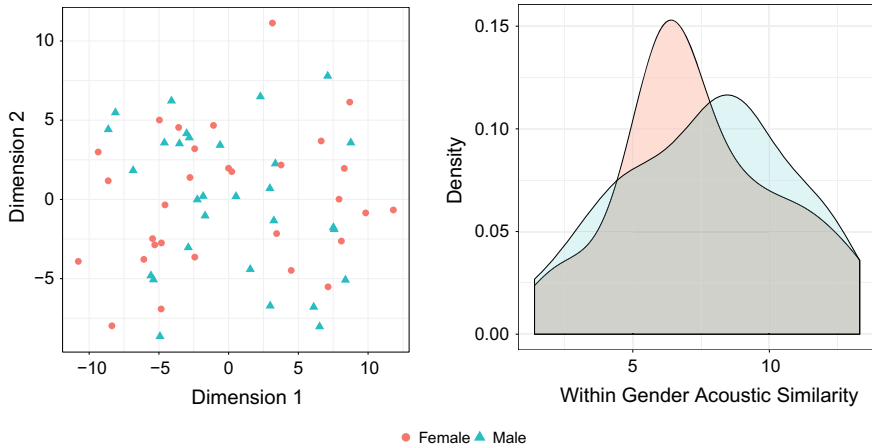


Fig. 6.3 The first two dimensions of the three-dimensional multidimensional-scaling solution for the perceptual similarity of the 60 voices (left panel) and a density plot showing the distribution of within-gender perceptual variability for the 60 voices (right panel), where higher values along the x-axis indicate more perceptually dissimilar voices. Female data are in red, and male in cyan

6.2.6.1 Participants

Sixty participants were recruited for explicit rating tasks from the student population of the University of California, Santa Cruz and received course credit or \$10 for their participation. Participants were divided into two groups of thirty (15 male, 15 female, each) and assigned to either the Stereotypicality rating group or the Attractiveness rating group.

6.2.6.2 Materials

The full set of 15 words for the 60 talkers were used in the tasks that elicited ratings of stereotypicality and attractiveness.

6.2.6.3 Procedure

For both experiments, subjects heard each voice say each of the 15 words followed by a pause where they were prompted to rate the voice using a 1–9 scale where 1 was “Very Unattractive” or “Very Atypical” and 9 was “Very Attractive” or “Very Typical.” All voices and words were presented in a randomized order. “Attractiveness” was not defined for the participant; they were free to evaluate the voice for sexual attractiveness or pleasantness.

Table 6.2 Mean and standard deviations of the Attractiveness and Stereotypicality Ratings for the male and female voices are shown in the leftmost columns. The Kendall's W values for the ratings are in the rightmost columns

	Female voices	Male voices	Female voices (W)	Male voices (W)
<i>Attractiveness</i>				
Female raters	5.05	4.67	0.274***	0.274***
Male raters	5.07	4.05	0.476***	0.185***
<i>Stereotypicality</i>				
Female raters	6.8	6.62	0.311***	0.255***
Male raters	6.54	6.52	0.325***	0.261***

Values marked with *** indicate p-values <0.001

6.2.6.4 Results

Female voices were overall rated as more Attractive and Stereotypical than male voices. Listeners' ratings were assessed for reliability using Kendall's W , and listeners showed a range of agreement levels. These values are given in Table 6.2.

6.2.7 Global Voice Assessment

While the six voice evaluation metrics are based on unique perception tasks posed to unique groups of listeners or, in the case of the acoustic similarity metric, an independent acoustic–auditory measurement, the metrics may indeed tap into common means of cognitively organizing voices. To assess this, we conducted a principal components analysis (PCA) on a centered and scaled data matrix using the averaged values for each talker's voice using a singular value decomposition strategy.⁶ The loadings of the PCA are shown in Table 6.3 and the model summary is presented in Table 6.4. The first principal component accounts for only about 32% of the variance in the data, and the loadings of this component illustrate the positive relationship between perceived attractiveness and stereotypicality along with the negative relationship of these two dimensions with categorization fluency (Babel & McGuire, 2015). The second principal component appears to show a negative relationship between acoustic similarity and intelligibility of the voices. The third component seems to be driven by perceptual similarity.

Somewhat surprisingly, it takes until the fifth principal component for the 95% of the variance to be accounted for. This suggests that not much is achieved through this process of dimensionality reduction and these dimensions, while not completely independent, are not wholly interconnected.

⁶This was done using the `prcomp()` command in base R.

Table 6.3 Rotation of the six voice evaluation metrics and the principal component loadings

	PC1	PC2	PC3	PC4	PC5	PC6
Attractiveness	0.5774	-0.3100	0.2280	-0.1354	0.2472	0.6625
Stereotypicality	0.6074	-0.03364	0.1849	-0.3875	-0.0630	-0.6645
Categorization fluency	-0.4454	-0.3349	0.4282	-0.2232	0.6449	-0.2021
Intelligibility	-0.2520	-0.6179	-0.1913	-0.4972	-0.5133	0.0863
Perceptual similarity	0.1189	-0.1101	-0.8328	-0.1379	0.5039	-0.0848
Acoustic similarity	0.1466	0.6298	0.0165	-0.7170	0.0456	0.253

Table 6.4 Summary of the PCA on the six voice evaluation metrics

	PC1	PC2	PC3	PC4	PC5	PC6
Standard deviation	1.3931	1.1541	1.0860	0.8931	0.6911	0.5222
Proportion of variance	0.3234	0.2220	0.1966	0.1329	0.0796	0.04546
Cumulative proportion	0.3234	0.5454	0.7420	0.8750	0.9545	1.000

368 Given this, these metrics will be used below to predict performance in the voice
 369 memory task.

370 6.3 Voice Memory Experiment

371 The previous sections summarized data evaluating voices using several subjective
 372 measures (Stereotypicality, Attractiveness), online processing measures (Categoriza-
 373 tion Fluency, Intelligibility), and similarity (Acoustic–Auditory, Perceptual). In this
 374 section, we turn to the original goal of the paper and use these measures to predict
 375 listeners' ability to recall individual voices. Following previous literature, we expect
 376 that less typical voices will be easier to recall than more typical voices, and the
 377 following experiment will elucidate which of our measures are best at predicting
 378 this.

379 6.3.1 Methods

380 6.3.1.1 Participants

381 There were 42 listeners in four counterbalanced groups. All were native speakers of
 382 American English and had lived in California since toddlerhood. They were recruited

383 at the University of California, Santa Cruz, and received partial course credit for their
384 participation.

385 **6.3.1.2 Procedure**

386 The voices were divided into two lists of 30 and two word sets for the purposes of
387 balancing. The two voice lists were designed to have an equal number of male and
388 female voices in each and to be roughly equivalent in stereotypicality. The words
389 were randomly assigned to two lists with the constraint that each list had two words
390 for each vowel. In the exposure phase, listeners were presented with one list of 30
391 voices each saying six words and asked to type each word as accurately as possible.
392 This was similar to Mullennix et al. (2011) in that the exposure phase was a linguistic
393 task rather than a talker-focused one. After a brief self-paced break listeners were
394 given a surprise memory task where they were again presented with voices. This
395 procedure was identical to the exposure phase except that (1) the full set of 60 voices
396 was used and (2) rather than type in the words spoken, subjects were asked to identify
397 each voice as either Old (i.e., previously heard) or New (i.e., not previously heard),
398 logging their response on labeled buttons on a serial response box. Participants were
399 run in groups of up to three at a time in a sound-attenuated booth.

400 **6.3.2 Results**

401 **6.3.2.1 Listener-Focused Analysis**

402 To model listeners' decisions regarding the voices, a mixed-effects logistic regression
403 model was used to analyze the probability that listeners could correctly identify the
404 voices as New or Old. Given that the dimensionality reduction of the PCA was not
405 particularly effective (e.g., it took five principal components to account for 95% of
406 the variance when six variables were entered into the model), we also assessed the
407 collinearity of the six voice evaluation metrics via condition number and a variance
408 in inflation (VIF) calculation prior to including these metrics in the model. The
409 condition number analysis, following Baayen (2008), gave a kappa statistic of 22,
410 and the highest VIF value was 2.5. These are both generally considered moderate in
411 terms of collinearity. Given this and the results of the PCA, we opted to include the
412 six metrics in the model. To assist in the interpretability of the model output, however,
413 the six metrics were entered into the model as fixed effects with interactions with
414 New/Old, but not as interactions with each other. New/Old was entered into the
415 model as a fixed effect with dummy coding; New was the reference level. There
416 were random slopes for listeners, along with the random intercepts for New/Old and

Table 6.5 Model output for the listener-focused voice memory analysis

	Estimate	Standard error	z-value	p-value
Intercept	-0.1841	0.1678	-1.097	0.2726
New/Old	0.72821	0.29796	2.444	0.01453*
Attractiveness	-0.1719	0.0999	-1.721	0.08517
Stereotypicality	-0.5680	0.09735	-5.835	< 0.001***
Categorization fluency	-0.0886	0.0781	-1.135	0.2563
Intelligibility	0.02133	0.07426	-0.287	0.7739
Perceptual similarity	0.0379	0.07043	0.539	0.5901
Acoustic similarity	-0.2146	0.07218	-2.97	0.0029**
New/Old:Attractiveness	0.1562	0.12655	1.234	0.2170
New/Old:Stereotypicality	0.7499	0.1279	5.863	< 0.001***
New/Old:Categorization fluency	-0.0071	0.10647	-0.067	0.9468
New/Old:Intelligibility	0.19416	0.0991	1.959	0.0501
New/Old:Perceptual similarity	-0.11858	0.0966	-1.226	0.2201
New/Old:Acoustic similarity	0.24846	0.0980	2.535	0.0603

P-values marked with * indicate values < 0.05, ** indicates values < 0.01, and *** indicates values < 0.001

417 the voice evaluation metrics. All of the voice evaluation metrics were centered and
418 scaled prior to the regression analysis.⁷

419 The results of this analysis are summarized in Table 6.5. The lack of a significant
420 intercept indicates that listeners were not very accurate at identifying previously
421 unheard or novel voices as New. The effect of New/Old illustrates that listeners
422 were more accurate at correctly recalling old voices as old than new voices as new.
423 In terms of the voice metrics, Stereotypicality was a significant predictor, and it
424 also surfaced in a significant interaction with New/Old. New voices that had been
425 independently rated as less stereotypical were more accurately identified as new than
426 more stereotypical new voices, and old voices which were more stereotypical were
427 more accurately identified as old than older voices that were less stereotypical. This
428 relationship is shown in the left panel of Fig. 6.4. Acoustic Similarity was also a
429 significant predictor. Listeners were less accurate on new voices that were further
430 from the Euclidean mean of the voice set. That is, listeners were more accurate
431 with voices that were more acoustically typical, somewhat in contradiction with the
432 Stereotypicality results. This relationship is shown in the right panel of Fig. 6.4.

⁷The following code was used: `glmer(Accuracy ~ New/Old + Attractiveness + Stereotypicality + Categorization Fluency + Intelligibility + Perceptual Similarity + Acoustic Similarity + New/Old:Attractiveness + New/Old:Stereotypicality + New/Old:Categorization Fluency + New/Old:Intelligibility + New/Old:Perceptual Similarity + New/Old:Acoustic Similarity + (1 + New/Old + Attractiveness + Stereotypicality + Categorization Fluency + Intelligibility + Perceptual Similarity + Acoustic Similarity |Listener))`.

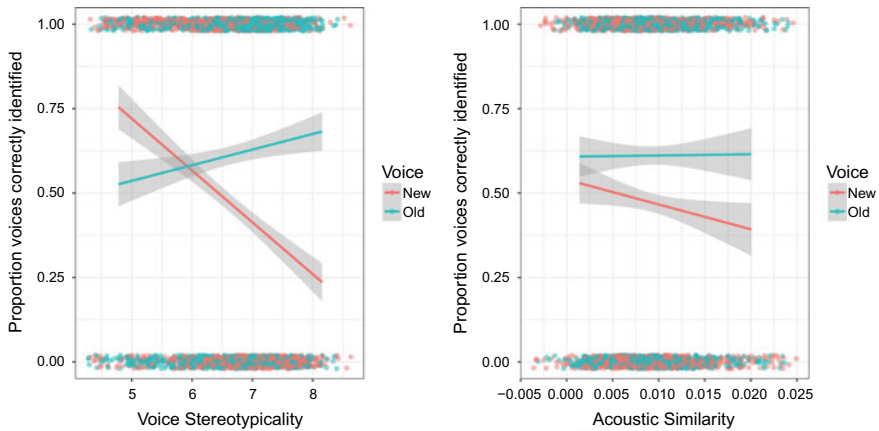


Fig. 6.4 The relationship between voice recall accuracy for Old and New voices and Stereotypicality (left panel) and Acoustic Similarity (right panel). The jittered points represent listener responses

433 6.3.2.2 Talker-Focused Analysis

434 To model voice memory with a focus on the talkers' voices, the signal detection
 435 theory measures of d' (sensitivity) and c (bias) were calculated across listeners for
 436 each voice (Macmillan & Creelman, 2004). For this analysis the data were averaged
 437 across listeners for each voice and correct responses to Old voices were assigned as
 438 hits and incorrect Old responses to New voices were assigned as false alarms. This
 439 calculation results in positive values of d' indicating that listeners correctly identified
 440 voices as Old or New, while negative values indicate listeners had more false alarms
 441 than hits and, thus, incorrectly classified the voices. The assignment of correct Old
 442 responses as hits also means that negative values of c , indicate a bias to respond Old
 443 and a positive number indicates a bias to respond New. These d' and c values were
 444 used as the dependent measures in simple linear regression models where each voice
 445 evaluation measure was entered as an independent variable along with talker gender.
 446 Because of the small number of observations one is left with in this style of analysis
 447 ($n = 60$, one data point per talker), we chose to run separate regression models for
 448 each voice evaluation metric.

449 Model results for the d' analysis are summarized in Table 6.6. They indicate
 450 voices which were lower in attractiveness and stereotypicality had higher d' values,
 451 indicating listeners were more sensitive to the New/Old decision for voices that
 452 were previously rated as less attractive or less stereotypical. The R^2 values indicate
 453 that this pattern was more robust along the Stereotypicality than the Attractiveness
 454 dimension. Figure 6.5 illustrates these patterns.

455 The c results complement these findings and are summarized in Table 6.7. There
 456 was a bias to respond Old to voices that had been rated as Attractive and Stereotypical.
 457 Again, there was a larger effect size for the Stereotypicality voice evaluation metric,
 458 compared to Attractiveness. These results are visualized in Fig. 6.6.

Table 6.6 Model summaries for the d' sensitivity talker-focused voice memory analysis. The Adjusted R^2 for each model's fit is reported in the final column

	Estimate	Standard error	z-value	p-value	Adjusted R^2
Intercept	1.22	0.32	3.83	<0.001***	
Attractiveness	-0.20	0.07	-3.11	0.003**	0.13
Intercept	2.41	0.48	5.06	<0.001***	
Stereotypicality	-0.33	0.07	-4.57	<0.001***	0.25
Intercept	-1.37	2.31	-0.59	0.56	
Categorization fluency	0.003	0.004	0.70	0.49	-0.009
Intercept	-1.98	1.30	-1.52	0.13	
Intelligibility	0.003	0.002	1.71	0.09	0.03
Intercept	-0.049	0.08	-1.01	0.08	
Perceptual similarity	2.508	0.10	0.89	0.32	-0.07
Intercept	0.48	0.17	2.87	0.0057**	
Acoustic similarity	-25.68	16.77	-1.531	0.13	0.022

P-values marked with ** indicate values <0.01 and *** indicates values <0.001

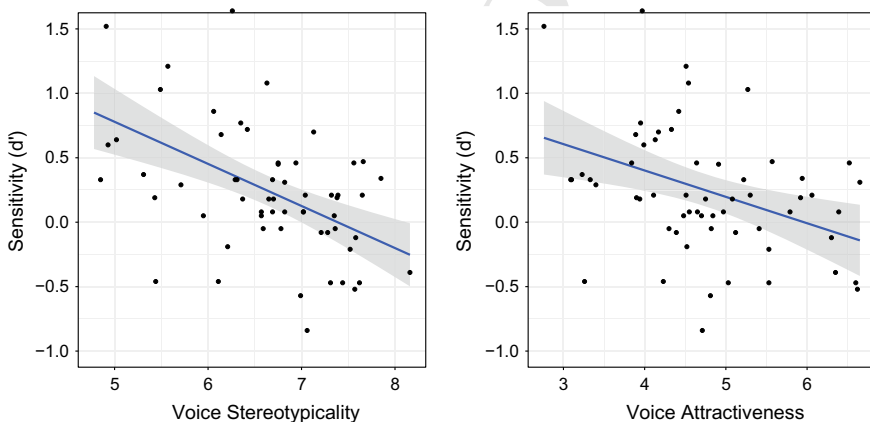


Fig. 6.5 Sensitivity by Stereotypicality (left panel) and Attractiveness (right panel) in the Voice Recall task. Each point represents a talker in the experiment

459 Together, these results indicate that listeners were more accurate in the voice
 460 memory task with voices that were less Attractive and Stereotypical, and there was
 461 a strong bias for listeners to respond Old to voices that were more Attractive and
 462 Stereotypical.

Table 6.7 Model summaries for the *c* bias talker-focused voice memory analysis. The Adjusted R^2 for the model fit is reported in the final column

	Estimate	Standard error	z-value	p-value	Adjusted R^2
Intercept	0.77	0.24	3.21	0.002 **	0.18
Attractiveness	-0.18	0.05	-3.68	<0.001***	
Intercept	1.89	0.34	5.54	<0.001***	0.36
Stereotypicality	-0.30	0.05	-5.86	<0.001***	
Intercept	-2.23	1.77	-1.26	0.21	0.008
Categorization fluency	0.004	0.003	1.21	0.23	
Intercept	0.02	1.03	0.02	0.99	-0.02
Intelligibility	-0.0002	0.002	-0.11	0.92	
Intercept	-0.092	0.412	-0.091	0.76	-0.01
Perceptual similarity	0.022	0.07	0.71	0.42	
Intercept	0.0095	0.133	0.071	0.943	-0.004
Acoustic similarity	-11.197	13.15	-0.85	0.39	

P-values marked with ** indicate values <0.01 and *** indicates values <0.001

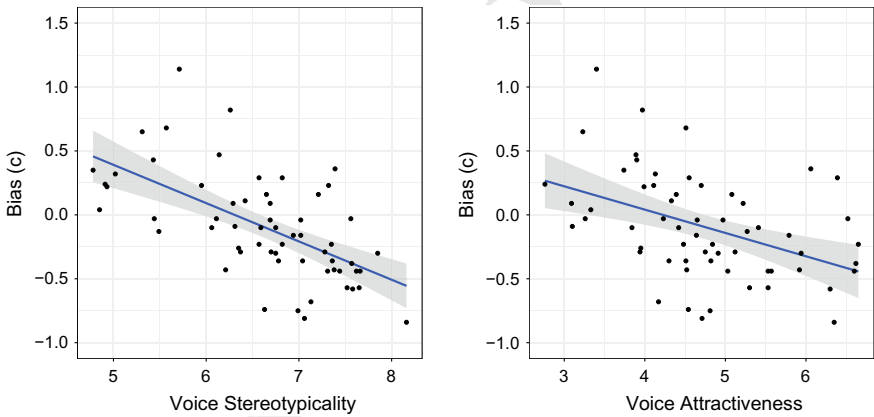


Fig. 6.6 Bias by Stereotypicality (left panel) and Attractiveness (right panel) in the Voice Recall task. Negative values indicate a bias to respond Old, while positive values indicate a bias to respond New

6.4 General Discussion

Listeners process the communicative linguistic signal of a voice while they evaluate it socially (Sumner et al., 2014). In this chapter, we used a combination of online intelligibility and processing measures, measures of acoustic–auditory and perceptual similarity, and subjective voice evaluations to predict voice memory. For decades, it has been established that voice evaluation related to distinctiveness or typicality was a strong predictor of listeners’ ability to recall voices (Papcun et al.,

Editor Proof

1989; Kreiman & Papcun, 1991; Mullennix et al., 2011). In line with these earlier claims, we find that our subjective measures of voice evaluation—perceived Stereotypicality and Attractiveness, two related dimensions for this set of voices (Babel & McGuire, 2015)—predict performance in a voice recall task, as did our measure of Acoustic–Auditory similarity. Notably, the more online measures of intelligibility and gender categorization fluency do not. Perceptual similarity also did not predict performance, but it is difficult to draw conclusions from one listener.

In this corpus of voices, we can conceive of voices that are more stereotypical and more attractive as being analogous to the voices that Papcun et al. (1989)’s listeners identified as hard-to-remember voices. In our study, these stereotypical and attractive voices are more accurately identified as old voices (i.e., voices previously heard in the experiment) when they are indeed old. Listener accuracy on stereotypical and attractive new voices that listeners were not exposed to was poor. The signal detection theoretic analyses illustrate that listeners had decreased sensitivity to more stereotypical and attractive voices and this was due to listeners having a strong bias to respond “old” to these stereotypical and aesthetically pleasing voices. Papcun, Kreiman, and colleagues (Papcun et al., 1989; Kreiman & Papcun, 1991) argue that their results support a prototype model of voice memory: voices that are typical are well-represented and thus trigger the illusion of experience. Our results complement these findings by providing insight into what voice attributes these prototypes are structured around. In the context of voice memory, it appears that more subjective voice evaluations are at the core of the prototype structure, particularly perceived stereotypicality, as opposed to more objective, online measures like intelligibility or categorization fluency or measures of voice similarity taken from the acoustic–auditory or perceptual space.

The results do raise a contradiction in that listeners were less accurate at identifying acoustically atypical voices as New while voices judged less stereotypical are more accurately identified. These two voice measures, Stereotypicality and Acoustic Similarity are not correlated for our data set [$r = -0.02$, $p = 0.25$]. Moreover, our measure of acoustic similarity is based on MFCCs, which while usefully exploited for automatic talker recognition systems, may not at all adequately capture the phonetic detail around which human listeners organize and distinguish voices. Our attempt to use an online measure of listener-derived voice similarity is stymied by the duration of the task, thus providing us with the perceptual space of a single listener. While the previous research aligns well with our results regarding stereotypicality and attractiveness, more research is necessary to understand the role of voice similarity in the acoustic and perceptual domains.

Sociocultural influences shape listeners’ interpretation and social assessment of voices and accents (Hay, Jennifer, Warren, Paul, & Drager, Katie, 2006; Babel & Russell, 2015), in addition to shaping the, for example, gender-specific realization of spoken language (Johnson, 2006; Foulkes, Docherty, & Watt, 2005). Listeners’ assessments of what is typical appear not to be based on veridical interpretation of the statistical distributions that listeners are exposed to, but rather are a reflection of a cognitive reorganization that is based on community standards and norms (Sumner et al., 2014; Babel & McGuire, 2015). The results of the voice memory task

515 reported here provide a concrete example of where this has implications: attractive
 516 and especially stereotypical voices are recalled less accurately because of a bias
 517 to assume they have been previously experienced. Individuals with more typical or
 518 attractive voices may thus receive a social benefit in terms of processing advantages
 519 that familiar accents experience.

520 6.5 Conclusion

521 These results generally support previous research that less typical and more unusual
 522 voices are more easily recalled from memory (Papcun et al., 1989; Kreiman & Pap-
 523 cun, 1991; Mullennix et al., 2011). Using several different evaluations of voices we
 524 find that stereotypicality and, to a lesser extent, attractiveness and acoustic similar-
 525 ity predict listeners' ability to recall voices, such that less stereotypical voices are
 526 recalled more easily, but there is a strong bias to determine that highly stereotypical
 527 voices have been previously heard. In contrast, online response time measures do
 528 not predict voice recall.

529 While further research is certainly necessary, a broader conclusion that can be
 530 gleaned from this study is that voices are organized and perceived fairly abstractly,
 531 with considerable reliance on social factors. This conclusion is a natural extension
 532 of the results. If online response time measures, which are typically diagnostic of
 533 experiential information and speed of processing, do not predict voice recall, then
 534 this negative result suggests that experience plays a more minimal role, or is dwarfed
 535 by the social factors that are tapped by asking listeners about attractiveness and
 536 stereotypicality. This is perhaps unsurprising as a voice is an aggregate of experiences
 537 and words. Many, if not most, exemplar models of speech (Pierrehumbert, 2001;
 538 Johnson, 1997) propose words as a basic unit of storage. In this study, participants
 539 were asked to recall voices holistically, after hearing six words produced by a voice,
 540 not respond "old"/"new" to individual words. Thus, when participants are asked
 541 about a voice as a whole, they rely more on abstracted, subjective information.

542 However, as is clear from diverse work in the speech sciences (Goldinger, 1998,
 543 1996; Nielsen, 2011; Palmeri, Goldinger, & Pisoni, 1993; Theodore & Miller, 2010;
 544 Dahan, Drucker, Sarah & Scarborough, 2008) individual instances in memory matter
 545 for speech perception. A full theory of voice organization will need to rectify such
 546 instances with more abstracted memories. Further research should elucidate this
 547 issue.

548 **Acknowledgments** We wish to thank Brianne Senior and Stephanie Chung for their contributions
 549 to this paper. Funding was, in part, provided by a Hampton Grant from the University of British
 550 Columbia and SSHRC to MB, and funding from the UCSC Humanities Institute to GM. Thanks to
 551 the audience at LabPhon14 in Tokyo for their feedback on this project.

References

- 552
- 553 Apicella, C. L., Feinberg, D. R. & Marlowe, F. W. (2007). Voice pitch predicts reproductive success
554 in male hunter-gatherers. *Biology Letters*, 3(6), 682–684.
- 555 Baayen, R. H. (2008). *Analyzing linguistics data: A practical introduction to statistics*. Cambridge:
556 CUP.
- 557 Babel, M., McGuire, G., & King, J. (2014b). Towards a more nuanced view of vocal attractiveness.
558 *PLoS One*, 9(2), e88616.
- 559 Babel, M., McGuire, G., Walters, S., & Nicholls, A. (2014a). Novelty and social preference in
560 phonetic accommodation. *Laboratory Phonology*, 5(1), 123–150.
- 561 Babel, M. (2012). Evidence for phonetic and social selectivity in spontaneous phonetic imitation.
562 *Journal of Phonetics*, 40(1), 177–189.
- 563 Babel, M., & McGuire, G. (2015). Perceptual fluency and judgments of vocal aesthetics and stereo-
564 typicality. *Cognitive Science*, 39(4), 766–787.
- 565 Babel, M., & Russell, J. (2015). Expectations and speech intelligibility. *The Journal of the Acoustical*
566 *Society of America*, 137(5), 2823–2833.
- 567 Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and
568 fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20(3–4), 255–272.
- 569 Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition*, 106(2),
570 707–729.
- 571 Bregman, M. R., & Creel, S. C. (2014). Gradient language dominance affects talker learning.
572 *Cognition*, 130(1), 85–95.
- 573 Clayards, M., Tanenhaus, M. K., Aslin, R. N., & Jacobs, R. A. (2008). Perception of speech reflects
574 optimal use of probabilistic speech cues. *Cognition*, 108(3), 804–809.
- 575 Clopper, C. G., Tamati, T. N., & Pierrehumbert, J. B. (2016). Variation in the strength of lexical
576 encoding across dialects. *Journal of Phonetics*, 58, 87–103.
- 577 Clopper, C. G. (2014). Sound change in the individual: Effects of exposure on cross-dialect speech
578 processing. *Laboratory Phonology*, 5(1), 69–90.
- 579 Clopper, C. G., & Pisoni, D. B. (2004). Effects of talker variability on perceptual learning of dialects.
580 *Language and Speech*, 47(3), 207–238.
- 581 Cristià, A., Mielke, J., Daland, R., & Peperkamp, S. (2013). Similarity in the generalization of
582 implicitly learned sound patterns. *Laboratory Phonology*, 4(2), 259–285.
- 583 Dahan, D., Drucker, S. J., & Scarborough, R. A. (2008). Talker adaptation in speech perception:
584 Adjusting the signal or the representations? *Cognition*, 108(3), 710–718.
- 585 Feinberg, D. R., DeBruine, L. M., Jones, B. C., & Perrett, D. I. (2008). The role of femininity and
586 averageness of voice pitch in aesthetic judgments of women’s voices. *Perception*, 37(4), 615–623.
- 587 Foulkes, P., Docherty, G., Watt, D. (2005). Phonological variation in child-directed speech. In
588 *Language*, pp. 177–206.
- 589 Goggin, J. P., Thompson, C. P., Strube, G., Simental, L. R. (1991). The role of language familiarity
590 in voice identification. *Memory & Cognition*, 19(5), 448–458.
- 591 Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and
592 recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*,
593 22(5), 1166.
- 594 Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological*
595 *Review*, 105(2), 251.
- 596 Hall, K. C., Allen, B., Fry, M., Mackie, S., McAuliffe, M. (2018). *Phonological CorpusTools*,
597 Version 1.2. <https://github.com/PhonologicalCorpusTools/CorpusTools/releases>.
- 598 Hay, J., Warren, P., & Drager, K. (2006). Factors influencing speech perception in the context of a
599 merger-in-progress. *Journal of Phonetics*, 34(4), 458–484.
- 600 Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In
601 *Talker Variability in Speech Processing*, pp. 145–165.
- 602 Johnson, K. (2006). Resonance in an exemplar-based lexicon: The emergence of social identity and
603 phonology. *Journal of Phonetics*, 34(4), 485–499.

- 604 Kinzler, K. D., & DeJesus, J. M. (2013). Northern = smart and Southern = nice: The development of
 605 accent attitudes in the United States. *The Quarterly Journal of Experimental Psychology*, 66(6),
 606 1146–1158.
- 607 Kreiman, J., & Papcun, G. (1991). Comparing discrimination and recognition of unfamiliar voices.
 608 *Speech Communication*, 10(3), 265–275.
- 609 Lippi-Green, R. (2012). *English with an accent: Language, ideology and discrimination in the*
 610 *United States*. Routledge.
- 611 Macmillan, N. A. & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology Press.
- 612 Ménard, L., Toupin, C., Baum, S.R., Drouin, S., Aubin, J., & Tiede, M. (2013). Acoustic and
 613 articulatory analysis of French vowels produced by congenitally blind adults and sighted adults.
 614 *The Journal of the Acoustical Society of America*, 134(4), 2975–2987.
- 615 Mielke, J. (2012). A phonetically based metric of sound similarity. *Lingua*, 122(2), 145–163.
- 616 Mullennix, J. W., Ross, A., Smith, C., Kuykendall, K., Conard, J., & Barb, S. (2011). Typicality
 617 effects on memory for voice: Implications for earwitness testimony. *Applied Cognitive Psychol-*
 618 *ogy*, 25(1), 29–34.
- 619 Newman, R. S., Clouse, S. A. & Burnham, J. L. (2001). The perceptual consequences of within-
 620 talker variability in fricative production. *The Journal of the Acoustical Society of America*, 109(3),
 621 1181–1196.
- 622 Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics*, 39(2),
 623 132–142.
- 624 Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception*
 625 *& Psychophysics*, 60(3), 355–376.
- 626 Orchard, T. L., & Yarmey, A. D. (1995). The effects of whispers, voice-sample duration, and voice
 627 distinctiveness on criminal speaker identification. *Applied Cognitive Psychology*, 9(3), 249–260.
- 628 Orena, A. J., Theodore, R. M., & Polka, L. (2015). Language exposure facilitates talker learning
 629 prior to language comprehension, even in adults. *Cognition*, 143, 36–40.
- 630 O'Toole, A.J., Deffenbacher, K. A., Valentin, D., McKee, K., Huff, D., & Abdi, H. (1998). The
 631 perception of face gender: The role of stimulus structure in recognition and classification. *Memory*
 632 *& Cognition*, 26(1), 146–160.
- 633 Palmeri, T. J., Goldinger, S. D. & Pisoni, D. B. (1993). Episodic encoding of voice attributes and
 634 recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory,*
 635 *and Cognition*, 19(2), 309.
- 636 Papcun, G., Kreiman, J., & Davis, A. (1989). Long-term memory for unfamiliar voices. *The Journal*
 637 *of the Acoustical Society of America*, 85(2), 913–925.
- 638 Perrachione, T. K., Chiao, J. Y., & Wong, P. C. (2010). Asymmetric cultural effects on perceptual
 639 expertise underlie an own-race bias for voices. *Cognition*, 114(1), 42–55.
- 640 Perrachione, T. K., Del Tufo, S. N. & Gabrieli, J. D. (2011). Human voice recognition depends on
 641 language ability. *Science*, 333(6042), 595–595.
- 642 Perrachione, T. K., Del Tufo, S. N. & Gabrieli, J. D. Human voice recognition depends on language
 643 ability'. In *Applied Psycholinguistics* (in press).
- 644 Perrachione, T. Recognizing speakers across languages. In Sascha Frühholz & Pascal Belin *The*
 645 *Oxford handbook of voice perception*. Oxford: Oxford University Press. (in press)
- 646 Perrachione, T. K., & Wong, P. C. M. (2007). Learning to recognize speakers of a non-native lan-
 647 guage: Implications for the functional organization of human auditory cortex. *Neuropsychologia*,
 648 45(8), 1899–1910.
- 649 Pierrehumbert, J. B. (2001). Exemplar dynamics: Word frequency, lenition and contrast. *Typological*
 650 *studies in language*, 45, 137–158.
- 651 Puts, D. A., Gaulin, S. J. & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism
 652 in human voice pitch. *Evolution and Human Behavior*, 27(4), 283–296.
- 653 Riding, D., Lonsdale, D., & Brown, B. (2006). The effects of average fundamental frequency and
 654 variance of fundamental frequency on male vocal attractiveness to women. *Journal of Nonverbal*
 655 *Behavior*, 30(2), 55–61.

- 656 Saxton, T. K., Caryl, P. G. & Craig Roberts, S. (2006). Vocal and facial attractiveness judgments of
 657 children, adolescents and adults: The ontogeny of mate choice. *Ethology*, 112(12), 1179–1185.
- 658 Senior, B., Hui, J., & Babel, M. (2018). Liu vs. Liu vs. Luke? Name influence on voice recall.
 659 *Applied Psycholinguistics*, 39(6), 1117–1146.
- 660 Stevenage, S. V., Clarke, G. & McNeill, A. (2012). The “other-accent” effect in voice recognition.
 661 *Journal of Cognitive Psychology*, 24(6), 647–653.
- 662 Strand, E. A. (1999). Uncovering the role of gender stereotypes in speech perception. *Journal of*
 663 *Language and Social Psychology*, 18(1), 86–100.
- 664 Sumner, M., & Kataoka, R. (2013). Effects of phonetically-cued talker variation on semantic encod-
 665 ing. In *The Journal of the Acoustical Society of America* 134(6), EL485–EL491.
- 666 Sumner, M., Kim, S. K., King, E., & McGowan, K. B. (2014). The socially weighted encoding of
 667 spoken words: a dual-route approach to speech perception. *Frontiers in Psychology*, 4, 1015.
- 668 Sumner, M., & Samuel, A. G. (2005). Perception and representation of regular variation: The case
 669 of final/t. *Journal of Memory and Language*, 52(3), 322–338.
- 670 Theodore, R. M., Myers, E. B. & Lomibao, J. A. (2015). Talker-specific influences on phonetic
 671 category structure. *The Journal of the Acoustical Society of America*, 138(2), 1068–1078.
- 672 Theodore, R. M., & Miller, J. L. (2010). Characteristics of listener sensitivity to talker-specific
 673 phonetic detail. *The Journal of the Acoustical Society of America*, 128(4), 2090–2099.
- 674 Thompson, C. P. (1987). A language effect in voice identification. *Applied Cognitive Psychology*,
 675 1(2), 121–131.
- 676 Van Bezooijen, R. (1995). Sociocultural aspects of pitch differences between Japanese and Dutch
 677 women. *Language and Speech*, 38(3), 253–265.
- 678 Vaughn, Vaughn, C., Baese-Berk, M., & Idemaru, K. Re-examining phonetic variability in native
 679 and non-native speech. In *Phonetica*, pp. 1–32 (in press).
- 680 Venables, W. N. & Ripley, B. D. (2002). *Modern applied statistics with S*. Fourth. ISBN 0-387-
 681 95457-0. New York: Springer. <http://www.stats.ox.ac.uk/pub/MASS4>.
- 682 Wade, T., Jongman, A., & Sereno, J. (2007). Effects of acoustic variability in the perceptual learning
 683 of non-native-accented speech sounds. *Phonetica*, 64(2–3), 122–144.
- 684 Winters, S. J., Levi, S. V. & Pisoni, D. B. (2008). Identification and discrimination of bilingual
 685 talkers across languages. *The Journal of the Acoustical Society of America*, 123(6), 4524–4538.
- 686 Xie, X., & Myers, E. (2015). The impact of musical training and tone language experience on talker
 687 identification. *The Journal of the Acoustical Society of America*, 137(1), 419–432.
- 688 Yarmey, A. D. (1991). Descriptions of distinctive and non-distinctive voices over time. *Journal of*
 689 *the Forensic Science Society*, 31(4), 421–428.
- 690 Zuckerman, M., & Miyake, K. (1993). The attractive voice: What makes it so? *Journal of Nonverbal*
 691 *Behavior*, 17(2), 119–135.

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Voice Attractiveness	
Series Title		
Chapter Title	Voice, Sexual Selection, and Reproductive Success	
Copyright Year	2020	
Copyright HolderName	Springer Nature Singapore Pte Ltd.	
Corresponding Author	Family Name	Suire
	Particle	
	Given Name	Alexandre
	Prefix	
	Suffix	
	Role	
	Division	Institut des Sciences de l'Evolution de Montpellier
	Organization	University of Montpellier, Centre National de la Recherche Scientifique, Institut pour la Recherche et le Développement, Ecole Pratique des Hautes Etudes – Place Eugène Bataillon
	Address	34095, Montpellier, France
	Email	alexandre.suire@umontpellier.fr
Author	Family Name	Raymond
	Particle	
	Given Name	Michel
	Prefix	
	Suffix	
	Role	
	Division	Institut des Sciences de l'Evolution de Montpellier
	Organization	University of Montpellier, Centre National de la Recherche Scientifique, Institut pour la Recherche et le Développement, Ecole Pratique des Hautes Etudes – Place Eugène Bataillon
	Address	34095, Montpellier, France
	Email	michel.raymond@umontpellier.fr
Author	Family Name	Barkat-Defradas
	Particle	
	Given Name	Melissa
	Prefix	
	Suffix	
	Role	
	Division	Institut des Sciences de l'Evolution de Montpellier
	Organization	University of Montpellier, Centre National de la Recherche Scientifique, Institut pour la Recherche et le Développement, Ecole Pratique des Hautes Etudes – Place Eugène Bataillon
	Address	34095, Montpellier, France
	Email	melissa.barkat-defradas@umontpellier.fr

Abstract

Beyond the linguistic content it conveys, voice is one of the fundamental aspects of human communication. It conveys an array of bio-psycho-social information about a speaker and enables the expression of a wide range of emotional and affective states so as to elicit a whole range of auditory impressions. Such aspects are of a great importance in determining the outcomes of competitive and courtship interactions as they influence the access to mating partners and thus reproduction. Sexual selection, the mechanism that promotes biological and social traits that confer a reproductive benefit, provides an interesting theoretical framework to understand the functional role of the human voice from an evolutionary perspective. This chapter aims to provide an overview of the research that lies at the crossroad of the human voice and evolutionary biology.

Keywords

Sexual selection - Reproductive success - Mate choice - Contest competition - Voice - Attractiveness

Chapter 7

Voice, Sexual Selection, and Reproductive Success



Alexandre Suire, Michel Raymond, and Melissa Barkat-Defradas

Abstract Beyond the linguistic content it conveys, voice is one of the fundamental aspects of human communication. It conveys an array of bio-psycho-social information about a speaker and enables the expression of a wide range of emotional and affective states so as to elicit a whole range of auditory impressions. Such aspects are of a great importance in determining the outcomes of competitive and courtship interactions as they influence the access to mating partners and thus reproduction. Sexual selection, the mechanism that promotes biological and social traits that confer a reproductive benefit, provides an interesting theoretical framework to understand the functional role of the human voice from an evolutionary perspective. This chapter aims to provide an overview of the research that lies at the crossroad of the human voice and evolutionary biology.

Keywords Sexual selection · Reproductive success · Mate choice · Contest competition · Voice · Attractiveness

7.1 Evolutionary Background

7.1.1 Sexual Selection

Sexual selection is an evolutionary process by which a specific trait, either biological or social, is selected depending on the advantages it confers to the individual that bears it in order to access sexual partners for reproduction (Darwin, 1871). Reproductive

A. Suire (✉) · M. Raymond · M. Barkat-Defradas
Institut des Sciences de l'Evolution de Montpellier, University of Montpellier,
Centre National de la Recherche Scientifique, Institut pour la Recherche et le Développement,
Ecole Pratique des Hautes Etudes – Place Eugène Bataillon, 34095 Montpellier, France
e-mail: alexandre.suire@umontpellier.fr

M. Raymond
e-mail: michel.raymond@umontpellier.fr

M. Barkat-Defradas
e-mail: melissa.barkat-defradas@umontpellier.fr

© Springer Nature Singapore Pte Ltd. 2020

B. Weiss et al. (eds.), *Voice Attractiveness*, Prosody, Phonology and Phonetics,
https://doi.org/10.1007/978-981-15-6627-1_7

131

19 success is thus a key aspect to assess. It describes an individual's capacity to pass its
 20 genes onto the next generation in a way that its descendants can pass it too. It can
 21 be estimated, given the situations, by one or several components, such as survival,
 22 fertility, or the number of offspring that are produced in the next generation. Sexual
 23 selection can be divided into two distinct selection processes: intra- and intersexual
 24 competition (Andersson, 1994).

25 On one hand, intrasexual competition refers to contest competition that occurs
 26 between same-sex individuals. When competition implies a physical confrontation,
 27 sexual selection will favor the evolution of any characteristic that strengthens the
 28 force and endurance of individuals, or any characteristic that diminishes the physical
 29 prowess of competitors. This leads to the evolution of specific "weapons" designed to
 30 repel and fight conspecifics. For instance, the antlers of male red deers are important
 31 physical attributes in duels during the mating season (Clutton-Brock, Guinness, &
 32 Albon, 1982), likewise the impressive body size of male sea lions, a key determinant
 33 of male–male fights to access harem of females (Ralls & Mesnick, 2009).

34 On the other hand, intersexual competition refers to the process of competition
 35 that depends on the choice made by opposite sex members, a mechanism commonly
 36 termed mate choice. This mechanism depends on sexual attractiveness (Sect. 7.2b
 37 deals with it). Evolutionary theory predicts that the sex that invests the more in
 38 reproduction (in the form of anisogamy and parental care) should have the scrutiny
 39 upon choosing a mate. This type of selection explains the origin of many extravagant
 40 characteristics, such as vivid colors, excessive plumage, and complex songs in male
 41 bird species (Bennett & Owens, 2002). Such traits are usually termed "ornaments".
 42 The most classical example is the tail of the blue peafowl, with its elongated upper
 43 tail which bears colorful eyespots.

44 In humans, many specific traits, such as height, the body size, and the immune
 45 system have been well studied under sexual selection theory and have provided a
 46 better understanding of their function within human mating systems (Miller, 1998;
 47 Puts, 2010). As we will see, contest competition and mate choice are two important
 48 evolutionary mechanisms that can also shed light on the evolution of the human
 49 voice.

50 7.1.2 *Vocal Dimorphism*

51 Humans display one of the most important vocal acoustic sexual dimorphism across
 52 anthropoids (Puts et al., 2016).

53 Differences in acoustic characteristics between the voices of men and women have
 54 long been recognized and studied (Titze, 1989). Men's vocal tract is about 15–20%
 55 longer than women because of their larger larynx and lower placement in the neck.
 56 Men's vocal chords are also about 50% longer and significantly more massive than
 57 those of women. These anatomical differences, which develop during puberty under
 58 the influence of the estrogen/testosterone ratio, explain the lower vocal resonant
 frequencies of male voices (Fitch & Giedd, 1999). Most notably, the fundamental

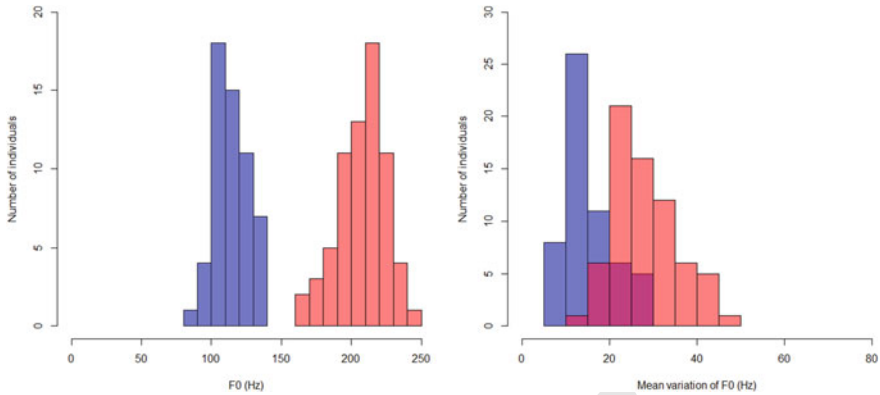


Fig. 7.1 Distribution of F0, F0-SD, and mean values of formant frequencies (F1–F4) for the vowels /a/, /i/ and /u/ for men (blue) and women (red). Purple values represent overlap between sexes. Acoustic data drawn from spontaneous speech; $n_{men} = 60$, $n_{women} = 68$ (Suire, unpublished data)

59 frequency shows relatively little to no overlap between the two sex: women's fun-
 60 damental frequency is typically double that of men (Fig. 7.1). Additionally, men
 61 display lower formant frequencies from F1 to F4 compared to women, and such differ-
 62 ences are consistent across different types of vowels (Simpson, 2009). Although
 63 less understood, the variation of F0 (generally noted as F0-SD) also appears to be
 64 sexually dimorphic, with men having a more monotonous voice than women (Puts
 65 et al., 2012).

66 Although the proximate mechanisms (i.e., physiology and anatomy) explain the
 67 observed difference between men and women, it does not tell which evolutionary
 68 factor has led to this phenomenon. When a trait shows a strong dimorphism between
 69 the two sex, it is reasonably well grounded to see sexual selection as a potential
 70 explaining factor. Although vocal attractiveness and dominance may be less relevant
 71 to human mating success in modern life than it has been during most of human evo-
 72 lution, the underlying logic of the following studies is that past contest competition
 73 and mate choice would have favored signals of threat potential and mate attraction
 74 (Puts, 2010).

75 7.2 The Functional Role of the Human Voice

76 7.2.1 Contest Competition and Vocal Dominance

77 Within same-sex competition, dominance is a key perception to assess. It can be
 78 defined as the capacity of one individual to repel competitors. Several studies have
 79 highlighted the importance of the fundamental and formant frequencies in the per-
 80 ception of both social and physical dominance, especially in men.

81 For instance, it has been regularly shown that men with a more masculine voice,
82 i.e., lower F0 and formant frequencies, are perceived as more dominant by same-sex
83 individuals, in both experimental settings (Feinberg, Jones, Little, Burt, & Perrett,
84 2005; Feinberg et al., 2006; Puts et al. 2006; Puts et al. 2007; Jones, Feinberg,
85 DeBruine, Little, & Vukovic, 2010; Watkins et al. 2010; Wolff & Puts, Wolff &
86 Puts 2010) and correlational studies (Aronovitch, 1976; Hodges-Simeon, Gaulin,
87 & Puts, 2010). Moreover, in a competitive setting, men who perceived themselves
88 as more dominant speak in a lower voice pitch and in a more monotonous manner
89 when speaking to competitors. Conversely, men who feel non-confident or more
90 “submissive” speak in a higher voice pitch (Puts et al., 2006). Interestingly, aggressive
91 and dominant communicative behavior can possibly go beyond simple acoustics, by
92 differentially producing phonetic variants relevant to the perception of masculinity
93 (Kempe, Puts, & Cárdenas, 2013). For instance, taller and more masculine men with
94 higher levels of circulating testosterone levels used less the alveolar stop consonant
95 /t/, as a mean to display threat potential. Effects of side observer or context-dependent
96 displays of aggression may be equally important to signal power and authority to an
97 audience, as it has been reported that observers seeing a man speaking aggressively
98 to other men are perceived as more dominant (Jones, DeBruine, Little, Watkins, &
99 Feinberg, 2011).

100 Another consequence of having a deeper voice is that it can lead to higher social
101 positions in men. For instance, it has been shown that people prefer to select a leader
102 with a more masculine voice (Anderson & Klofstad, 2012; Klofstad, Anderson,
103 & Peters, 2012), which can also influence voting behaviors (Tigue et al., 2012)
104 and predict actual presidential election outcomes (Klofstad, 2016; Banai, Banai,
105 & Bovan, 2017). Interestingly, voice pitch can be linked to leadership’s positions
106 within companies: CEOs with lower pitch voices managed larger companies, earned
107 more money, and enjoy longer tenures (Mayew et al., 2013). More generally, voice
108 pitch and formant frequencies seem to signal potential threat and aggression, higher
109 social status (including social dominance), all of which may have been particularly
110 important in past human environments (Puts, 2010).

111 For women, there are relatively few studies that have looked at the acoustic cor-
112 relates of dominance. One study from Borkowska & Pawlowski (2011) showed that
113 men and women perceived women with lower voice pitch as more dominant, with
114 women being more sensitive to this vocal cue than men. Another study showed that
115 feminine voices were perceived as more flirtatious and more attractive to men, and
116 women were most sensitive to formant dispersion (i.e., the relative distance of two
117 adjacent formants) than the fundamental frequency, suggesting that women may track
118 competitors’ femininity using this vocal cue (Puts et al., 2011).

119 The lack of studies for women’s vocal dominance can be partly explained by
120 the fact that past research has shown that competition among women, at least dur-
121 ing human evolutionary history, relies very little on physical combat or aggression;
122 women are assumed to be more prone to use indirect aggression. Such attempts may
123 include social manipulation, for instance, by spreading false information about one’s
124 reputation or interfering with friendships and group inclusion of competitors (Fisher,
125 2015). Therefore, this kind of competition does not lead to larger, taller, and stronger

126 statures in women, and thus women do not need to convey impressions of dominance
 127 or largeness through their vocal features against competitors.

128 Several authors have recently argued that intrasexual competition has mainly
 129 driven the evolution of several morphological traits in men, including voice pitch
 130 and its resonant frequencies (Puts, 2010; Hill et al., 2013; Kordsmeyer, Hunt, Puts,
 131 Ostner, & Penke, 2018), but mate choice should not be regarded as an insignificant
 132 evolutionary force in shaping vocal acoustic features (Suire et al., 2018).

133 7.2.2 *Mate Choice and Vocal Attractiveness*

134 Attractiveness, which can be defined as the capacity of one individual to attract oppo-
 135 site sex members, is an important component of voice perception in seductive and
 136 romantic settings. Other perceptions, such as the propensity to fidelity or trustwor-
 137 thiness, are also possibly important indexical cues to assess (Vukovic et al., 2011;
 138 O'Connor, Pisanski, Tigue, Fraccaro, & Feinberg, 2014a).

139 In men, consensus toward the attractiveness of relatively more masculine voices
 140 has been well established, that is, a relatively lower voice pitch (Collins, 2000; Fein-
 141 berg et al., 2005, 2006, 2008; Ridings et al., 2006; Jones et al. 2010, but see Shirazi,
 142 Puts, & Escasa-Dorne, 2018). Additionally, simultaneously masculinizing pitch and
 143 formant frequencies increases men's vocal attractiveness (Feinberg et al. 2005, 2006;
 144 Puts, 2005). However, preferences for vocal monotonicity are contradictory (Ridings
 145 et al., 2006; Hodges-Simeon et al., 2010) and further studies are needed. Nonethe-
 146 less, women's visual object memory seems to increase after hearing masculine male
 147 voices, but not after hearing feminine male voices or female voices, suggesting that
 148 women may be particularly attuned to masculine voices (Smith, Jones, Feinberg,
 149 & Allan, 2012). Voice pitch and formants are well-studied acoustic correlates of
 150 voice attractiveness, but multiple components of voice quality have not been studied
 151 within an evolutionary context and are known to potentially affect vocal attractive-
 152 ness, such as vocal roughness and breathiness (Suire et al., 2018). In addition, as for
 153 vocal dominance, attractiveness can go beyond the acoustics' limits, as it appears that
 154 specific sociolinguistic dialects, combined with a lower voice pitch, are preferentially
 155 selected by women (O'Connor et al., 2014b).

156 Interestingly, women's preferences for vocal masculinity seem to shift during
 157 the ovulatory cycle. Given that hormonal profiles (i.e., levels of progesterone and
 158 estradiol) vary during the ovulatory cycle, women may prefer less masculinized
 159 voices in men during the luteal phase as opposed to preferring masculinized voices
 160 in men toward ovulation peak (Puts, 2005; Feinberg et al. 2006). This result can be
 161 interpreted by the fact that women observe a trade-off when choosing a partner: a more
 162 cooperative and submissive individual during the luteal phase, with relatively lower
 163 testosterone levels, and a strong, testosterone-filled masculine men when approaching
 164 ovulation. Choosing the former can be understood by the fact that a more cooperative
 165 men is preferred when a woman seeks a long-term partner, particularly important so
 166 as to provide shelter and resources, and choosing the latter may be important when

167 a woman seeks a short-term partner (i.e., one-night stand) to maximize reproductive
 168 success (Buss & Schmitt, 1993). However, recent evidence has found no significant
 169 shift of women's preferences over the ovulatory cycle for both vocal and facial
 170 masculinity (Jones et al., 2018; Jünger, Kordsmeyer, Gerlach, & Penke, 2018).

171 Regarding men's preferences for women's voices, both experimental and corre-
 172 lational studies have found a consistent positive relationship between attractiveness
 173 and F0, that is, men are attracted in average to relatively higher voice pitch (Collins
 174 & Missing, 2003; Feinberg et al., 2008; Jones et al., 2010; Borkowska & Pawlowski,
 175 2011; Puts et al., 2011, however, see Tuomi & Fisher, 1979; Hughes et al., 2010,
 176 2014). However, this relationship might not be linear (Borkowska & Pawlowski,
 177 2011), suggesting a possible optimum for women's vocal attractiveness. Moreover,
 178 relatively higher formant dispersion (i.e., Df, the relative distance between two con-
 179 secutive formants, which correlates to the vocal tract length and perceived timbre) is
 180 also perceived as more attractive by men (Puts et al., 2011; Babel et al., 2014). Addi-
 181 tionally, the variation of the F0 has also been hypothesized to play upon the perception
 182 of indexical cues relevant in human competing and mating contexts (Leongómez et
 183 al., 2014; Hogdes-Simeon et al., 2010, 2011) but has so far received scant atten-
 184 tion. Although sexually dimorphic, it has only been tested for women's preferences
 185 (Bruckert et al., 2006; Hodges-Simeon et al., 2010), but one study suggests that men
 186 may be attracted to higher F0-SD profiles in women as it may be a cue of femininity
 187 (Leongómez et al., 2014).

188 Nonetheless, it is possible that vocal preferences for both men and women may not
 189 be culturally universal. As a matter of fact, physiological and anatomical differences
 190 do not explain the full variation in mean F0 between men and women, as individuals of
 191 both sexes exhibit considerable variation from one language to another (Rose, 1991;
 192 Traunmüller & Eriksson, 1995; Yamazawa & Hollien, 1992; Keating & Kuo, 2012;
 193 Andreeva et al., 2014; Pépiot, 2014). For instance, even under the same speaking
 194 conditions and balanced in age, American women exhibit a lower F0 than Japanese
 195 women (mean F0: 211 versus 224 Hz, Yamazawa & Hollien, 1992), while Bulgarian
 196 and Polish women exhibit a higher F0 than German and English women (mean F0:
 197 272 and 266 Hz versus 210 and 217 Hz, Andreeva et al., 2014). As males and females
 198 vary in mean F0 across various languages, this strongly suggests that some of the
 199 differences must be accounted for learned behavior or specific sociocultural practices
 200 (Simpson, 2009, e.g., Loveday, 1981). For instance, Dutch women display a lower
 201 F0 than Japanese women, and interestingly, Dutch and Japanese men tend to prefer
 202 female voices that exhibit culturally congruent vocal heights that is: low female
 203 voices versus high female voices for Dutch versus Japanese men, respectively (Van
 204 Bezooijen, 1995). Even in men, vocal attractiveness may not be solely predicted by
 205 voice pitch. For instance, the harmonics-to-noise ratio (a proxy of vocal breathiness)
 206 can predict Namibian men's vocal attractiveness (Šebesta et al., 2017).

7.3 Reproductive and Mating Successes

7.3.1 *Its Quantification*

Giving such observations, it is interesting to know how much variance can voice explain for an individual's overall reproductive success.

Investigating reproductive success within hunter-gatherer societies is of a particular interest because it is argued that such societies better reflect past human environments, practices, and cultures. However, studies are scarce. Hadza men with relatively lower F0 had higher reproductive success (Apicella et al., 2007). However, it has been recently reported that this relationship does not hold when controlling for reputation (Smith et al., 2017). In women, it has been shown that F0 significantly predicted several measures of reproductive success in a group of Namibian females: higher voice pitch was associated with overall higher reproductive success (Atkinson et al., 2012).

An easier measure of reproductive success is to measure mating success, and mostly the number of past-year sexual partners. Although less powerful, this measure is interesting because it represents a time window over which participants' recollections are expected to be accurate (contrary to asking the lifetime number of sexual partners) and the measured acoustics' characteristics are likely to be stable (Hodges-Simeon et al., 2011). Moreover, human mating success should be an important component of expected reproductive success in past environments, as it represents their potential fertility (Perusse, 1993).

Through a simulated dating game, lower F0 negatively correlated to men's mating success (Puts 2005), but another study found that it was not significant (Puts et al., 2006). Using a similar approach, men who spoke in a more monotonous manner (i.e., lower F0-SD) and faster when confronted to a competitor declared more sexual partners over the past year (Hodges-Simeon et al., 2011; Suire et al., 2018). Lastly, it has been reported that female and male vocal attractiveness (when rated by members of the opposite sex) could predict their mating success, their declared number of extra-pair copulations, and their age at first sexual intercourse (Hughes et al., 2004).

However, methodologies varied concerning speech samples used in previous studies; some studies used the recordings of spoken vowels and read speech without any contextual background (Apicella et al., 2007; Atkinson et al., 2012; Hughes et al., 2004; Smith et al., 2017). This approach of read speech has been also intensively used in perceptual studies when attractiveness and dominance need to be judged. This is problematic as it does not properly reflect how an individual vocally behave in ecological settings. Indeed, it has been regularly shown that studies conducted on read/reciting versus spontaneous speech produce quite different results (Howell & Kadi-Hanifi, 1991; Blaauw, 1992; Daly & Zue, 1992). As spontaneous speech is more difficult to analyze experimentally, it has been little used. Nonetheless, the simulated dating game studies have attempted to use it. These studies also provide an interesting way to quantify the relative contribution of both types of sexual selection in shaping vocal acoustic features (Hodges-Simeon et al., 2011; Suire et al., 2018).

249 7.3.2 *The Underlying Biological Quality of Voice*

250 To understand the ultimate reasons behind the correlations between vocal acoustic
 251 features, attractiveness, dominance, and reproductive success, the “honest signaling
 252 theory” offers an interesting explanation.

253 Regarding communication systems (i.e., the exchange of information through
 254 different mechanisms involving at least two parties), this theory posits that, giving
 255 conflicting interests between and within sexes, an individual should give an honest
 256 signal to the receiver rather than cheating. This is due to the fact that cheating will
 257 select over time for skeptical individuals who, in turn, have no benefits in “listening.”
 258 Thus, a communication system cannot emerge if false or manipulative information
 259 is exchanged actively. Here, voice has long been considered as an honest signal
 260 of overall biological quality, given the physiological and anatomical constraints in
 261 speech production (Feinberg et al., 2005; Evans et al., 2006; Puts et al., 2006). This
 262 means that voice should reflect another trait particularly relevant in contest and mate
 263 choice competitions, which are correlated to the aforementioned perceptions.

264 It has first been suggested that voice should be a reliable signal of body size, a
 265 feature particularly important in physical competitions between same-sex individu-
 266 als. Correlations between vocalizations’ frequencies have been well established in
 267 numerous species (Bowling et al., 2017) but surprisingly, such correlations are very
 268 weak within the human species. A meta-analysis showed that F0 did not explain
 269 more than 2% of the variation in body size, and formant frequencies only explained
 270 up to 10% (Pisanski et al., 2014a). This is interesting as both men and women still
 271 perceptually associate lower pitch voices to larger and taller individuals, and con-
 272 versely higher pitch voices to thinner and smaller individuals (Rendall, Vokey, &
 273 Nemeth, 2007, but see Pisanski et al., Pisanski et al. 2014b).

274 An alternative hypothesis is the immuno-handicap principle (Zahavi, 1975). It
 275 has been suggested that voice should reflect immuno-competence of individuals.
 276 As testosterone is a sexual hormone that is immunosuppressive, individuals with
 277 higher testosterone levels could bear the costs of impacting their immune system,
 278 and are thus supposedly in a better biological shape. Although lower F0 may be
 279 linked to higher testosterone circulating levels (Dabbs & Mallinger, 1999; Evans,
 280 Neave, Wakelin, & Hamilton, 2008), the immuno-handicap principle has yielded
 281 mixed results in humans (Roberts, Buchanan, & Evans, 2004; Boonekamp et al.,
 282 2008). Nonetheless, it has been reported that men plasma testosterone levels were
 283 positively correlated with sexual language and the use of swear words in the presence
 284 of their partners (Mascaro et al., 2018). Additionally, bioavailable testosterone was
 285 also found to be associated with the sound pressure level of the normal speaking
 286 voice in men and the softest speaking voice in women (Jost et al. 2018). The most
 287 convincing study to date has shown that some masculinized vocal characteristics
 288 were correlated to a specific antibody (Arnocky et al., 2018). The authors showed
 289 that men with lower voice pitch and formant position had higher concentrations of
 290 immunoglobulin A, an antibody produced by the mucus and constituting the first
 291 line of immune defense against toxins and infectious agents.

7.4 Conclusion and Future Perspectives

Since the beginning of the 2000s, research has provided a better understanding on the functional role of the human voice from an evolutionary perspective. Although considerable efforts have been dedicated, further studies are needed to understand understudied aspects.

For instance, although acoustic features seem to be heritable (Przybyla, Horii, & Crawford, 1992; Debruyne, 2002) and possibly related to the prenatal and/or pre-pubertal androgen exposure (Fouquet, Pisanski, Mathevon, & Reby, 2016), little is still known of its biological foundations. Other understudied acoustic components part of voice quality, such as roughness and breathiness, have also been little studied and are known to potentially affect attractiveness perceptions (Šebesta et al., 2017; Suire et al., 2018). Sociocultural variation in vocal preferences is also one important avenue for research. Additional efforts should also be devoted to study the interaction between linguistic material and vocal acoustic features to project indexical cues relevant to mating and competing contexts. Lastly, another interesting avenue for further research is to investigate vocal modulation, a capacity described as a volitional control of nonverbal vocal features evolutionarily linked to traits important in the context of sexual selection. However, context-dependent vocal modulation patterns have been little relatively studied so far, but provides evidence that individuals of both sexes alter several acoustic characteristics to signal traits relevant to contest competition and mate choice (see Pisanski et al. 2016 for a review).

References

- Anderson, R. C., & Klofstad, C. A. (2012). Preference for leaders with masculine voices holds in the case of feminine leadership roles. *PloS One*, 7(12), e51216.
- Andersson, M. B. (1994). *Sexual selection*. Princeton University Press.
- Andreeva, B., Demenko, G., Möbius, B., Zimmerer, F., Jügler, J., & Oleskowicz-Popiel, M. (2014). Differences of pitch profiles in Germanic and Slavic languages. In *Fifteenth Annual Conference of the International Speech Communication Association*.
- Apicella, C. L., Feinberg, D. R., & Marlowe, F. W. (2007). Voice pitch predicts reproductive success in male hunter-gatherers. *Biology Letters*, 3(6), 682–684.
- Arnocky, S., Hodges-Simeon, C., Ouellette, D., & Albert, G. (2018). Do men with more masculine voices have better immunocompetence? *Evolution and Human Behavior*.
- Aronovitch, C. D. (1976). The voice of personality: Stereotyped judgments and their relation to voice quality and sex of speaker. *The Journal of Social Psychology*, 99(2), 207–220.
- Atkinson, J., Pipitone, R. N., Sorokowska, A., Sorokowski, P., Mberira, M., Bartels, A., et al. (2012). Voice and handgrip strength predict reproductive success in a group of indigenous African females. *PloS One*, 7(8), e41811.
- Babel, M., McGuire, G., & King, J. (2014). Towards a more nuanced view of vocal attractiveness. *PloS one*, 9(2), e88616.
- Banai, I. P., Banai, B., & Bovan, K. (2017). Vocal characteristics of presidential candidates can predict the outcome of actual elections. *Evolution and Human Behavior*, 38(3), 309–314.
- Bennett, P. M., & Owens, I. P. (2002). Evolutionary ecology of birds: Life histories, mating systems and extinction.

- 335 Blaauw, E. (1992). Phonetic differences between read and spontaneous speech.
- 336 Boonekamp, J. J., Ros, A. H., & Verhulst, S. (2008). Immune activation suppresses plasma testosterone level: A meta-analysis. *Biology Letters*, 4(6), 741–744.
- 337 Borkowska, B., & Pawlowski, B. (2011). Female voice frequency in the context of dominance and
338 attractiveness perception. *Animal Behaviour*, 82(1), 55–59.
- 339
340 Bowling, D. L., Garcia, M., Dunn, J. C., Ruprecht, R., Stewart, A., Frommolt, K. H., et al. (2017).
341 Body size and vocalization in primates and carnivores. *Scientific Reports*, 7, 41070.
- 342 Bruckert, L., Lienard, J. S., Lacroix, A., Kreuzer, M., & Leboucher, G. (2006). Women use voice
343 parameters to assess men's characteristics. *Proceedings of the Royal Society B: Biological Sciences*,
344 273(1582), 83–89.
- 345 Buss, D. M., & Schmitt, D. P. (1993). Sexual strategies theory: An evolutionary perspective on
346 human mating. *Psychological Review*, 100(2), 204.
- 347 Charles, D. (1871). *The descent of man and selection in relation to sex*. London: Murray.
- 348 Clutton-Brock, T. H., Guinness, F. E., & Albon, S. D. (1982). *Red deer: Behavior and ecology of
349 two sexes*. University of Chicago press.
- 350 Collins, S. A. (2000). Men's voices and women's choices. *Animal behaviour*, 60(6), 773–780.
- 351 Collins, S. A., & Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal
352 Behaviour*, 65(5), 997–1004.
- 353 Dabbs, J. M., Jr., & Mallinger, A. (1999). High testosterone levels predict low voice pitch among
354 men. *Personality and Individual Differences*, 27(4), 801–804.
- 355 Daly, N. A., & Zue, V. (1992). Statistical and linguistic analyses of F0 in read and spontaneous
356 speech.
- 357 Debruyne, F., Decoster, W., Van Gijsel, A., & Vercammen, J. (2002). Speaking fundamental frequency
358 in monozygotic and dizygotic twins. *Journal of Voice*, 16(4), 466–471.
- 359 Evans, S., Neave, N., & Wakelin, D. (2006). Relationships between vocal characteristics and body
360 size and shape in human males: An evolutionary explanation for a deep male voice. *Biological
361 Psychology*, 72(2), 160–163.
- 362 Evans, S., Neave, N., Wakelin, D., & Hamilton, C. (2008). The relationship between testosterone
363 and vocal frequencies in human males. *Physiology & Behavior*, 93(4–5), 783–788.
- 364 Feinberg, D. R., Jones, B. C., Smith, M. L., Moore, F. R., DeBruine, L. M., Cornwell, R. E., et al.
365 (2006). Menstrual cycle, trait estrogen level, and masculinity preferences in the human voice.
366 *Hormones and Behavior*, 49(2), 215–222.
- 367 Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M., & Perrett, D. I. (2005). Manipulations of
368 fundamental and formant frequencies influence the attractiveness of human male voices. *Animal
369 Behaviour*, 69(3), 561–568.
- 370 Feinberg, D. R., DeBruine, L. M., Jones, B. C., & Little, A. C. (2008). Correlated preferences for
371 men's facial and vocal masculinity. *Evolution and Human Behavior*, 29(4), 233–241.
- 372 Fisher, M. L. (2015). Women's competition for mates: Experimental findings leading to ethological
373 studies. *Human Ethology Bulletin*, 30(1), 53–70.
- 374 Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study
375 using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, 106(3),
376 1511–1522.
- 377 Fouquet, M., Pisanski, K., Mathevon, N., & Reby, D. (2016). Seven and up: Individual differences
378 in male voice fundamental frequency emerge before puberty and remain stable throughout
379 adulthood. *Royal Society Open Science*, 3(10), 160395.
- 380 Hill, A. K., Hunt, J., Welling, L. L., Cárdenas, R. A., Rotella, M. A., Wheatley, J. R., et al. (2013).
381 Quantifying the strength and form of sexual selection on men's traits. *Evolution and Human
382 Behavior*, 34(5), 334–341.
- 383 Hodges-Simeon, C. R., Gaulin, S. J., & Puts, D. A. (2010). Different vocal parameters predict
384 perceptions of dominance and attractiveness. *Human Nature*, 21(4), 406–427.
- 385 Hodges-Simeon, C. R., Gaulin, S. J., & Puts, D. A. (2011). Voice correlates of mating success in
386 men: examining “contests” versus “mate choice” modes of sexual selection. *Archives of sexual
387 behavior*, 40(3), 551–557.

- 388 Hodges-Simeon, C. R., Gaulin, S. J., & Puts, D. A. (2010). Different vocal parameters predict
389 perceptions of dominance and attractiveness. *Human Nature*, 21(4), 406–427.
- 390 Hodges-Simeon, C. R., Gaulin, S. J., & Puts, D. A. (2011). Voice correlates of mating success in
391 men: Examining “contests” versus “mate choice” modes of sexual selection. *Archives of Sexual*
392 *Behavior*, 40(3), 551–557.
- 393 Howell, P., & Kadi-Hanifi, K. (1991). Comparison of prosodic properties between read and spon-
394 taneous speech material. *Speech Communication*, 10(2), 163–169.
- 395 Hughes, S. M., Farley, S. D., & Rhodes, B. C. (2010). Vocal and physiological changes in response
396 to the physical attractiveness of conversational partners. *Journal of Nonverbal Behavior*, 34(3),
397 155–167.
- 398 Hughes, S. M., Mogilski, J. K., & Harrison, M. A. (2014). The perception and parameters of
399 intentional voice manipulation. *Journal of Nonverbal Behavior*, 38(1), 107–127.
- 400 Hughes, S. M., Dispenza, F., & Gallup, G. G., Jr. (2004). Ratings of voice attractiveness predict
401 sexual behavior and body configuration. *Evolution and Human Behavior*, 25(5), 295–304.
- 402 Jones, B. C., Hahn, A. C., Fisher, C. I., Wang, H., Kandrik, M., Han, C., et al. (2018). No compelling
403 evidence that preferences for facial masculinity track changes in women’s hormonal status. *Psy-*
404 *chological Science*, 29(6), 996–1005.
- 405 Jones, B. C., Feinberg, D. R., DeBruine, L. M., Little, A. C., & Vukovic, J. (2010). A domain-
406 specific opposite-sex bias in human preferences for manipulated voice pitch. *Animal Behaviour*,
407 79(1), 57–62.
- 408 Jones, B. C., DeBruine, L. M., Little, A. C., Watkins, C. D., & Feinberg, D. R. (2011). ‘Eavesdrop-
409 ping’ and perceived male dominance rank in humans. *Animal Behaviour*, 81(6), 1203–1208.
- 410 Jost, L., Fuchs, M., Loeffler, M., Thiery, J., Kratzsch, J., Berger, T., et al. (2018). Associations of
411 sex hormones and anthropometry with the speaking voice profile in the adult general population.
412 *Journal of Voice*, 32(3), 261–272.
- 413 Jünger, J., Kordsmeyer, T. L., Gerlach, T. M., & Penke, L. (2018). Fertile women evaluate male
414 bodies as more attractive, regardless of masculinity. *Evolution and Human Behavior*.
- 415 Keating, P., & Kuo, G. (2012). Comparison of speaking fundamental frequency in English and
416 Mandarin. *The Journal of the Acoustical Society of America*, 132(2), 1050–1060.
- 417 Kempe, V., Puts, D. A., & Cárdenas, R. A. (2013). Masculine men articulate less clearly. *Human*
418 *Nature*, 24(4), 461–475.
- 419 Klothstad, C. A., Anderson, R. C., & Peters, S. (2012). Sounds like a winner: Voice pitch influences
420 perception of leadership capacity in both men and women. *Proceedings of the Royal Society of*
421 *London B: Biological Sciences*, rspb20120311.
- 422 Klothstad, C. A. (2016). Candidate voice pitch influences election outcomes. *Political Psychology*,
423 37(5), 725–738.
- 424 Kordsmeyer, T. L., Hunt, J., Puts, D. A., Ostner, J., & Penke, L. (2018). The relative importance of
425 intra- and intersexual selection on human male sexually dimorphic traits. *Evolution and Human*
426 *Behavior*.
- 427 Leongómez, J. D., Binter, J., Kubicová, L., Stolarová, P., Klapilová, K., Havlíček, J., & Roberts,
428 S. C. (2014). Vocal modulation during courtship increases perceptivity even in naive listeners.
429 *Evolution and Human Behavior*, 35(6), 489–496.
- 430 Loveday, L. (1981). Pitch, politeness and sexual role: An exploratory investigation into the pitch
431 correlates of English and Japanese politeness formulae. *Language and Speech*, 24(1), 71–89.
- 432 Mascaro, J. S., Rentscher, K. E., Hackett, P. D., Lori, A., Darcher, A., Rilling, J. K., & Mehl, M. R.
433 (2018). Preliminary evidence that androgen signaling is correlated with men’s everyday language.
434 *American Journal of Human Biology*, e23136.
- 435 Mayew, W. J., Parsons, C. A., & Venkatachalam, M. (2013). Voice pitch and the labor market
436 success of male chief executive officers. *Evolution and Human Behavior*, 34(4), 243–248.
- 437 Miller, G. F. (1998). How mate choice shaped human nature: A review of sexual selection and human
438 evolution. *Handbook of evolutionary psychology: Ideas, issues, and applications* (pp. 87–129).

- 439 O'Connor, J. J., Pisanski, K., Tigue, C. C., Fraccaro, P. J., & Feinberg, D. R. (2014a). Perceptions of
440 infidelity risk predict women's preferences for low male voice pitch in short-term over long-term
441 relationship contexts. *Personality and Individual Differences*, *56*, 73–77.
- 442 O'Connor, J. J., Fraccaro, P. J., Pisanski, K., Tigue, C. C., O'Donnell, T. J., & Feinberg, D. R.
443 (2014b). Social dialect and men's voice pitch influence women's mate preferences. *Evolution
444 and Human Behavior*, *35*(5), 368–375.
- 445 Pépiot, E. (2014). Male and female speech: A study of mean f0, f0 range, phonation type and speech
446 rate in Parisian French and American English speakers. In *Speech Prosody 7* (pp. 305–309).
- 447 Perusse, D. (1993). Cultural and reproductive success in industrial societies: Testing the relationship
448 at the proximate and ultimate levels. *Behavioral and Brain Sciences*, *16*(2), 267–283.
- 449 Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J., & Feinberg, D. R., (2014b). Return to
450 Oz: Voice pitch facilitates assessments of men's body size. *Journal of Experimental Psychology:
451 Human Perception and Performance*, *40*(4), 1316.
- 452 Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J., Röder, S., Andrews, P. W., et al. (2014a).
453 Vocal indicators of body size in men and women: A meta-analysis. *Animal Behaviour*, *95*, 89–99.
- 454 Pisanski, K., Cartei, V., McGettigan, C., Raine, J., & Reby, D. (2016). Voice modulation: A window
455 into the origins of human vocal control? *Trends in Cognitive Sciences*, *20*(4), 304–318.
- 456 Przybyla, B. D., Horii, Y., & Crawford, M. H. (1992). Vocal fundamental frequency in a twin
457 sample: Looking for a genetic effect. *Journal of Voice*, *6*(3), 261–266.
- 458 Puts, D. A. (2005). Mating context and menstrual phase affect women's preferences for male voice
459 pitch. *Evolution and Human Behavior*, *26*(5), 388–397.
- 460 Puts, D. A., Hill, A. K., Bailey, D. H., Walker, R. S., Rendall, D., Wheatley, J. R., et al. (2016). Sexual
461 selection on male vocal fundamental frequency in humans and other anthropoids. *Proceedings
462 of the Royal Society of London B: Biological Sciences*, *283*(1829), 20152830.
- 463 Puts, D. A. (2010). Beauty and the beast: Mechanisms of sexual selection in humans. *Evolution
464 and Human Behavior*, *31*(3), 157–175.
- 465 Puts, D. A., Gaulin, S. J., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism
466 in human voice pitch. *Evolution and Human Behavior*, *27*(4), 283–296.
- 467 Puts, D. A., Hodges, C. R., Cárdenas, R. A., & Gaulin, S. J. (2007). Men's voices as dominance
468 signals: Vocal fundamental and formant frequencies influence dominance attributions among
469 men. *Evolution and Human Behavior*, *28*(5), 340–344.
- 470 Puts, D. A., Barnrdt, J. L., Welling, L. L., Dawood, K., & Burriss, R. P. (2011). Intrasexual competi-
471 tion among women: Vocal femininity affects perceptions of attractiveness and flirtatiousness.
472 *Personality and Individual Differences*, *50*(1), 111–115.
- 473 Puts, D. A., Apicella, C. L., & Cárdenas, R. A. (2012). Masculine voices signal men's threat potential
474 in forager and industrial societies. *Proceedings of the Royal Society of London B: Biological
475 Sciences*, *279*(1728), 601–609.
- 476 Ralls, K., & Mesnick, S. (2009). Sexual dimorphism. In *Encyclopedia of marine mammals* (Second
477 Edition) (pp. 1005–1011).
- 478 Rendall, D., Vokey, J. R., & Nemeth, C. (2007). Lifting the curtain on the Wizard of Oz: Biased
479 voice-based impressions of speaker size. *Journal of Experimental Psychology: Human Perception
480 and Performance*, *33*(5), 1208.
- 481 Riding, D., Lonsdale, D., & Brown, B. (2006). The effects of average fundamental frequency and
482 variance of fundamental frequency on male vocal attractiveness to women. *Journal of Nonverbal
483 Behavior*, *30*(2), 55–61.
- 484 Roberts, M. L., Buchanan, K. L., & Evans, M. R. (2004). Testing the immunocompetence handicap
485 hypothesis: A review of the evidence. *Animal Behaviour*, *68*(2), 227–239.
- 486 Rose, P. (1991). How effective are long term mean and standard deviation as normalisation param-
487 eters for tonal fundamental frequency? *Speech Communication*, *10*(3), 229–247.
- 488 Šebesta, P., Kleisner, K., Tureček, P., Kočnar, T., Akoko, R. M., Třebický, V., et al. (2017). Voices of
489 Africa: Acoustic predictors of human male vocal attractiveness. *Animal Behaviour*, *127*, 205–211.

- 490 Shirazi, T. N., Puts, D. A., & Escasa-Dorne, M. J. (2018). Filipino women's preferences for male
491 voice pitch: Intra-individual, life history, and hormonal predictors. *Adaptive Human Behavior*
492 *and Physiology*, 4(2), 188–206.
- 493 Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and Lin-*
494 *guistics Compass*, 3(2), 621–640.
- 495 Smith, K. M., Olkhov, Y. M., Puts, D. A., & Apicella, C. L. (2017). Hadza men with lower voice pitch
496 have a better hunting reputation. *Evolutionary Psychology*, 15(4), 1474704917740466. Suire,
497 A., Raymond, M., Barkat-Defradas, M. (2018). *Human vocal behavior within competitive and*
498 *courtship contexts and its relation to mating success*. Accepted: *Evolution and Human Behavior*
499 (in press).
- 500 Smith, D. S., Jones, B. C., Feinberg, D. R., & Allan, K. (2012). A modulatory effect of male
501 voice pitch on long-term memory in women: evidence of adaptation for mate choice? *Memory*
502 *& Cognition*, 40(1), 135–144.
- 503 Suire, A., Raymond, M., & Barkat-Defradas, M. (2018). Human vocal behavior within competitive
504 and courtship contexts and its relation to mating success. *Evolution and Human Behavior*, 39(6),
505 684–691.
- 506 Tigue, C. C., Borak, D. J., O'Connor, J. J., Schandl, C., & Feinberg, D. R. (2012). Voice pitch
507 influences voting behavior. *Evolution and Human Behavior*, 33(3), 210–216.
- 508 Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *The*
509 *Journal of the Acoustical Society of America*, 85(4), 1699–1707.
- 510 Traunmüller, H., & Eriksson, A. (1995). The frequency range of the voice fundamental in the speech
511 of male and female adults. Unpublished manuscript.
- 512 Tuomi, S. K., & Fisher, J. E. (1979). Characteristics of simulated sexy voice. *Folia Phoniatrica et*
513 *Logopaedica*, 31(4), 242–249.
- 514 Van Bezooijen, R. (1995). Sociocultural aspects of pitch differences between Japanese and Dutch
515 women. *Language and Speech*, 38(3), 253–265.
- 516 Vukovic, J., Jones, B. C., Feinberg, D. R., DeBruine, L. M., Smith, F. G., Welling, L. L., et al.
517 (2011). Variation in perceptions of physical dominance and trustworthiness predicts individual
518 differences in the effect of relationship context on women's preferences for masculine pitch in
519 men's voices. *British Journal of Psychology*, 102(1), 37–48.
- 520 Watkins, C. D., Fraccaro, P. J., Smith, F. G., Vukovic, J., Feinberg, D. R., DeBruine, L. M., et al.
521 (2010). Taller men are less sensitive to cues of dominance in other men. *Behavioral Ecology*,
522 21(5), 943–947.
- 523 Wolff, S. E., & Puts, D. A. (2010). Vocal masculinity is a robust dominance signal in men. *Behavioral*
524 *Ecology and Sociobiology*, 64(10), 1673–1683.
- 525 Yamazawa, H., & Hollien, H. (1992). Speaking fundamental frequency patterns of Japanese women.
526 *Phonetica*, 49(2), 128–140.
- 527 Zahavi, A. (1975). Mate selection—a selection for a handicap. *Journal of theoretical Biology*, 53(1),
528 205–214.

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Voice Attractiveness
Series Title	
Chapter Title	On Voice Averaging and Attractiveness
Copyright Year	2020
Copyright HolderName	Springer Nature Singapore Pte Ltd.
Corresponding Author	Family Name Belin Particle Given Name Pascal Prefix Suffix Role Division Institut de Neurosciences de La Timone Organization CNRS et Aix-Marseille Université Département de Psychologie, Université de Montréal Address Montreal, Canada Email pascal.belin@univ-amu.fr
Abstract	Several experiments investigating the perceptual, acoustical and neural bases of the ‘voice attractiveness averaging phenomenon’ are briefly summarized. We show that synthetic voice composites generated by averaging multiple (same gender) individual voices (short syllables) are perceived as increasingly attractive with the number of voices averaged. This phenomenon, independent of listener or speaker gender and analogous to a similar effect in the visual domain for face attractiveness, is explained in part by two acoustical correlates of averaging: reduced ‘Distance-to-Mean’, as indexed by the Euclidean distance between a voice and its same-gender population average in f0-F1 space and increased voice ‘texture smoothness’ as indexed by increased harmonics-to-noise ratio (HNR). These two acoustical parameters co-vary with perceived attractiveness and manipulating them independently of one another also affects attractiveness ratings. The neural correlates of implicitly perceived attractiveness consist in a highly significant negative correlation between attractiveness and fMRI signal in large areas of bilateral auditory cortex, largely overlapping with the Temporal Voice Areas, as well as inferior prefrontal cortex: more attractive voices elicit less activity in these regions. While the correlations in auditory areas were largely explained by distance-to-mean and HNR, inferior prefrontal areas bilaterally were observed even after co-varying out variance explained by these acoustical parameters, suggesting a role as abstract voice attractiveness evaluators.
Keywords	Averageness - Aperiodicity - Distance-to-mean - Distinctiveness - Pitch - Formant dispersion

Chapter 8

On Voice Averaging and Attractiveness



Pascal Belin

Abstract Several experiments investigating the perceptual, acoustical and neural bases of the ‘voice attractiveness averaging phenomenon’ are briefly summarized. We show that synthetic voice composites generated by averaging multiple (same gender) individual voices (short syllables) are perceived as increasingly attractive with the number of voices averaged. This phenomenon, independent of listener or speaker gender and analogous to a similar effect in the visual domain for face attractiveness, is explained in part by two acoustical correlates of averaging: reduced ‘Distance-to-Mean’, as indexed by the Euclidean distance between a voice and its same-gender population average in f0-F1 space and increased voice ‘texture smoothness’ as indexed by increased harmonics-to-noise ratio (HNR). These two acoustical parameters co-vary with perceived attractiveness and manipulating them independently of one another also affects attractiveness ratings. The neural correlates of implicitly perceived attractiveness consist in a highly significant negative correlation between attractiveness and fMRI signal in large areas of bilateral auditory cortex, largely overlapping with the Temporal Voice Areas, as well as inferior prefrontal cortex: more attractive voices elicit less activity in these regions. While the correlations in auditory areas were largely explained by distance-to-mean and HNR, inferior prefrontal areas bilaterally were observed even after co-varying out variance explained by these acoustical parameters, suggesting a role as abstract voice attractiveness evaluators.

Keywords Averageness · Aperiodicity · Distance-to-mean · Distinctiveness · Pitch · Formant dispersion

P. Belin (✉)

Institut de Neurosciences de La Timone, CNRS et Aix-Marseille
Université Département de Psychologie, Université de Montréal, Montreal, Canada
e-mail: pascal.belin@univ-amu.fr

© Springer Nature Singapore Pte Ltd. 2020

B. Weiss et al. (eds.), *Voice Attractiveness*, Prosody, Phonology and Phonetics,
https://doi.org/10.1007/978-981-15-6627-1_8

145

8.1 Introduction

The faces shown in Fig. 8.1a are computer generated: they are the pixel-wise average of a large number of pictures of different faces after conformation to a same configuration (eyes and mouth in the same position). Observers typically find these faces more attractive than most of the individual constituting faces. This so-called ‘averaging attractiveness phenomenon’ has been observed since the nineteenth century and the beginnings of photography when experimenters such as Sir Francis Galton noticed that by superimposing portraits of different individuals on a same photographic plate one obtained a quite attractive picture (Galton, 1878; Jastrow, 1885). Since those pioneering times the averaging attractiveness phenomenon has been replicated many times with more sophisticated computer graphics techniques such as in Fig. 8.1a (Langlois et al., 2000; Langlois & Roggman, 1990; Perrett, May, & Yoshikawa, 1994; Thornhill & Gangestad, 1999).

There are two main accounts for the face averaging attractiveness phenomenon. One account from evolutionary psychology—the ‘good genes’ explanation—proposes that we tend to prefer averaged faces because if they were real faces they would signal a potential mate with particularly high fitness. Indeed, facial features such as proximity to the population average, facial symmetry, or face texture smoothness appear to signal high fitness in real faces (Grammer, Fink, Moller, & Thornhill, 2003; Langlois & Roggman, 1990; Thornhill & Gangestad, 1999). The averaging procedure enhances all three of these features, artificially signalling high fitness in a synthetic face, and hence their attractiveness. Another account from cognitive psychology—the ‘perceptual fluency’ account—proposes that observers prefer averaged faces because they are closer in face space, i.e. more similar to a central face prototype based on which all face identities are coded, and so they are easier to process, and hence more attractive (Winkielman, Halberstadt, Fazendeiro, & Catty, 2006). These two accounts are not mutually exclusive: the ‘perceptual fluency’ account can be viewed as an explanation at the proximate level, in terms of cognitive mechanisms implementing the effect, while the ‘good genes’ account is an explanation at a more ultimate level, in terms of the selective evolutionary pressures that gave rise to such a phenomenon in our ancestors.

Crucially, both the cognitive and evolutionary accounts suggest that a similar phenomenon could exist for voices. Thanks to the development of voice morphing technology, and the excellent and generous contribution of Professor Hideki Kawahara at Wakayama University, we were able to test that hypothesis for the first time in Bruckert et al. (2010).

8.2 Voice Attractiveness Increases with Averaging

To start addressing the complex problem of voice averaging, we decided to focus on the simpler problem, more manageable in an experimental setting, of averaging of brief, quasi-stationary vocalizations, and opted to use short syllables as stimuli.

63 We reasoned that such stimuli, for which time plays minimal role, would be easier
 64 to process through averaging than longer, more complex and variable utterances.
 65 Quasi-static syllables are also analogous to the static photographs with which most
 66 face attractiveness research has been performed so far.

67 We selected from a database of high-quality recordings of English syllables
 68 (Hillenbrand, Getty, Clark, & Wheeler, 1995) as set of recordings of the syllable
 69 /had/ spoken in isolation by 32 different male and 32 female American speakers
 70 (duration: mean \pm s.d.: female voices: 320 ± 51 ms; male voices: 267 ± 42 ms). We
 71 then identified in each stimulus a set of spectro-temporal landmarks to be put in
 72 correspondence across speakers during averaging. As shown by the black dots in
 73 Fig. 8.1b, these landmarks consisted of first three formant frequencies at onset and
 74 offset of phonation, and at the beginning of the formant transition of the final /d/. We
 75 then used the Straight software (Kawahara & Matsui, 2003) to generate voice com-
 76 posites consisting of an interpolation of the aperiodic and spectral temporal density
 77 components of varying numbers of individual voices of the same gender (arbitrarily
 78 chosen, such that not all possible composites have been generated). For each speaker

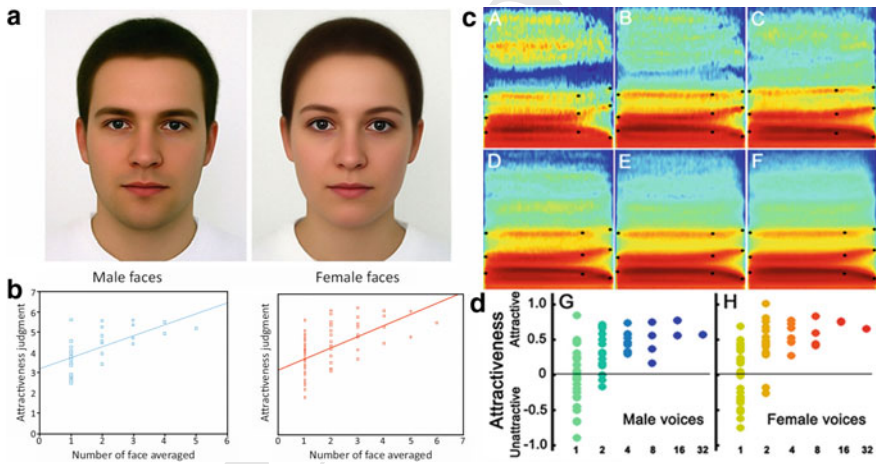


Fig. 8.1 Face and voice attractiveness judgments as a function of averaging. **a** Face composites generated by averaging 32 male faces (left) and 64 female faces (right). **b** Attractiveness ratings as a function of number of faces averaged. Note the steady increase in attractiveness ratings with increasing number of averaged faces, for both male (left) and female (right) faces. Reproduced with permission from Braun et al. (2001). **c** Spectrograms of voice composites generated by averaging an increasing number of voices of the same gender (different speakers uttering the syllable ‘had’). Top left panel: 1-voice composite; middle top panel: 2-voice composite; right top panel: 4-voice composite; bottom left panel: 8-voice composite; bottom middle panel: 16-voice composite; bottom right panel: 32-voice composite. **d** Attractiveness ratings as a function of number of voices averaged in the composites (individual points). Note the steady increase in attractiveness ratings with increasing number of averaged voices, for both male (left) and female (right) voices. Reproduced with permission from Bruckert et al. (2010)

79 gender, this procedure resulted in thirty two 1-voice composites (the individual voices
80 resynthesized), sixteen 2-voice composites, eight 4-voice composites, four 8-voice
81 composites, two 16-voice composites, and a single average of all voices of the same
82 gender, the 32-voice composite. Example composite stimuli are shown in Fig. 8.1c.

83 We then played these stimuli in a pseudorandom order to 25 listeners (13 females)
84 who were asked to rate the perceived attractiveness of each stimulus using a visual
85 analogue scale ranging from ‘not at all’ to ‘extremely’ attractive. Analysis of the
86 data provided striking results (Bruckert et al., 2010). As shown in Fig. 8.1d, for the
87 1-voice composites (the resynthesized original voices) we found as expected a normal
88 distribution of attractiveness ratings around the average: for both male and female
89 speakers, most of the voices were rated with average attractiveness while a few voices
90 were perceived as more attractive than average and others as less. But as soon as two or
91 more voices were averaged together, we witnessed a marked progressive increase
92 of average attractiveness ratings, similar for the male and female voices. 4-voice
93 composites were already perceived as markedly above average and 16- and 32-voice
94 composites all resulted in very high ratings. The correlation between attractiveness
95 z-scores and number of voices in the composite was highly significant ($p < 0.001$)
96 for both male and female voices (Bruckert et al., 2010).

97 Thus, we could observe for the first time a ‘voice averaging attractiveness phe-
98 nomenon’ that was directly predicted by analogous studies in face perception: the
99 more speakers are included by averaging in a synthetic voice, the more attractive it
100 is perceived. Two implications of these results are worth discussing.

101 First, there is a highly striking similarity between the attractiveness ratings
102 obtained in face and voice averaging experiments. Despite the very different nature
103 of the sensory input (vibrations of the tympanic membrane versus light on the retina)
104 the effects of averaging gave rise in the two sensory modalities to highly similar and
105 gender-independent averaging-induced attractiveness increases (compare Fig. 8.1b
106 and d). This beautifully illustrates the notion of similar functional architectures for
107 face and voice processing in the human brain. Indeed, many sources of evidence from
108 patient observation to neuroimaging studies converge to the notion that the compu-
109 tational problems posed by face and voice processing, being of very similar nature,
110 and subjected in our ancestors to comparable evolutionary pressures, are addressed
111 by the brain using similar neurophysiological solutions (Yovel & Belin, 2013).

112 Second, and more relevant to voice attractiveness, the voice averaging attractive-
113 ness phenomenon opens an exciting window onto the acoustical underpinnings of
114 this complex percept. Indeed, the averaging procedure had at least two independ-
115 ent acoustic effects on the synthesized composites. Including an increasing number
116 of different voices in the composite’s resulted in: (i) a progressive decrease in the
117 distance-to-mean (increased similarity to the average) and (ii) a progressive decrease
118 in the amount of aperiodicity (increased harmonicity or voice texture smoothness).

8.3 Effects of Distance-to-Mean

Because of the linear combination of individual spectral temporal landmarks in the composites during averaging, at each successive averaging step the resulting composites mathematically became closer in acoustical space to the average: their fundamental frequency and formant frequency values became increasingly similar to those of the 32-voice average, resulting at each step in decreasing average ‘distance-to-mean’, as defined by the Euclidean distance between a voice and the same-gender average in f_0 - F_1 (first formant frequency) space. In other words, the more voices are averaged together, the more the resulting composite sounds like the population average. This suggests that distance-to-mean could potentially provide an acoustical parameter relevant for voice attractiveness. We tested that hypothesis in two different ways: (i) by examining the relationship between distance-to-mean and perceived attractiveness in our set of natural and synthetic voices and (ii) by explicitly manipulating distance-to-mean (but not other parameters such as aperiodicity) in synthetic voices.

We first tested, independently for male and female voices, whether distance-to-mean in our set of voice composites would correlate with their average perceived attractiveness. For both voice genders, we found highly significant negative correlations between distance-to-mean and attractiveness: the higher the distance-to-mean, the lower the perceived attractiveness. As including composites of all levels in this analysis, known to be both closer to the average and more attractive involves some level of circularity, we repeated the analysis by only including the 1-voice composites, resynthesized versions of the original recordings (indistinguishable by ear): the results remained strongly significant, for both male and female voices. Thus, in our set of 32 male and 32 female voices, those that were naturally closer to the same-gender average were also perceived as more attractive (Bruckert et al., 2010)—a result that should be tested on larger samples (Fig. 8.2).

Does modifying distance-to-mean also modify perceived attractiveness? We tested the hypothesis by using morphing to generate, for each of the 32 individual voices of each gender, a pair of synthetic voices that differed from the original by having been moved either towards the average or away from the average by the exact similar amount of acoustical change (50% of the natural distance-to-mean). We predicted that although the new synthetic voices were acoustically equally dissimilar to the original, the one closer to the average would be perceived as more attractive. Results confirmed that prediction for both voice genders (Bruckert et al., 2010).

Thus, not only are voices naturally closer to the same-gender average perceived as more attractive, but acoustically modifying voices to move them closer to the average also makes them more attractive than moving them away. Distance-to-mean thus appears as one important acoustical correlate of voice attractiveness. Interestingly, distance-to-mean can be consciously modified, if not by altering formant frequencies (largely dependent on vocal tract size) but by consciously modifying one’s pitch of voice so that our average fundamental frequency is closer to the gender-typical value

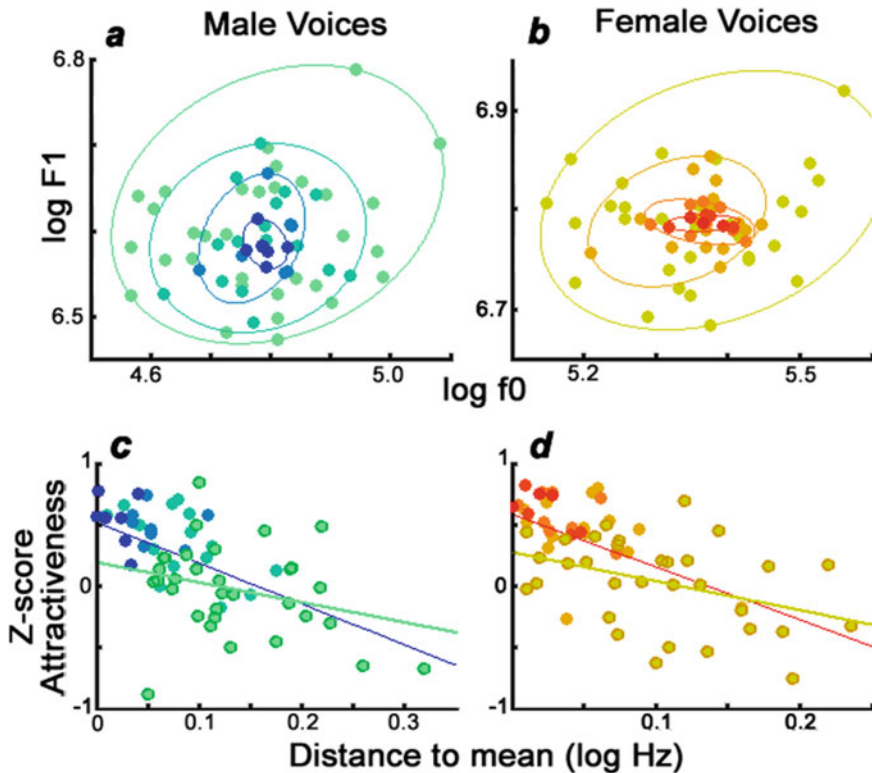


Fig. 8.2 Effects of distance-to-mean. **a.** Male voice composites are represented as coloured dots in $\log f_0$ - $\log F_1$ space. Colour indicates degree of averaging with darker colours indicating more voices in the composite. Lines indicate the smallest elliptic contours containing all composites of a same degree of averaging. Note how composites progressively become closer to the average with increasing number of constituting voices. **b.** Female voice composites, legend as in **(a)**. **c.** Relation between distance-to-mean and attractiveness ratings for each male voice composite (coloured dots). Lines indicate the regression line when all composites are considered (blue line) or when only the 1-voice composites are considered (green line). **d.** Relation between distance-to-mean and attractiveness for female voice composites. Legend as in **(c)**

161 (about 125 Hz for men and 215 Hz for women Hillenbrand et al., 1995), not too
 162 low and not too high, as a means of ‘vocal make-up’ to enhance one’s perceived
 163 attractiveness.

164 8.4 Effects of Voice Texture Smoothness

165 Another important effect of averaging on the acoustical structure of voices, largely
 166 independent from the effect of distance-to-mean in f_0 -formant frequency space, is a

167 progressive decrease in the amount of aperiodicity with the number of voices aver-
 168 aged, as the morphing procedure averages out aperiodic noise in the signal. This effect
 169 can be plainly seen in Fig. 8.1c as the spectrograms become progressively smoother
 170 with the increasing number of voices in the composite from the top left panel (1-voice
 171 composite, showing much spectro-temporal irregularities) to the bottom right panel
 172 (32-voice composite) with a very smooth structure. This effect is analogous to the
 173 increase in face texture smoothness (see Fig. 8.1a) caused by averaging as individual
 174 local variations in luminance and reflection (the ‘villainous irregularities’ of Galton,
 175 1878) are averaged out across individual faces. This effect of smoothing of the ‘voice
 176 texture’ can be quantified using measures such as the harmonics-to-noise ratio (HNR)
 177 that captures the amount of regularity in the sound. When the harmonic-to-noise ratio
 178 of each composite is plotted as a function of its number of constituent voices (Bruck-
 179 ert et al., 2010), there is a clear and highly significant progressive increase in HNR
 180 along with number of voices in the composite that nearly mirrors the increase in
 181 attractiveness ratings. Thus, the amount of energy in the aperiodic component of
 182 voice could constitute another acoustical correlate of voice attractiveness.

183 We tested this hypothesis by generating for each of the 32 male and 32 female
 184 voices of our sample, a ‘smoother’ and ‘rougher’ version of each voices. Those were
 185 generated by moving stimuli away or closer to the average by equal amounts of
 186 acoustic change, as for the manipulation of distance-to-mean above, but this time
 187 modifying only the aperiodic component of voice. We verified that the ‘smoother’
 188 synthetic voices had greater harmonics-to-noise ratio than the ‘rougher’ for both
 189 voice genders. We then presented listeners with voice pairs made of the smoother
 190 and rougher version of a same original voice and asked them to decide the one
 191 they found the more attractive. Subjects overwhelmingly preferred the smoother
 192 version with reduced periodicity and increased HNR to the rougher version (Bruckert
 193 et al., 2010).

194 Overall, the increase in voice attractiveness induced by averaging highlights
 195 distance-to-mean and voice textures smoothness as two largely independent and
 196 important acoustical correlates of voice attractiveness. They can potentially be used
 197 to predict listeners’ ratings and can be manipulated in synthetic, but also in natural
 198 voices, to artificially increase perceived attractiveness. Note, however, that while
 199 distance-to-mean already correlated with attractiveness ratings in natural, unaver-
 200 aged voices, this was not the case for HNR that showed essentially no relation with
 201 attractiveness ratings for the natural voices. This suggest that, while both parameters
 202 contribute to the attractiveness averaging effect, distance-to-mean is more important
 203 than HNR in determining the attractiveness of natural voices.

204 8.5 Neural Correlates of Perceived Voice Attractiveness

205 We then turned to the question of the neural correlates of voice attractiveness. Indeed
 206 neuroimaging studies have shown linear or quadratic relations between perceived
 207 facial attractiveness and neural activity in orbitofrontal cortex as well as in amygdala

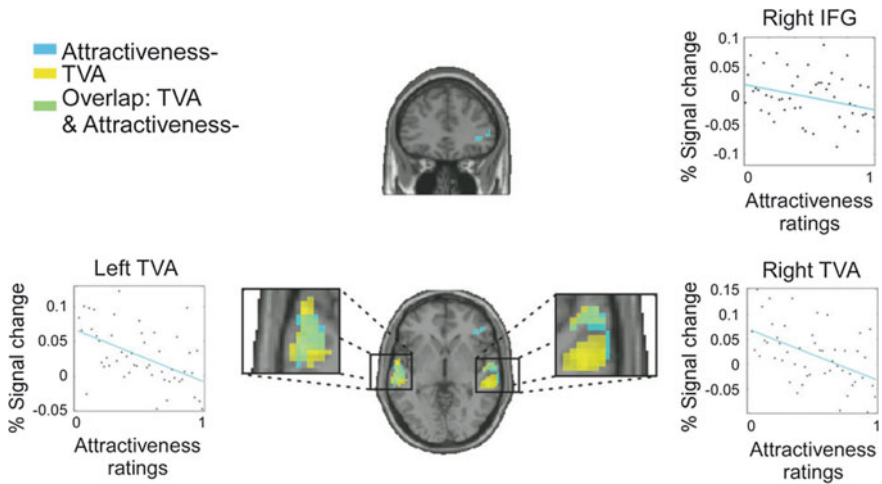


Fig. 8.3 Neural correlates of perceived attractiveness. Cerebral regions modulated by implicitly perceived attractiveness during passive listening to voices. Cortical areas in blue showed significant negative correlation between BOLD signal and attractiveness (graphs in insets showing regression lines for three regions of interest): more attractive voices elicited less neural activity in those regions. They largely overlap with the voice-sensitive temporal voice areas (in yellow) but also involve right inferior prefrontal cortex (top central panel)

(Winston, O’Doherty, Kilner, Perrett, & Dolan, 2007). To address this question in the domain of voice perception, we performed a functional magnetic resonance imaging (fMRI) study in normal participants (Bestelmeyer et al., 2012). They were scanned while passively listening to our set of voice composites presented in a pseudorandom order. We used a so-called ‘cluster volume acquisition’ fMRI protocol with brief silent intervals during fMRI volume acquisitions allowing the presentation of voice stimuli during silent periods for optimal stimulation. Subjects were not informed of our focus on voice attractiveness and were simply instructed to listen to the voices and press the button when they would hear an infrequent pure-tone stimulus.

In the fMRI analyses, we first asked whether there would be regions of the brain in which stimulus-induced activity would co-vary with the average attractiveness rating obtained offline for each voice. Indeed a well-defined network of cortical region showed significant correlations between fMRI signal and attractiveness ratings (Fig. 8.3). Most prominently, negative correlations were observed in large areas of bilateral superior temporal gyrus and sulci, overlapping with the voice-selective temporal voice areas (TVA) of auditory cortex (Belin, Zatorre, Lafaille, Ahad, & Pike, 2000; Pernet, Charest, Belizaire, Zatorre, & Belin, 2007) of secondary auditory cortex. But such negative correlation was also observed in the inferior frontal gyrus (IFG) of the right hemisphere, outside of voice-sensitive regions (Bestelmeyer et al., 2012).

We asked whether part of these strong negative correlations could be partly explained by one or the other acoustic parameters highlighted above—distance-

230 to-mean or texture smoothness (as measured by HNR). We ran new analyses in
 231 which we searched for correlations between the brain activity elicited by each voice
 232 and their distance-to-mean on one hand and their HNR on the other. Both analyses
 233 revealed auditory cortical regions in which voice-elicited fMRI signal significantly
 234 correlated with these measures, although in different locations. Areas of secondary
 235 auditory cortex overlapping with the TVAs bilaterally showed a positive correlation
 236 with distance-to-mean (voices farther away from the mean—also less attractive on
 237 average—eliciting greater signal). This phenomenon has since been replicated and
 238 extended in subsequent work (Latinus & Belin, 2011; Latinus, McAleer, Bestelmeyer,
 239 & Belin, 2013). The positive correlation between distance-to-mean and neural activ-
 240 ity constitutes a hallmark of ‘norm-based coding’ as evidenced in visual cortex for
 241 face identity processing (Leopold, Bondar, & Giese, 2006): individual voices appear
 242 to be coded in the TVAs as a function of their difference with the average voice:
 243 whether the negative correlation with attractiveness in those areas is a consequence
 244 of, or drives, the positive correlation with distance-to-mean remains to be established.
 245 Other, more anterior parts of the TVAs instead showed a negative correlation with
 246 HNR, with more aperiodic voices eliciting higher activity. Thus, the large negative
 247 correlation between attractiveness and fMRI signal is in part explained by a sensitiv-
 248 ity of auditory cortex to the two underlying acoustical features shown as determinant
 249 for perceived attractiveness.

250 But could we detect attractiveness-related changes that would be independent of
 251 the underlying acoustics? We addressed that question by performing another anal-
 252 ysis in which measures of distance-to-mean HNR were included in the model and
 253 regressed out to examine variance not accounted for by these parameters. Results
 254 showed that the large negative correlation in the auditory cortex had disappeared,
 255 confirming that it was largely explained by the HNR and distance-to-mean of the
 256 voices. However, two bilateral regions of inferior prefrontal cortex, pars triang-
 257 ularis, survived after removing variance accounted for by acoustics: these regions still
 258 showed the negative relation with attractiveness. This region is part of Broca’s area
 259 (Anwander, Tittgemeyer, von Cramon, Friederici, & Knosche, 2007) and is strongly
 260 connected to sensory cortex (Petrides & Pandya, 2009). In addition to its involve-
 261 ment in language perception, bilateral activity in Broca’s area has been linked to
 262 auditory working memory in which increased task demands correlate with increased
 263 activity (Martinkauppi, Rama, Aronen, Korvenoja, & Carlson, 2000; Arnott, Grady,
 264 Hevenor, Graham, & Alain, 2005). Our results thus may suggest that increasingly
 265 unattractive voices demand larger processing resources and may point towards the
 266 role of the IFG pars triangularis as being involved not only in the processing of
 267 language and affective prosody but also in integrating acoustic information received
 268 from bilateral TVA into a unified percept of attractiveness.

269 A clear limitation of the above findings is that they were obtained with the use of
 270 brief vowels and hence cannot be easily generalized to realistic speaking situations
 271 in which a number of additional cues are present, including intonation, speaking
 272 rate, etc. Therefore, our results concern only one component that contributes to
 273 perceived voice attractiveness in realistic settings. Nonetheless, these findings have
 274 important potential implications for voice-based technology, suggesting simple ways

of enhancing the attractiveness of synthetic voices at a time when automated voice production systems become ubiquitous.

Acknowledgments I gratefully acknowledge my co-authors on publications discussed above: Laetitia Bruckert, Patricia E.G. Bestelmeyer, Marianne Latinus, Julien Rouger, Ian Charest, Guillaume A. Rousselet, Hideki Kawahara and Frances Crabbe. P.B. was supported by the French Fondation pour la Recherche Médicale (AJE201214) and Agence Nationale de la Recherche (PRIMA VOICE), and by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI).

References

- Anwander, A., Tittgemeyer, M., von Cramon, D. Y., Friederici, A. D., & Knosche, T. R. (2007). Connectivity-based parcellation of Broca's area. *Cerebral Cortex*, *17*(4), 816–825.
- Arnott, S. R., Grady, C. L., Hevenor, S. J., Graham, S., & Alain, C. (2005). The functional organization of auditory working memory as revealed by fMRI. *Journal of Cognitive Neuroscience*, *17*, 819–831.
- Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*, 309–312.
- Bestelmeyer, P. E., Latinus, M., Bruckert, L., Rouger, J., Crabbe, F. & Belin, P. (2012). Implicitly perceived vocal attractiveness modulates prefrontal cortex activity. *Cereb Cortex* **22**, 1263–1270, <https://doi.org/10.1093/cercor/bhr204>
- Braun, C., Gruendl, M., Marberger, C., & Scherber, C. (2001). Beautycheck—Ursachen und Folgen von Attraktivitaet.
- Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., et al. (2010). Vocal attractiveness increases by averaging. *Current Biology*, *20*(2), 116–120.
- Galton, F. (1878). Composite portraits. *Journal of the Anthropological Institute*, *8*, 132–144.
- Grammer, K., Fink, B., Moller, A. P., & Thornhill, R. (2003). Darwinian aesthetics: Sexual selection and the biology of beauty. *Biological Reviews*, *78*(3), 385–407.
- Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*, 3099–3111.
- Jastrow, J. (1885). Composite portraiture. *Science*, *6*, 165.
- Kawahara, H., & Matsui, H. (2003). Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (pp. 256–259).
- Langlois, J. H., & Roggman, L. A. (1990). Attractive faces are only average. *Psychological Science*, *1*(2), 115–121.
- Langlois, J. H., Kalakanis, L., Rubenstein, A. J., Larson, A., Hallam, M., & Smoot, M. (2000). Maxims or myths of beauty? A meta-analytic and theoretical review. *Psychological Bulletin*, *126*(3), 390–423.
- Latinus, M., & Belin, P. (2011). Anti-voice adaptation suggests prototype-based coding of voice identity. *Frontiers in Psychology*, *2*, 175. <https://doi.org/10.3389/fpsyg.2011.00175>.
- Latinus, M., McAleer, P., Bestelmeyer, P. E., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, *23*(12), 1075–1080.
- Leopold, D. A., Bondar, I. V., & Giese, M. A. (2006). Norm-based face encoding by single neurons in the monkey inferotemporal cortex. *Nature*, *442*(7102), 572–575.
- Martinkauppi, S., Rama, P., Aronen, H. J., Korvenoja, A., & Carlson, S. (2000). Working memory of auditory localization. *Cereb Cortex*, *10*, 889–898.
- Pernet, C., Charest, I., Belizaire, G., Zatorre, R. J., & Belin, P. (2007). The temporal voice area (TVA): Spatial characterization and variability. *Neuroimage*, *36*, S1–S168.

- 321 Perrett, D. I., May, K. A., & Yoshikawa, S. (1994). Facial shape and judgements of female attrac-
322 tiveness. *Nature*, 368, 239–242.
- 323 Petrides, M., & Pandya, D. N. (2009). Distinct parietal and temporal pathways to the homologues
324 of Broca’s area in the monkey. *PLOS Biology*, 7, e1000170.
- 325 Thornhill, R., & Gangestad, S. W. (1999). Facial attractiveness. *Trends in Cognitive Sciences*, 3(12),
326 452–460.
- 327 Winkielman, P., Halberstadt, J., Fazendeiro, T., & Catty, S. (2006). Prototypes are attractive because
328 they are easy on the mind. *Psychological Science*, 17(9), 799–806.
- 329 Winston, J. S., O’Doherty, J., Kilner, J. M., Perrett, D. I., & Dolan, R. J. (2007). Brain systems
330 for assessing facial attractiveness. *Neuropsychologia*, 45(1), 195–206. [https://doi.org/10.1016/j.
331 neuropsychologia.2006.05.009](https://doi.org/10.1016/j.neuropsychologia.2006.05.009).
- 332 Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in
333 Cognitive Sciences*, 17(6), 263–271.

Part III
Prosody

1
2

UNCORRECTED PROOF

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Voice Attractiveness	
Series Title		
Chapter Title	Attractiveness of Male Speakers: Effects of Pitch and Tempo	
Copyright Year	2020	
Copyright HolderName	Springer Nature Singapore Pte Ltd.	
Corresponding Author	Family Name	Quené
	Particle	
	Given Name	Hugo
	Prefix	
	Suffix	
	Role	
	Division	Utrecht institute of Linguistics
	Organization	Utrecht University
	Address	Trans 10, 3512 JK, Utrecht, The Netherlands
	Email	h.quene@uu.nl
Author	Family Name	Boomsma
	Particle	
	Given Name	Geke
	Prefix	
	Suffix	
	Role	
	Division	Utrecht institute of Linguistics
	Organization	Utrecht University
	Address	Trans 10, 3512 JK, Utrecht, The Netherlands
	Email	gekeboomsma@hotmail.com
Author	Family Name	van Erning
	Particle	
	Given Name	Roméé
	Prefix	
	Suffix	
	Role	
	Division	Utrecht institute of Linguistics
	Organization	Utrecht University
	Address	Trans 10, 3512 JK, Utrecht, The Netherlands
	Email	romeevanerning@gmail.com

Abstract Men with lower pitched voices tend to be rated as more attractive by female listeners; this tendency has been attributed to female sexual selection. Males not only speak with a lower pitch than females, however, but they also tend to speak at a faster tempo. Therefore, this study investigates whether speech tempo also affects the subjective attractiveness of male speakers for female listeners. To this end, sentences read by 24 male speakers were changed in relative tempo (factors 0.85, 1.00, and 1.15) and in overall pitch (−1.5, 0, +1.5 semitone), and were presented with and without fictitious portraits of the speakers. Ratings of

speakers' attractiveness by female heterosexual listeners show significant effects of both tempo and pitch, in that voices with increased pitch and with decreased tempo are rated as significantly less attractive. In conclusion, female listeners rate a male speaker as less attractive if his voice pitch is increased (higher) and if his speech tempo is decreased (slower). Therefore, both tempo and pitch may be relevant for speech-based sexual selection of males by females.

Keywords

Sexual selection - Voice pitch - Speech tempo - Speaking rate - Attractiveness - Experiment - Proportional odds model

Chapter 9

Attractiveness of Male Speakers: Effects of Pitch and Tempo



Hugo Quené, Geke Boomsma, and Romée van Erning

Abstract Men with lower pitched voices tend to be rated as more attractive by female listeners; this tendency has been attributed to female sexual selection. Males not only speak with a lower pitch than females, however, but they also tend to speak at a faster tempo. Therefore, this study investigates whether speech tempo also affects the subjective attractiveness of male speakers for female listeners. To this end, sentences read by 24 male speakers were changed in relative tempo (factors 0.85, 1.00, and 1.15) and in overall pitch (−1.5, 0, +1.5 semitone), and were presented with and without fictitious portraits of the speakers. Ratings of speakers' attractiveness by female heterosexual listeners show significant effects of both tempo and pitch, in that voices with increased pitch and with decreased tempo are rated as significantly less attractive. In conclusion, female listeners rate a male speaker as less attractive if his voice pitch is increased (higher) and if his speech tempo is decreased (slower). Therefore, both tempo and pitch may be relevant for speech-based sexual selection of males by females.

Keywords Sexual selection · Voice pitch · Speech tempo · Speaking rate · Attractiveness · Experiment · Proportional odds model

9.1 Introduction

Male and female speakers differ in their average fundamental frequency (F₀, perceived as pitch), viz., typically about 110 Hz for males and 205 Hz for females (Holmberg, Hillman, & Perkell, 1988; Simpson, 2009, Puts, Apicella, & Cárdenas, 2012).

H. Quené (✉) · G. Boomsma · R. van Erning
Utrecht institute of Linguistics, Utrecht University, Trans 10, 3512 JK
Utrecht, The Netherlands
e-mail: h.quene@uu.nl

G. Boomsma
e-mail: gekeboomsma@hotmail.com

R. van Erning
e-mail: romeevanering@gmail.com

© Springer Nature Singapore Pte Ltd. 2020
B. Weiss et al. (eds.), *Voice Attractiveness*, Prosody, Phonology and Phonetics,
https://doi.org/10.1007/978-981-15-6627-1_9

159

21 This large and significant difference in F0 develops in conjunction with primary and
22 secondary sexual characteristics, during puberty. This suggests that the pitch differ-
23 ence may be related to some sexual function. The voice pitch of an adult male
24 speaker is indeed reportedly related to the speaker's level of testosterone (Dabbs
25 & Mallinger, 1999; Puts et al., 2012) and to the speaker's self-reported number of
26 children (Apicella, Feinberg, & Marlowe, 2007) (but see Smith, Olkhov, Puts, & Api-
27 cella, 2017 for the mediating effect of hunting reputation). Thus, a male speaker's
28 pitch may indicate his health and physical dominance, by virtue of the intercorrela-
29 tions between male speakers' pitch and testosterone level (Dabbs & Mallinger, 1999;
30 Puts et al., 2012), body height (Pisanski et al., 2014), physical strength (Puts et al.,
31 2012), and masculinity (Clark & Henderson, 2003; Archer, 2006). Female listen-
32 ers may, therefore, use voice pitch to assess the male speaker's physical suitability
33 for producing and protecting offspring, i.e., in sexual selection via female choice
34 of mate (Andersson, 1994). Indeed, ratings of attractiveness by female listeners are
35 (negatively) correlated with the male speaker's F0 (Collins, 2000; Bruckert, Liénard,
36 Lacroix, Kreutzer, & Leboucher, 2006), and experiments have confirmed that manip-
37 ulations of F0 influence these attractiveness ratings (Feinberg, Jones, Little, Burt, &
38 Perrett, 2005). In addition, voice pitch may be used to indicate health and dominance
39 among male competitors, i.e., in sexual selection via male–male competition (Puts,
40 Gaulin, & Verdolini, 2006), a mechanism which may be more important than female
41 choice (Hill et al., 2013; Kordsmeyer, Hunt, Puts, Ostner, & Penke, 2018).

42 Males not only speak with a lower F0 than females, however, but they also tend to
43 speak at a faster speech rate or tempo than females (about 5% faster) (Quené, 2008;
44 Jacewicz, Fox, & Wei, 2010). This difference too may be related to male dominance,
45 as the faster tempo presumably indicates the speaker's cognitive abilities and motor
46 skills through his speaking. The faster tempo requires more physical energy (Moon &
47 Lindblom, 2003), even more so because the male speech organs have somewhat more
48 mass than the females', and it also requires more cognitive effort in linguistic planning
49 and motor control. Indeed, faster speakers tend to be rated as more convincing,
50 reliable, empathic, serious, active, and competent (Apple, Streeter, & Krauss, 1979;
51 Smith, Brown, Strong, & Rencher, 1975). Presumably, then, female listeners also
52 use a male speaker's tempo, to assess his motor skills and cognitive suitability as a
53 potential mate.

54 This study aims primarily to replicate previous findings on female preference for
55 male voices with lower *pitch*, and secondly to extend that work by investigating the
56 presumed female preference for male speakers speaking at a faster *tempo*. Thirdly,
57 we are interested in the interaction between the two factors. From a sexual selec-
58 tion perspective, a speaker who combines a low pitch with a fast tempo may be
59 most attractive (and vice versa), because this combination would suggest a healthy
60 physique as well as good motor and cognitive capabilities, a combination which is
61 presumably more rare in potential male partners than the separate capabilities and
62 characteristics.

63 The experiment reported below addresses these questions by manipulating Dutch
 64 sentences in tempo and pitch, and then asking Dutch female listeners to rate the
 65 attractiveness of the speaker. This attractiveness rating is regarded here as a proxy
 66 for the female listener's degree of preference for that male speaker in sexual selection,
 67 although vocal attractiveness also affects other social attributions (Babel, McGuire,
 68 & King, 2014).

69 In addition, this study also investigates whether these hypothesized effects of pitch
 70 and tempo are moderated by the presence of visual cues about the speaker in a portrait
 71 photo (see details below). On the one hand, humans have evolved to assess speakers
 72 not only by ear but also by eye,¹ so the task of rating a speakers' attractiveness may be
 73 more ecologically valid when a portrait is available. On the other hand, the presence
 74 of visual cues may well dampen the effects of prosodic cues. The fourth aim of this
 75 study was, therefore, to establish whether and how the presence of a portrait photo
 76 would affect a listeners' ratings of attractiveness of the speaker.

77 9.2 Methods

78 The experiment consisted of two sessions, in which the same speech stimuli were
 79 used. In the first session, speaker's voices were presented without a simultaneous
 80 portrait photo. In the second session, which included listeners who participated in
 81 the first session as well as new listeners, the same speech stimuli were presented *with*
 82 a portrait photo, in order to assess the effects of the portrait on listeners' responses.
 83 Listeners' task was to rate the attractiveness of the speaker.

84 During each session, a listener rated two different sentences spoken by the same
 85 speaker. One sentence was unchanged from the original, and the other sentence
 86 was manipulated orthogonally in pitch and/or in tempo, as described below. (This
 87 single-interval rating paradigm was chosen, instead of a two-interval forced-choice
 88 paradigm, because the latter would have highlighted the phonetic manipulations
 89 in one of the two speech intervals, and thus would have introduced biases in the
 90 responses subsequent to a listener noticing the manipulations).

91 The within-listener and within-speaker design allows for testing our primary pre-
 92 dictions regarding the hypothesized effects of manipulated pitch and manipulated
 93 tempo on the subjective voice attractiveness of male speakers. Listeners' judgements
 94 are predicted to be affected by the phonetic manipulations, with higher ratings for
 95 lowered pitch and faster tempo, and with lower ratings for higher pitch and slower
 96 tempo, as argued above. The effects of phonetic manipulations may interact, and
 97 may be moderated by the photo conditions.

¹ Although present-day listeners may be used to hear speakers without seeing them, this is presumably not how speech has evolved in humans.

9.2.1 Participants

Listeners were 208 students or employees at Utrecht University, from 8 different undergraduate course groups taught in Dutch. In order to conceal the research topic (knowledge of which might have biased responses), targeted participants as well as other persons were tested and subsequently presented with a questionnaire asking about gender, sexual orientation, age, speech/hearing problems, and guess about the purpose of the experiment. Data from 58 persons were excluded for various reasons listed in Table 9.1.

Subsequent analysis was based on data from 150 remaining targeted participants: all female, self-identified other than lesbian, within age range 17–29 (median age 20, median absolute deviation 1.5, at second session; this was done to select participants from approximately the same age range as the speakers, to improve ecological validity), without self-identified speech/hearing problems, and not aware of the purpose of the study. All participants were highly proficient in Dutch, as their native language or as a non-native language attested at an advanced academic level (B2 or higher).

9.2.2 Materials

Stimulus sentences were taken from Dutch spontaneous monologues by 24 male speakers (age $M = 18.0$, $s = 0.7$, range 16–19 years), who spoke about an informal topic of their own choice. These monologues had been previously recorded for a different study at 44.1 kHz (for further details, see Quené & Orr, 2014; Quené, Orr, & van Leeuwen, 2017). Two sentences were selected from each speaker's interview. Selected sentences were between 2.5 and 3.5 s in duration, which were spoken fluently and without a long pause, with neutral content, comprehensible without context, and not elliptic (i.e., contained both a subject and an inflected verb). Thus the sentences

Table 9.1 Numbers of participants, with reasons for exclusion from data analysis. Multiple reasons may apply to a single participant

Description	Female	Male	Total
All participants tested	≥ 155	≤ 53	208
Aborted prematurely	≤ 3	≤ 1	3
Already knew purpose of study	6	3	9
Speech/hearing problems	7	4	11
No valid responses	0	1	1
Gender male or unspecified	0	≤ 35	35
Orientation lesbian	3	0	3
Age < 16 or > 30	3	0	3
Participants remaining	150	0	150

122 should provide listeners with enough speech material to rate voice attractiveness,
 123 without requiring listeners' inference of context or grammar.

124 For each of the 24×2 selected stimulus sentences, average syllable duration
 125 (excluding pauses, Quené, 2008 and average F0 (over voiced portions) were measured
 126 using Praat (Boersma & Weenink, 2015). These measurements were then analyzed
 127 by means of linear mixed models (Quené & van den Bergh, 2004; 2008; Bates et
 128 al., 2015; R Core Team, 2018) with only the intercept as a fixed predictor, and with
 129 speakers as random intercepts. The estimated average syllable duration was 0.188 s
 130 ($s_u = 0.015$, $s_e = 0.026$, ICC = 0.25, i.e., with most variance between sentences
 131 within speakers), and the estimated average F0 was 116 Hz ($s_u = 16$, $s_e = 7$, ICC =
 132 0.82, i.e., with most variance between speakers).

133 In order to once again conceal the research topic, similar filler stimuli, but spoken
 134 by female speakers, were also included in the experiment. These filler sentences
 135 were taken from recorded monologues of 24 female speakers (each contributing one
 136 sentence) from the same corpus and using the same selection criteria as for male
 137 speakers. Neither the filler sentences themselves nor any responses to these fillers
 138 were further analyzed.

139 For the second session, each individual speaker (male or female voice) was
 140 matched to an individual portrait photo. These photos were taken from 3 public
 141 databases of facial portraits (Hancock, 2008; Nefian, 1999; Spacek, 2008) and did
 142 *not* portray the actual speakers. The selected photos of 24 males and 24 females
 143 each showed one person in the target age range (18–25 years) with a neutral facial
 144 expression. All selected photos were cropped and/or resized to the same size.

145 9.2.3 Speech Manipulations

146 One of the two sentences of each male speaker was retained as a baseline stimulus
 147 with unchanged tempo and unchanged pitch. The other sentence of each male speaker
 148 was varied in tempo (factors 0.85, 1.00, 1.15) and in overall pitch (-1.5 , 0, $+1.5$
 149 semitone), yielding 8 manipulated versions of each sentence. The changes are well
 150 above the respective just noticeable differences (Quené, 2006; 'THart, Collier, &
 151 Cohen, 1990) and they correspond to approximately $\pm 1s_e$ for both manipulations,
 152 while the resulting sentences still sound very natural to us. Filler sentences by female
 153 speakers were not varied. Tempo and pitch were manipulated by means of sox
 154 (Bagwell, 2013). Finally, stimulus and filler sentences were all scaled to -0.5 dB
 155 relative to the maximum amplitude.

156 9.2.4 Procedure

157 The 8 manipulated versions of each sentence were distributed over 8 experimental
 158 lists, counterbalanced over the 24 male speakers. The 24 unchanged male-spoken

159 sentences and 24 female-spoken filler sentences were added to each experimental
 160 list. Hence, the unchanged sentences of all speakers were presented to all listen-
 161 ers, whereas the changed sentences were partitioned over lists so that each listener
 162 heard only a single changed version of a particular sentence. This design allowed
 163 subsequent within-speaker and within-listener comparisons of baseline and changed
 164 versions. The 72 sentences were presented in quasi-random order² (which was how-
 165 ever the same across the 8 lists).

166 The experiment was conducted in a classroom setting, with each experimental
 167 list presented to a separate undergraduate course group. In the first session, speech
 168 stimuli were presented (using PowerPoint) over the classroom sound system. In the
 169 second session, typically a few days later, the same speech stimuli were presented
 170 with simultaneous portraits visible, using the same sound system and the classroom
 171 computer projector. The inter-stimulus interval was 3 s in both sessions, as determined
 172 in pilot tests.

173 Of the remaining 150 participants, 76 participated only in the first session (absent
 174 from the second session), 20 only in the second session (absent from the first ses-
 175 sion), and 54 participated in both sessions, the latter group allowing within-subject
 176 comparisons.

177 Participants were instructed to rate the attractiveness of the speaker on a 7-point
 178 Likert scale (1 extremely unattractive, 7 extremely attractive) on a printed response
 179 sheet. For the first session, their instruction was as follows (in translation):

180 ... In a moment you will hear 72 sound fragments of people saying something. We'd like to
 181 ask you to indicate for every sound fragment how attractive you find the speaker. You have
 182 about 3 s to respond for each person.

183 For the second session, participants' instruction was as follows (in translation):

184 ... In a moment you will see 72 photos of people. With every face you will also hear a sound
 185 fragment. We'd like to ask you to indicate for every person how attractive [Dutch: "hoe
 186 aantrekkelijk"] you find that person. You have about 3 s to respond for each person.

187 After the rating sessions, participants were invited to answer a brief questionnaire
 188 about their gender, age, native language(s), hearing problems, speech problems, dex-
 189 terity, and sexual orientation as heterosexual or homosexual or bisexual or unknown
 190 (including unwilling to answer); see Sect. 9.2.1.

191 9.3 Results

192 The average ratings by the targeted listeners observed in the listening experiment are
 193 summarized in Table 9.2. The lower standard error in the baseline condition is due
 194 to the larger number of responses in this condition, because all listeners have judged
 195 the unchanged sentences of all speakers (see Sect. 9.2.4).

²Between stimuli involving the same speaker, at least 5 different test or filler sentences were pre-
 sented.

Table 9.2 Mean responses (by targeted listeners only) of subjective attractiveness on a 7-point scale, broken down by manipulations of tempo and pitch, with standard errors in parentheses

		Pitch		
		Lower	Unchanged	Higher
Tempo	Slower	2.78 (0.06)	2.94 (0.06)	2.47 (0.05)
	Unchanged	3.28 (0.06)	3.30 (0.02)	2.55 (0.05)
	Faster	3.15 (0.06)	3.39 (0.06)	2.55 (0.06)

The separate responses given by each of the 150 remaining listeners to each of the 24 unchanged and 24 manipulated speech stimuli were analyzed by means of a cumulative-link mixed-effects model (CLMM) (Quené & van den Bergh, 2004; Christensen, 2015). This family of models (also known as proportional odds models) regards the dependent variable as ordinal, and coefficients represent the changes in log odds of a response falling in the *j*th category or higher. In other words, a CLMM as used here is somewhat similar to a GLMM (Quené & Van den Bergh, 2008), but with multiple ordered response categories. Fixed predictors in the CLMM were the 8 manipulated conditions of tempo and pitch (using dummy coding, with the unchanged condition as baseline), the centered trial number,³ and the absence (baseline code 0) or presence (contrast code 1) of a portrait photo. Two-way interactions between photo and manipulation conditions were also included as fixed predictors. Random predictors in the CLMM were listeners (*n* = 150), speakers (*n* = 24), and sentences (*n* = 48) as three crossed random intercepts. The main effect of the photo condition was also included as a random slope at the speaker level, thus allowing for nonuniform effects of the portrait photo across speakers.

The fixed regression coefficients, random variances and correlations, and category thresholds estimated by the CLMM described above are listed in Table 9.3.

The **fixed** part of the CLMM shows several interesting effects. In the conditions without a photo (first session), the conditions with slower tempo, as well as the conditions with higher pitch, all yield a significant negative effect: slower tempo is *less* attractive than the unchanged baseline, and so is higher pitch. However, none of the opposite conditions with faster tempo (conditions FU and FL), and none of the conditions with lower pitch, yields a positive effect: faster tempo is equally attractive as the unchanged baseline, and so is lower pitch.

Second, the photo condition yielded a large and significant negative main effect, with considerably lower ratings in the second session (with photo) as compared to the first session (without photo). The interactions suggest that the negative effect of adding a photo is significantly mitigated, in particular, in those phonetic conditions yielding the most negative ratings without a photo. As discussed below, this interaction pattern may suggest a floor effect.

³This centered trial number was scaled by factor 0.1 for computational reasons.

Editor Proof

Table 9.3 Estimated coefficients of the CLMM, for intercepts and effects of conditions of tempo (S: slower, U: unchanged, F: faster) and pitch (L: lower, U: unchanged, H: higher), trial number (centered and scaled), and photo condition. Random effects are reported in units of variance of log odds (logit), with standardized correlation among random effects; a significant correlation is marked with an asterisk ($p < 0.05$ according to bootstrapped 95% confidence interval of the correlation, over 200 bootstrap replications). Fixed effects are reported in log odds (logit) units; significant coefficients are marked with an asterisk ($p < 0.05$)

Random: listeners	Variance			
(Intercept)	0.9569			
Random: speakers	Variance	Correlation		
(Intercept)	0.8032			
Photo	0.4856	-0.57*		
Random: sentences	Variance			
(Intercept)	0.4076			
Fixed	Estimate	Std. Error	z value	p value
cond.SH	-1.75	0.22	-8.11	<0.0001*
cond.SU	-0.78	0.21	-3.65	0.0003*
cond.SL	-0.97	0.21	-4.54	<0.0001*
cond.UH	-1.29	0.21	-6.05	<0.0001*
cond.UL	-0.10	0.21	-0.49	0.6224
cond.FH	-1.84	0.22	-8.55	<0.0001*
cond.FU	0.15	0.21	0.71	0.4806
cond.FL	-0.07	0.21	-0.32	0.7457
photo	-1.38	0.16	-8.84	<0.0001*
cond.SH:photo	0.59	0.16	3.36	0.0008*
cond.SU:photo	0.26	0.17	1.53	0.1257
cond.SL:photo	0.45	0.17	2.63	0.0087*
cond.UH:photo	0.17	0.18	0.98	0.3261
cond.UL:photo	0.12	0.17	0.71	0.4789
cond.FH:photo	1.14	0.17	6.53	<0.0001*
cond.FU:photo	-0.20	0.17	-1.22	0.2225
cond.FL:photo	-0.01	0.17	-0.04	0.9700
trial	-0.09	0.06	-1.59	0.1120
Category thresholds	Estimate	Std. Error	z value	
1 2	-3.25	0.24	-13.36	
2 3	-1.43	0.24	-5.91	
3 4	-0.04	0.24	-0.17	
4 5	1.19	0.24	4.94	
5 6	2.75	0.24	11.28	
6 7	4.82	0.26	18.55	

227 Finally, the coefficients in the fixed part of the CLMM did not show an effect
228 of the trial number on listeners' judgements: listeners did not tend to increase or
229 decrease their ratings during a session.

230 The **random** part of the CLMM shows that speakers' intercepts correlate with
231 speakers' slope of the photo condition ($r = -0.57$): speakers whose voices were
232 judged as more attractive tended to "lose" less when combined with an alleged
233 portrait, or in other words, the negative main effect of the photo portrait was relatively
234 stronger (more negative) for less-attractive voices.

235 9.4 Discussion

236 First, the results confirm previously reported effects of **pitch** manipulations on attrac-
237 tiveness (Collins, 2000; Feinberg et al., 2005): male voices with increased pitch are
238 rated as less attractive by heterosexual female listeners. While previous studies used
239 only short vowel stimuli, these findings are partially replicated here with sentence-
240 length stimuli. This result further corroborates the evidence for the role of male voice
241 pitch in sexual selection through female choice of mate. In spite of this effect in the
242 manipulated stimuli, however, the corresponding effect was not observed for voices
243 with decreased pitch.

244 Second, the results confirm our prediction that manipulations of **tempo** also affect
245 the speaker's attractiveness, with slower speech being less attractive. Slower speakers
246 may be regarded as less attractive because speech tempo may indicate the speaker's
247 (relatively poor) motor skills and cognitive capabilities. Again, the corresponding
248 effect was not observed for voices with increased tempo.

249 In comparison, the detrimental effect of slower tempo appears to be somewhat
250 smaller than that of higher pitch (cf. Table 9.3). This difference in effect size for pitch
251 and tempo may be explained in three ways. One explanation could be that pitch con-
252 stitutes a more salient cue in sexual selection than tempo, because pitch varies more
253 between speakers (and less within speakers) than tempo does (cf. Sect. 9.2.2 for vari-
254 ations in our stimuli), so that pitch may be a more reliable indicator of the speaker's
255 individual characteristics than tempo. Another plausible explanation could be that
256 our pitch manipulations were perceptually larger than our tempo manipulations, rela-
257 tive to the individual differences between speakers. The prosodic measurements
258 and manipulations described above (Sects. 9.2.2–9.2.3), however, do not support this
259 latter explanation: the pitch manipulations are about $\pm \frac{1}{2}s_u$ whereas the tempo manip-
260 ulations are relatively larger, about $\pm 2s_u$ (for comparison, both manipulations were
261 about $\pm 1s_e$ in magnitude). A third explanation was proposed by Babel et al. (2014)
262 who argue that attractive voice properties may not be universal, but dependent on cul-
263 tural preferences; the weights of tempo and pitch properties on voice attractiveness
264 may thus be culturally constrained. Further research, with different sizes of phonetic
265 manipulations and with listeners sampled from different cultures, would be required
266 to rule out one or more of these explanations.

267 Third, the results do not support the hypothesized interaction between pitch and
268 tempo cues on speakers' attractiveness. In the first session (without photo), neither
269 lower pitch, nor faster tempo, nor the combination of these two manipulations yielded
270 a positive effect on voice attractiveness. Moreover, lowest ratings were obtained
271 in conditions with increased pitch, irrespective of the tempo manipulations. This
272 suggests that the combined traits of physical and cognitive capabilities are somehow
273 assessed independently, contrary to the expectations outlined in Sect. 9.1.

274 Finally, the results suggest that the **photo** portraits may have introduced floor
275 effects in this experiment. Coefficients in the fixed part of the CLMM suggest that
276 conditions yielding the lowest ratings without a photo (session 1) also decrease
277 less with a photo (session 2), which may be because the conditions involving less-
278 attractive speech cannot "lose" as many points when combined with a photo. In
279 addition, speakers who are rated as more attractive tend to "lose" more when accom-
280 panied by a photo ($r = -0.57$, Table 9.3), which may again be because less-attractive
281 speakers cannot be rated below the floor of the Likert scale. The photos were included
282 in the experimental design in order to investigate the effects of (ecologically valid)
283 visual cues on voice attractiveness ratings. However, the unexpected negative effect
284 of adding a portrait photo may have resulted in ratings that were too low to show
285 the effects of phonetic properties. One possible explanation is that the photos were
286 taken from relatively old sources (portraits were at least 8 years old at the time of
287 testing) and may have contained outdated visual cues regarding style, hairdress, etc.,
288 for the target listeners in our study. More speculatively, there may have been some
289 unknown mismatch between (non-Dutch) portraits (Hancock, 2008; Nefian, 1999;
290 Spacek, 2008) and (Dutch) voices, yielding a negative effect on the ratings in the
291 with-portrait condition. For further phonetic research into listeners' attractiveness
292 judgements, we recommend to refrain from randomly matched portraits accompa-
293 nying the voice stimuli.

294 9.5 Conclusions

295 Female listeners rate a male speaker as less attractive if his voice pitch is increased
296 and if his speech tempo is decreased, relative to a baseline sentence with unchanged
297 pitch and tempo. These effects suggest that both pitch and tempo play a role in
298 speech-based sexual selection of males by females, although our results suggest that
299 the underlying mechanisms for pitch and tempo may well be different. Voice pitch
300 indicates the speaker's health and physical dominance (Dabbs & Mallinger, 1999;
301 Puts et al., 2012; Collins, 2000; Feinberg et al., 2005), while speech tempo may
302 indicate the speaker's motor skills and cognitive competence (Apple et al., 1979;
303 Smith et al., 1975). The effect of voice pitch on attractiveness is larger than that of
304 speech tempo, perhaps because pitch varies relatively more between speakers than
305 within speakers, in contrast to tempo, so that pitch may constitute a more reliable
306 cue to a speaker's individual characteristics.

307 **Acknowledgments** Results from a different, related study (using the same audio stimuli, always
 308 presented with photos, with different participants) were reported at the Speech Prosody 2016 confer-
 309 ence (Boston, U.S.A.). We thank Nivja de Jong, Gerrit Bloothoof, Huub van den Bergh, the
 310 audience at Speech Prosody 2016, and three anonymous reviewers, for helpful comments and sug-
 311 gestions.

312 References

- 313 Andersson, M. B. (1994). *Sexual selection*. Princeton: Princeton University Press.
- 314 Apicella, C. L., Feinberg, D. R., & Marlowe, F. W. (2007). Voice pitch predicts reproductive success
 315 in male hunter-gatherers. *Biology Letters*, *3*, 682–684.
- 316 Apple, W., Streeter, L. A., & Krauss, R. M. (1979). Effects of pitch and speech rate on personal
 317 attributions. *Journal of Personality and Social Psychology*, *37*, 715–727.
- 318 Archer, J. (2006). Testosterone and human aggression: An evaluation of the challenge hypothesis.
 319 *Neuroscience and Biobehavioral Reviews*, *30*, 319–345.
- 320 Babel, M., McGuire, G., & King, J. (2014). Towards a more nuanced view of vocal attractiveness,
 321 *PLoS ONE*, *9*(2), e88616. <https://doi.org/10.1371/journal.pone.0088616>
- 322 Bagwell, C. (2013). Sound eXchange (SOX), version 14-4-1. <http://sourceforge.net/projects/sox/>.
- 323 Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). lme4: Linear mixed-effects models using
 324 Eigen and S4. R package version 1.1-9. <http://CRAN.R-project.org/package=lme4>
- 325 Boersma, P., & Weenink, D. (2015) Praat: Doing phonetics by computer, version 6.0 <http://www.praat.org>
- 326
- 327 Bruckert, L., Liénard, J.-S., Lacroix, A., Kreutzer, M., & Leboucher, G. (2006). Women use voice
 328 parameters to assess men's characteristics. *Proceedings of the Royal Society B: Biological Sci-*
 329 *ences*, *273*(1582), 83–89.
- 330 Christensen, R. H. B. (2015). ordinal—Regression Models for Ordinal Data. R package version
 331 2015.6-28. <http://www.cran.r-project.org/package=ordinal/>.
- 332 Clark, A. S., & Henderson, L. P. (2003). Behavioral and physiological responses to anabolic-
 333 androgenic steroids. *Neuroscience and Biobehavioral Reviews*, *27*, 413–436.
- 334 Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour*, *60*, 773–780.
- 335 Dabbs, J. M., & Mallinger, A. (1999). High testosterone levels predict low voice pitch among men.
 336 *Personality and Individual Differences*, *27*, 801–804.
- 337 Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M., & Perrett, D. I. (2005). Manipulations of
 338 fundamental and formant frequencies influence the attractiveness of human male voices. *Animal*
 339 *Behaviour*, *69*, 561–568.
- 340 Hancock, P. (2008). Utrecht ECVP: Psychological Image Collection at Stirling. <http://pics.stir.ac.uk/>.
- 341
- 342 'tHart, J. T., Collier, R., & Cohen, A. (1990). *A perceptual study of intonation: An experimental-*
 343 *phonetic Approach to speech perception*. Cambridge: Cambridge University Press.
- 344 Hill, A. K., Hunt, J., Welling, L. L. M., Cárdenas, R. A., Rotella, M. A., Wheatley, J. R., et al.
 345 (2013). Quantifying the strength and form of sexual selection on men's traits. *Evolution and*
 346 *Human Behavior*, *34*, 334–341.
- 347 Holmberg, E. B., Hillman, R. E., & Perkell, J. S. (1988). Glottal airflow and transglottal air pressure
 348 measurements for male and female speakers in soft, normal, and loud voice. *The Journal of the*
 349 *Acoustical Society of America*, *84*, 511–529.
- 350 Kordsmeyer, T. L., Hunt, J., Puts, D. A., Ostner, J., & Penke, L. (2018). The relative importance of
 351 intra- and intersexual selection on human male sexually dimorphic traits. *Evolution and Human*
 352 *Behavior*, *39*, 424–436.
- 353 Jacewicz, E., Fox, R. A., & Wei, L. (2010). Between-speaker and within-speaker variation in speech
 354 tempo of American English. *The Journal of the Acoustical Society of America*, *128*, 839–850.

- 355 Moon, S. -J., & Lindblom, B. (2003). Two experiments on oxygen consumption during speech
 356 production: Vocal effort and speaking tempo. In *Proceedings of XVth International Congress of*
 357 *Phonetic Sciences* (pp. 3129–3132), Barcelona, Spain.
- 358 Nefian, A. V. (1999). Georgia Tech face database. http://www.anefian.com/research/face_reco.htm
- 359 Orr, R., Quené, H., van Beek, R., Diefenbach, T., van Leeuwen, D. A., Huijbregts, M. (2011). An
 360 International English speech corpus for longitudinal study of accent development. In *InterSpeech*
 361 *2011, 27–31 Aug, Florence, Italy, Proceedings* (pp. 1889–1892).
- 362 Puts, D. A., Gaulin, S. J. C., & Verdolini, K. (2006). Dominance and the evolution of sexual
 363 dimorphism in human voice pitch. *Evolution and Human Behavior*, *27*, 283–296.
- 364 Puts, D. A., Apicella, C. L., & Cárdenas, R. A. (2012). Masculine voices signal men's threat potential
 365 in forager and industrial societies. *Proceedings of the Royal Society B: Biological Sciences*, *279*,
 366 601–609.
- 367 Pisanski, K., Fraccaro, P. J., Tigue, C. C., O'Connor, J. J. M., Röder, S., Andrews, P. W., et al.
 368 (2014). Vocal indicators of body size in men and women: A meta-analysis. *Animal Behaviour*,
 369 *95*, 89–99.
- 370 Quené, H. (2006). On the just noticeable difference for tempo in speech. *Journal of Phonetics*, *35*,
 371 353–362.
- 372 Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spon-
 373 taneous speech tempo. *The Journal of the Acoustical Society of America*, *123*, 1104–1113.
- 374 Quené, H., Orr, R., Long-term convergence of speech rhythm in L1 and L2 English. In *Speech*
 375 *Prosody 2014, 20–23 May, Dublin, Ireland, Proceedings* (pp. 342–345).
- 376 Quené, H., Orr, R., & van Leeuwen, D. (2017). Phonetic similarity of /s/ in native and second
 377 language: Individual differences in learning curves. *Journal of the Acoustical Society of America*,
 378 *142*(6), EL519–EL524. <https://doi.org/10.1121/1.5013149>.
- 379 Quené, H., & van den Bergh, H. (2004). On multi-level modeling of data from repeated measures
 380 designs: A tutorial. *Speech Communication*, *43*(1–2), 103–121.
- 381 Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random
 382 effects and with binomial data. *Journal of Memory and Language*, *59*(4), 413–425.
- 383 R Core Team (2018) R: A language and environment for statistical computing. R Foundation for
 384 Statistical Computing, Vienna, Austria, version 3.4.4. <http://www.R-project.org>.
- 385 Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and Lin-*
 386 *guistics Compass*, *3*, 621–640.
- 387 Smith, B. L., Brown, B. L., Strong, W. J., & Rencher, A. C. (1975). Effects of speech rate on
 388 personality perception. *Language & Speech*, *18*, 145–152.
- 389 Smith, K. M., Olkhov, Y. M., Puts, D. A., & Apicella, C. L. (2017). Hadza men with lower
 390 voice pitch have a better hunting reputation. *Evolutionary Psychology*. <https://doi.org/10.1177/1474704917740466>.
- 391 Spacek, L. (2008). Collection of facial images (faces94, faces95). <http://cswww.essex.ac.uk/mv/allfaces/>.
- 392
393

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Voice Attractiveness
Series Title	
Chapter Title	The Contribution of Amplitude Modulations in Speech to Perceived Charisma
Copyright Year	2020
Copyright HolderName	Springer Nature Singapore Pte Ltd.
Corresponding Author	Family Name Bosker Particle Given Name Hans Rutger Prefix Suffix Role Division Organization Max Planck Institute for Psycholinguistics Address P. O. Box 310, 6500 AH, Nijmegen, The Netherlands Division Psychology of Language Department Organization Donders Institute for Brain Cognition and Behaviour, Radboud University Address Kapittelweg 29, 6525 EN, Nijmegen, The Netherlands Email HansRutger.Bosker@mpi.nl
Abstract	Speech contains pronounced amplitude modulations in the 1–9 Hz range, correlating with the syllabic rate of speech. Recent models of speech perception propose that this rhythmic nature of speech is central to speech recognition and has beneficial effects on language processing. Here, we investigated the contribution of amplitude modulations to the subjective impression listeners have of public speakers. The speech from presidential candidates Hillary Clinton and Donald Trump in the three TV debates of 2016 was acoustically analyzed by means of modulation spectra. These indicated that Clinton’s speech had more pronounced amplitude modulations than Trump’s speech, particularly in the 1–9 Hz range. A subsequent perception experiment, with listeners rating the perceived charisma of (low-pass filtered versions of) Clinton’s and Trump’s speech, showed that more pronounced amplitude modulations (i.e., more ‘rhythmic’ speech) increased perceived charisma ratings. These outcomes highlight the important contribution of speech rhythm to charisma perception.
Keywords	Amplitude modulations - Speech rhythm - Modulation spectrum - Charisma perception - Temporal envelope - Political debates

Chapter 10

The Contribution of Amplitude Modulations in Speech to Perceived Charisma



Hans Rutger Bosker

Abstract Speech contains pronounced amplitude modulations in the 1–9 Hz range, correlating with the syllabic rate of speech. Recent models of speech perception propose that this rhythmic nature of speech is central to speech recognition and has beneficial effects on language processing. Here, we investigated the contribution of amplitude modulations to the subjective impression listeners have of public speakers. The speech from presidential candidates Hillary Clinton and Donald Trump in the three TV debates of 2016 was acoustically analyzed by means of modulation spectra. These indicated that Clinton’s speech had more pronounced amplitude modulations than Trump’s speech, particularly in the 1–9 Hz range. A subsequent perception experiment, with listeners rating the perceived charisma of (low-pass filtered versions of) Clinton’s and Trump’s speech, showed that more pronounced amplitude modulations (i.e., more ‘rhythmic’ speech) increased perceived charisma ratings. These outcomes highlight the important contribution of speech rhythm to charisma perception.

Keywords Amplitude modulations · Speech rhythm · Modulation spectrum · Charisma perception · Temporal envelope · Political debates

10.1 Introduction

Any spoken utterance, regardless of talker, language, or linguistic content, contains fast-changing spectral information (e.g., vowel formants, consonantal frication, etc.) as well as slower changing temporal information. The temporal information in speech is particularly apparent in the temporal envelope of speech, which includes the fluctuations in amplitude from consonants (constricted vocal tract, lower amplitude) to vowels (unconstricted vocal tract, higher amplitude), from stressed (prominent) to

H. R. Bosker (✉)

Max Planck Institute for Psycholinguistics, P. O. Box 310, 6500 AH Nijmegen, The Netherlands
e-mail: HansRutger.Bosker@mpi.nl

Psychology of Language Department, Donders Institute for Brain Cognition and Behaviour, Radboud University, Kapittelweg 29, 6525 EN Nijmegen, The Netherlands

© Springer Nature Singapore Pte Ltd. 2020

B. Weiss et al. (eds.), *Voice Attractiveness*, Prosody, Phonology and Phonetics,
https://doi.org/10.1007/978-981-15-6627-1_10

171

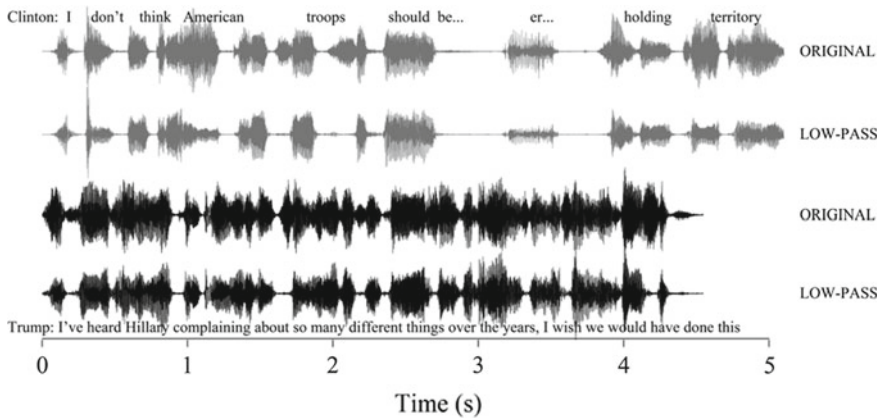


Fig. 10.1 Excerpts of Clinton's speech (in gray) with a notable syllabic rhythm around 3 Hz and Trump's speech (in black) with a notable lack of consistent slow-amplitude modulations. Below each waveform are the low-pass filtered versions of the excerpts, demonstrating that the original slow-amplitude modulations are maintained to a large degree

24 unstressed syllables (less prominent), etc. For instance, the top example in Fig. 10.1
 25 has pronounced fluctuations in amplitude (also known as amplitude modulations)
 26 occurring at around 3 Hz, related to the syllabic rate of the utterance (i.e., roughly
 27 three syllables per second).

28 The temporal dynamics of speech (e.g., energy patterns and syllable durations
 29 in speech) are semi-regular at multiple (segmental, syllabic, sentential) timescales
 30 (Poeppl, 2003; Rosen, 1992). Hence, speech is an intrinsically rhythmic signal, with
 31 'rhythmic' referring to the semi-regular recurrence over time of waxing and waning
 32 prominence profiles in the amplitude signature of speech (for other conceptualiza-
 33 tions of speech rhythm, see Kohler, 2009; Nolan & Jeon, 2014). Naturally produced
 34 syllable rates typically do not exceed a rate of 9 Hz (Ghitza, 2014; Jacewicz, Fox,
 35 & Wei, 2010; Pellegrino, Coupé, & Marsico, 2011; Quené, 2008; Varnet, Ortiz-
 36 Barajas, Erra, Gervain, & Lorenzi, 2004). As such, most of the energy in the ampli-
 37 tude modulations in the speech signal is found below 9 Hz (Ghitza & Greenberg,
 38 2009; Greenberg & Arai, 1999, 2004), across a range of typologically distant lan-
 39 guages (Ding et al., 2017; Varnet, Ortiz-Barajas, Erra, Gervain, & Lorenzi, 2017),
 40 with the most prominent modulation frequencies near the average syllable rate of
 41 3–4 Hz (Delgutte 1998).

42 In recent models of speech perception (Ghitza 2011; Giraud & Poeppel, 2012;
 43 Peelle & Davis, 2012), this rhythmic nature of speech is said to play a central role in
 44 speech recognition. For instance, speakers who are intrinsically more intelligible than
 45 others show more pronounced low-frequency modulations in the amplitude envelope
 46 (Bradlow, Torretta, & Pisoni, 1996). In fact, when the slow amplitude fluctuations
 47 in speech are degraded or filtered out, intelligibility drops dramatically (Drullman,
 48 Festen, & Plomp, 1994; Ghitza, 2012; Houtgast & Steeneken, 1973), while speech

49 with only minimal spectral information remains intelligible as long as low-frequency
50 temporal modulations are preserved (Shannon, Zeng, Kamath, Wygonski, & Ekelid,
51 1995). Similarly, speech stream segregation (understanding speech in noise; Aikawa
52 & Ishizuka, 2002), word segmentation (resolving continuous speech into words;
53 Cutler, 1994; Cutler & Butterfield, 1992; Cutler & Norris, 1988), and phoneme
54 perception (Bosker, 2017a; Bosker & Ghitza, 2018; Quené, 2005) are all influenced
55 by regular energy fluctuations in speech.

56 A powerful demonstration of the contribution of regular amplitude modulations
57 to speech comprehension is the finding that otherwise unintelligible speech can be
58 made intelligible by imposing an artificial rhythm (Bosker & Ghitza, 2018; Doelling,
59 Arnal, Ghitza, & Poeppel, 2014; Ghitza, 2012, 2014). For instance, Bosker and
60 Ghitza (2018) took Dutch recordings of seven-digit telephone numbers (e.g., “215–
61 4653”) and compressed these by a factor of 5 (i.e., make the speech five times as
62 fast while preserving spectral properties such as pitch and formants). This heavy
63 compression manipulation made the intelligibility of the telephone numbers drop
64 from the original 99% to about 39% digits correct. However, Bosker and Ghitza then
65 imposed an artificial rhythm onto the heavily compressed speech, by taking 66 ms
66 windows of compressed speech and spacing these apart by 100 ms of silence (i.e.,
67 inserting 100-ms silent intervals). This ‘repackaged’ condition did not contain any
68 additional linguistic or phonetic information compared to the heavily compressed
69 speech; it only differed in having a very pronounced amplitude modulation around
70 6Hz. The authors found that imposing this artificial rhythm onto the compressed
71 speech boosted intelligibility (from 39 to 71%) digits correct, demonstrating that
72 regular amplitude modulations play a central role in speech perception.

73 Rhythmic amplitude modulations in speech not only affect speech intelligibility
74 but they also play a role in spoken communication more generally. For instance, syn-
75 tactic processing (Roncaglia-Denissen, Schmidt-Kassow, & Kotz, 2013), semantic
76 processing (Rothermich, Schmidt-Kassow, & Kotz, 2012), and recognition memory
77 (Essens & Povel 1985) are all facilitated by regular meter. Moreover, there are even
78 suggestions in the literature that listeners explicitly prefer listening to speech with a
79 clear rhythmic structure. For instance, Obermeier et al. (2013) took four-verse stan-
80 zas from old German poetry and independently manipulated the rhyme and meter of
81 these poetry fragments. Rhyme was manipulated by substituting rhyming sentence-
82 final words with non-rhyming words with the same metrical structure (maintaining
83 meter), while meter was manipulated by substituting a sentence-medial word with a
84 word with mismatching metrical structure (e.g., “Nacht” > “Dunkelheit”; maintain-
85 ing rhyme in sentence-final words). Native German participants rated the original and
86 manipulated fragments of poetry on liking and perceived intensity. Results indicated
87 that non-rhyming and non-metrical stanzas received lower ratings on both the liking
88 and perceived intensity scales, suggesting that the presence of rhythmical structure
89 induces greater esthetic liking and more intense emotional processing (Obermeier et
90 al., 2013, 2016).

91 Here, we examined the contribution of rhythmic amplitude modulations to the
 92 perception of charisma in public speakers' voices. Charisma and charismatic lead-
 93 ership are intensively studied topics, with clear implications for public speakers,
 94 politics, religion, and society at large. There seems to be a consensus in the literature
 95 that being a charismatic speaker is a necessary precondition for being a charismatic
 96 leader. In fact, how one speaks (i.e., performance characteristics, such as pitch, loud-
 97 ness, prosody, etc.) has been argued to contribute to charisma perception more than
 98 what one says (i.e., the linguistically formulated communicative message; Awamleh
 99 & Gardner, 1999; Rosenberg & Hirschberg, 2009). Several studies have, therefore,
 100 attempted to find acoustic correlates of charisma in public speakers' voices (see
 101 also in this volume; Rosenberg & Hirschberg, this volume; Brem & Niebuhr, this
 102 volume). For instance, pausing behavior (D'Errico, Signorello, Demolin & Poggi,
 103 2013), speech rate (D'Errico, et al. & Poggi, 2013), overall intensity (Niebuhr, Voße &
 104 Brem, 2016), number and type of disfluencies (Novák-Tóth, Niebuhr, & Chen, 2017),
 105 and timbre (Weiss and Burkhardt, 2010) have all been identified as contributing to
 106 perceived charisma and personality. However, although there are suggestions in the
 107 literature that greater variability in pitch and intensity contours increases perceived
 108 charisma (D'Errico et al., 2013; Niebuhr et al., 2016; Rosenberg & Hirschberg, 2009),
 109 it is unclear what the role of the rhythm of speech is in charisma perception. There-
 110 fore, the present research goal was to investigate how political debaters make use of
 111 variation in the amplitude envelope in speech production and how this variation, in
 112 turn, may affect speech perception.

113 Regarding rhythm in speech production, we report an acoustic comparison of
 114 the temporal amplitude modulations in the speech produced by two presidential
 115 candidates in the American elections of 2016: Hillary Clinton and Donald Trump.
 116 Recordings from three national presidential debates were collected and the speech
 117 produced by both candidates was first matched for overall intensity. Thereafter, their
 118 speech was analyzed by means of modulation spectra (Bosker & Cooke, 2018; Ding
 119 et al., 2017; Krause & Braida, 2004). These modulation spectra quantify the power
 120 of individual modulation frequency components present in a given signal (e.g., see
 121 Fig. 10.2), with power on the y-axis and modulation frequency on the x-axis. They
 122 can be used to assess which modulation frequencies are most prominent in differ-
 123 ent signals (e.g., speech and music show well-separated peaks around 5 and 2 Hz,
 124 respectively; Ding et al. 2017) but also to compare the overall power (in differ-
 125 ent frequency bands) across talkers or speech registers (Krause & Braida, 2004). For
 126 instance, Bosker and Ghitza 2018 calculated modulation spectra of spoken sentences
 127 produced in quiet (plain speech) and the same sentences produced in noise (Lom-
 128 bard speech). Results showed greater power in Lombard speech compared to plain
 129 speech, particularly in the 1–4 Hz range, demonstrating that talkers produce more
 130 pronounced amplitude modulations when talking in noise, presumably to aid speech
 131 comprehension.

132 Similarly, the present acoustic analysis compared the power of different mod-
 133 ulation frequency bands across the two talkers. Greater power in the modulation
 134 spectrum of one speaker over another would reveal a more pronounced temporal

135 envelope in that particular candidate's speech (i.e., greater amplitude modulations).
136 Specifically, we expect power differences to occur within the frequency range of
137 typical speech rates, namely below 9 Hz because (1) this modulation range is most
138 characteristic of spontaneous speech (Ding et al., 2017); and (2) previous research
139 indicates that differences between speech registers (plain vs. Lombard speech) are
140 apparent in the lower modulation range (Bosker and Ghitza 2018). Power differ-
141 ences in this 1–9 Hz modulation range would be indicative of a more regular syllabic
142 rhythm. Moreover, the locations of peaks in the modulation spectrum would reveal
143 which modulation frequencies are most pronounced in that speaker's amplitude enve-
144 lope, being indicative of a specific rhythm preference. By contrast, differences in the
145 power of modulation frequencies between 9–15 Hz are expected to be smaller (if
146 present at all) since this modulation range is less pronounced in speech and is not
147 straightforwardly related to particular acoustic or perceptual units in speech.

148 When it comes to quantifying rhythm in speech, modulation spectra have several
149 advantages over other rhythm metrics that have been introduced in the literature,
150 such as %V (percentage over which speech is vocalic; Ramus et al. (1999)), *ThetaC*
151 (standard deviation of consonantal intervals; Ramus et al. (1999)), PVI (pairwise
152 variability index; Grabe and Low (2002)), or normalized metrics such as VarcoV
153 and VarcoC (Dellwo, 2006; White and Mattys, 2007). These metrics assess dura-
154 tional variability (Loukina et al., 2011), not necessarily periodicity. That is, both
155 isochronous and isochronous distributions of vowels and consonants can have the
156 same %V. Moreover, such measures are influenced by between-language differences,
157 whereas modulation spectra are not (Ding et al., 2017).

158 Going beyond merely identifying differences in the use of rhythm between speak-
159 ers in speech production, we also tested the contribution of pronounced amplitude
160 modulations to speech perception. Specifically, a rating experiment was carried out
161 with low-pass filtered versions of (a subset of) the speech from both speakers. Fil-
162 tering was applied to reduce the contribution of lexical-semantic information to
163 participants' judgments while maintaining the temporal structure of the acoustic sig-
164 nal (see Fig. 10.1), forcing listeners to base their judgments primarily on temporal
165 characteristics. In line with the introduced beneficial effects of rhythmic regular-
166 ity on speech intelligibility and esthetic liking, we hypothesized that the perceived
167 charisma ratings would correlate with the speech rhythm in the signals. That is,
168 speech fragments with more pronounced amplitude modulations in the 1–9 Hz range
169 would be expected to be rated as more charismatic than speech fragments with less
170 pronounced amplitude modulations. If corroborated, this would indicate that speech
171 rhythm not only contributes to intelligibility and the qualitative appreciation of the
172 linguistic message but also to the subjective impression listeners have of a (public)
173 speaker.

174 10.2 Acoustic Analysis

175 10.2.1 Method

176 10.2.1.1 Materials

177 Recordings of all three presidential debates between Hillary Clinton and Donald
 178 Trump were retrieved from Youtube. The first debate (NBC News 2016) took place
 179 at Hofstra University, Hempstead, NY, USA, on September 26, 2016, and had the
 180 form of a traditional debate: the two candidates responded to questions posed by a
 181 moderator. The second debate (ABC News, 2016a) was broadcasted from Washing-
 182 ton University in St. Louis, St. Louis, MO, USA, on October 9, 2016. This debate
 183 was structured as a ‘town hall discussion’ with the candidates responding mostly to
 184 audience member questions. To illustrate, Fig. 10.1 shows two excerpts of Clinton’s
 185 and Trump’s speech in the second debate. The presence of a 3 Hz syllabic ‘beat’ is
 186 clearly visible in Clinton’s waveform, whereas Trump’s speech notably lacks slow-
 187 amplitude modulations. Finally, the third debate (ABC News, 2016b) took place at
 188 the University of Nevada, Las Vegas, Las Vegas, NV, USA, on October 19, 2016,
 189 and had the form of a traditional debate again.

190 All monologue speech from either candidate was manually annotated. That is,
 191 only those speech fragments in which one talker and one talker alone was speaking
 192 (uninterrupted monologue including all pauses, corrections, hesitations, etc.) was
 193 analyzed. Speech fragments that included crosstalk, laughter, applause, questions
 194 posed by the moderator, etc., were excluded from analyses. Monologues longer than
 195 approximately 35 s were cut into smaller fragments of <35 s at sentence boundaries.
 196 For the first debate, these annotations resulted in 93 speech fragments produced by
 197 Clinton (duration: $M = 24$ s; $SD = 7$ s; $range = 5-36$ s; $total = 2263$ s) and 98
 198 speech fragments produced by Trump (duration: $M = 25$ s; $SD = 7$ s; $range = 6-$
 199 35 s; $total = 2514$ s). For the second debate, these annotations resulted in 77 speech
 200 fragments produced by Clinton (duration: $M = 29$ s; $SD = 5$ s; $range = 8-36$ s;
 201 $total = 2243$ s) and 82 speech fragments produced by Trump (duration: $M = 27$ s;
 202 $SD = 6$ s; $range = 7-35$ s; $total = 2241$ s). For the third debate, these annotations
 203 resulted in 93 speech fragments produced by Clinton (duration: $M = 24$ s; $SD = 7$
 204 s; $range = 5-35$ s; $total = 2245$ s) and 76 speech fragments produced by Trump
 205 (duration: $M = 23$ s; $SD = 8$ s; $range = 5-34$ s; $total = 1779$ s).

206 10.2.1.2 Procedure

207 Before analysis of the speech fragments, the overall power (root mean square; RMS)
 208 in each fragment was normalized (set to an arbitrary fixed value), thus matching
 209 the overall power of the speech from both speakers. Following this normalization
 210 procedure, the speech fragments from each debate were analyzed separately.

211 First, the modulation spectrum of each individual speech fragment produced by
 212 Clinton was calculated, using a method adapted from (Bosker and Cooke 2018).
 213 It involved filtering the speech fragment by a band-pass filter spanning the 500–
 214 4000 Hz range and deriving the envelope of the filter’s bandlimited output (i.e.,
 215 Hilbert envelope). The envelope signal was zero-padded to the next power of 2
 216 higher than the length of the longest fragment of that particular speaker to achieve
 217 the same frequency resolution across recordings. This signal was then submitted
 218 to a Fast Fourier Transform (FFT), resulting in the modulation spectrum of that
 219 particular speech fragment. Finally, the average power in two frequency bands was
 220 calculated: average power in the 1–9 Hz range and average power in the 9–15 Hz
 221 range, resulting in two different observations for each of the speech fragments. Note
 222 that natural speech rates typically fall below 9 Hz. The same steps were then repeated
 223 for Trump’s speech fragments.

224 This analysis procedure was followed for each of the three debates and formed
 225 the two dependent variables (average power below and above 9 Hz) for statistical
 226 analyses reported below. In order to visualize the average rhythmicity in the speech
 227 of one speaker in one debate, all individual modulation spectra of one speaker in one
 228 debate were downsampled by a factor of 25 and thereafter averaged.

229 10.2.2 Results

230 Data from the three debates are reported separately to allow for comparison across
 231 debates. Note, however, that follow-up analyses did not reveal large qualitative dif-
 232 ferences between the outcomes of the three debates.

233 10.2.2.1 First Debate

234 The average modulation spectra of the speech produced by both speakers in each of
 235 the three debates is given in Fig. 10.2.

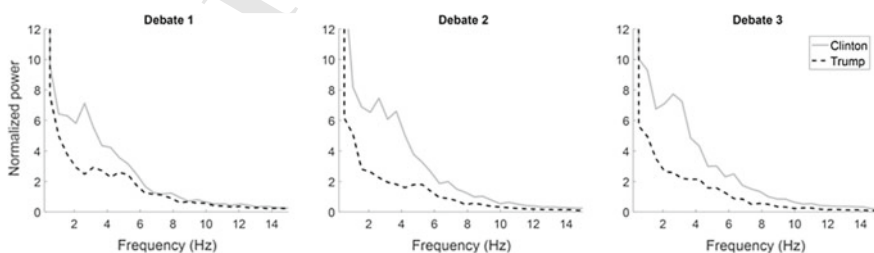


Fig. 10.2 Average modulation spectra of the speech produced by Hillary Clinton (gray solid lines) and Donald Trump (black dashed lines), separately for the three presidential debates

236 A simple linear model was built in R (R Development Core Team, 2012) separately
 237 for each of the two frequency bands (1–9 and 9–15 Hz), predicting the average power
 238 for each of the two speakers. The first model, predicting power in the 1–9 Hz range,
 239 showed a significant effect of Speaker ($b = 1.265$, $F(1, 189) = 90.91$, $p < 0.001$,
 240 $adjustedR^2 = 0.321$), indicating that Clinton’s speech contained more power in the
 241 lower frequencies compared to Trump’s speech. The other model, predicting power
 242 in the 9–15 Hz range, also showed a significant difference between the two speakers,
 243 only with a much smaller effect size ($b = 0.164$, $F(1, 189) = 42.75$, $p < 0.001$,
 244 $adjustedR^2 = 0.180$). These findings reveal that, in the first presidential debate,
 245 Clinton’s speech contained more power in the 1–9 Hz range, and also slightly more
 246 power in the frequency range above 9 Hz.

247 10.2.2.2 Second Debate

248 The average modulation spectra of all speech produced by the two speakers in the
 249 second debate are given in Fig. 10.2.

250 Again, simple linear models were built separately for each of the two frequency
 251 bands (1–9 Hz and 9–15 Hz). The first model, predicting power in the 1–9 Hz range,
 252 showed a significant effect of Speaker ($b = 2.322$, $F(1, 157) = 434.5$, $p < 0.001$,
 253 $adjustedR^2 = 0.733$), as did the second model, predicting power in the 9–15 Hz
 254 range, only with a considerably smaller effect size ($b = 0.263$, $F(1, 157) = 250.9$,
 255 $p < 0.001$, $adjustedR^2 = 0.613$). These findings reveal that, in the second presi-
 256 dential debate, Clinton’s speech contained considerably more power in the 1–9 Hz
 257 range, and also somewhat more power in the frequency range above 9 Hz.

258 Note that, similar to the first debate, there is a clear peak in the modulation
 259 spectrum of Clinton around 3 Hz. This peak indicates a pronounced syllabic rhythm
 260 around 3 Hz in the amplitude envelope of Clinton’s speech (cf. Fig. 10.1).

261 10.2.2.3 Third Debate

262 The average modulation spectra of the speech produced by both speakers in the third
 263 debate are given in Fig. 10.2.

264 Once more, simple linear models were built separately for each of the two fre-
 265 quency bands (1–9 Hz and 9–15 Hz). The first model, predicting power in the 1–
 266 9 Hz range, showed a significant effect of Speaker ($b = 2.427$, $F(1, 167) = 207.5$,
 267 $p < 0.001$, $adjustedR^2 = 0.551$), as did the second model, predicting power in the
 268 9–15 Hz range, only with a considerably smaller effect size ($b = 0.350$, $F(1, 167) =$
 269 197.6 , $p < 0.001$, $adjustedR^2 = 0.539$). These findings from the third debate mirror
 270 those from the second debate: Clinton’s speech contained considerably more power
 271 in the 1–9 Hz range, and also slightly more power in the frequency range above 9 Hz.

10.3 Perception Experiment

10.3.1 Participants

Native Dutch participants ($N = 20$; 17 females, 3 males; $M_{age} = 25$) with normal hearing were recruited from the Max Planck Institute's participant pool. Participants in all experiments reported here gave informed consent as approved by the Ethics Committee of the Social Sciences department of Radboud University (project code: ECSW2014-1003-196).

10.3.2 Material

Only speech fragments from the third debate were included in the perception experiment because (1) it was impossible to include the speech from all debates in a single rating experiment for reasons of length and (2) the third debate showed the largest difference between the two talkers in the power of amplitude modulations in the 1–9 Hz range.

Speech fragments from the third debate were first scaled to 70 dB using Praat (Boersma & Boersma, 2016). We did not want raters to base their judgments on the linguistic content of the speech since this was not controlled across the two speakers. Therefore, all speech was low-pass filtered (450 Hz cutoff, using a Hann window with a roll-off width of 25 Hz as implemented in Praat) to avoid lexical-semantic interference, while preserving sufficient ecological validity (being like naturally filtered speech, as if overhearing a person in another room). This manipulation crucially leaves the amplitude fluctuations present in the original speech signals relatively intact (cf. Fig. 10.1). After low-pass filtering, the speech was scaled to 70 dB.

10.3.3 Procedure

Participants in the experiment listened to the low-pass filtered speech fragments from either Clinton or Trump (counter-balanced across participants) in random order. Participants were instructed to rate the items for charisma, basing their judgments on the sound of the speech. They were explicitly pointed to the speaker's identity (but remained unaware that ratings of the other speaker were also collected). Nevertheless, they were told not to let any potential political or personal preferences influence their ratings. The use of a between-participants design reduced the contrast between the two speakers, thus further minimizing potential biases due to speaker sex, pitch, political stance, etc. Participants were instructed to rate the items for charisma using an Equal Appearing Interval Scale (Thurstone, 1928), including seven stars with labeled extremes (not charismatic on the left; very charismatic on the right).

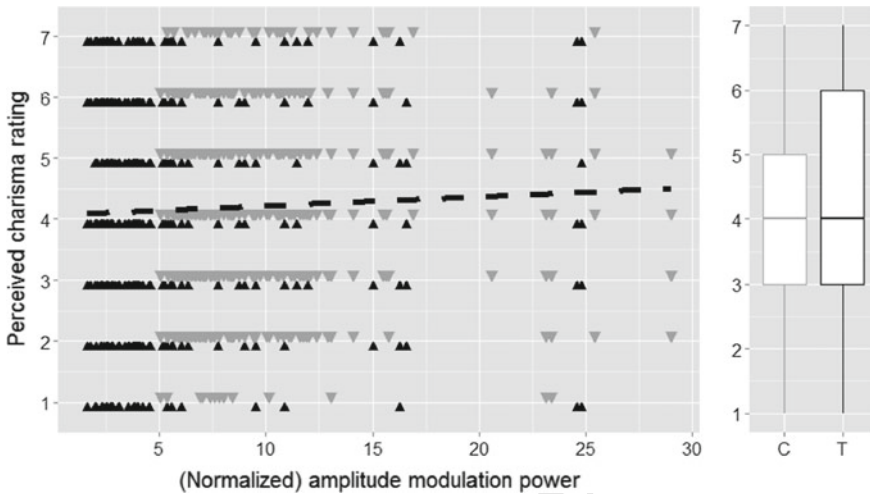


Fig. 10.3 *Left panel:* Individual perceived charisma ratings (on a scale from 1 “not charismatic” to 7 “very charismatic”) of each speech fragment as a function of the (normalized) average power of amplitude modulations in the 1–9 Hz range. Gray triangles indicate speech fragments from Clinton and black triangles those from Trump. The black dashed line shows a (simple) linear regression line across all data points. *Right panel:* Boxplots showing the charisma ratings split for the two speakers (C = Clinton; T = Trump)

10.3.4 Results

The average perceived charisma rating of the speech of Clinton was 4.1, while Trump received an average rating of 4.3. Speech fragments with outlier values for the average power of amplitude modulations in the 1–9 Hz range (i.e., $> 2 * SD$; $n = 8$) were excluded to avoid the heavy weight of these outliers on the correlation analyses reported below. Figure 10.3 shows the individual perceived charisma ratings of speech fragments as a function of the average power of amplitude modulations in the 1–9 Hz range.

The right panel of Fig. 10.3 suggests that, on average, Trump (black) received higher charisma ratings than Clinton (gray). The left panel suggests that the charisma ratings seem to be a function of the average power of amplitude modulations in the 1–9 Hz range, with greater power of the amplitude modulations leading to higher charisma ratings.

Perceived charisma ratings were entered into a simple linear model, including the predictor’s Speaker (categorical predictor; deviation coding, with Trump coded as -0.5 and Clinton as $+0.5$), Modulation Power Below 9 Hz (continuous predictor; z-scored), Modulation Power Above 9 Hz (continuous predictor; z-scored), and interactions between Speaker and the two Modulation Power predictors. This model, first, revealed a significant effect of Modulation Power Below 9 Hz ($b = 0.318$, $F(5, 1664) = 2.245$, $p = 0.041$). This indicates that, across the two talkers, speech

326 with greater power in the 1–9 Hz range led to higher charisma ratings. Second, we
327 found a main effect of Speaker ($b = -0.209$, $F(5, 1664) = 2.245$, $p = 0.014$), sug-
328 gesting that Trump’s speech was rated as more charismatic overall than Clinton’s
329 speech. No effect of Modulation Power Above 9 Hz was observed ($p = 0.151$), nor
330 was their statistical evidence for either interaction term.

331 10.3.5 General Discussion

332 The present research goal was to investigate the role of temporal amplitude modula-
333 tions in charisma perception in political debates. An acoustic analysis of the speech
334 from two presidential candidates, Hillary Clinton and Donald Trump, in three dif-
335 ferent debates was carried out by means of modulation spectra, revealing the spec-
336 tral content of the amplitude envelopes. Also, a perception experiment investigated
337 whether judgments of perceived charisma would be sensitive to the speech rhythm
338 in the acoustic signal.

339 Comparison of the amplitude spectra of Hillary Clinton’s and Donald Trump’s
340 speech revealed considerably greater power in the modulation spectra of Clinton’s
341 speech than in those of Trump’s speech. This power difference cannot be due to
342 overall intensity differences between the two speakers since all speech was normal-
343 ized in overall power prior to analysis, matching the overall intensity of Clinton’s
344 and Trump’s speech fragments. Also, the power difference cannot be attributed to
345 differences in habitual speech rate since such differences would be expected to lead
346 to peaks at different frequencies in the modulation spectra, rather than differences
347 in overall power. Instead, this finding indicates that there was a more pronounced
348 temporal envelope in Clinton’s speech (compared to Trump’s speech).

349 Note that this power difference was concentrated (i.e., largest) in the 1–9 Hz range,
350 the range of typical syllable rates (Ding et al., 2017; Ghitz & Greenberg, 2009;
351 Greenberg & Arai, 1999, 2004). This suggests that the power difference between
352 Clinton and Trump is driven by more pronounced syllabic amplitude fluctuations
353 in the speech of Clinton. Moreover, across the three debates, there seems to be a
354 relatively consistent peak around 3 Hz in Clinton’s modulation spectra, suggesting
355 a preferred syllabic rate. In contrast, Trump’s modulation spectra lack pronounced
356 peaks, indicating particularly flat, that is, unmodulated amplitude envelope contours.

357 Whether or not Clinton used this particular speaking style (with regular ampli-
358 tude modulations) purposefully and strategically remains unknown. In this regard,
359 one may note that speakers, in general, tend to produce greater amplitude modulations
360 when instructed to produce clear speech (Krause & Braid, 2004) or when talking
361 in noise (Bosker & Cooke, 2018), presumably for reasons of achieving greater intel-
362 ligibility. As such, Clinton’s speaking style during the three debates examined here
363 may be the result of her extensive experience with making herself understood during
364 public addresses. We may speculate that the influence of the enhanced modulation
365 signature of Clinton’s speech did not influence charisma perception alone. Regular
366 energy fluctuations have been shown to benefit speech recognition (Doelling et al.,

367 2014; Ghitza, 2012, 2014), particularly in noisy listening conditions (Aikawa and
368 Ishizuka, 2002), and, as such, may have improved Clinton's intelligibility in the noisy
369 environment of a live debate. This seems particularly relevant considering the large
370 number of interruptions (i.e., overlapping speech) that Clinton encountered during
371 the three debates (Trump: $N = 106$ vs. Clinton : $N = 27$). Also, rhythmic ampli-
372 tude modulations facilitate recognition memory (Essens & Povel 1985), potentially
373 serving Clinton's political aims at the time.

374 One may also speculate about the absence of amplitude modulations in Trump's
375 speech. Tian's recent analysis (Tian, 2017) of Trump's disfluency patterns during
376 these presidential debates indicated that Trump was considerably more disfluent
377 than Clinton. Trump was found to use particularly many repetitions, repairs, and
378 abandoned utterances (Tian, 2017); all types of disfluencies that signal less extensive
379 utterance planning and self-monitoring. As such, Tian suggested that Trump used
380 less rehearsed utterances compared to Clinton. This difference in utterance planning
381 can well be thought to underlie the difference in rhythmic structure between the two
382 speakers: putting more effort in cognitive planning would also allow the speaker to
383 better temporally organize the syllabic structure of the utterance, and especially so
384 with increased public-speaking experience.

385 The outcomes of the perception experiment supported two conclusions. First,
386 more pronounced amplitude modulations biased raters toward higher perceived
387 charisma ratings. Across all speech fragments from both talkers, we observed that
388 those items with a higher power of amplitude modulations in the 1–9 Hz range also
389 received higher perceived charisma ratings—independent from the main speaker
390 effect. This suggests that the rhythm of speech contributes to perceived charisma,
391 with implications for public speakers in general.

392 The second conclusion is that Trump's speech was, on the whole, rated as more
393 charismatic than Clinton's. Although this may seem at odds with the observation
394 that less pronounced amplitude modulations result in lower perceived charisma rat-
395 ings, it is important to realize that listeners could base their judgments on a larger
396 set of acoustic characteristics than just rhythm. It is unlikely that participants in the
397 study based their perceived charisma ratings solely on the amplitude modulation
398 signatures of the speech signals. Many other (acoustic) characteristics are likely to
399 have contributed to participants' judgments—even in the case of low-pass filtered
400 speech (i.e., without access to linguistic content). One potential acoustic cue that
401 was available to listeners and that may account for the main effect of Speaker is
402 pitch. The low-pass filter applied to the speech only filtered out spectral informa-
403 tion above 450 Hz, leaving fundamental frequencies relatively intact. As such, the
404 low-pass filtered stimuli still contained acoustic cues to talker gender (distinction
405 male vs. female cued by pitch). Indeed, talker gender is known to bias charisma
406 ratings (and the perception of other personality traits), with male talkers generally
407 being perceived as more charismatic than female talkers (Brooks, Huang, Kearney,
408 & Murray, 2014; Niebuhr, Skarnitzl, & Tylecková, 2018; Novák-Tát, 2017). There-
409 fore, the main effect of Speaker is likely driven by a range of acoustic and social
410 factors that were not controlled for. Still, it is important to note that the correlation
411 between more pronounced amplitude modulations and higher perceived charisma

412 ratings held across talkers (no interaction between modulation power and speaker).
413 This means that, despite an overall difference between the male and female voice,
414 enhanced amplitude modulations in speech equally affected the ratings of Trump's
415 and Clinton's speech.

416 Another possible explanation for the overall effect of Speaker could be related to
417 the concept of 'effectiveness windows' in charisma perception (Niebuhr, Tegtmeier,
418 & Brem, 2017). It has been proposed that public speakers, in attempting to per-
419 suade their audiences, should use charisma-relevant acoustic cues within particular
420 functional ranges, avoiding, for instance, exaggerated vocal characteristics. Maybe
421 Clinton's consistent use of regular amplitude modulations was perceived as an "over-
422 dose" of charismatic vocal cues, thus at some point hurting, rather than serving, the
423 subjective impression listeners had of her. However, such an interpretation would
424 also predict an inverse U-curve in the relationship between modulation power and
425 charisma perception, such that greater rhythmicity would be beneficial only up to
426 a certain point. However, follow-up statistical analyses (i.e., testing for a quadratic
427 effect of Modulation Power Below 9 Hz) and visual inspection of Fig. 10.3 do not
428 support the presence of such a U-shaped relationship, arguing against this particular
429 explanation.

430 The fact that we used low-pass filtered speech may be seen as both a strength as
431 well as a limitation of the current study. It is a strength of the methodology of the
432 experiment because this allowed us to isolate the (temporal) acoustics of the speech
433 from the linguistic content. In this fashion, potential interference from the linguistic
434 message was reduced. At the same time, one may argue that it limits the generaliz-
435 ability of the present findings since in most natural communicative situations we hear
436 unfiltered speech. For our current purposes, we valued experimental control higher
437 than ecological validity and future studies may investigate whether the rhythm of
438 speech also influences charisma perception in more natural settings.

439 Another limitation of this study is that we only performed correlational analyses.
440 Even though we are unaware of possible confounds, we acknowledge that the present
441 empirical evidence does not necessarily warrant the conclusion that more pronounced
442 amplitude modulations causally influence perceived charisma. Future investigations
443 may, for instance, examine this causal relationship by directly manipulating the
444 modulation depth of speech fragments—while keeping all other (acoustic, linguistic,
445 social) cues present in the signal constant.

446 Finally, one further highly relevant issue in the field of charisma research is the
447 role of listener variation in charisma perception. Most empirical studies of charisma
448 perception have used subjective ratings collected from young university students. In
449 fact, some studies, like the present one, recruited non-native speakers of the language
450 under study (e.g., Brem & Niebuhr, this volume). It remains unclear how variation
451 among raters might impact charisma perception and the perceptual weight assigned to
452 various vocal characteristics. Is charisma perception language- or culture-dependent
453 (cf. D'Errico, 2013)? Do non-native speakers of a language weight the acoustic cues
454 to charisma differently from native speakers, possibly through influences from their
455 L1? Do male and female raters differ in how they judge male versus female public
456 speakers (cf. Brem & Niebuhr, this volume)? What is the role of one's own speech

457 production patterns on the perception of others (cf. Bosker, 2017b)? For instance, do
 458 fast talkers find fast speech more attractive or persuasive than others? These questions
 459 regarding inter-individual variation in charisma perception are promising avenues for
 460 future research.

461 10.4 Conclusion

462 The present outcomes shed light on the use and function of speech rhythm in polit-
 463 ical debates, specifically comparing the speech produced by Hillary Clinton and
 464 Donald Trump in three presidential debates in 2016. Clinton’s speech was observed
 465 to contain more power in the modulation spectra, particularly in the 1–9 Hz range,
 466 suggesting more pronounced amplitude modulations in her speech (compared to
 467 Trump). This may be argued to indicate that Clinton planned her utterances more
 468 extensively, allowing more opportunity to temporally organize the syllabic structure
 469 of her utterances. At the same time, the lack of rhythmic amplitude modulations in
 470 Trump’s speech may indicate a level of spontaneity in his speech production, with
 471 little attempt to pre-plan certain utterances.

472 Perceptual data revealed a positive correlation between the strength of amplitude
 473 modulations in the syllabic range (1–9 Hz), on the one hand, and perceived charisma
 474 ratings, on the other hand. This suggests that greater rhythm in the speech of a public
 475 speaker positively influences listeners’ impressions of the speaker charisma. Thus,
 476 it highlights the important contribution of speech rhythm to charisma perception.

477 **Acknowledgments** The author was supported by a Gravitation grant from the Dutch Government
 478 to the Language in Interaction Consortium. Parts of the acoustic analysis have been presented at
 479 Interspeech 2017, Stockholm, Sweden. Thanks go to YouTube and the various news agencies for
 480 making the digital recordings of the presidential debates freely available. Thanks also to Annelies
 481 van Wijngaarden for coordinating the perception experiment, to Joe Rodd for help with visualizing
 482 the data, and to the student-assistants in the Psychology of Language Department of the Max Planck
 483 Institute for Psycholinguistics for help with annotating the speech recordings.

484 References

- 485 ABC News. (2016). *FULL VIDEO: Donald Trump vs Hillary Clinton—2nd Presidential Debate*.
 486 Retrieved October 9, 2016, https://www.youtube.com/watch?v=h-gkBUbU_F4.
- 487 ABC News (2016). *FULL VIDEO: Donald Trump vs Hillary Clinton—3rd Presidential Debate*.
 488 Retrieved October 9, 2016, from <https://www.youtube.com/watch?v=LsA6Gj8y8rU>.
- 489 Aikawa, K., & Ishizuka, K. (2002). Noise-robust speech recognition using a new spectral estimation
 490 method “PHASOR”. In *Proceedings of Acoustics, Speech, and Signal Processing (ICASSP)* (pp.
 491 397–400).
- 492 Awamleh, R., & Gardner, W. L. (1999). Perceptions of leader charisma and effectiveness: The
 493 effects of vision content, delivery, and organizational performance. *The Leadership Quarterly*,
 494 10(3), 345–373.
- 495 Boersma, P., & Weenink, D. (2016). Praat: Doing phonetics by computer. Computer program.

- 496 Bosker, H. R. & Cooke, M. (2018). Talkers produce more pronounced amplitude modulations when
 497 speaking in noise. *Journal of the Acoustical Society of America*, 143(2), EL121-EL126.
- 498 Bosker, H. R. (2017a). Accounting for rate-dependent category boundary shifts in speech perception.
 499 *Perception & Psychophysics*, 79(1), 333–343.
- 500 Bosker, H. R. (2017b). How our own speech rate influences our perception of others. *Journal of*
 501 *Experimental Psychology: Learning, Memory, and Cognition*, 43(8), 1225–1238.
- 502 Bosker, H. R., & Ghitza, O. (2018). Entrained theta oscillations guide perception of subsequent
 503 speech: Behavioural evidence from rate normalisation. *Language, Cognition and Neuroscience*,
 504 33(8), 955–967.
- 505 Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and
 506 fine-grained acoustic-phonetic talker characteristics. *Speech Communication*, 20(3–4), 255–272.
- 507 Brooks, A. W., Huang, L., Kearney, S. W., & Murray, F. E. (2014). Investors prefer entrepreneurial
 508 ventures pitched by attractive men. *Proceedings of the National Academy of Sciences*, 111(12),
 509 4427–4431.
- 510 Cutler, A. (1994). Segmentation problems, rhythmic solutions. *Lingua*, 92, 81–104.
- 511 Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture
 512 misperception. *Journal of Memory and Language*, 31(2), 218–236.
- 513 Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access.
 514 *Journal of Experimental Psychology: Human Perception and Performance*, 14(1), 113–121.
- 515 Delgutte, B., Hammond, B., & Cariani, P. (1998). Neural coding of the temporal envelope of speech:
 516 relation to modulation transfer functions. Psychophysical and physiological advances in hearing,
 517 595–603.
- 518 Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for ΔC . *Language and language-*
 519 *processing* (pp. 231–241). Frankfurt a. M.: Peter Lang.
- 520 D'Errico, F., Signorello, R., Demolin, D., & Poggi, I. (2013). The perception of charisma from
 521 voice: A cross-cultural study. In *Proceedings of Affective Computing and Intelligent Interaction*
 522 *(ACII)* (552–557).
- 523 Ding, N., Patel, A., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in
 524 speech and music. *Neuroscience and Biobehavioral Reviews*, 14(1), 113–121.
- 525 Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta-theta
 526 oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, 85,
 527 761–768.
- 528 Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of reducing slow temporal modulations on
 529 speech recognition. *The Journal of the Acoustical Society of America*, 95(5), 2670–2680.
- 530 Essens, P. J., & Povel, D.-J. (1985). Metrical and nonmetrical representations of temporal patterns.
 531 *Perception & Psychophysics*, 37(1), 1–7. <https://doi.org/10.3758/bf03207132>
- 532 Ghitza, O. (2011). Linking speech perception and neurophysiology: Speech decoding guided by
 533 cascaded oscillators locked to the input rhythm. *Frontiers in Psychology*, 2, 130.
- 534 Ghitza, O. (2012). On the role of theta-driven syllabic parsing in decoding speech: Intelligibility of
 535 speech with a manipulated modulation spectrum. *Frontiers in Psychology*, 3, 238.
- 536 Ghitza, O. (2014). Behavioral evidence for the role of cortical Θ oscillations in determining auditory
 537 channel capacity for speech. *Frontiers in Psychology*, 5, 652.
- 538 Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception:
 539 Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Pho-*
 540 *netica*, 66(1–2), 113–126.
- 541 Giraud, A.-L., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging com-
 542 putational principles and operations. *Nature Neuroscience*, 15(4), 511–517.
- 543 Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis.
 544 *Laboratory Phonology*, 7, 515–546.
- 545 Greenberg, S., & Arai, T. (2004). What are the essential cues for understanding spoken language?
 546 *IEICE Transactions on Information and Systems*, E87-D(5), 1059–1070.
- 547 Greenberg, S., & Arai, T. (1999). Speaking in shorthand—A syllable-centric perspective for under-
 548 standing pronunciation variation. *Speech Communication*, 29(2), 159–176.

- 549 Houtgast, T., & Steeneken, H. J. (1973). Modulation transfer-function in room acoustics as a predictor of speech intelligibility. *Acustica*, 28(1), 66–73.
- 550
- 551 Jacewicz, E., Fox, R. A., & Wei, L. (2010). Between-speaker and within-speaker variation in speech tempo of American English. *The Journal of the Acoustical Society of America*, 128(2), 839–850.
- 552
- 553 Kohler, K. J. (2009). Rhythm in speech and language. *Phonetica*, 66(1–2), 29–45.
- 554 Krause, J. C., & Braid, L. D. (2004). Acoustic properties of naturally produced clear speech at normal speaking rates. *The Journal of the Acoustical Society of America*, 115(1), 362–378.
- 555
- 556 Loukina, A., Kochanski, G., Rosner, B., Keane, E., & Shih, C. (2011). Rhythm measures and dimensions of durational variation in speech. *The Journal of the Acoustical Society of America*, 129(15), 3258–3270.
- 558
- 559 NBC News. (2016). *FULL VIDEO: The First Presidential Debate: Hillary Clinton and Donald Trump (Full Debate)*. Retrieved September 28, 2016, from <https://www.youtube.com/watch?v=855Am6ovK7s>.
- 560
- 561
- 562 Niebuhr, O., Skarnitzl, R., & Tylecková, L. (2018). The acoustic fingerprint of a charismatic voice - Initial evidence from correlations between long-term spectral features and listener ratings. In *Proceedings of Speech Prosody* (pp. 359–363).
- 563
- 564 Niebuhr, O., Voße, J., & Brem, A. (2016). What makes a charismatic speaker? A computer-based acoustic-prosodic analysis of Steve Jobs tone of voice. *Computers in Human Behavior*, 64, 366–382.
- 565
- 566
- 567
- 568 Niebuhr, O., Tegtmeier, S., & Brem, A. (2017). Advancing research and practice in entrepreneurship through speech analysis—from descriptive rhetorical terms to phonetically informed acoustic charisma metrics. *Journal of Speech Sciences*, 6(3), 3–26.
- 569
- 570 Nolan, F., & Jeon, H.-S. (2014). Speech rhythm: A metaphor? *Philosophical Transactions of the Royal Society B-Biological Sciences*, 369(1658).
- 571
- 572
- 573 Novák-Tót, E., Niebuhr, O., & Chen, A. (2017). A gender bias in the acoustic-melodic features of charismatic speech? In *Proceedings of Interspeech* (pp. 2248–2252).
- 574
- 575 Obermeier, C., Menninghaus, W., von Koppenfels, M., Raettig, T., Schmidt-Kassow, M., Otterbein, S., & Kotz, S. A. (2013). Aesthetic and emotional effects of meter and rhyme in poetry. *Frontiers in Psychology*, 4(10).
- 576
- 577
- 578 Obermeier, C., Kotz, S. A., Jessen, S., Raettig, T., von Koppenfels, M., & Menninghaus, W. (2016). Aesthetic appreciation of poetry correlates with ease of processing in event-related potentials. *Cognitive, Affective, & Behavioral Neuroscience*, 16(2), 362–373.
- 579
- 580
- 581 Peelle, J. E., & Davis, M. H. (2012). Neural oscillations carry speech rhythm through to comprehension. *Frontiers in Psychology*, 3(10).
- 582
- 583 Pellegrino, F., Coupé, C., & Marsico, E. (2011). Across-language perspective on speech information rate. *Language*, 87(3), 539–558.
- 584
- 585 Peter, J., & Povel, D. -J. (1985). Metrical and nonmetrical representations of temporal patterns. *Perception & Psychophysics*, 37(1), 1–7.
- 586
- 587 Poeppel, D. (2003). The analysis of speech in different temporal integration windows: Cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication*, 41(1), 245–255.
- 588
- 589 Quené, H., & Port, R. (2005). Effects of timing regularity and metrical expectancy on spoken-word perception. *Phonetica*, 62(1), 1–13.
- 590
- 591 Quené, H. (2008). Multilevel modeling of between-speaker and within-speaker variation in spontaneous speech tempo. *The Journal of the Acoustical Society of America*, 132(2), 1104–1113.
- 592
- 593 R Development Core Team. (2012). R: A Language and Environment for Statistical Computing. Computer program.
- 594
- 595 Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73, 265–292.
- 596
- 597 Roncaglia-Denissen, M. P., Schmidt-Kassow, M., & Kotz, S. A. (2013). Speech rhythm facilitates syntactic ambiguity resolution: ERP evidence. *PloS One*, 8(2), e56000.
- 598
- 599 Rosen, S. (1992). Temporal information in speech—acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London Series B-Biological Sciences*, 336(1278), 367–373.
- 600
- 601

- 602 Rosenberg, A., & Hirschberg, J. (2009). Charisma perception from text and speech. *Speech Com-*
603 *munication*, 51(7), 640–655.
- 604 Rothermich, K., Schmidt-Kassow, M., & Kotz, S. A. (2012). Rhythm’s gonna get you: Regular
605 meter facilitates semantic sentence processing. *Neuropsychologia*, 50(2), 232–244.
- 606 Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition
607 with primarily temporal cues. *Science*, 270(5234), 303.
- 608 Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology*, 33, 529–554.
- 609 Tian, Y. (2017). Disfluencies in Trump and Clinton first presidential debate. In *Proceedings of the*
610 *conference Fluency and Disfluency Across Languages and Language Varieties* (pp. 106–109).
- 611 Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J., & Lorenzi, C. (2017). A cross-linguistic
612 study of speech modulation spectra. *The Journal of the Acoustical Society of America*, 142(4),
613 1976–1989.
- 614 Verhoeven, J., De Pauw, G., & Kloots, H. (2004). Speech rate in a pluricentric language: A com-
615 parison between Dutch in Belgium and the Netherlands. *Language and Speech*, 47(3), 279–308.
- 616 Weiss, B., & Burkhardt, F. (2010). Voice attributes affecting likability perception. In *Proceedings*
617 *of Interspeech* (pp. 2014–2017).
- 618 White, L., & Mattys, S. L. (2007). Calibrating rhythm: First language and second language studies.
619 *Journal of Phonetics*, 35(4), 501–522.

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Voice Attractiveness
------------	----------------------

Series Title	
--------------	--

Chapter Title	Dress to Impress? On the Interaction of Attire with Prosody and Gender in the Perception of Speaker Charisma
Copyright Year	2020
Copyright HolderName	Springer Nature Singapore Pte Ltd.

Author	Family Name	Brem
	Particle	
	Given Name	Alexander
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Innovation and Technology Management, Friedrich-Alexander-Universität Erlangen-Nürnberg
	Address	Erlangen, Germany
	Email	alexander.brem@fau.de

Corresponding Author	Family Name	Niebuhr
	Particle	
	Given Name	Oliver
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Mads Clausen Institute, Centre for Electrical Engineering, University of Southern
	Address	Odense, Denmark
	Email	olni@sdu.dk

Abstract	<p>Understanding charismatic speech becomes a highly relevant issue in times of globalized markets and mobile on-demand mass media that strengthen the influence of individuals. Pushing phonetic research further into the realm of non-lexical charisma triggers, the present study is the first to investigate the combined effects of variation in attire and prosody on the perception of male and female speaker charisma. A perception experiment was carried out with Attire and Prosody as independent variables, each with two manipulation steps and embedded in a 2×2 orthogonal design. A total of 53 participants took part in the experiment and rated eight senior business leaders of well-known US American companies, four males and four females, on three approved charisma-related scales: convincing, passionate, charming. The audio-visual stimuli consisted of a keynote-speech excerpt of a speaker in combination with a matching photograph. Results clearly show that both Attire and Prosody had significant effects on the speakers' perceived charisma. The charisma effects of Attire and Prosody are additive, but in gender-specific ways and with gender-specific effect sizes. A bipartite results pattern among the female speakers further suggests that it depends on their physical attractiveness whether Attire and Prosody conditions have a charisma-supporting or charisma-reducing effect. The results are discussed in terms of their practical implications for the daily business life of men and women.</p>
----------	---

Keywords

Charisma - Passion - Charm - Persuasion - Attire - Public speaking - prosody - English speech - Perception
- Expressive - Speech

Chapter 11

Dress to Impress? On the Interaction of Attire with Prosody and Gender in the Perception of Speaker Charisma



Alexander Brem and Oliver Niebuhr

Abstract Understanding charismatic speech becomes a highly relevant issue in times of globalized markets and mobile on-demand mass media that strengthen the influence of individuals. Pushing phonetic research further into the realm of non-lexical charisma triggers, the present study is the first to investigate the combined effects of variation in attire and prosody on the perception of male and female speaker charisma. A perception experiment was carried out with Attire and Prosody as independent variables, each with two manipulation steps and embedded in a 2×2 orthogonal design. A total of 53 participants took part in the experiment and rated eight senior business leaders of well-known US American companies, four males and four females, on three approved charisma-related scales: convincing, passionate, charming. The audio-visual stimuli consisted of a keynote-speech excerpt of a speaker in combination with a matching photograph. Results clearly show that both Attire and Prosody had significant effects on the speakers' perceived charisma. The charisma effects of Attire and Prosody are additive, but in gender-specific ways and with gender-specific effect sizes. A bipartite results pattern among the female speakers further suggests that it depends on their physical attractiveness whether Attire and Prosody conditions have a charisma-supporting or charisma-reducing effect. The results are discussed in terms of their practical implications for the daily business life of men and women.

Keywords Charisma · Passion · Charm · Persuasion · Attire · Public speaking · prosody · English speech · Perception · Expressive · Speech

A. Brem
Innovation and Technology Management, Friedrich-Alexander-Universität Erlangen-Nürnberg,
Erlangen, Germany
e-mail: alexander.brem@fau.de

O. Niebuhr (✉)
Mads Clausen Institute, Centre for Electrical Engineering, University of Southern, Odense,
Denmark
e-mail: olni@sdu.dk

© Springer Nature Singapore Pte Ltd. 2020

B. Weiss et al. (eds.), *Voice Attractiveness*, Prosody, Phonology and Phonetics,
https://doi.org/10.1007/978-981-15-6627-1_11

189

11.1 Introduction

11.1.1 Charisma and Delivery

We live in times in which individual politicians are increasingly able to determine people's opinion and voting behavior (whether for better or for worse), in which managers become an integral part of a company's brand image (like Steve Jobs was for Apple and Elon Musk is for Tesla), and in which entrepreneurship, i.e., the motivation, passion, and persuasive power of individuals, becomes a mainstay of national prosperity in international competition networks. In these times, it is of high societal and economical importance to understand in detail what good speakers actually do and how, in which way, and to what degree they influence listeners. Good speakers draw us under their spell. We cannot help but listen to them, we believe in what they tell us, and we are willing to adopt their opinions, attitudes, and/or agendas. Attracting attention as well as gaining and persuading followers without having to use force or referring to formal authority is the essence of charisma. In the more speaker- than listener-oriented words of Antonakis, Fenley & Liechti (2016: 304), charisma is defined as "values-based, symbolic, and emotion-laden leader signaling".

Charisma leads to more successful brainstorming outputs and salary negotiations (Pentland, 2008), results in better learning outcomes of students and, generally, in more satisfied subordinates (Towler, 2003; Lee, 2014), helps raise more start-up funding (Davis, Hmieleski, Webb, & Coombs, 2017), changes people's opinions and decisions (Brilman, 2015), and makes a product or service appear more credible and likable to customers (Gélinas-Chebat, Chebat, & Vaninsky, 1996). Previous studies also demonstrated that charisma is not a mysterious talent of a few gifted people, as was originally claimed by Weber (1947), but a tangible skill that anyone can learn and improve (Antonakis et al., 2011, 2012).

However, this learning and improving requires that we understand how the mechanisms work that makes a speaker sound charismatic, in particular the mechanisms of the so-called "delivery" that consists of everything a speaker conveys beyond the words themselves. Delivery includes auditory components like the speaker's speech prosody as well as visual components like body language and attire, and cross-modal components like age and gender.¹ Results of experimental studies repeatedly suggested that these components of delivery are—alone or in combination—more important than words for a speaker's charismatic impact (Holladay & Coombs, 1994; Awamleh & Gardner, 1999; Chen et al., 2014; Brilman, 2015).

¹Note that "gender" is very often used also to refer to the biological concept of "sex", even in the scientific literature and across disciplines (cf. Brooks, Huang, Kearney, & Murray, 2014). Therefore, in order to be easily understandable for a broad, interdisciplinary readership, we decided to use "gender" in the sense of "sex" in the present paper.

11.1.2 *The Roles of Prosody and Attire*

As the transdisciplinary science “whose goal is the description, modeling and explanation of speech communication in the languages of the world” (Kohler, 2000: 1) and whose areas of instrumental-experimental research range from physiology through acoustics to cognition and perception, phonetics is perhaps in the best position of all scientific disciplines to decipher, objectively quantify, and ultimately understand how and by means of which signal cues charisma is created in the perceiver’s brain. In fact, the intensive exploration of charismatic speech in phonetic production and perception experiments has already greatly expanded our knowledge of the acoustic-phonetic indicators of perceived speaker charisma. We know today that acoustic parameters such as the level, range, and dynamics of pitch² and intensity patterns, the durations of pauses and utterances, the number of emphatically emphasized words, the vocal tract’s resonance frequencies (lower levels of the first three formants), and the timbre of the voice (e.g., in terms of HNR or the Hammarberg index) are all involved in the signaling of speaker charisma (Touati, 1993; Rosenberg & Hirschberg, 2009; Signorello, D’Errico, Poggi, & Demolin, 2012; Scherer, Layher, Kane, Neumann, & Campbell, 2012; D’Errico, Signorello, Demolin, & Poggi, 2013; Chen et al., 2014; Brillman, 2015; Shim et al., 2015; Hiroyuki & Rathcke, 2016; Bosker, 2007; Niebuhr, Thumm, & Michalsky, 2018a, b). In addition, we know that the general relevance of these parameters for speaker charisma does not differ between politics and business (Niebuhr, Brem, Novák-Tót, & Voße 2016; Novák-Tót, Niebuhr, & Chen 2017); maybe not even across cultures.

However, what does differ is which parameter level is appropriate and how strongly each parameter contributes to making a speaker sound charismatic. Not only culture, situation, industry sector, and listener age are relevant factors in this connection (Biadys, Rosenberg, Carlson, Hirschberg, & Strangert, 2008; Abidi & Gumpert, 2018; Jokisch, Iaroshenko, Maruschke, & Ding, 2018), but also speaker gender. Most prosodic parameters have an identical effect on the charisma of male and female speakers and differ solely in the magnitude of this effect. Two parameters are different, though. These two parameters are pitch level and speaking rate. While men need to raise their pitch levels to sound more charismatic, women need to lower the pitch level (Berger, Niebuhr, & Peters, 2017; Niebuhr et al., 2018b); and while it is beneficial for the charisma effect of male speakers to increase the speaking rate, women must reduce their speaking rate to sound more charismatic (Bachsleitner & Popp, 2018). The gender-specific effect of speaking rate may be due to the fact that women already sound subjectively faster than men at the same objectively measured speaking rate (Weirich & Simpson, 2014).

For the visual components of charismatic delivery, and attire in particular, there are far less solid empirical findings from controlled experimental studies. For male

²Similarly as for “gender”, we use the term “pitch” here as it is easily understandable to a broad, interdisciplinary (and non-expert) readership. What we actually mean is the acoustic fundamental frequency (F0), from which pitch is derived in the perception of speech signals, see Terhardt (1974) for further information.

95 speakers, things seem pretty straightforward, though. Compared to any form of casual
 96 or smart-casual attire, formal business attire supports the perception of charisma in
 97 terms of charisma-related attributes such as competence, credibility, and assertive-
 98 ness. In addition, the formal business attire of men is quite clearly and narrowly
 99 defined as a dark-colored suit, see Furnham and Petrova (2010), Furnham, Chan,
 100 and Wilson (2014). In contrast, for women things are not that straightforward. On
 101 the one hand, women “have less freedom to wear more comfortable or casual attire”
 102 in the workplace (Franz & Norton, 2001: 88, see also Behling & Williams, 1991;
 103 Furnham et al., 2014). That is, wearing casual attire is less harmful for men than for
 104 women. On the other hand, the appropriate standard for the formal business attire of
 105 women is less clearly defined than for men. While every little detail counts for the
 106 perception of male charisma (up to the pattern of the tie and the garment of the suit,
 107 cf. Howlett, Pine, Orakçioğlu, & Fletcher, 2013), even somewhat salient differences
 108 in female attire like that between a skirt suit and a pantsuit seem to play a lesser role
 109 in the perception of charisma-related attributes of women (Morris, Gorham, Cohen,
 110 & Huffman, 1996), perhaps because female attire is subject to much greater and
 111 faster fashion variation than male attire (Auty & Elliot, 1998; Molloy, 1996).

112 Furthermore, contradicting the traditional dress-code instructions for women in
 113 brochures and guidebooks (cf. Molloy, 1977; McEwan & Agno, 2011; Hoover, 2013),
 114 recent papers advise female leaders to “think color” (Karabell, 2016). More specif-
 115 ically, concerning the color range of a proper female business attire, these papers
 116 recommend wearing “all shades of red” (Karabell, 2016), i.e., all colors from blue-
 117 red to pink, as they are supposed to represent signals of power and charismatic qual-
 118 ities like “confidence and leadership” (Silverberg, 2017). The experimental study of
 119 Radeloff (1990) showed that red can compete with traditional business colors like
 120 (dark) blue and black when it comes to proper female business attire. Molloy’s (1996)
 121 more practical research agrees with Radeloff’s experimental data. Additionally, he
 122 points out that things have changed since the 1980s and that today “using color
 123 correctly can give businesswomen an advantage over men” (p. 157). In this context,
 124 Radeloff (1990) especially highlights the value of red for businesswomen, whereas
 125 for businessmen, the range of wearable colors is typically restricted to black or dark
 126 blue and, beyond that, hardly addressed in the literature; or, as in the case of grey and
 127 earth tones, associated in brochures and guidebooks with very different statements
 128 and recommendations that clearly reflect the lack of a solid empirical basis.

129 So, while everything from red to pink seems to be more effective than dark blue
 130 or black for female speakers’ charisma, this color range certainly is a charisma killer
 131 for male speakers (e.g., the Financial Post³ regards red as one of the three worst
 132 colors for men to wear in the office). In the opposite direction, while male leaders
 133 can at least dare to speak to an audience in jeans, T-shirt or hody, this kind of casual
 134 business attire seems to be an absolute no-go for female leaders.

³<https://business.financialpost.com/business-insider/the-best-and-worst-colours-to-wear-to-the-office>.

135 **11.1.3 Aims and Assumptions**

136 In summary, with respect to the key charisma factor of delivery, intensive phonetic
137 research provided us with a fairly detailed empirical picture of the charisma-relevant
138 parameters of speech prosody and their context-specific phonetic variation. However,
139 this does not apply to the same degree to the visual communication signals of
140 speaker charisma, especially not to the factor attire. Moreover, we still know practi-
141 cally nothing about the interplay of attire and prosody in the perception of speaker
142 charisma, which is interesting not only because both factors make a major contribu-
143 tion to speaker charisma, but also because of the gender-specific differences in each
144 factor.

145 Therefore, our goal is to expand the empirical knowledge of the non-verbal ingre-
146 dients of speaker charisma beyond prosody into the visual components of delivery.
147 Continuing our previous studies (see Niebuhr et al., 2017), the focus of this line of
148 research is not on political leaders but on business leaders. The first step presented
149 here addresses the attire of speakers and their interaction with prosody. We report
150 the results of a perception experiment with special emphasis on the gender-specific
151 aspects of prosody and attire. The factor prosody was represented by a two-step
152 manipulation of speaking rate and pitch level in male and female speech stimuli.
153 The factor of attire was also represented by a two-step variation. However, unlike for
154 prosody, this variation was not carried out analogously for male and female speakers,
155 but took into account the fact that for men it is the style of attire that is most relevant
156 in everyday business life, while for women it is primarily the color of attire.

157 Our study is able to test three basic assumptions. The present experiment:

- 158 (1) replicates the known gender-specific effects of pitch level and speaking rate on
159 perceived speaker charisma;
- 160 (2) finds an additional gender-specific effect of attire on perceived speaker charisma,
161 with male and female speakers being supported by a dark-colored suit or a red
162 attire, respectively;
- 163 (3) finds the gender-specific effects of attire and prosody to be additive in the
164 perception of speaker charisma.

165 **11.2 Method**

166 **11.2.1 Speakers**

167 Instead of using specifically designed and staged laboratory data, we opted for an
168 approach with genuine, ecologically valid field data. This was for two reasons. First,
169 for complex concepts like charisma whose multi-faceted perceptual nature is still
170 too poorly understood to replicate it properly and consistently in the laboratory, the
171 practical relevance of research findings critically relies on the authenticity of the
172 analyzed data. We wanted our results to have as much practical value as possible,

173 and one obvious way to enhance the relevance of our findings *for* practitioners was to
 174 take real data *from* practitioners. Second, both pilot tests and our own experience from
 175 previous studies indicated that subjects in perception experiments take the assessment
 176 of speaker charisma the more seriously (i.e., they make more reflective, elaborate
 177 judgments) the more well-known and influential the speakers are (cf. also Pearce
 178 & Brommel, 1972). This can only be credibly achieved with data of real speakers.
 179 All our speakers (or the companies they represent) can be considered similarly well
 180 known and influential. A high degree of popularity and influence also had the positive
 181 side effect that enough speech and image material of our speakers was available on
 182 the internet.

183 Since we worked with publicly available materials (i.e., field data), we had to
 184 choose our speakers such as to minimize the influence of potentially confounding
 185 between-speaker variables. On this basis, we chose the following eight speakers, four
 186 females (F1–F4) and four males (M1–M4):

187 (F1) Margret Whitman, born August 4, 1956, in Cold Spring Harbor, New York,
 188 USA; CEO and President of Hewlett Packard Enterprise (until January 31, 2018).

189 (F2) Virginia Marie Rometty, born July 29, 1957, in Chicago, Illinois, USA; CEO
 190 and President of IBM.

191 (F3) Sara Blakely, born February 27, 1971, in Clearwater, Florida, USA; Founder
 192 and CEO of Spanx Inc.

193 (F4) Sheryl Kara Sandberg, born August 28, 1969, in Washington D.C., USA;
 194 COO of Facebook Inc.

195 (M1) Reid Hoffman, born August 5, 1967, in Stanford, California, USA; Co-
 196 Founder of LinkedIn, former manager of PayPal.

197 (M2) Satya Nadella, born August 19, 1967, in Hyderabad, India; CEO of
 198 Microsoft.

199 (M3) Sundar Pichai, born 1972, in Madurai, India; CEO of Google LLC.

200 (M4) Mark Zuckerberg, born May 14, 1984, in White Plains, New York, USA;
 201 CEO of Facebook Inc.

202 All selected speakers are leading senior managers (CEOs or COOs) of well-known
 203 US American companies and were either born in the US or came from other English-
 204 speaking countries and then lived in the US for decades. Accordingly, all selected
 205 managers were native speakers of English and fluent speakers of American English,
 206 albeit with different regional and dialectal characteristics. However, these charac-
 207 teristics can be considered irrelevant to the questions of the present study, not least
 208 because—as became apparent from the metadata and participant feedback collected
 209 after the perception experiment—our participants were unable to either detect these
 210 characteristics or to associate them consistently with a specific geographical origin.
 211 Thus, it is unlikely that dialectal or regional stereotypes, their related socio-economic
 212 associations, or similar socio-phonetic effects were able to bias the participants' judg-
 213 ments of speaker charisma in a systematic and consistent way, cf. Ladegaard (1998),
 214 Bayard, Weatherall, Gallois, and Pittam (2001), Bailey (2003), and Andersson (2009)
 215 for the relationships between varieties of English and the judgment of their speakers.

216 All speakers are from the educated upper social class of the USA; and all were
 217 between 30 and 60 years old when they gave those speeches whose excerpts we used
 218 to create our stimuli. In this middle biological age range, we can assume all speakers
 219 to have the same basic physiological prerequisites with regard to the production of
 220 speech prosody (e.g., Schötz, 2006), except for some gender-specific differences,
 221 of course (Xu & Sun, 2002; Pépiot, 2013). Similarly, our speakers' age range was
 222 chosen small enough to prevent any potential age-related charisma differences from
 223 masking the actually investigated main effects of Attire and Prosody. Results of
 224 empirical studies suggest that perceived age has a separate influence on speaker
 225 charisma (e.g., Jokisch et al., 2018). Speaker charisma increases with age, but not
 226 linearly.

227 All speakers are leading IT executives. This restriction was added because initial
 228 results from another study (Abidi & Gumpert, 2018) suggest that speaker charisma
 229 is produced and assessed in an industry-specific fashion. For example, it seems that
 230 different ideas of charismatic presentations exist in the automotive sector as compared
 231 to the IT sector, which, in turn, seems to have different expectations concerning
 232 charismatic speeches than the financial sector. Although these results are still very
 233 preliminary, we nevertheless wanted to control this factor by keeping our dataset
 234 homogenous by focussing on the IT sector. A further advantage of this decision is
 235 that the IT sector is the same sector from which also the participants of the perception
 236 experiment were recruited. This had the advantage that our participants had already
 237 dealt with the eight selected speakers in one way or another, for example, by reading
 238 or writing about them in their course of studies or in related journals or newspapers.
 239 That is, all participants were similarly familiar with the speakers and well aware of
 240 their top positions in market-leading companies (see Rosenberg & Hirschberg, 2009
 241 for the charisma-increasing effect of speaker familiarity and Pearce & Brommel,
 242 1972 for the charisma-increasing effect of a higher social status).

243 ***11.2.2 Image Material and the Independent Variable Attire***

244 The independent variable Attire is represented in the experiment by presenting the
 245 eight male and female speakers on different photographs. Two photographs were
 246 selected for each speaker. One shows the speaker in a more formal or conservative
 247 attire. The other one shows the speaker in a more casual or expressive attire. The
 248 full set of photographs can be made available to interested persons upon individual
 249 request.

250 Like in the selection of speakers (2.1), a maximum of comparability and control
 251 of potentially confounding factors was a major criterion for choosing suitable
 252 photographs. This was true within and across each speaker's pair of photographs.
 253 For example, all selected photographs showed the eight speakers from a similar
 254 angle (frontal view), in a similar posture (standing upright), and against the similar
 255 background of a large exhibition hall. Furthermore, all photographs showed the eight

256 speakers with open and slightly rounded lip positions, which indicated that the photo-
 257 graph was taken while giving a speech. Head postures and hand and arm gestures on
 258 each photograph additionally characterized their speech as animated and passionate.
 259 In addition, we made sure that the two points in time at which the photographs of a
 260 speaker had been taken were less than 24 months apart (so as to prevent differences
 261 in a speaker's visual age across attire conditions, cf. Grd, 2013) and that the two
 262 photographs showed the speaker similarly large, i.e., up to the hips with the legs not
 263 being visible. The latter was important as the size of a person on a photograph (or
 264 screen) influences the subjective spatial distance of this person to the viewer. This
 265 distance, in turn, determines the level of social and emotional connection that the
 266 viewer feels for the person on the photograph (Reeves & Nass, 1996). As this connection
 267 is obviously related to concepts like perceived charisma, we had to control for
 268 this factor in the experiment.

269 Figure 11.1 shows, as an example, pairs of photographs for one female speaker
 270 (Sheryl Sandberg) and for one male speaker (Mark Zuckerberg) similar to those
 271 used in the actual experiment. As can be seen in Fig. 11.1, and as was mentioned
 272 Sect. 11.1.2, the biggest difference between the pairs of photographs was that, in
 273 the case of the male speakers, the independent variable Attire was operationalized



Fig. 11.1 Examples of photographs showing female and male speakers giving a keynote speech in conservative (left) and expressive (right) business attire. Top left photo taken by Pete Souza (2015), top right photo taken by Anthony Quintano (2018), bottom left photo taken by Moritz Hager (2012, photo edited by 1st author), bottom right photo taken by Remy Steinegger (2016). All photo are under CC-BY license

274 as the difference between a dark-colored business suit and a light-blue T-shirt. In
 275 the case of the female speakers, it was operationalized as the difference between a
 276 dark-colored and a red or pink pantsuit. Thus, in the case of the male speakers, attire
 277 concerns the *style* of clothing, whereas, in the case of female speakers, it concerns
 278 the *color* of clothing. For lack of better generic terms that equally apply to both
 279 types of attire variation (formal vs. casual is considered inappropriate), we refer the
 280 attire variation in both gender conditions as *conservative* versus *expressive*. Note
 281 that, based on the literature summarized in the introduction, it is the *conservative*
 282 Attire condition that is assumed to support the charisma perception of *male* speakers,
 283 whereas the *expressive* Attire condition is assumed to make *female* speakers more
 284 charismatic.

285 Mixing up style and color in the Attire variable follows the charisma-related
 286 statements in the literature on gender and attire. However, we also had no other
 287 option. It was no problem to find photographs of the male speakers wearing a T-
 288 shirt, even for similar public-speaking situations as in the business-suit condition.
 289 The same was not possible for the female speakers, though. In fact, for none of our
 290 female speakers, we were able to find one single photograph on which the speaker
 291 does not wear a formal dress or pantsuit. That is, photographs showing our female
 292 executive leaders in a T-shirt, sweater, hoody, jeans, or a similar casual clothing do
 293 not exist on the internet, no matter which occasion or which monologue or dialogue
 294 situation. We think that this fact resonates well with the literature in Sect. 11.1.2,
 295 in that it tells a lot about the different socio-cultural demands on male and female
 296 business attire, and about the leeway that male and female executive leaders have
 297 for choosing their attire in the workplace (or for public speeches as in the present
 298 experiment). Thus, although the two *expressive* Attire conditions of men and women
 299 obviously differ at the surface level (style vs. color), the variable Attire is nevertheless
 300 appropriately and homogeneously implemented in the experiment, because the two
 301 variable levels *conservative* and *expressive* equally cover for both genders the real full
 302 range of possible attire variation in the workplace. Yet, an obvious task of subsequent
 303 studies is, of course, to repeat the present experiment with staged photographs of fake
 304 executive leaders in order to implement the variable Attire in a consistent way across
 305 both genders, i.e., as the difference between business suit and T-shirt.

306 **11.2.3 *Speech Material and the Independent Variable*** 307 ***Prosody***

308 We chose one YouTube clip per speaker from one of his/her major public keynotes
 309 held in front of a large audience. Since the durations of speech stimuli are known
 310 to correlate positively with the perception of speaker charisma (i.e., the longer the
 311 stimulus the higher the speaker charisma, see Biadys et al., 2008; Rosenberg &
 312 Hischberg, 2009), a similarly long speech section of 19–20 s was extracted from
 313 all eight YouTube clips. The onsets and offsets of these speech excerpts coincided

314 in all cases with major intonational phrase boundaries (see AE-ToBI, Beckman,
 315 Hirschberg, & Shattuck-Hufnagel, 2005) at the beginnings and ends of syntactically
 316 complete utterance. Furthermore, all eight speech excerpts were free from disflu-
 317 encies like hesitational lengthening, hesitation particles, overlong silent pauses (for
 318 turn-internal standards, Ten Bosch, Oostdijk, & de Ruiter, 2004), etc. The speech
 319 excerpts also contained no applause, music inserts, and other background noises.

320 Studies by Antonakis et al. (2011, 2012) showed on an experimental basis that, in
 321 addition to prosody, traditional morphosyntactic and lexical instruments of rhetoric
 322 have an influence on the perceived charisma of a speaker as well (an effect that
 323 manifests itself in both speaker ratings and participant behavior). Antonakis et al.
 324 summarized these effective rhetorical instruments under the umbrella term “Charis-
 325 matic Leadership Tactics” (CLTs). These CLTs include, for example, metaphors,
 326 analogies, contrasts, rhetorical questions, and three-part lists (marked either explic-
 327 itly/verbally or implicitly/prosodically). Also, the use of the 1st person (instead of
 328 the 3rd person) singular or plural contributes to perceived speaker charisma (Biadys
 329 et al., 2008).

330 We controlled our speech excerpts such that they all contained a similar total
 331 number of CLT items and were dominated by verbs of the 1st person singular or plural.
 332 There were 3–4 CLT items within the 19–20 s excerpt of each speaker. The range of
 333 CLT items ranged from rhetorical questions (“How do you communicate authentically?”,
 334 Sheryl Sandberg) through metaphors and analogies (“... we will unlock new
 335 platforms”, Mark Zuckerberg; “We could not think of our users as wallets”, Margret
 336 Whitman) or syntagmatic contrast constructions (“We have talked about machine
 337 learning [...], but it also important to think about ...”, Sundar Pichai) to three-part
 338 lists (“it’s black, it’s invisible, it’s not understood—sight, sound, music ...”, Virginia
 339 Marie Rometty; “It’s the same amount of blood, sweat, and tears when you start a
 340 company”, Reid Hoffman). In addition, all eight speech excerpts are similar in that
 341 they outline an inspiring new idea in the context of a visionary future perspective
 342 (“You actually do not know inside of it, what it is—and that’s what’s changing in this
 343 new era”, Virginia Marie Rometty; “Aiming for something large is really important”,
 344 Reid Hoffman; “but it’s also important to think about how to do this technology can
 345 have an immediate impact on people’s lives”, Sundar Pichai).

346 Using an online script, the selected waveform signal was extracted from each
 347 YouTube clip and stored as an uncompressed audio file (.wav) in mono with a sound
 348 quality of 48 kHz and 24-bit. Each speech excerpt was characterized by a moderate
 349 speaking rate of on average about 5 syllables per second (syll/s) and a moderate
 350 pitch level of on average about 140 Hz (male speakers) or 205 Hz (female speakers).
 351 These moderate levels are suitable for performing a parameter manipulation without
 352 creating audible artifacts or extreme values of speaking rate and pitch.

353 The manipulation was done by means of the PSOLA resynthesis algorithm imple-
 354 mented in PRAAT (Mouliner & Charpentier, 1990; Boersma, 2001). For each speech
 355 excerpt, two combined PSOLA manipulations were performed and presented to the
 356 participants in the perception experiment instead of the original speech excerpt. That
 357 is, the perception experiment included only resynthesized audio stimuli. In this way,
 358 we ensured that all audio stimuli had the same sound quality.

359 The first stimulus condition of the independent variable Prosody was created by
 360 decreasing the speaking rate by 10% and the pitch level by 2 semitones (st) for each
 361 speaker. The pitch level was manipulated in st (i.e., along a logarithmic scale) so
 362 that the changes in acoustic F0 were perceptually equal for men and women. The
 363 size of the manipulation (2 st) was above the Just Noticeable Difference (JND) and
 364 hence audible for participants (Jongman, Qin, Zhang, & Sereno, 2017), but still small
 365 enough not to affect the naturalness of the speech. The speaking-rate manipulation
 366 was performed linearly across consonants and vowels. This is a simplification. In
 367 actual speech, vowel durations would change more as a function of speaking rate
 368 than consonant durations (van Santen, 1994); rate changes would also be paralleled
 369 by changes in speech reduction that cannot be imitated in resynthesized speech (see
 370 Ernestus & Smith, 2018). However, the resulting PSOLA output still sounded natural;
 371 also because 10% was, like for pitch, above the JND for speaking-rate changes at
 372 the utterance level (Quené, 2004), but small enough for the simplification and other
 373 manipulation artifacts to not become salient.

374 The second stimulus condition of the independent variable Prosody was created
 375 exactly inversely to the first one. That is, the speaking rate was increased by 10%,
 376 and the pitch level by 2 st compared to the original parameter values.

377 Note that we manipulated speaking rate and pitch level in combination rather than
 378 independently of each other because our focus was not on the interplay of the two
 379 prosodic parameters in charisma perception. Both parameters are well investigated
 380 in this respect already (Berger et al., 2017). Our aim was to create a strong and
 381 reliable variation in prosody-induced charisma and determine its interplay with a
 382 variation in attire-induced charisma. To that end, it was an advantage to co-vary two
 383 prosodic parameters, especially those whose effects on charisma are consistent and
 384 well investigated, also with respect to speaker gender.

385 At the end of the manipulation procedure, we had two versions of the same
 386 19–20 s speech excerpt for each speaker: one with higher parameter values (+10%
 387 speaking rate, +2 st pitch level) and one with lower parameter values (–10% speaking
 388 rate, –2 st pitch level). In connection with the independent variable Prosody, the
 389 former version is henceforth called the *high* condition; the latter version is referred
 390 to as the *low* condition. Note that, like for Attire, the two variable levels have a
 391 gender-specific implication for charisma perception. Based on previous findings,
 392 male speakers should sound more charismatic in the *high* Prosody condition, whereas
 393 female speakers should sound more charismatic in the *low* Prosody condition.

394 11.2.4 Experiment Design

395 The more and less charismatic speech excerpts (audio stimuli) of a speaker were
 396 combined with the conservatively and expressively dressed photographs (visual
 397 stimuli) of that speaker. Thus, all stimuli of the perception experiment were multi-
 398 modal. Per speaker, there were $2 \times 2 = 4$ different audio-visual stimulus conditions:

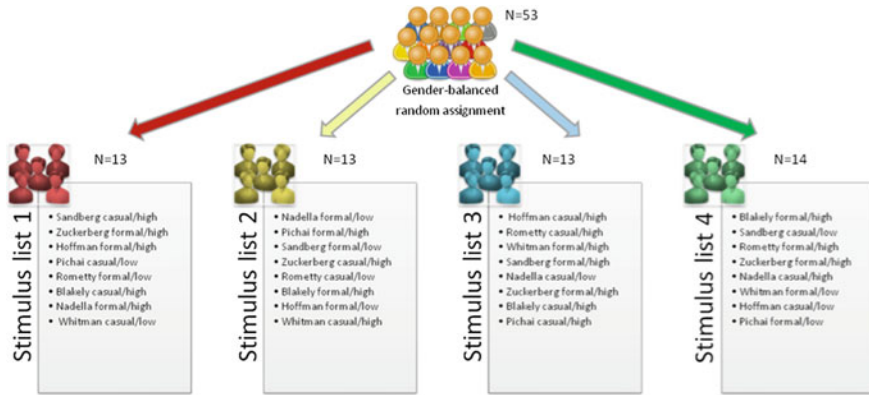


Fig. 11.2 Assignment of the 53 participants to four stimulus lists. Note that the order of the stimuli in each list is an example. Stimulus orders were individually randomized in the experiment

399 *conservative/high, conservative/low, expressive/high, and expressive/low*. For eight
400 speakers, this gave a total of 32 stimuli.

401 In order to keep the experiment short and interesting, the 32 stimuli were not all
402 presented to each participant. Rather, four different stimulus lists were compiled.
403 The four audio-visual stimulus conditions of each speaker were distributed across
404 these four lists such that each participant saw, heard, and rated each speaker only
405 once, see Fig. 11.2. This made it impossible for individual participants to uncover the
406 independent variables and their manipulations and infer from them the actual goal of
407 the experiment. Participants only received eight differently dressed and differently
408 speaking leading managers of US American companies, men and women, whose
409 speaker charisma was to be assessed by them. Note that due to distributing the four
410 audio-visual stimulus conditions across the four lists, both Attire and Prosody became
411 between-subject factors in the experiment design.

412 Charisma is a complex, multi-faceted concept. Accordingly, our experience from
413 pilot testing suggests in agreement with previous studies that participants respond
414 insecurely and/or heterogeneously when being asked to rate the charisma of a speaker
415 directly on a scale. For this reason, we decomposed charisma into three attributes
416 that participants could rate separately for each audio-visual stimulus: “The speaker is
417 ...” (1) convincing (German: *überzeugend*); (2) passionate (German: *enthusiastisch*);
418 (3) charming (German: *ansprechend*). This decomposition creates a clear frame of
419 reference and provides participants with a concrete idea of what they are supposed to
420 rate. In this way, the ratings become simpler and more consistent. The three attributes
421 were chosen, because they are known from previous studies to be highly correlated
422 with perceived speaker charisma (Rosenberg & Hirschberg, 2009), and because they
423 are equally applicable to both attire and speech prosody.

424 11.2.5 Participant Sample and Experimental Procedure

425 The experiment was conducted as an online experiment (based on SoSci Survey).
 426 A total of 53 participants took part in the experiment; 23 men and 30 women who
 427 were between 21 and 48 years old. The average age of the participant sample was
 428 24.7 years. All participants were native speakers of German and undergraduate or
 429 graduate students of social-science disciplines (“Innovation and Business” or “Inno-
 430 vation and Technology Management”). All had a very good command of English,
 431 i.e., either level B2 or C1 according to university-internal student entry tests. Never-
 432 theless, their skills as non-native speakers were not sufficient to consistently identify
 433 regional or dialectal differences between speakers and associate them with positive
 434 or negative stereotypes or speaker attributes (Bailey, 2003). The 53 participants were
 435 distributed almost equally over the four stimulus lists. The 13 or 14 participants per
 436 list included about equal numbers of men and women. Except for controlling these
 437 basic factors, the participant-to-list assignment was entirely random.

438 Each online session of SoSci Survey started with the information that the exper-
 439 iment would be about the assessment of perceived speaker charisma. The compo-
 440 sitional concept of charisma was briefly outlined to the participants with reference
 441 to Antonakis et al. (2016) who defined charisma as “values-based, symbolic, and
 442 emotion-laden leader signaling” (p. 304). In addition, the participants were given
 443 some names of particularly charismatic speakers for further illustration. These names
 444 included, for example, Steve Jobs, Barak Obama, and Martin Luther King Jr. In order
 445 to increase the spontaneity and impartiality of assessments, it was emphasized to the
 446 participants that assessments of perceived speaker charisma are inevitably subjective
 447 and that there is no right or wrong in subjective assessments.

448 Subsequently, participants were told that they would successively see and hear
 449 eight fairly popular and influential male and female managers (CEOs or COOs) of
 450 leading US companies. Each audio-visual stimulus would consist of a photograph
 451 of one of these eight managers at an important keynote speech and an approxi-
 452 mately 20-second audio clip from that keynote speech. On this basis, their task
 453 would simply be to listen to each of the eight audio-visual stimuli separately, i.e.,
 454 without drawing comparisons between the speakers, and each time as if being part
 455 of the speaker’s keynote audience. Then, ratings were to be made about how the
 456 speaker was experienced in terms of perceived charisma on three scales

- 457 • Convincing,
- 458 • Passionate,
- 459 • Charming.

460 Participants received the 6-level system of German school grades from “1” (=very
 461 good) to “6” (=not good at all) for their assessments, as this is a system that all
 462 participants were well familiar with. Judgments were made by clicking, for each
 463 charisma attribute, the respective button of a 6-point Likert scale whose endpoints,
 464 “very good” and “not good at all”, were displayed above the three scales. An example
 465 of one judgment trial of the experiment is shown in Fig. 11.3.

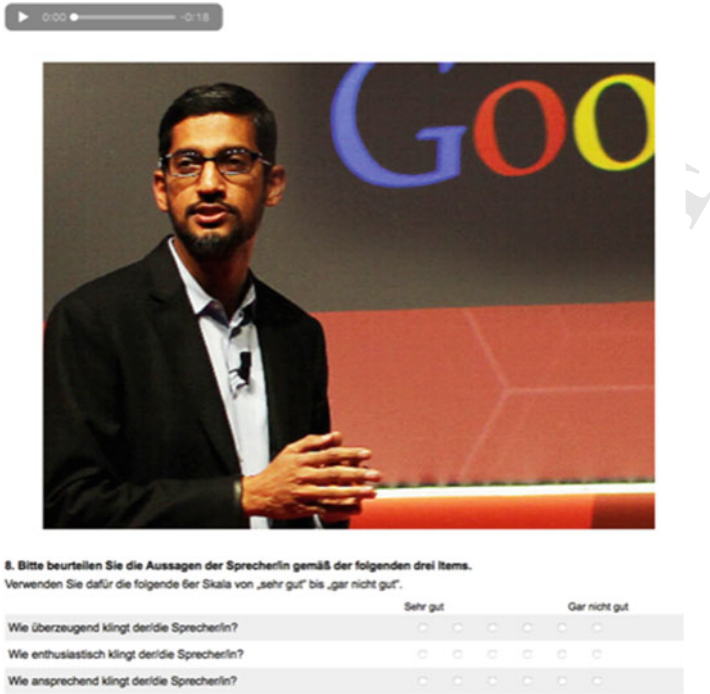


Fig. 11.3 Screenshot of one judgment trial of the male speaker Sundar Pichai in the experiment. Photo of Sundar Pichai taken by Maurizio Pesce (2015, edited under CC-BY license by 1st author)

466 Following the initial instructions, the participants were presented with the eight
 467 audio-visual stimuli of their respective stimulus list. The experiment was performed
 468 in a self-paced fashion. Each participant received the eight audio-visual stimuli of
 469 his/her list in an individually randomized order.

470 After the experiment was over, a few metadata of the participants were queried.
 471 These included age, gender, level of English, familiarity with the eight managers, and
 472 further speaker-oriented judgments on estimated age, perceived physical attractive-
 473 ness, and estimated leadership experience. Furthermore, the participants were asked
 474 to specify their foreign or second language skills (besides English) and to give some
 475 feedback on the difficulty and the assumed purpose of the experiment as well as on
 476 the applicability of the rating scales. Together with this final metadata questionnaire,
 477 the entire experiment took about 10–12 min.

478 11.3 Results

479 The statistical processing of the data was performed separately for the two quadru-
 480 plets of male and female speakers, taking into account that we expect the Attire

Editor Proof

481 and Prosody manipulations to affect speaker charisma in diametrically opposed
 482 ways depending on speaker gender. The gender-specific results are presented in
 483 Sects. 11.3.1 and 11.3.2. One of the three charisma-related scales, charming, did
 484 not yield conclusive results, and in the feedback questionnaire after the experiment
 485 participants also reported problems with applying this scale to the stimuli (we will
 486 address these problems in more detail in the discussion). For this reason, only the other
 487 two scales—convincing and passionate—were taken into account in the analysis and
 488 presentation of the results.

489 In accord with the use of convincingsness and passion scales in previous studies,
 490 we found that the two scales are good representatives of charisma and suitable for
 491 asking participants to assess perceived speaker charisma. First, the participants rated
 492 the application of the scales to the stimuli and the general concept of charisma as
 493 simple and intuitive. Second, matching with the participants' report, we found no
 494 contradicting ratings in our results data, i.e., no cases in which the convincingsness
 495 and passion ratings of a single stimulus go in opposite directions. On the contrary,
 496 the convincingsness and passion ratings are correlated with each other in an order of
 497 magnitude that matches with how strongly they correlated with charisma itself in
 498 previous studies ($r[200] = 0.55$, $p < 0.001$ for the male speakers' stimuli and $r[200]$
 499 $= 0.69$, $p < 0.001$ for the female speakers' stimuli). That is, convincingsness and
 500 passion both represent perceived speaker charisma equally well, but are nevertheless
 501 related to different facets of the phenomenon. Reflecting this fact, the results section
 502 presents the convincingsness and passion ratings separately, but at the same time
 503 interprets them coherently in terms of perceived speaker charisma.

504 For the statistical analysis, a three-way General Linear Model (GLM) was used,
 505 with the two independent variables Attire and Prosody being fixed factors. As the
 506 third fixed factor, Speaker was additionally included in the model (four levels for the
 507 four male or female speakers). For supplementary t-tests and multiple comparisons
 508 between factor levels (e.g., of the fixed factor Speaker), alpha-error levels were
 509 adjusted using the Sidak method. Dependent variable was the rating score 1–6 on the
 510 respective grading scale per participant. The participant him/herself was taken into
 511 account as a random factor in the GLM. Participant as a random factor was appropriate
 512 here for two reasons. First, the participants were randomly selected, and, secondly,
 513 we were not interested in identifying possible differences between participants as a
 514 previous inspection of the data already indicated no separate systematic effects of
 515 participant age, gender, and international/linguistic background. In contrast, in the
 516 case of Speaker, we were interested in possible differences among the male or female
 517 speakers. For this reason, we made Speaker a fixed factor. However, note that we
 518 would arrive at the same conclusions with (male or female) Speaker being a second
 519 random factor. Further aspects of the generalization of the findings are addressed in
 520 Sect. 11.4.5.

521 Separate statistical analyses (GLMs) were conducted for the two assessment scales
 522 convincing and passionate. Each of these analyses was based on 212 participant
 523 ratings, 106 for the variable Attire, and 106 for the variable Prosody. All individual
 524 t-tests comparisons were conducted with 52–56 participant ratings in each sample.

525 We use bar plots below for illustrating the statistical patterns and summarizing
 526 the results descriptively. As it would be confusing for many readers that higher rising
 527 bars mean worse and lower rising bars better ratings of speaker charisma (as 6 =
 528 best and 1 = worst), we plotted the bars upside down. So, the lower a bar reaches
 529 the more negative is the charisma-related rating.

530 **11.3.1 Male Speakers**

531 The bar plot in Fig. 11.4 shows the effects of the variable Attire on the rating of the
 532 four male speakers. The individual bars display, in a different color for each speaker,
 533 the cumulative mean value of the difference between the two Prosody conditions
 534 *low* and *high* across all 53 participants. So, for example, if the mean rating of a
 535 speaker on the convincingness scale were 3.4 in the Prosody condition *low* and 2.4
 536 in the Prosody condition *high*, then Fig. 11.4 would show a value of +1 for this
 537 speaker (recall that higher numbers in the German school grading system mean a
 538 worse performance).

539 The results shown in Fig. 11.4 can be summarized as follows. In terms of the
 540 two attributes convincing and passionate, speaker charisma is perceived to be higher
 541 for the *conservative* Attire condition than for the *expressive* Attire condition. In
 542 other words, wearing a conservative attire supports the speakers to the extent that it
 543 doubles their perceived charisma The scale values halved accordingly: For perceived
 544 convincingness, we can see a decrease in the overall assessment across the four

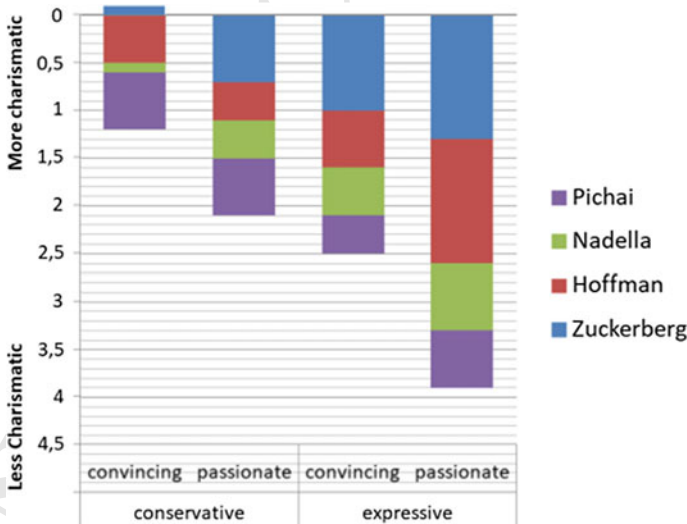


Fig. 11.4 Results of the Attire conditions conservative and expressive on the male-speaker assessments

Editor Proof

545 speakers from 2.5 in the *expressive* Attire condition to about 1.2 in the *conservative*
 546 Attire condition. For perceived passion, the cumulative mean value of just under 4.0
 547 in the *expressive* condition is halved to only 2.1 in the *conservative* condition.

548 With respect to Prosody, Fig. 11.4 shows further that, with one exception, all
 549 the mean differences between the two Prosody conditions *low* and *high* give a
 550 positive value. This means that each speaker was judged to be more convincing and
 551 and passionate—i.e., overall more charismatic—for the higher than for the lower
 552 parameter values of speaking rate and pitch level.

553 In the corresponding GLMs the results of Fig. 11.4 manifest themselves in signif-
 554 icant main effects of Attire (convincing: $F[1,196] = 440.70, p < 0.001, \eta_p^2 = 0.69$;
 555 passionate: $F[1,196] = 687.20, p < 0.001, \eta_p^2 = 0.78$) as well as in similar, but
 556 in terms of partial Eta-squared slightly weaker significant main effects of Prosody
 557 (convincing: $F[1,196] = 219.68, p < 0.001, \eta_p^2 = 0.53$; passionate: $F[1,196] = 350.75,$
 558 $p < 0.001, \eta_p^2 = 0.64$). The fixed factor Speaker had significant main effects as well
 559 (convincing: $F[3,196] = 307.48, p < 0.001, \eta_p^2 = 0.83$; passionate: $F[3,196] = 629.17,$
 560 $p < 0.001, \eta_p^2 = 0.91$). Moreover, there were, for both assessment scales, significant
 561 interactions between Speaker and Attire (convincing: $F[3,196] = 33.52, p < 0.001,$
 562 $\eta_p^2 = 0.34$; passionate: $F[3,196] = 11.85, p < 0.001, \eta_p^2 = 0.15$). The three-way
 563 interaction was not significant.

564 Figure 11.5 shows in more detail how the speaker rating changes as a result of the
 565 Prosody variable, pooled across the two scales convincing and passionate. There is a
 566 significant interaction of the variable Prosody with the variable Attire (convincing:
 567 $F[1,196] = 6.79, p < 0.01, \eta_p^2 = 0.11$; passionate: $F[1,196] = 41.72, p < 0.001,$

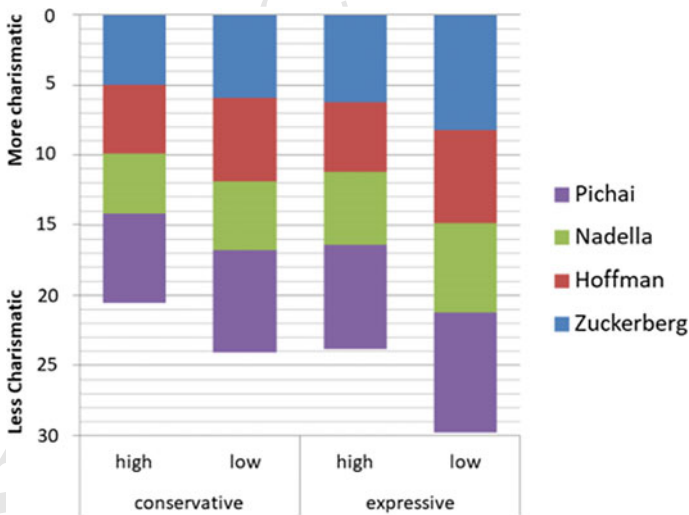


Fig. 11.5 Results of the Prosody conditions *high* and *low* in each Attire condition on the male-speaker assessments

Editor Proof

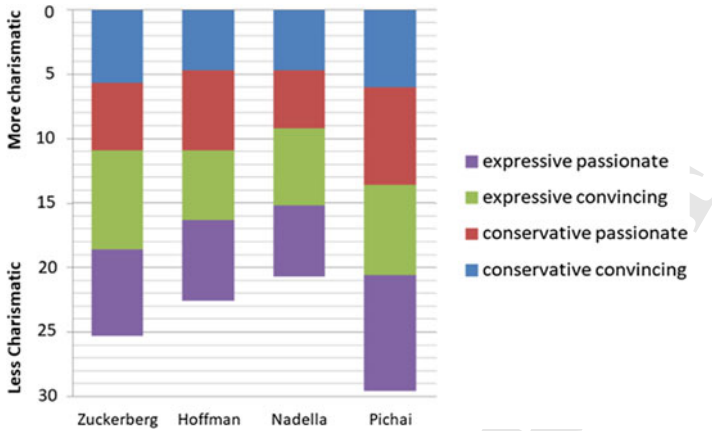


Fig. 11.6 Total assessment of the male speakers on the two charisma scales convincing and passionate

568 $\eta_p^2 = 0.18$). For a *conservative* attire, the charisma-supporting or reducing effect of
 569 Prosody is smaller than for an *expressive* attire. This means that, for a participant's
 570 rating of a speaker's charisma, the factor Prosody counts more if the speaker wears
 571 an expressive attire. In other words, those who wear a expressive attire (as a man)
 572 have to focus more on producing a charismatic speech prosody than those who wear a
 573 conservative attire. In fact, the two Attire-Prosody combinations *conservative/low*
 574 and *expressive/high* came out as statistically equivalent ($p > 0.05$) in separate t-tests
 575 for all 4 male speakers.

576 Figure 11.5 also shows that some speakers consistently contributed more than
 577 others to the cumulative mean values of each bar. That is, some speakers were consistently
 578 rated worse than others. Figure 11.6 illustrates this finding more clearly. Across
 579 the Attire and Prosody conditions and the two scales convincing and passionate,
 580 Zuckerberg and Pichai yielded the highest sums of mean ratings and hence the overall
 581 worst charisma ratings, with Pichai being slightly worse than Zuckerberg. Nadella
 582 performed best. Reid Hoffman's performance was, in the overall assessment of the 53
 583 participants, somewhere in between Pichai and Nadella. Multiple t-test comparisons
 584 between the four speakers showed accordingly that all speakers differed from each
 585 other at $p < 0.001$, except for Zuckerberg and Pichai on the convincingness scale (p
 586 > 0.05) and Nadella and Hoffman on the same scale ($p > 0.05$).

587 11.3.2 Female Speakers

588 The results of the four female speakers are different. Unlike for the male speakers,
 589 the main effects of Attire and Prosody are not significant. Figure 11.7a, i.e., the
 590 counterpart of the male speakers' Fig. 11.5, shows very clearly that the cumulative

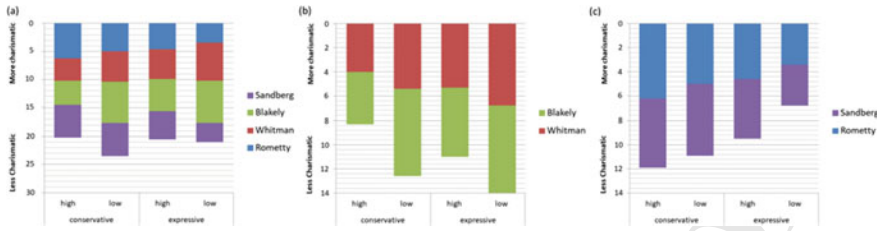


Fig. 11.7 Results of the Prosody conditions *high* and *low* in each Attire condition **a** for all four female speakers, and separately for **b** the Blakely–Whitman speaker pair and **c** the Sandberg–Rometty speaker pair

591 charisma ratings of the 53 participants are roughly the same for all independent vari-
 592 able conditions. The reason for this becomes obvious in Figs. 11.7b–c and 11.8 (the
 593 counterpart of the male speakers’ Fig. 11.4): The female speaker sample contains
 594 two pairs of speakers whose Attire and Prosody conditions were rated in a diamet-
 595 rically opposed fashion by the 53 participants. This manifests itself in the GLMs
 596 in a significant main effect of Speaker (convincing: $F[3,196] = 221.01, p < 0.001,$
 597 $\eta_p^2 = 0.77$; passionate: $F[3,196] = 100.50, p < 0.001, \eta_p^2 = 0.61$) and in significant
 598 interactions of Speaker and Attire (convincing: $F[3,196] = 199.78, p < 0.001, \eta_p^2 =$
 599 0.75 ; passionate: $F[3,196] = 169.39, p < 0.001, \eta_p^2 = 0.72$) and of Speaker and
 600 Prosody (convincing: $F[3,196] = 148.63, p < 0.001, \eta_p^2 = 0.70$; passionate: $F[3,196]$
 601 $= 276.08, p < 0.001, \eta_p^2 = 0.81$), each with high Eta-squared effect sizes. The three-
 602 way interaction is significant as well (convincing: $F[3,196] = 21.49, p < 0.001, \eta_p^2$

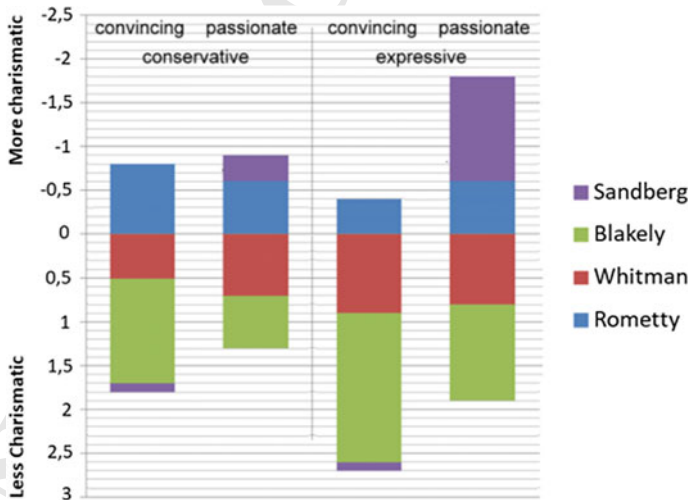


Fig. 11.8 Results of the Attire conditions *conservative* and *expressive* on the female speaker assessments

Editor Proof

603 = 0.25; passionate: $F[3,196] = 27.12$, $p < 0.001$, $\eta_p^2 = 0.29$). Multiple-comparisons
 604 tests within the factor Speaker showed further that the participants' ratings of Blakely
 605 and Whitman differ on neither of the two scales. The same negative result was found
 606 for Sandberg and Rometty. At the same time, the latter two speakers differ signifi-
 607 cantly from the former two speakers on both scales at $p < 0.001$. These test statistics
 608 support that Blakely and Whitman on the one hand and Sandberg and Rometty on
 609 the other really formed two different pairs of speakers.

610 In order to look at the two speaker pairs in more detail, we ran separate additional
 611 GLMs for the Blakely–Whitman pair and for the Sandberg–Rometty pair.

612 The results of the two female speakers Blakely and Whitman largely agree with
 613 those of the male speakers. That is, the *conservative* Attire condition (dark-colored
 614 pantsuits) supports the two speakers' perceived charisma relative to the *expressive*
 615 Attire condition (red or pink pantsuits). The corresponding main effects are signif-
 616 icant (convincing: $F[1,102] = 59.23$, $p < 0.001$, $\eta_p^2 = 0.45$; passionate: $F[1,102] =$
 617 39.86 , $p < 0.001$, $\eta_p^2 = 0.23$). Likewise, the Prosody condition *high*—characterized
 618 by increases in speaking rate and pitch level—supports the charisma perception of
 619 the two speakers relative to the Prosody condition *low*. The corresponding main
 620 effects are significant as well (convincing: $F[1,102] = 61.71$, $p < 0.001$, $\eta_p^2 = 0.51$;
 621 passionate: $F[1,102] = 363.44$, $p < 0.001$, $\eta_p^2 = 0.66$).

622 In contrast, for the two female speakers Sandberg and Rometty, the effects of
 623 Attire are exactly inverse and hence also run counter to those of the four male speakers
 624 (convincing: $F[1,102] = 121.26$, $p < 0.001$, $\eta_p^2 = 0.80$; passionate: $F[1,102] = 411.41$,
 625 $p < 0.001$, $\eta_p^2 = 0.94$). The same applies to Prosody (convincing: $F[1,102] = 50.58$,
 626 $p < 0.001$, $\eta_p^2 = 0.37$; passionate: $F[1,102] = 123.77$, $p < 0.001$, $\eta_p^2 = 0.82$). Unlike
 627 for Blakely and Whitman and the four male speakers, it is the Prosody condition
 628 *low* rather than *high* in which Sandberg and Rometty sound more charismatic in the
 629 ears of the 53 participants. Moreover, it is the *expressive* rather than the *conservative*
 630 attire condition that makes Sandberg and Rometty look more charismatic in the eyes
 631 of the 53 participants.

632 What all four female speakers have in common is that the overall effect of Attire is
 633 smaller than for the male speakers. While the choice between a conservative and an
 634 expressive attire was able to increase male speaker charisma by about 50%, female
 635 speaker charisma could only be increased by about 20%. A t-test based on abso-
 636 lute difference values between the Attire conditions in the male and female speaker
 637 samples shows that this gender-specific effect size of Attire is significant ($p < 0.01$).
 638 For the effect of Prosody, it were the female speakers for whom the difference between
 639 the two conditions *low* and *high* had an overall larger effect on perceived charisma
 640 than for the male speakers. Going from low to high (for Blakely and Whitman)
 641 or from high to low (for Sandberg and Rometty) enhanced the female speakers'
 642 charisma level by up 50%, independently of the Attire condition. In contrast, for the
 643 male speakers, the ability of Prosody to increase speaker charisma was between 10
 644 and 20% and depended on the Attire condition. A t-test based on absolute differ-
 645 ence values between the Prosody conditions in the male and female speaker samples
 646 shows that this gender-specific effect size of Prosody is also significant ($p < 0.001$).

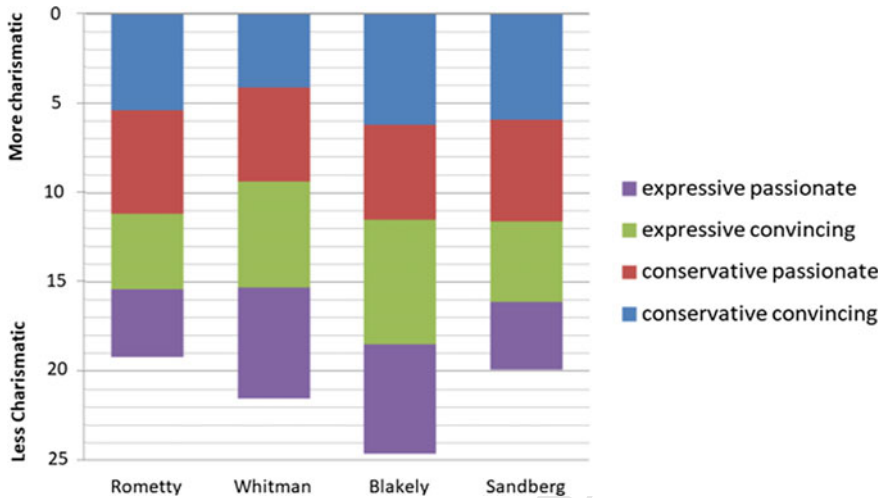


Fig. 11.9 Total assessment of the female speakers' charisma on the two scales convincing and passionate

647 Regarding the reason why the female speaker sample included two differently
 648 rated pairs of speakers, we discovered a parallel between the rating of participants
 649 and the speakers' perceived physical attractiveness. These attractiveness ratings (on
 650 a scale of 0–10) were made by participants in the feedback questionnaire after the
 651 experiment. An analysis of these judgments revealed that Rometty and Sandberg
 652 obtained physical-attractiveness values that were, according to within-subjects t-
 653 tests, statistically equivalent ($p > 0.05$), but clearly and significantly lower than those
 654 obtained by Blakely and Whitman (\bar{x} 4.1 vs. 6.6, $t[105] = -12.76$, $p < 0.01$), whose
 655 physical-attractiveness judgments were again statistically equivalent ($p > 0.05$).
 656 Further questionnaire analyses and even additional acoustic-prosodic measurements
 657 and analyses of the keynote-speech excerpts (for the charisma-relevant parameters
 658 specified in Niebuhr et al. (2017) showed that attractiveness was the only factor whose
 659 statistical results pattern runs exactly parallel to that of the two differently rated
 660 female speaker pairs. The total charisma scores shown in Fig. 11.9 yielded a similar,
 661 but not exactly parallel results picture as there is a significant difference between
 662 Blakely and Whitman ($p < 0.05$) on the one hand, but no significant differences
 663 different Whitman and Sandberg and Rometty on the other hand.

664 11.4 Discussion

665 The present study investigated the interaction effects of variation in attire and prosody
 666 on the perception of male and female speaker charisma. A total of 53 participants

667 took part in the experiment and rated, in individually randomized orders, audio-
 668 visual stimuli of eight senior business leaders, four males and four females, on three
 669 charisma-related scales that were successfully tried and tested in many previous
 670 studies. In the debriefing questionnaire, the participants described the experiment as
 671 pleasant and easy, and judged the charisma ratings on the two scales of convincingness
 672 and passion as intuitive and applicable (the problematic scale “charming” is discussed
 673 in Sect. 11.4.4). Therefore, we view the significant effects of our results as (internally
 674 and externally) valid and reliable. This view is also corroborated by the fact that Mark
 675 Zuckerberg turned out to be a fairly uncharismatic speaker, which is consistent with
 676 previous studies (Niebuhr et al., 2016b). The following discussion is based on this
 677 validity and reliability.

678 11.4.1 Assumptions

679 We tested three assumptions with our experiment. The first one was whether or
 680 not the experiment replicates the known gender-specific effects of pitch level and
 681 speaking rate on perceived speaker charisma. This assumption is partially supported
 682 by the results of the experiment. The male speakers were rated more charismatic if
 683 they spoke with increased pitch and speaking-rate levels (compared to the original
 684 prosodic setting of the corresponding speaker). Changes toward lower pitch and
 685 speaking-rate levels significantly negatively impacted the charisma of male speakers.
 686 For two of the female speakers, Sandberg and Rometty, this influence of prosody
 687 on the perceived charisma was exactly inverse. That is, it was the lower pitched,
 688 slower way of speaking that was more charismatic, not the higher pitched, faster
 689 way of speaking. This gender-specific difference meets the first assumption and is
 690 consistent with the results of Berger et al. (2017) and Bachsleitner and Popp (2018).
 691 For the other two female speakers, Blakely and Whitman, however, the results were
 692 diametrically opposed (i.e., in line with those the male speakers again). Thus, they
 693 run counter to what we expected from our assumption (1) for female speakers. In
 694 Sect. 11.4.2, we offer an explanation for why the bipartition of our female speakers’
 695 results have occurred and why the deviating results for Blakely and Whitman are
 696 probably only in apparent contradiction to assumption (1) and the findings of Berger
 697 et al. (2017), Bachsleitner and Popp (2018).

698 Our second assumption was that the experiment would find a gender-specific
 699 effect of attire on perceived speaker charisma. This assumption is clearly supported
 700 by the findings. Male speaker charisma was enhanced by the conservative style of
 701 a dark-colored suit rather than by the expressive style of t-shirt, jeans, and similar
 702 casual clothes. The attire effect on female speaker charisma differed from that of
 703 the male speakers and was overall more complex. For two female speakers, the
 704 assumption was met that an expressive red, as opposed to a conservative dark color,
 705 had a charisma-supporting effect. For the other two speakers, it was the other way
 706 around.

707 The third assumption was that the experiment would find the gender-specific
 708 effects of attire and prosody to be additive in the perception of speaker charisma.
 709 Additive means that an unfavorable attire condition and an unfavorable prosody
 710 condition together reduce the perceived speaker charisma more than each unfavorable
 711 condition alone. In the opposite direction, a favorable attire condition and a favorable
 712 prosody condition together should enhance perceived speaker charisma more than
 713 each favorable condition alone. Combinations of favorable and unfavorable attire
 714 and prosody conditions should neutralize each other or result in minimally positive
 715 or negative charisma effects only. Exactly this overall pattern was found in the exper-
 716 iment for all our eight male and female speakers. For example, it is clearly visible
 717 for the male speakers in Fig. 11.5 that *conservative/low* was less charismatic than
 718 *conservative/high*, and that *expressive/low* was less charismatic than *expressive/high*.
 719 The two extreme pairs of conditions, i.e., the maximally favorable *conservative/high*
 720 combination and the maximum unfavorable *expressive/low* combination, yielded the
 721 largest overall difference in perceived speaker charisma. The two cross-over combi-
 722 nations *conservative/low* and *expressive/high* neutralized each other statistically. The
 723 third assumption was thus clearly met by the data.

724 11.4.2 The Bipartition of the Female Speaker Group

725 As was reported in the results section, the bipartition of the female speaker group
 726 in terms of charisma ratings runs parallel to the attractiveness ratings of the female
 727 speakers. The speaker pair Sandberg/Rometty was perceived most charismatic in
 728 the *expressive/low* stimuli and received at the same time relatively low physical-
 729 attractiveness ratings ($\bar{\mu}$ 4.1, between-speaker difference <0.5 , n.s.). The speaker
 730 pair Blakely/Whitman received significantly higher physical-attractiveness ratings ($\bar{\mu}$
 731 6.6, between-speaker difference <0.5 , n.s.) and was perceived most charismatic in the
 732 *conservative/high* stimuli. No other differences in speaker judgments, metadata, or
 733 personal characteristics (like hair color, age, size, or estimated leadership experience),
 734 and no uncontrolled acoustic-prosodic parameter differences matched equally well
 735 with the bipartition of the female speaker group as the attractiveness rating. Although
 736 it is “a myth that you have to be attractive to be charismatic” (Fox Cabane, 2012:
 737 102), charisma and physical attractiveness are still to some degree related perceptual
 738 concepts (Grabo, Spisak, & van Vugt, 2017). Furthermore, it is known that charisma
 739 can also be exaggerated and, thus, reversed by an overdose of acoustic or visual
 740 triggers. For this reason, Niebuhr et al. (2017) determined so-called “effectiveness
 741 windows” that charisma-relevant parameters should neither fall below nor exceed.
 742 Against this background we suggest the following explanation for why the bipartition
 743 of the female speaker group occurred.

744 If physically more attractive female speakers already start from an inherently
 745 higher perceived charisma level than physically less attractive female speakers, then
 746 adding further charisma-enhancing stimuli like a red attire and a slow, low-pitched
 747 prosody can result in an overdose and hence in a reversed effect of attire and prosody

748 on perceived charisma. This could have happened for the speaker pair Blakely and
 749 Whitman. In contrast, if physically less attractive female speakers start from an
 750 inherently lower perceived charisma level, then they can still benefit from adding
 751 further charisma-enhancing stimuli like a red attire and a slow, low-pitched prosody
 752 to the overall charisma they convey. This could be true of the speaker pair Sandberg
 753 and Rometty.

754 The advantage of this explanation is that it would be consistent with both the
 755 assumed charisma-enhancing effect of a red attire and the previously found gender-
 756 specific prosodic effects of pitch level and speaking rate in the studies of Berger et al.
 757 (2017), Bachsleitner and Popp (2018). Moreover, the provided explanation would
 758 also mean that the Attire and Prosody conditions did actually have the same effect on
 759 *all* females speakers. It would only be due to the interaction with attractiveness that
 760 this uniform effect surfaces differently for the two speaker pairs Blakely/Whitman
 761 and Sandberg/Rometty. On this basis, assumption (1) would be fully supported by
 762 the present results. It is further in accord with the provided explanation that no
 763 attractiveness differences showed up for the four male speakers (all received average
 764 ratings between 5.5 and 6.5 on the 10-point scale). Thus, Attire and Prosody were
 765 able to influence charisma ratings in a uniform way for the male speakers. In fact, it
 766 seems that men are generally rated less critically in terms of physical attractiveness
 767 than women, especially in the context of business, leadership, and perceived charisma
 768 (Friedman, Riggio, & Casella, 1988). Note in this context that rater gender did not
 769 play a significant role in the physical attractiveness ratings of our speakers. Female
 770 raters behaved in the same way as male raters.

771 An alternative but related explanation refers to the experiment of Pearce and
 772 Brommel (1972). They found that non-lexical charisma triggers only have a positive
 773 effect on attributes of perceived speaker charisma if the audience assesses the speaker
 774 as credible and competent. If the same charisma signals are conveyed by a less
 775 credible and competent speaker, then they have no effect or even a negative effect
 776 on the speaker's charisma. In the light of these findings, the bipartition of the female
 777 speaker group in the present experiment could also mean that the 53 participants (i.e.,
 778 both males and females) assessed the physically more attractive female speakers
 779 Blakely and Whitman to be less credible and competent than the less attractive
 780 speakers Sandberg and Rometty.

781 Subsequent studies must continue to investigate which of the two explanations (or
 782 maybe a third one) underlies the bipartition of the female speaker group in the present
 783 experiment. However, regardless of the explanation, the present findings already have
 784 an important practical implication: Female speakers need to pay more attention than
 785 men to how many and strong audio-visual charisma triggers they convey, and it is
 786 likely that physical attractiveness is an important factor to take into account in this
 787 context. More physically attractive women should perhaps rather try to downgrade
 788 their remaining charisma triggers, for example, by using a conservative dark-colored
 789 outfit and clearly also a less charismatic prosody, whereas for physically less attractive
 790 women the opposite can be recommended, i.e., using a more expressively colored
 791 outfit and definitely a more charismatic prosody. Why we stress prosody in this
 792 connection is stated in 11.4.3, together with further practical implications.

793 11.4.3 Further Practical Implications

794 Our results show that the charisma rating of male speakers can be increased or
 795 decreased by about 50% through the attire choice alone. The effect of prosody on
 796 the charisma rating was smaller and depended on the attire (at least for the two
 797 parameters pitch level and speaking rate manipulated here). For women, the effect
 798 of prosody was larger than the effect of attire. Like for men, there was an interaction
 799 with the choice of attire. However, as we discussed in detail in Sect. 11.4.2, this
 800 interaction did not affect the size of the prosody effect, but its direction. The size of
 801 the prosody effect was independent of the choice of attire.

802 Two practical implications can be derived from these findings. First, women
 803 benefit more from using the right prosody, while men benefit more from choosing
 804 the right attire. Second, in a charisma-supporting conservative attire style (dark suit),
 805 men may well afford smaller weaknesses in prosodic charisma performance. In an
 806 expressive, casual attire style, on the other hand, men have to take care to deliver a
 807 very charismatic prosodic performance if they still want to make a strong charismatic
 808 impression. So, anyone who (as a man) has confidence in his excellent delivery can
 809 basically also perform in an expressive, casual style of clothing in front of his audi-
 810 ence (although a conservative attire would still be better). For those who are insecure
 811 and unskilled in their speech performance, a conservative dress style should be a
 812 must.

813 11.4.4 The Scale “Charming”

814 The inconclusive results of the scale charming and the application problems reported
 815 by the participants in the debriefing questionnaire came unexpected. The scale
 816 charming was selected, as Rosenberg and Hirschberg (2009) showed that this attribute
 817 is even higher correlated with charisma than convincing and passionate and can also
 818 be applied more consistently to charisma than convincing and passionate. However,
 819 the key difference between our study and that of Rosenberg and Hirschberg is that we
 820 presented not just audio stimuli, but multi-modal audio-visual stimuli. It is obvious
 821 that charming—unlike convincing and passionate—has both an auditory and a visual
 822 rating dimension (to a limited degree, this is also true of passionate, but all passion-
 823 related signals of body language were carefully controlled and kept homogeneous in
 824 the photographs). In accord with the participants’ comments in the debriefing ques-
 825 tionnaire, we assume that it was this modality-based ambiguity of the term charming
 826 that caused the inconclusive results of the corresponding scale. For example, it turned
 827 out that some participants interpreted charming in the sense of a purely visual physical
 828 attractiveness and then used it automatically in the sense of sex appeal/attractiveness
 829 rather than in the intended sense of speaker charisma.

830 In summary, the correlated, consistent use of the scales convincing and passionate
 831 on the one hand shows that, with the multi-dimensional scaling method, we have a

832 valid and sensitive instrument for the evaluation of speaker charisma. Pilot studies
 833 show that charisma is a too complex concept to be directly rated by participants in a
 834 consistent way, see Sect. 11.2.4. By breaking down the concept into scales that are
 835 highly correlated with each other and with charisma, we can make the rating task
 836 easier and more consistent—and still measure the same “thing”. However, on the
 837 other hand, the inconclusive, inconsistent use of the scale charming also reveals and
 838 stresses the current limitations of this instrument. We have not fully understood as
 839 yet which facets of charisma are covered by each scale and how complementary and
 840 exhaustive this coverage is. Moreover, we have not enough knowledge today to put
 841 together a set of scales that are specifically tailored to measuring perceived speaker
 842 charisma for different types and modalities of stimuli. Also note in this context that
 843 the male speakers were generally rated worse on the passionate scale than on the
 844 convincing scale in the present study. For women it was the other way around. That
 845 is, independently of the set of scales and the stimuli, special care should be taken
 846 when comparing absolute scale levels between experimental conditions.

847 11.4.5 Generalization

848 As for all other experiments, our results apply primarily to the conditions under which
 849 they were obtained. The simpler and more controlled these conditions are, the lower
 850 is the potential generalization of the findings. As we said in the beginning, we selected
 851 photos and speech materials from the “field” and, moreover, used multiple speakers
 852 per gender to maximize generalizability within the experimental setup. Therefore, we
 853 think that our findings are sufficiently generalizable to have a practical use and to give
 854 male and female speakers guidance in public-speaking and presentation scenarios.
 855 We show with regard to perceived speaker charisma that prosody has an effect, that
 856 attire has an effect, that the effect of attire can also be negative (like that of prosody),
 857 and that the effects compensate, cancel out, or enhance each other and, in the latter
 858 case, can probably also cause overdoses. These facts will be valid in the real world
 859 regardless of the current experimental setting.

860 But, of course, there are many other auditory and visual sources of perceived
 861 speaker charisma that play a role, but are not considered or varied here. That is, we
 862 expect the strength of the present effects to be shaped by a number of other variables,
 863 which themselves may have favorable or unfavorable charisma effects. On the part of
 864 the recipients (i.e., the raters), these are, for example, variables from which norms and
 865 stereotypes emerge, such as educational attainment, age, cultural background (Power
 866 & Galvin, 1997), and the zeitgeist (50 or 100 years ago, a different way of speaking
 867 may have been considered more charismatic, cf. Madill, 2015 and the term “vocal
 868 zeitgeist” in McCabe & Altman, 2017; also business fashion changes constantly,
 869 especially for women, see Sect. 11.1.1). On the part of the speakers, relevant further
 870 variables are those that determine competency and prestige attributions, such as race,
 871 age, gender, attractiveness, occupation, and social status. Additionally, on the part of
 872 both recipient and speaker, there are the linguistic (including dialectal) background

873 and the communication medium, which in our opinion represent secondary variables.
874 These variables do not interact directly with speaker charisma, but indirectly through
875 an influence on primary variables such as competence, stereotypes, etc.

876 With the exception of some indications on attractiveness, our study cannot make
877 any new conclusions about these additional variables. However, as our male speakers
878 were all rated consistently—despite showing a greater racial diversity than our female
879 speakers—it appears that the factor race plays a subordinate role in speaker charisma,
880 at least among educated raters (like students) and for speakers with a high status
881 and prestige (like business leaders). Age and gender have an effect on charisma. In
882 tendency, those speakers are considered more charismatic who have a similar age as
883 the audience; and men tend to be inherently more charismatic than women (Jokisch
884 et al., 2018; Brooks et al., 2014), in both women’s and men’s ratings (recall that we,
885 too, have found no gender-specific rating differences). Our own data from a different
886 study (Abidi & Gumpert, 2018) further suggest that the factor second language (L2)
887 does not have to have a negative effect on charisma. Direction and strength of the
888 L2 effect seem to depend less on the comprehensibility of the foreign accent or the
889 command of the foreign language than on the prestige of foreign and native language.
890 Regarding the communication channel, Gallardo and Weiß (2017) found a positive
891 correlation between the signal-compression rate in (mobile) phone communication
892 and listener ratings of charisma-related features. Despite initial emergent answers,
893 there is still a plethora of open questions for all of the factors mentioned above. These
894 open questions must be answered step by step, successively involving more factors.
895 On this basis, we offer a brief outlook.

896 11.5 Conclusion and Outlook

897 The present experiment further supports the results of earlier research by identi-
898 fying attire and prosody as relevant factors in the perception of speaker charisma. In
899 addition, given the considerable effects of the two factors, our findings also support
900 the conclusion of earlier studies that non-lexical factors such as attire and prosody
901 are particularly influential for the perception of speaker charisma; probably more
902 important than the words of a speaker. The paper started with a question: Dress to
903 impress? The answer must clearly be “yes”, especially in the case of men. Unlike
904 women, it seems that men are less able to compensate for a charismatically unfavor-
905 able attire with prosodic means. Women, in turn, should probably be more careful in
906 combining attire and prosody with other factors such as their own physical attractive-
907 ness. Regardless of the gender-specific interactions of attire and prosody, the effects
908 of the two factors in the perception of charismatic speakers are largely additive, both
909 in the positive and in the negative direction.

910 Based on these new findings, the task of follow-up studies must be to further
911 refine and differentiate the very roughly varied attire and prosody conditions of the
912 present study, and to homogenize the attire variable, for which we had to mix-up style
913 and color in order to be able to use authentic field data of real senior leaders. Using

(staged) lab data or field data of less popular speakers (entrepreneurs) could be ways to achieve a greater control of the independent variable conditions. Follow-up studies could also work with A/V videos instead of combinations of photographs and speech, especially if more and richer body-language factors are to be addressed. Two findings are conceivable with using A/V videos. Either the richer body language of videos distracts the raters from the factor attire so much that the latter becomes less relevant than with the photos in this study; or, through the attribution of speaker competence (Pearce & Brommel, 1972), attire functions as a limiting factor, so that any charisma-supporting effects of a richer body language cannot unfold without a favorable attire. In this context, it is also essential to check the charisma attributes used in multi-dimensional rating tasks for their multi-modal suitability. In fact, we believe that the exploration and development of methods for the assessment of speaker charisma or similar socio-communicative concepts is a field of research in its own right. Methods need a solid empirical foundation and have to meet certain standards in terms of their internal validity, exhaustiveness, contextual vulnerability, and sensitivity. Regarding the contributions in this volume as well as the recent developments in human-machine interaction and the growing intercultural and digital communication, it is obvious that the experimental investigation of charisma and similar socio-communicative concepts becomes a topic of growing relevance and urgency.

References

- Abidi, M., & Gumpert, K. (2018). *Cross-cultural comparison of speeches and pitches*. Seminar thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, School of Business and Economics, Department of Technology Management, Germany.
- Agno, J., & McEwen, B. (2011). Decoding the executive woman's dress code. Kestly. <http://kestlydevelopment.com/hosted/ExecutiveWomansDressCodeRevised.pdf>.
- Andersson, N. (2009). *Stereotypes of English in hollywood movies. A case study of the use of different varieties of english in Star Wars, The lord of the rings and transformers*. BA thesis, University of Stockholm, Sweden.
- Antonakis, J., Fenley, M., & Liechti, S. (2011). Can charisma be taught? Tests of two interventions. *The Academy of Management Learning and Education*, 10, 374–96.
- Antonakis, J., Liechti, S., & Fenley, M. (2012). Learning charisma. *Harvard Business Review*. <https://hbr.org/2012/06/learning-charisma-2>
- Antonakis, J., Bastardo, N., & Jacquart, P. (2016). Charisma: An ill-defined and ill-measured gift. *Annual Review of Organizational Psychology and Organizational Behavior*, 3, 293–319.
- Auty, S., & Elliott, R. (1998). Fashion involvement, self-monitoring and the meaning of brands. *Journal of Product & Brand Management*, 7, 109–123.
- Awamleh, R., & Gardner, W. L. (1999). Perceptions of leader charisma and effectiveness: The effects of vision content, delivery, and organizational performance. *The Leadership Quarterly*, 10, 345–373.
- Bachsleitner, N., & Popp, U. (2018). *Gender-related impact of the speech rate on the perception of charisma*. Seminar thesis, Friedrich-Alexander-Universität Erlangen-Nürnberg, School of Business and Economics, Department of Technology Management, Germany.
- Bailey, R. W. (2003). Ideologies, attitudes, and perceptions. In D. R. Preston (Ed.), *Needed research in American Dialects* (pp. 123–150). Durham: Duke University Press.

- 958 Bayard, D., Weatherall, A., Gallois, C., & Pittam, J. (2001). Pax Americana: Accent attitudinal
959 evaluations in New Zealand, Australia, and America. *Journal of Sociolinguistics*, 5, 22–49.
- 960 Beckman, M. E., Hirschberg, J., & Shattuck-Hufnagel, S. (2005). The original ToBI system and
961 the evolution of the ToBI framework. In S.-A. Jun (Ed.), *Prosodic typology—the phonology of*
962 *intonation and phrasing* (pp. 9–54). Oxford: Oxford University Press.
- 963 Behling, D. U., & Williams, E. A. (1991). Influence of dress on perception of intelligence and
964 expectations of scholastic achievement. *Clothing and Textiles Research Journal*, 9, 1–7.
- 965 Berger, S., Niebuhr, O., & Peters, B. (2017, March). Winning over an audience—A perception-based
966 analysis of prosodic features of charismatic speech. In Proceedings of 43rd Annual Conference
967 of the German Acoustical Society, Kiel, Germany (pp. 1454–1457).
- 968 Biadys, F., Rosenberg, A., Carlson, R., Hirschberg, J., Strangert, E. (2008). A cross-cultural compar-
969 ison of American, Palestinian, and Swedish perception of charismatic speech. In *Proceedings of*
970 *4th International Conference of Speech Prosody, Campinas, Brazil* (pp. 579–582).
- 971 Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glott International*, 5, 341–345.
- 972 Bosker, H. R. (2007). The role of temporal amplitude modulations in the political arena: Hillary
973 Clinton vs. Donald Trump. In *Proceedings of 18th Interspeech Conference, Stockholm, Sweden*
974 (pp. 1–5).
- 975 Brillman, M. (2015). *A multimodal predictive model of successful debaters—Or how i learned to*
976 *sway votes*. MA thesis, University of Twente, The Netherlands.
- 977 Brooks, A. W., Huang, L., Kearney, S. W., & Murray, F. E. (2014). Investors prefer entrepreneurial
978 ventures pitched by attractive men. *Proceedings of National Academy of Sciences of the United*
979 *States of America (PNAS)*, 111, 4427–4431.
- 980 Chen, L., Feng, G., Joe, J., Leong, C. W., Kitchen, C., & Lee, C. M. (2014). Towards automated
981 assessment of public speaking skills using multimodal cues. In *Proceedings of 16th International*
982 *Conference on Multimodal Interaction* (pp. 200–203).
- 983 Davis, B. C., Hmieleski, K. M., Webb, J. W., & Coombs, J. E. (2017). Funders’ positive affective
984 reactions to entrepreneurs’ crowd-funding pitches: The influence of perceived product creativity
985 and entrepreneurial passion. *Journal of Business Venturing*, 32, 90–106.
- 986 D’Errico, F., Signorello, R., Demolin, D., & Poggi, I. (2013). The perception of charisma from
987 voice: A cross-cultural study. In *Proceedings of Affective Computing and Intelligent Interaction*
988 (pp. 552–557).
- 989 Ernestus, M., & Smith, R. (2018). Qualitative and quantitative aspects of phonetic variation in Dutch
990 eigenlijk. In F. Cangemi, M. Clayards, O. Niebuhr, B. Schuppler, & M. Zellers (Eds.), *Rethinking*
991 *reduction: Interdisciplinary perspectives on conditions, mechanisms, and domains for phonetic*
992 *variation* (pp. 129–163). Berlin/Boston: De Gruyter Mouton.
- 993 Fox Cabane, O. (2012). *The charisma myth: How anyone can master the art and science of personal*
994 *magnetism*. New York: Penguin.
- 995 Franz, T., & Norton, D. S. (2001). Investigating business casual dress policies: Questionnaire
996 development and exploratory research. *Applied HRM Research*, 6, 79–94.
- 997 Friedman, H. S., Riggio, R. E., & Casella, D. F. (1988). Nonverbal skill, personal charisma, and
998 initial attraction. *Personality and Social Psychology Bulletin*, 14, 203–211.
- 999 Furnham, A., & Petrova, E. (2010). *Body language in business: Decoding the signals*. Grand Street:
1000 Palgrave Macmillan.
- 1001 Furnham, A., Chan, P. S., & Wilson, E. (2014). What to wear? The influence of attire on the perceived
1002 professionalism of dentists and lawyers. *Journal of Applied Psychology*, 43, 1838–1885.
- 1003 Gallardo, L. F., & Weiß, B. (2017). Towards speaker characterization: Identifying and predicting
1004 dimensions of person attribution. In *Proceedings of 18th Interspeech Conference, Stockholm,*
1005 *Sweden* (pp. 904–908).
- 1006 Gélinas-Chebat, C., Chebat, J. C., & Vaninsky, A. (1996). Voice and advertising: Effects of intonation
1007 and intensity of voice on source credibility, attitudes and the intend to buy. *Perceptual and Motor*
1008 *Skills*, 83, 243–262.
- 1009 Grabo, A., Spisak, B., & van Vugt, M. (2017). Charisma as signal: An evolutionary perspective on
1010 charismatic leadership. *The Leadership Quarterly*, 28, 473–485.

- 1011 Grd, P. (2013). Introduction to age estimation using face images. *Research Papers Faculty of*
 1012 *Material Science and Technology Slovak University Bratislava* (vol. 21, pp. 24–30)
- 1013 Hiroyuki, T., & Rathcke, T. (2016). Then, what is charisma? The role of audio-visual prosody in
 1014 L1 and L2 political speeches. In *Proceedings of Phonetik & Phonologie im deutschsprachigen*
 1015 *Raum, Munich, Germany* (pp. 1–3).
- 1016 Holladay, S. J., & Coombs, W. T. (1994). Speaking of visions and visions being spoken an explo-
 1017 ration of the effects of content and delivery on perceptions of leader charisma. *Management*
 1018 *Communication Quarterly*, 8, 165–189.
- 1019 Hoover, M. (2013). *Dressing to impress: The secrets of proper attire*. Manuscript, the Florida State
 1020 University, Career Center.
- 1021 Howlett, N., Pine, K. J., Orakçioğlu, I., & Fletcher, B. (2013). The influence of clothing on first
 1022 impressions: Rapid and positive responses to bespoke features in male attire. *Journal of Fashion,*
 1023 *Marketing and Management*, 17, 38–48.
- 1024 Jokisch, O., Iaroshenko, V., Maruschke, M., & Ding, H. (2018). Influence of age, gender and sample
 1025 duration on the charisma assessment of German speakers. In *Proceedings of 29th Konferenz für*
 1026 *Elektronische Sprachsignalverarbeitung, Ulm, Germany* (pp. 1–8).
- 1027 Jongman, A., Qin, Z., Zhang, J., & Sereno, J. A. (2017). Just noticeable differences for pitch
 1028 direction, height, and slope for Mandarin and English listeners. *The Journal of the Acoustical*
 1029 *Society of America*, 142, EL163–EL169.
- 1030 Karabell, S. (2016). Dressing like a leader: Style tips for women in the spotlight. *Forbes*
 1031 *Magazin*. [https://www.forbes.com/sites/shelliekarabell/2016/01/16/dressing-like-a-leader-style-](https://www.forbes.com/sites/shelliekarabell/2016/01/16/dressing-like-a-leader-style-tips-for-women-in-the-spotlight/)
 1032 [tips-for-women-in-the-spotlight/](https://www.forbes.com/sites/shelliekarabell/2016/01/16/dressing-like-a-leader-style-tips-for-women-in-the-spotlight/).
- 1033 Kohler, K. J. (2000). The future of phonetics. *Journal of the International Phonetic Association*,
 1034 30, 1–24.
- 1035 Ladegaard, H. J. (1998). National stereotypes and language attitudes: The perception of British,
 1036 American and Australian language and culture in Denmark. *Language Communication*, 18, 251–
 1037 274.
- 1038 Lee, M. (2014). Transformational leadership: is it time for a recall? *International Journal Of*
 1039 *Management and Applied Research*, 1(1), 17–29.
- 1040 McCabe, D., & Altman, K. W. (2017). Prosody: An overview and applications to voice therapy.
 1041 *Global Journal of Otolaryngology*, 7, 1–8.
- 1042 Madill, C. (2015). Keep an eye on vocal fry—It’s all about power, status and gender. The conver-
 1043 sation. Retrieved October 31, 2018, from [http://theconversation.com/keep-an-eye-on-vocal-fry-](http://theconversation.com/keep-an-eye-on-vocal-fry-its-all-about-power-status-and-gender-45883)
 1044 [its-all-about-power-status-and-gender-45883](http://theconversation.com/keep-an-eye-on-vocal-fry-its-all-about-power-status-and-gender-45883).
- 1045 Molloy, J. (1977). *The women’s dress for success book*. Chicago: Follet.
- 1046 Molloy, J. T. (1996). *New women’s dress for success*. New York: Warner.
- 1047 Morris, T. L., Gorham, J., Cohen, S. H., & Huffman, D. (1996). Fashion in the classroom: Effects of
 1048 attire on student perceptions of instructors in college classes. *Communication Education*, 45(2),
 1049 135–148.
- 1050 Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for
 1051 text-to-speech synthesis using diphones. *Speech Communication*, 9, 453–467.
- 1052 Niebuhr, O., Brem, A., Novák-Tóth, E., & Voße, J. (2016). Prosodic constructions of charisma
 1053 in business speeches—A contrastive acoustic analysis of Steve Jobs and Mark Zuckerberg. In
 1054 *Proceedings of 8th International Conference of Speech Prosody, Boston, USA* (pp. 1–3).
- 1055 Niebuhr, O., Voße, J., & Brem, A. (2016). What makes a charismatic speaker? A computer-based
 1056 acoustic prosodic analysis of Steve Jobs tone of voice. *Computers and Human Behavior*, 64,
 1057 366–382.
- 1058 Niebuhr, O., Tegtmeier, S., & Brem, A. (2017). Advancing research and practice in entrepreneurship
 1059 through speech analysis—From descriptive rhetorical terms to phonetically informed acoustic
 1060 charisma metrics. *Journal of Speech Sciences*, 6, 3–26.
- 1061 Niebuhr, O., Thumm, J., & Michalsky, J. (2018a). Shapes and timing in charismatic speech—
 1062 Evidence from sounds and melodies. In *Proceedings of 9th International Conference of Speech*
 1063 *Prosody, Poznan, Poland* (pp. 582–586).

- 1064 Niebuhr, O., Skarnitzl, R., & Tylečková, L. (2018b). The acoustic fingerprint of a charismatic voice—
 1065 Initial evidence from correlations between long-term spectral features and listener ratings. In
 1066 *Proceedings of 18th International Conference of Speech Prosody, Poznań, Poland* (pp. 359–363).
- 1067 Novák-Tót, E., Niebuhr, O., & Chen, A. (2017). A gender bias in the acoustic-melodic features
 1068 of charismatic speech? In *Proceedings of Annual Conference of the International Speech*
 1069 *Communication Association* (vol. 18, pp. 2248–2252).
- 1070 Pearce, W. B., & Brommel, B. J. (1972). Vocalic communication in persuasion. *Quarterly Journal*
 1071 *of Speech*, 58(3), 298–306.
- 1072 Pentland, A. (2008). *Honest signals—How they shape our world*. Cambridge: MIT Press.
- 1073 Pépiot, E. (2013). Voice, speech and gender: male-female acoustic differences and cross-language
 1074 variation in English and French speakers. In *Proceedings of 15th Rencontres Jeunes Chercheurs*
 1075 *de l'ED 268, Paris, France* (pp. 1–13).
- 1076 Power, M. R., & Galvin, C. (1997). The culture of speeches: Public speaking across cultures. *Culture*
 1077 *Mandala: The Bulletin of the Centre for East-West Cultural and Economic Studies*, 2, 2.
- 1078 Quené, H. (2004). What is the just noticeable difference for tempo in speech? In H. Quené & V.
 1079 van Heuven (Eds.), *On Speech and Language: Studies for Sieb G. Nooteboom* (pp. 149–158).
 1080 Utrecht: Netherlands Graduate School of Linguistics. LOT Occasional Series 2.
- 1081 Radeloff, D. J. (1990). Role of color in perception of attractiveness. *Perceptual and Motor Skills*,
 1082 71, 151–160.
- 1083 Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and*
 1084 *new media like real people and places*. New York: Cambridge University Press.
- 1085 Rosenberg, A., & Hirschberg, J. (2009). Charisma perception from text and speech. *Speech*
 1086 *Communication*, 51, 640–655.
- 1087 Scherer, S., Layher, G., Kane, J., Neumann, H., Campbell, N. (2012). An audiovisual political speech
 1088 analysis incorporating eye-tracking and perception data. In *Proceedings of 8th International*
 1089 *Conference on Language Resources and Evaluation* (pp. 1114–1120).
- 1090 Schötz, S. (2006). *Perception, Analysis, and Synthesis of Speaker Age*. Ph.D. thesis, Lund University,
 1091 Sweden.
- 1092 Shim, H. S., Park, S., Chatterjee, M., Scherer, S., Sagae, K., Morency, L. -P. (2015). Acoustic
 1093 and paraverbal indicators of persuasiveness in social multimedia. In *Proceedings of IEEE*
 1094 *International Conference on Acoustics, Speech and Signal Processing* (pp. 1–8).
- 1095 Signorello, R., D'Errico, F., Poggi, I., Demolin, D. (2012). How charisma is perceived from
 1096 speech: A multidimensional approach. Privacy, security, risk and trust (PASSAT). In *International*
 1097 *Conference on Social Computing (SocialCom), Amsterdam, The Netherlands* (pp. 435–440).
- 1098 Silverberg, D. (2017). Do the colours you wear at work matter? *BBC Business*. <https://www.bbc.co.uk/news/business-41003867>
- 1100 Ten Bosch, L., Oostdijk, N., & de Ruiter, J. P. (2004). Turn-taking in social dialogues: Temporal,
 1101 formal and functional aspects. In *Proceedings SPECOM, St. Petersburg*.
- 1102 Terhardt, E. (1974). Pitch, consonance, and harmony. *Journal of the Acoustic Society of America*,
 1103 55, 1061–1069.
- 1104 Touati, P. (1993). Prosodic aspects of political rhetoric. In *Proceedings of ESCA Workshop on*
 1105 *Prosody, Lund, Sweden*, 168–171.
- 1106 Towler, A. J. (2003). Effects of charismatic influence training on attitudes, behavior, and
 1107 performance. *Personnel Psychology*, 56, 363–381.
- 1108 van Santen, J. P. H. (1994). Assignment of segmental duration in text-to-speech synthesis. *Computer*
 1109 *Speech & Language*, 8, 95–128.
- 1110 Weber, M. (1947). *The theory of social and economic organization*. New York: The Free Press of
 1111 Glencoe.
- 1112 Weirich, M., & Simpson, A. P. (2014). Differences in acoustic vowel space and the perception of
 1113 speech tempo. *Journal of Phonetics*, 43, 1–10.
- 1114 Xu, Y., & Sun, X. (2002). Maximum speed of pitch change and how it may relate to speech. *Journal*
 1115 *of the Acoustical Society of America*, 111, 1399–1413.

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Voice Attractiveness	
Series Title		
Chapter Title	Birds of a Feather Flock Together But Opposites Attract! On the Interaction of F0 Entrainment, Perceived Attractiveness, and Conversational Quality in Dating Conversations	
Copyright Year	2020	
Copyright HolderName	Springer Nature Singapore Pte Ltd.	
Corresponding Author	Family Name	Michalsky
	Particle	
	Given Name	Jan
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	University of Oldenburg
	Address	Ammerländer Heerstraße 114-118, 26129, Oldenburg, Germany
	Email	j.michalsky@uni-oldenburg.de
Author	Family Name	Schoormann
	Particle	
	Given Name	Heike
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	University of Oldenburg
	Address	Ammerländer Heerstraße 114-118, 26129, Oldenburg, Germany
	Email	heike.schoormann@uni-oldenburg.de
Abstract	<p>Dating conversations are especially influenced by the interlocutors' perceived attractiveness. As visual attractiveness determines the course and nature of the interaction, the perceived overall quality of the conversation may also be influenced by the perceived attractiveness and simultaneously also affect the further development of the conversation. Accordingly, perceived attractiveness and conversational quality constantly interact in dating conversations. Studies focusing on the effects of both impressions on a speaker's vocal behavior in terms of prosodic entrainment, i.e., the adaptation of a speaker's prosodic features relative to his/her interlocutor, suggest that higher visual attractiveness leads to a greater divergence in f0 in mixed-sex pairs, while greater conversational quality results in larger degrees of f0 entrainment. In this paper, we further investigate the effects of both perceived attractiveness and conversational quality on prosodic entrainment of f0 in dating conversations with a special focus on their interaction. We conducted a dating experiment with 20 young heterosexual singles who engaged in 100 short spontaneous mixed-sex dating conversations. The results suggest that f0 entrainment correlates with both perceived attractiveness and conversational quality. Prosodic entrainment decreased with higher ratings of perceived attractiveness and increased with higher ratings of perceived conversational quality. Additionally, the results indicate that f0 entrainment not only depends on the impressions of attractiveness and conversational quality but also affects them. Furthermore, seemingly conflicting effects may be resolved by emphasizing one effect over the other, e.g., quality over attractiveness. This emphasis seems to depend on speaker sex and may also change during the course of the conversation. The details of this</p>	

complex interaction, their interdependence, the importance of speaker sex, as well as possible implications are discussed.

Keywords

Attractiveness - Conversational quality - Likability - Dating - Entrainment - Accommodation - Adaptation
- F0

Chapter 12

Birds of a Feather Flock Together But Opposites Attract! On the Interaction of F0 Entrainment, Perceived Attractiveness, and Conversational Quality in Dating Conversations



Jan Michalsky and Heike Schoormann

Abstract Dating conversations are especially influenced by the interlocutors' perceived attractiveness. As visual attractiveness determines the course and nature of the interaction, the perceived overall quality of the conversation may also be influenced by the perceived attractiveness and simultaneously also affect the further development of the conversation. Accordingly, perceived attractiveness and conversational quality constantly interact in dating conversations. Studies focusing on the effects of both impressions on a speaker's vocal behavior in terms of prosodic entrainment, i.e., the adaptation of a speaker's prosodic features relative to his/her interlocutor, suggest that higher visual attractiveness leads to a greater divergence in f0 in mixed-sex pairs, while greater conversational quality results in larger degrees of f0 entrainment. In this paper, we further investigate the effects of both perceived attractiveness and conversational quality on prosodic entrainment of f0 in dating conversations with a special focus on their interaction. We conducted a dating experiment with 20 young heterosexual singles who engaged in 100 short spontaneous mixed-sex dating conversations. The results suggest that f0 entrainment correlates with both perceived attractiveness and conversational quality. Prosodic entrainment decreased with higher ratings of perceived attractiveness and increased with higher ratings of perceived conversational quality. Additionally, the results indicate that f0 entrainment not only depends on the impressions of attractiveness and conversational quality but also affects them. Furthermore, seemingly conflicting effects may be resolved by emphasizing one effect over the other, e.g., quality over attractiveness. This emphasis seems to depend on speaker sex and may also change during the course of the conversation. The details of this complex interaction, their interdependence, the importance of speaker sex, as well as possible implications are discussed.

J. Michalsky (✉) · H. Schoormann
University of Oldenburg, Ammerländer Heerstraße 114-118, 26129 Oldenburg, Germany
e-mail: j.michalsky@uni-oldenburg.de

H. Schoormann
e-mail: heike.schoormann@uni-oldenburg.de

© Springer Nature Singapore Pte Ltd. 2020
B. Weiss et al. (eds.), *Voice Attractiveness*, Prosody, Phonology and Phonetics,
https://doi.org/10.1007/978-981-15-6627-1_12

221

25 **Keywords** Attractiveness · Conversational quality · Likability · Dating ·
 26 Entrainment · Accommodation · Adaptation · F0

27 12.1 Introduction

28 12.1.1 Prosodic Entrainment and Its Role in Interaction

29 Most if not all prosodic features bear a high functional load on several commu-
 30 nicative levels. Pitch, intensity, or speaking rate, for example, can convey linguistic
 31 functions such as focus (cf. Ladd 2008), paralinguistic meanings such as a speaker's
 32 emotions or attitudes (cf. Scherer, Ladd, & Silverman, 1984; Ladd, Silverman, Tolkmitt,
 33 Bergmann, & Schere, 1985), while simultaneously providing extra-linguistic
 34 information such as the sex or age of a speaker (cf. Linville, 1996) within the same
 35 stretch of speech. A change in prosodic features, such as increasing the speaking rate,
 36 reducing intensity, or raising the f0 mean, can reflect the social relationship of two
 37 speakers, e.g., in terms of social status (cf. Gregory, 1996) or dominance (cf. Puts,
 38 Gaulin, & Verdolini, 2006), while signaling attitudes and emotions that in turn affect
 39 and influence the interpersonal relationship. However, a phenomenon that has been
 40 linked to signaling and influencing interpersonal relationships has not been observed
 41 in the way prosodic features vary by themselves, i.e., in absolute terms, but in the
 42 way they change relative to the correspondent prosodic features of the interlocutor.

43 Entrainment, also often referred to as accommodation, convergence, or adapta-
 44 tion among others, describes this observed interdependence, i.e., speakers adjusting
 45 their linguistic features to those of the interlocutor particularly by becoming more
 46 similar (cf. Levitan, 2014). Entrainment can occur on all linguistic levels and may
 47 lead to an adaptation of the lexical choice (Brennan & Clark, 1996) and the syntac-
 48 tic structure (Reitter & Moore, 2007) but it can also influence prosodic features by
 49 matching speaking rate (Schweitzer, Lewandowski & Duran, 2017), intensity (cf.
 50 Levitan, 2014), or aspects of the fundamental frequency (cf. Levitan, 2014). Edlund,
 51 Heldner, and Hirschberg (2009) as well as Levitan (2014; see also Sect. 11.2.3) distin-
 52 guish three types of prosodic entrainment which need to be differentiated. Proximity
 53 describes two interlocutors becoming similar with respect to a prosodic feature dur-
 54 ing a conversation, convergence describes two interlocutors becoming increasingly
 55 more similar during the course of a conversation, and synchrony describes a relative
 56 adaptation to the dynamics of an interlocutor's prosodic feature without necessarily
 57 becoming more similar in absolute terms.

58 There are two explanatory approaches to the occurrence of entrainment in human
 59 communication. Although they are often considered to be competing and mutually
 60 exclusive, we suggest that both approaches complement each other. According to
 61 the *communication model* (Natale, 1975) as well as the *perception behavior link*
 62 (Chartrand & Bargh, 1999), entrainment can be regarded as a device to enhance
 63 intelligibility by matching speaking styles and thus facilitating the identification of

64 phonological categories by reducing phonetic variability. Accordingly, entrainment
65 is a more or less automatic human behavior. This approach is supported by the fact
66 that we also find entrainment in non-social interaction with synthetic voices used by
67 machine applications (cf. Gessinger et al., 2018). The *communication accommoda-*
68 *tion theory* (Giles, Coupland, & Coupland, 1991) among others, however, assumes
69 an iconic relationship between entrainment and social distance with smaller linguist-
70 ic differences signaling closeness on a social level. This approach thus suggests that
71 entrainment is not a mere automatism in interaction but it is dependent on the social
72 relationship.

73 The focus of this paper lies on the role of f_0 in signaling the relationship between
74 interlocutors and a speaker's perceived attitude toward an interlocutor, respectively.
75 Specifically, we study the connection between f_0 entrainment and social distance
76 in the situational setting of dating conversations. One factor that has a stronger
77 influence in the current setting compared to other communicative situations is the
78 perceived visual attractiveness of the interlocutor as dating conversations involves
79 mating intention. Although especially important in mating contexts, the perceived
80 attractiveness affects most if not every kind of conversation from everyday small talk
81 to business communication (cf. Cialdini, 2009; Brooks, Huang, Kearney, & Murray,
82 2014). As of yet, it is largely unknown how perceived attractiveness interacts with
83 prosodic entrainment and the perceived pleasantness of a conversation, henceforth
84 referred to as conversational quality. The connection between the perceived visual
85 attractiveness of the interlocutor, the perceived conversational quality, and a speaker's
86 change in fundamental frequency constitutes the objective of the study at hand.

87 ***12.1.2 Prosodic Entrainment and Perceived Conversational*** 88 ***Quality***

89 Assuming a link between prosodic entrainment and social distance, the question
90 arises how social distance was measured in previous studies. Rather than measured
91 directly, social distance was approached as a construct derived from a wide variety
92 of social features associated with closeness such as mutual liking (Levitan et al.,
93 2012), support (Street, 1984; Levitan et al., 2012), giving encouragement (Nenkova,
94 Gravano, & Hirschberg, 2008), or higher degrees of collaboration and cooperation
95 (Lubold & Pon-Barry, 2014). We can assume that social closeness is reflected in the
96 perceived quality of the conversation and will thus regard conversational quality as
97 a predictor for social distance in the framework of this study. As we assess conver-
98 sational quality through the subjective evaluation of the interlocutors, any further
99 mention of conversational quality refers to the perceived conversational quality.

100 First evidence for the connection between entrainment and conversational quality
101 stems from it signaling a closer connection between interlocutors and resulting in
102 higher degrees of communicative success. Entrainment as an indicator for task suc-
103 cess has been described for several different tasks. Thomason, Nguyen, and Litman
104 (2013) report that student engineering groups that showed higher degrees of entrain-

105 ment also showed better task results. Similar observations also hold for map task
 106 experiments (Reitter & Moore, 2007) as well as student tutoring programs (Fried-
 107 berg, Litman, & Paletz, 2012). According to the theory of alignment (Pickering &
 108 Garrod, 2006), entrainment is a crucial contributor to communicative success in
 109 general. Lubold and Pon-Barry (2014) suggest that entrainment is connected to col-
 110 laboration and rapport in learning tasks which also positively affects communicative
 111 success and thereby task success. Similarly, Taylor (2009) and Beňuš (2014) pro-
 112 pose that task success greatly depends on the establishment of a common situational
 113 model, a process which is facilitated by the coordination of behavior. Accordingly,
 114 becoming closer with respect to verbal and non-verbal behavior might facilitate the
 115 construction of a common situational model.

116 How does task success relate to conversational quality? In other words, what
 117 is the goal of a non-task-oriented conversation? Although this is a rather difficult
 118 question to answer extensively, we can regard the establishment of a social bond
 119 as a major goal of verbal communication (cf. Dunbar, 2020). This is even more
 120 apparent in dating conversations where the establishment of a social bond serves as
 121 the basis for a romantic relationship that can be regarded as an explicit rather than an
 122 implicit goal (Hewstone, Stroebe, & Jonas, 2012: 870ff).¹ We can assume that the
 123 quality of a conversation greatly affects a conversations' ability to establish and/or
 124 improve the social relationship of two interlocutors. Accordingly, conversational
 125 quality can be regarded as the non-task-oriented equivalent of collaboration, affecting
 126 communicative success by affecting the social relationship.

127 Although previous studies on entrainment have for the most part been linked
 128 to either a speaker's perception of his/her interlocutor or the previously mentioned
 129 task success, there are some studies on meanings more closely related to conversa-
 130 tional quality. Gonzales, Hancock, and Pennebaker (2009) found entrainment to be
 131 correlated to overall dialogue quality. Ireland et al. (2011) report that entrainment
 132 predicts the probability of initiating romantic relationships as well as the stability
 133 of existing relationships. In marriage counseling dialogues, Lee et al. (2010) found
 134 higher degrees of entrainment when couples were talking about positive rather than
 135 negative topics. Furthermore, entrainment was reported to result in smoother conver-
 136 sation with respect to turn latencies and fewer interruptions which can be regarded as
 137 attributes of high quality in conversations (Nenkova et al., 2008). Lastly, Michalsky
 138 et al. (2018) also found conversational quality and entrainment to be connected in
 139 dating conversations with smaller differences in f0 occurring in conversations that
 140 were perceived as more pleasant.

141 In conclusion, although conversational quality has rarely been assessed explicitly
 142 within the respective studies, we expect conversations that are perceived as better
 143 or more pleasant to show a higher degree of prosodic entrainment. This expectation
 144 applies to conversations in general and specifically to dating conversations.

¹However, this is only true if we restrict our investigation to dating conversations which aim at finding a potential partner, which of course is not true for every kind of dating conversation.

12.1.3 Prosodic Entrainment and Perceived Attractiveness

The topic of vocal attractiveness received a lot of attention not only from a phonetic or even linguistic perspective but also from a sociological and psychological perspective. Furthermore, the immediate connection between attraction and aspects of evolutionary biology has generated assumptions that lead to specific linguistic hypotheses. Although this paper focuses on how speakers react prosodically to perceived attractiveness, i.e., the attracted voice, the underlying assumption is that we react to attractiveness by trying to sound more attractive (cf. Hughes, Farley, & Rhodes, 2010; Fraccaro et al., 2011). Accordingly, speakers would try to imitate features of attractive voices when perceiving their interlocutor as more attractive. To this end, a short overview on the prosodic features of attractive voices will be provided.

What prosodic features contribute to the impression of vocal attractiveness is a complex topic and cannot be solely and maybe not even primarily attributed to voice pitch. However, fundamental frequency as the acoustic correlate of voice pitch is the commonly studied feature of vocal attractiveness. The main reason for this can be found in the frequency code (Ohala, 1983, 1984) which assumes an evolutionary connection between pitch and attractiveness. In animal mating behavior, female individuals show the general tendency to select bigger and stronger male individuals to ensure protection as well as survival of their offspring. Accordingly, size is a biological factor in natural selection. While many species developed strategies to project size visually, others employ strategies to signal largeness through vocal features. Since due to physiological reasons larger individuals generally have a lower fundamental frequency, certain species such as wolves use lower pitch to suggest largeness. As largeness plays a role in selecting a partner for female individuals rather than males, it is associated with masculinity while smallness and high pitch are associated with femininity.

In general, studies confirmed these findings for human communication. Female listeners were found to evaluate male voices as significantly more attractive when they were realized with a lower f_0 mean (Collins, 2000; Feinberg, Debruine, Jones, & Perrett, 2005; Hodges-Simeon, Gaulin, & Puts, 2010; Jones Feinberg, Debruine, Little, & Vukovic, 2010; Xu, Lee, Wu, Liu, & Birkholz, 2013) while male listeners judged female voices with a higher f_0 mean as more attractive (Collins & Missing, 2003; Feinberg et al., 2008; Jones et al., 2010; Xu et al., 2013). However, the results for male listeners and thus female voices were not consistent. Oguchi and Kikuchi (1997) as well as Leaderbrand et al. (2008) suggest that female voices are perceived as more attractive when realized with a lower f_0 mean. One explanation for this contradiction is provided by Karpf (2006) who proposed two different types of female attractiveness. Following Karpf's (2006) distinction, lower pitch is associated with the concept of sexiness and seductiveness while high pitch is associated with femininity. However, both are perceived as attractive female voices in general. Another explanation may be found in the communicative setting and thus the communicative intent. There are several goals such as intimacy goals, identity goals, or status goals

sought in a relationship (cf. Zimmer-Gembeck, Hughes, Kelly, & Connoly, 2011) that can affect whether individuals are looking for anything from short-term flings to long-term relationships as well as different qualities sought in a partner associated with different goals which may lead to different concepts of attractiveness. However, this assumption has never been incorporated into experimental studies on vocal attractiveness.

In addition to the general features of male and female vocal attractiveness, the following findings are of relevance to the study at hand. Firstly, Vukovic et al. (2010) report that the perception of pitch as a cue to attractiveness not only depends on the speaker's absolute pitch but also on the listener's own average pitch. Furthermore, Borokowska and Pawlowski (2011) found a threshold at which an increase or decrease in mean fundamental frequency, respectively, does not increase perceived attractiveness any further. Lastly, Fraccaro et al. (2011) point toward the importance of naturalness and context when investigating perceived attractiveness as this feature seems to be especially susceptible to artificiality.

Most studies on vocal attractiveness commonly avoid defining the concept of attractiveness altogether. As evident from the inconsistent findings for female voices, listeners may employ a variety of different concepts of attractiveness when judging vocal attractiveness. However, we propose that investigating the prosodic effects of perceived visual attractiveness allow us to dispense with this problem and the need for defining the concept. Although listeners may still have a variety of reasons to perceive another person as more or less attractive, the result of the perceived attractiveness should always be attractive which can be connected to a physiological reaction and should therefore be more or less consistent across individuals (cf. Fraccaro et al., 2011). Although speakers may still employ different vocal strategies to express attraction, those differences are most likely not caused by differences in the concept of attractiveness that caused said attraction. Accordingly, the concept of attractiveness should be largely independent of the effects found for perceived attractiveness.

The effects of perceived attractiveness of an interlocutor on a speaker's f_0 seem to confirm the assumption that speakers react to perceived attractiveness by mimicking the features of attractive voices and thereby trying to sound more attractive themselves. Male speakers who interacted with more attractive female interlocutors were found to lower their f_0 mean (Hughes et al., 2010). For female speakers, however, we again find contradicting results. Female speakers were found to lower their f_0 mean (Hughes et al., 2011) when talking to a more attractive male interlocutor as well as to raise their f_0 mean under the same conditions (Fraccaro et al., 2011). According to Fraccaro et al. (2011) this may be explained through different experimental settings with varying degrees of contextual naturalness. In addition, the differences could again be related to the two different concepts of female attractiveness suggested by Karpf (2006). However, this assumption not only implies that male listeners have two different concepts of attractiveness associated with female voices but also that female speakers readily employ these two different concepts when signaling attraction.

How the prosodic effects of perceived visual attractiveness relate to prosodic entrainment has not been studied prior to Michalsky and Schoormann (2017) but there

233 are some conclusions to be drawn from the research described above. Studies suggest
234 that male speakers lower their f_0 while female speakers, at least in some cases, raise
235 their f_0 when interacting with a more attractive interlocutor. Since male speakers
236 on average have a lower f_0 mean than female speakers for physiological reasons,
237 both effects result in the speakers increasing the distance in f_0 and thus showing
238 what is called *prosodic disentrainment*. Michalsky and Schoormann (2017, 2018)
239 suggest that this connection of prosodic disentrainment and perceived attractiveness
240 can indeed be found in spontaneous dating conversations. A recent study by Beňuš
241 et al. (2018) suggests that disentrainment can lead to the impression of dominance,
242 which, according to the frequency code (Ohala, 1983, 1984), can be associated with
243 masculinity. Yet, these results obtained from human–machine interaction would only
244 support the hypothesis for the female listeners and not for the male listeners. A study
245 by Schweitzer et al. (2017) suggest that there might also be effects of entrainment
246 related to attractiveness. However, their findings are restricted to speaking rate and
247 not f_0 and furthermore focused on the concept of social attractiveness in same-sex
248 dialogues.

249 In conclusion, we expect perceived attractiveness to result in prosodic disentrain-
250 ment, directly contradicting the effects we expect for conversational quality.

251 **12.1.4 The Dilemma: Good Conversations with Attractive** 252 **Interlocutors**

253 Regarding the effects of conversational quality and perceived attractiveness on
254 prosodic entrainment we arrive at the preliminary expectation that higher conversa-
255 tional quality would result in social closeness and thus smaller differences in prosodic
256 features, i.e., prosodic entrainment, while perceived attractiveness results in larger
257 prosodic differences and hence prosodic disentrainment. This contradiction poses a
258 challenge since conversational quality and perceived attractiveness not only operate
259 on the same prosodic feature (f_0 mean) while pointing in opposite directions but also
260 because we expect both social parameters to highly influence dating conversations
261 and thus to frequently co-occur and even interact. As such, we need to ask what hap-
262 pens with a speaker's f_0 in conversations with high conversational quality and high
263 perceived attractiveness, i.e., in conversations in which we would expect prosodic
264 entrainment as well as prosodic disentrainment?

265 One hypothesis is that one effect overrules the other, i.e., signaling either conversa-
266 tional quality or perceived attractiveness is more important in dating conversations
267 and thus only one of the contradicting effects is observable in this conversational
268 setting.

269 A second hypothesis would be that the effects of perceived attractiveness are sensi-
270 tive to the naturalness and context of the interaction. Higher perceived attractiveness
271 may result in disentrainment only when investigated specifically in a mating context
272 with scripted messages as done by Fraccaro et al. (2011) while possibly enhancing

273 the effects of conversational quality by strengthening the social bond and thereby
 274 leading to more entrainment in spontaneous dating scenarios. However, this would
 275 contradict our previous finding on perceived attractiveness in dating conversations
 276 (Michalsky & Schoormann 2017, 2018).

277 Thirdly, the effects of perceived attractiveness and conversational quality may
 278 cancel each other out. This assumption would, however, entail a very uneconomic
 279 use of social signals. Accordingly, where f_0 fails to signal conversational quality and
 280 perceived attractiveness simultaneously, other prosodic parameters may assume this
 281 function. Unfortunately, the scope of this paper is limited to f_0 .

282 Lastly, although effects of perceived attractiveness and conversational quality
 283 may co-occur within the same conversation, they need not occur simultaneously.
 284 One possibility is that perceived attractiveness is based on a first impression and the
 285 signaling of attraction hence decisive for the initiation of a conversation. Accordingly,
 286 the effects may be restricted to the first part of a conversation. Conversational quality
 287 on the other hand, develops over time and peaks during the course of the conversation.
 288 The effects of conversational quality may, therefore, be the strongest in the later
 289 parts of the conversation when the effects of perceived attractiveness have declined.
 290 Another distribution might regard different topics or even different intentions during
 291 the conversation. There may be phases where interlocutors are predominantly flirting,
 292 showing stronger effects of perceived attractiveness, and others where interlocutors
 293 are bonding, showing stronger effects of conversational quality. While we investigate
 294 only the former question by looking at different time points of the conversation, the
 295 latter remains for future research.

296 The study at hand was designed to improve on three shortcomings encountered
 297 in the previous research. Firstly, most studies investigate either perceived attractive-
 298 ness or conversational quality. We suggest that if not explicitly asked to separate the
 299 two, speakers are inclined to let the two notions influence each other. Accordingly,
 300 we expect the judgement on perceived attractiveness to be influenced by the over-
 301 all conversational quality and in return the impression of conversational quality to
 302 be compromised by the attractiveness of the interlocutor. Although this interdepend-
 303 ence can never be totally excluded, explicitly instructing participants to judge both
 304 impressions on different scales is a first approach to telling them apart by raising
 305 awareness of the potential conflict.

306 Secondly, perception ratings are often taken from external observers rather than
 307 from the subjects participating in the study. Since the perception of attractiveness as
 308 well as conversational quality can and will greatly vary between participants actually
 309 partaking in the respective conversations and external observers, all judgements in
 310 this study are taken directly from the interlocutors.

311 Lastly, there are two possible perspectives regarding the connection of prosodic
 312 entrainment and social variables with respect to causality that are frequently separ-
 313 ated and rarely both investigated within the same studies. On the one hand, the social
 314 relationship of two interlocutors can manifest itself in prosodic entrainment which
 315 thus serves as an indicator for the social relationship. On the other hand, prosodic
 316 entrainment may in return affect the social relationship and even facilitate the estab-
 317 lishment of social bonds. Accordingly, we can either ask how the relationship affects

318 prosodic entrainment but also how prosodic entrainment affects a social relation-
 319 ship. In this study, we incorporate both views to shed some light on the question of
 320 correlation and causality, although a definite answer to that question is categorically
 321 impossible.

322 This study is based on the same corpus as some of our previous work on the topic
 323 (cf. Michalsky, 2017; Michalsky & Schoormann, 2016, 2017, 2018; Michalsky et al.,
 324 2018, 2018). We would like to inform the reader about the possibility of conflicting
 325 information. Our previous research on the topic constitutes work in progress on a
 326 growing corpus with changing normalization methods and shifting focus regarding
 327 the f0 parameters in question. Since the results presented in this paper constitute the
 328 final state of the analysis, the information presented in this paper explicitly replaces
 329 older information.

330 *Perceived visual attractiveness*

- 331 1. Does the perception of visual attractiveness in an opposite-sex interlocutor sys-
 332 tematically correlate with a speaker's f0 entrainment in accordance with previous
 333 findings?
- 334 2. Do changes in a speaker's f0 entrainment correlate with an interlocutor's percep-
 335 tion of the speaker in terms of visual attractiveness?
- 336 3. Are these two effects connected in a systematic way?

337 *Perceived conversational quality*

- 338 1. Does the perceived conversational quality systematically correlate with a speaker's
 339 f0 entrainment in accordance with previous findings?
- 340 2. Do changes in a speaker's f0 entrainment correlate with an interlocutor's percep-
 341 tion of the conversational quality?
- 342 3. Are these two effects connected in a systematic way?

343 *Perceived visual attractiveness and conversational quality*

- 344 1. Do the effects of perceived attractiveness and conversational quality interact in
 345 their influence on f0 entrainment?
- 346 2. Does f0 entrainment show contradicting or complementing effects on the percep-
 347 tion of attractiveness and conversational quality?

348 **12.2 Method**

349 *12.2.1 Subjects*

350 The study was conducted with 20 participants, 10 female, and 10 male, all paid
 351 volunteers and at the time of the experiment students at the University of Oldenburg.
 352 All subjects were aged between 19 and 28 and monolingual speakers of High German
 353 who spent the majority of their lives in Lower Saxony. Furthermore, all subjects

354 reported to be heterosexual as well as single during the whole course of the study.
 355 With the exception of two speakers, whose conversation was excluded from the
 356 experiment, all subjects were previously unacquainted. All subjects were informed
 357 about the nature of the experiment as a dating situation.

358 **12.2.2 Procedure**

359 All participants were informed about the dating setting of the experiment prior to
 360 their recordings. Female and male participants waited in separated rooms and were
 361 led to the recording rooms via separate staircases to avoid any interaction prior to
 362 their actual conversations. Each participant was paired with every other participant of
 363 the opposite sex resulting in a total of 100 opposite-sex conversations, all recorded
 364 in two parallel recording sessions in two quiet separate university office rooms.
 365 All recordings were done within two weeks during spring break. The use of the
 366 phonetic laboratory was explicitly avoided to ensure a more natural setting based on
 367 the importance of naturalness in evaluating attractiveness by Fraccaro et al. (2011).
 368 The participants were encouraged to engage in spontaneous conversations of 15–
 369 20 min without any restrictions or guidelines regarding the choice of conversational
 370 topics. However, example topics were provided in case conversations were stalling
 371 and subjects needed inspiration.

372 Immediately before each conversation all participants judged their respective
 373 interlocutor on a 10-point Likert scale with respect to their perceived visual attractive-
 374 ness and general likability. The participants were separated by a screen that allowed
 375 them to see each other's faces but concealed the questionnaire so that the evaluations
 376 were not revealed to the interlocutor. The screen was removed at the beginning of
 377 the conversation. Directly after each conversation, the participants received another
 378 questionnaire and repeated the covert evaluation of perceived visual attractiveness
 379 and general likability. Furthermore, a third scale was added to this second question-
 380 naire to evaluate how pleasant the subjects perceived the conversation as a whole to
 381 assess conversational quality.

382 Recordings were made in stereo using head-mounted microphones (DPA 4065
 383 FR) to ensure an optimal balance between recording quality and naturalness. We
 384 used a portable digital recorder (Tascam HD P2) at a sampling rate of 48 kHz and
 385 24-bit resolution.

386 **12.2.3 Types and Measurements of Entrainment**

387 According to Edlund et al. (2009) and Levitan (2014) we can distinguish three dif-
 388 ferent types of entrainment: *proximity*, *convergence*, and *synchrony*.

389 *Proximity* covers what is usually referred to as entrainment, accommodation, or
390 adaptation and describes two speakers being more similar with respect to a linguistic
391 feature when talking to each other than when not talking to each other. Accordingly,
392 proximity needs some sort of reference value by either operating on a local level and
393 comparing the differences of prosodic features at adjacent turns with non-adjacent
394 turns (Levitan, 2014) or globally by comparing the differences during a conversation
395 with differences to other speakers or conversations. In this study we combine the two
396 by comparing the general differences at adjacent turns in correlation with perceived
397 attractiveness as well as conversational quality across conversations.

398 *Convergence* describes increasing proximity over time during the course of a single
399 conversation. Accordingly, we again measure the difference in a linguistic feature at
400 adjacent turns but with respect to their changes during the conversation. Convergence
401 can either be assessed locally by tracking changes from turn to turn or globally, e.g.
402 by comparing the first and second half of a conversation.

403 *Synchrony* constitutes a categorically different type of entrainment that is either not
404 considered at all or assumed as the primary type of entrainment. Synchrony describes
405 the relative adaptation of a speaker's linguistic features to the respective feature of
406 his/her interlocutor by adjusting their values relative to each other without necessarily
407 becoming more similar. For example, a speaker may react to a raised f0 mean of
408 his/her interlocutor by raising his/her own f0 mean by the same amount, thus imitating
409 his/her interlocutor's prosodic behavior without a decrease in the differences between
410 the two as it is the case for proximity or convergence. To measure synchrony, we
411 check for correlations between the prosodic feature of the turn-taking speaker and
412 the turn-passing speaker in adjacent turns of a speaker change inducing turn break. A
413 positive correlation generally points toward synchrony while a negative correlation
414 is often linked to an effect of increased or decreased proximity.

415 12.2.4 Acoustic Analysis

416 For the acoustic analysis we used Praat (Boersma & Weenink, 2016). Since the
417 recordings were made in stereo, we separated the audio tracks for each speaker.
418 The audio tracks were manually annotated for interpausal units (IPU, cf. Levitan,
419 2014). We analyzed all IPUs adjacent to a turn break inducing speaker change.
420 IPUs were defined mechanically by stretches of speech preceded or followed by a
421 pause with a pause defined as an interruption of speech by silence or non-speech
422 noise of at least 500 ms. Accordingly, we made no difference between pauses at
423 phrase boundaries and hesitation pauses in favor of interlabeler reliability. The corpus
424 consists of 14,687 IPUs from 98 conversations. One conversation had to be excluded
425 due to prior acquaintance of the participants another one was lost to a recording
426 error. We extracted the f0 mean from the interpausal units as we suggest that the
427 f0 mean captures the register better than the median (cf. Michalsky & Schoormann,
428 2016; Michalsky et al., 2018). Furthermore, range features at phrase final boundaries

429 are heavily distorted by pragmatic functions and therefore unreliable in capturing
 430 entrainment in this specific data set (cf. Michalsky, 2014, 2015). For synchrony, we
 431 measured the f_0 mean of the IPU of the turn-passing and the turn-taking speaker and
 432 converted it to semitones to a reference value of 50 Hz. We z-transformed the data by
 433 speaker by subtracting the IPU's f_0 mean values from the average f_0 mean value of all
 434 IPUs of each speaker across all conversations and dividing it by the standard deviation
 435 of the same set. For proximity and convergence, we calculated the absolute difference
 436 between the f_0 mean of IPUs adjacent to turn breaks in semitones. Furthermore, we
 437 tagged the IPUs occurring in the first five minutes as well as the last five minutes of
 438 the conversations.

439 12.2.5 Statistical Analysis

440 For the statistical analysis, we conducted linear mixed effects models using *R* (R
 441 Core Team 2017) with the *lme4*-package (Bates, Maechler, Bolker & Walker, 2015)
 442 and the *lmerTest*-package (Kuznetsova, Brockhoff, & Christensen, 2016). Model fit
 443 was determined by maximum likelihood ratio tests. P-values were calculated using
 444 the Satterthwaite approximation. We calculated different models for the effects of
 445 prosodic entrainment on the investigated social variables and the effects of these
 446 social variables on prosodic entrainment.

447 For the effects of perceived ATTRACTIVENESS and CONVERSATIONAL
 448 QUALITY on prosodic entrainment we used different dependent variables with
 449 respect to the type of prosodic entrainment. For *synchrony*, we calculated Pear-
 450 son correlation coefficients (*f0 correlation*) between the f_0 mean of the turn-passing
 451 speaker and the turn-taking speaker, which resembles the degree of synchrony (cf.
 452 Edlund et al., 2009; Levitan, 2014), and used it as the dependent variable. As fixed
 453 factors, we used perceived visual attractiveness (ATTRACTIVENESS), perceived
 454 conversational quality (QUALITY), speaker sex (SEX) and all interactions. In both
 455 the proximity model and the convergence model, we used the difference between the
 456 f_0 means of the IPUs adjacent to turn breaks (*f0 difference*) as dependent variables.
 457 For the proximity model, fixed factors were identical to the synchrony model. For
 458 the convergence model, we added the TIME (in seconds) of the respective turn break
 459 as a fixed factor.

460 For the effects of prosodic entrainment on perceived *attractiveness* and *conver-*
 461 *sational quality*, we calculated two different models for each of the three types of
 462 entrainment with either perceived attractiveness (*attractiveness*) or perceived conver-
 463 sational quality (*conversational quality*) as dependent variables with the respective
 464 counterpart serving as fixed factor (ATTRACTIVENESS or QUALITY). In the syn-
 465 chrony model, we used the correlation coefficients (F0 CORRELATION) described
 466 above as a fixed factor as well as SEX and all interactions. In the proximity model
 467 we used the difference between the IPUs adjacent to turn breaks (F0 DIFFERENCE)
 468 as well as SEX and their interaction as fixed factors. For the convergence model, we

again expanded the proximity model by the fixed factor TIME and the additional possible interactions.

Since we suggest that the effects of perceived attractiveness and conversational quality may affect different parts of the conversation to different degrees, we conducted post hoc tests for every model described above separated by conversational part. The variable conversational part splits the data set into turns occurring within the first five minutes of each conversation and turns occurring within the last five minutes of each conversation to see whether the effects are restricted to certain conversational parts. Note that this leads to a substantial reduction of the data set and may result in statistically insignificant effects due to insufficient data points. However, this was only done for proximity as effects for convergence were already absent from the entire conversation and the Pearson correlation coefficients calculated for synchrony were not robust enough for splitting the data set into thirds.

12.3 Results

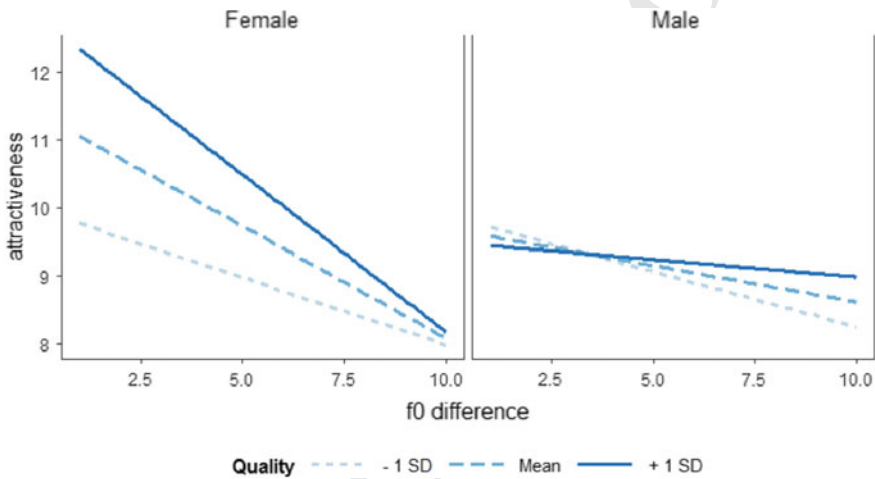
12.3.1 Effects of Perceived Attractiveness and Conversational Quality on Prosodic Entrainment

12.3.1.1 Proximity

Table 12.1 presents the results for the effects of perceived ATTRACTIVENESS and conversational QUALITY on proximity. We find significant interactions between the two factors as well as a three-way-interaction with SPEAKER SEX illustrated in Fig. 12.1. Accordingly, we conducted post hoc tests separated by SPEAKER SEX to investigate the nature of this three-way-interaction. Tables 12.2 and 12.3 present the post hoc results for the female and the male speakers, respectively. Table 12.2 shows that the male speakers show significant effects for both perceived ATTRACTIVENESS and QUALITY without interactions although marginal effects are suggested by Fig. 12.1. Male speakers decrease their *f0 differences* between turns with increasing conversational QUALITY and increase these differences with increasing visual ATTRACTIVENESS of the interlocutor. For the female speakers we find a significant interaction between ATTRACTIVENESS and conversational QUALITY (s. Table 12.3). Female speakers also decrease their *f0 differences* with increasing CONVERSATIONAL QUALITY and increase *f0 differences* with visual ATTRACTIVENESS. However, as shown in Fig. 12.1, the effects for ATTRACTIVENESS become smaller with increasing conversational QUALITY. This means that female speakers do react less to the perceived ATTRACTIVENESS of their interlocutor when the conversation is perceived as highly positive. In conversations with below average QUALITY, however, ATTRACTIVENESS significantly correlates with the degree of *proximity*.

Table 12.1 Significant main effects and interactions of perceived ATTRACTIVENESS and CONVERSATIONAL QUALITY on *proximity*

Fixed factors	<i>b</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
ATTRACTIVENESS	0.74	0.09	14560.00	8.32	<0.001
SEX	3.11	0.99	129.90	3.16	<0.01
ATTRACTIVENESS * QUALITY	-0.07	0.01	14570.00	-5.27	<0.001
ATTRACTIVENESS * SEX	-0.84	0.15	14570.00	-5.72	<0.001
QUALITY * SEX	-0.34	0.12	14570.00	-2.79	<0.01
ATTRACTIVENESS * QUALITY * SEX	0.10	0.02	14570.00	4.83	<0.001

**Fig. 12.1** Interaction of the effects of perceived ATTRACTIVENESS, QUALITY, and SEX on *f0 difference***Table 12.2** Post hoc significant main effects and interactions of perceived ATTRACTIVENESS and CONVERSATIONAL QUALITY on *proximity* for the male speakers

Fixed factors	<i>b</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
ATTRACTIVENESS	0.11	0.04	7375.06	2.86	<0.01
QUALITY	-0.10	0.04	7372.77	-2.65	<0.01

Table 12.3 Post hoc significant main effects and interactions of perceived ATTRACTIVENESS and CONVERSATIONAL QUALITY on *proximity* for the female speakers

Fixed factors	<i>b</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
ATTRACTIVENESS	0.74	0.09	7100.57	8.10	<0.001
ATTRACTIVENESS * QUALITY	-0.07	0.01	7179.94	-5.17	<0.001

Table 12.4 Post hoc significant main effects and interactions of perceived ATTRACTIVENESS and CONVERSATIONAL QUALITY on *proximity* for the male speakers for the first five minutes of a conversation

Fixed factors	<i>b</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
QUALITY	-0.41	0.16	2661.52	-2.49	<0.05
QUALITY * ATTRACTIVENESS	0.05	0.03	2661.27	1.97	<0.05

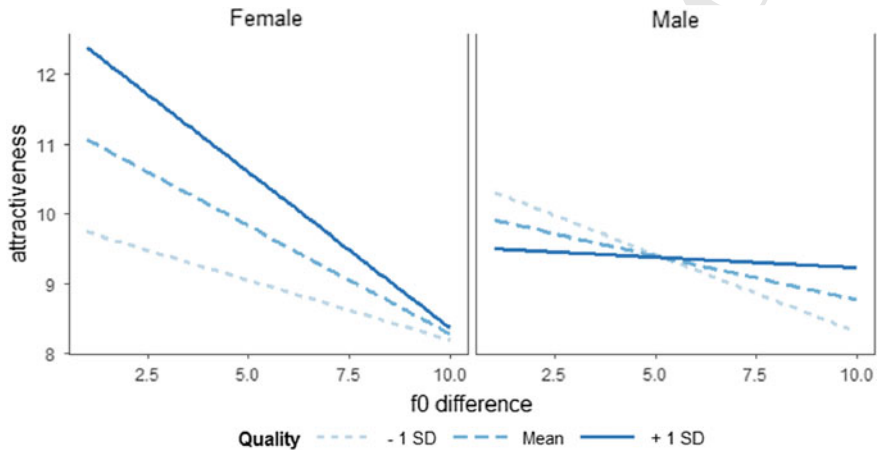


Fig. 12.2 Interaction of the effects of perceived ATTRACTIVENESS, QUALITY, and SEX on *f0 difference* in the first five minutes

506 The post hoc investigation of the conversational parts reveals that female speakers
 507 behave consistently throughout the entire conversation and show the same effects as
 508 reported above for the first as well as the last five minutes. For the male speakers, how-
 509 ever, the effects change over the course of the conversation. While the effects for the
 510 last five minutes are identical to the effects we found for the whole conversation, we
 511 find a deviation in the first five minutes. As shown in Table 12.4, male speakers show
 512 a significant interaction between ATTRACTIVENESS and conversational QUAL-
 513 ITY in the first five minutes. Although this interaction also seems to be present in the
 514 whole conversation when comparing Figs. 12.1 and 12.2, it only reaches significance
 515 for the first five minutes. This interaction is different to the one found for the female
 516 speakers. Figure 12.2 shows that the effects of conversational QUALITY become
 517 smaller with increasing perceived ATTRACTIVENESS. Accordingly, while for the
 518 female speakers conversational QUALITY overruled ATTRACTIVENESS through-
 519 out the whole conversation, for the male speakers, ATTRACTIVENESS overrules
 520 conversational QUALITY. In other words, male speakers do entrain less in pleasant
 521 conversations with attractive women. However, this effect is restricted to the first five
 522 minutes and is found for neither the last 5 min nor the conversational as a whole.

Table 12.5 Significant main effects for *convergence*

Fixed factors	<i>b</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
TIME	-0.0005	0.0001	14550.0000	-4.2530	<0.001

523 12.3.1.2 Convergence

524 Table 12.5 shows that there is a significant effect for TIME on the *f0 differences* at turn
 525 breaks. Accordingly, we find a general effect for *convergence* with speakers becoming
 526 more similar to each other over time. However, this effect shows no interaction with
 527 either perceived ATTRACTIVENESS or conversational QUALITY. Hence, while
 528 we find effects of conversational QUALITY and perceived ATTRACTIVENESS on
 529 entrainment, the observed general convergence is not enhanced by the social variables
 530 investigated.

531 12.3.1.3 Synchrony

532 Table 12.6 reports the effects of perceived ATTRACTIVENESS and conversational
 533 QUALITY on *synchrony*. In contrast to *proximity*, we find no significant effects for
 534 SPEAKER SEX or any interaction with SEX. Accordingly, we find main effects for
 535 ATTRACTIVENESS and conversational QUALITY as well as their interaction for
 536 both sexes. Figure 12.3 illustrates the interaction between ATTRACTIVENESS and
 537 conversational QUALITY. We find that increasing ATTRACTIVENESS is correlated
 538 with greater synchrony if conversational QUALITY is low but correlates with less
 539 synchrony if conversational QUALITY is high. The same is true for the opposite per-
 540 spective. Increasing conversational QUALITY is correlated with stronger *synchrony*
 541 if the perceived ATTRACTIVENESS is low but is correlated with lower synchrony
 542 if the perceived ATTRACTIVENESS is high.

Table 12.6 Significant main effects and interactions of perceived ATTRACTIVENESS and CONVERSATIONAL QUALITY on *synchrony*

Fixed factors	<i>b</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
ATTRACTIVENESS	0.05	0.02	174.50	2.90	<0.01
QUALITY	0.05	0.01	190.18	3.44	<0.001
ATTRACTIVENESS * QUALITY	-0.01	0.00	187.27	-3.04	<0.01

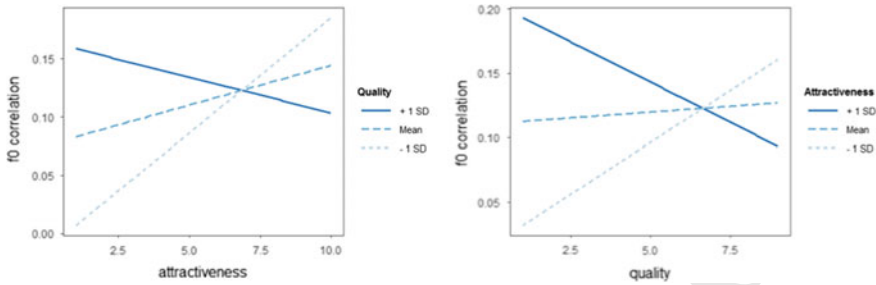


Fig. 12.3 Interaction of the effects of perceived ATTRACTIVENESS and QUALITY on *f0 correlation*

12.3.2 Effects of Prosodic Entrainment on Perceived Attractiveness and Conversational Quality

12.3.2.1 Proximity

Table 12.7 presents the effects of PROXIMITY on perceived *attractiveness*. We find significant effects for the three-way-interaction between F0 DIFFERENCE, QUALITY, and SEX. Figure 12.4 illustrates this three-way-interaction while Tables 12.8 and 12.9 report the post hoc results separated by SEX. For both sexes, we find a significant interaction between F0 DIFFERENCE and QUALITY. In general, both female and male speakers show a tendency to judge speakers as more *attractive* if they show greater F0 DIFFERENCES and hence a greater degree of disentrainment. However, female speakers show strong effects of F0 DIFFERENCE for *attractiveness* if QUALITY is low or average but close to no effects if QUALITY is high. Male speakers on the other hand show noticeable effects for F0 DIFFERENCE if QUALITY is high and less pronounced effects if QUALITY is average or low.

Table 12.10 presents the effects of PROXIMITY on perceived *quality*. Comparable to *attractiveness*, we find significant effects for the three-way-interaction between F0 DIFFERENCE, ATTRACTIVENESS, and SEX. Figure 12.5 illustrates this three-

Table 12.7 Significant main effects and interactions of PROXIMITY on perceived *attractiveness*

Fixed factors	<i>b</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
F0 DIFFERENCE	0.10	0.01	14570.00	9.39	<0.001
QUALITY	0.58	0.02	14580.00	34.78	<0.001
SEX	1.03	0.32	38.96	3.19	<0.01
F0 DIFFERENCE * QUALITY	-0.01	0.00	14570.00	-6.98	<0.001
F0 DIFFERENCE * SEX	-0.12	0.02	14570.00	-7.06	<0.001
F0 DIFFERENCE * QUALITY * SEX	0.02	0.00	14570.00	6.58	<0.001

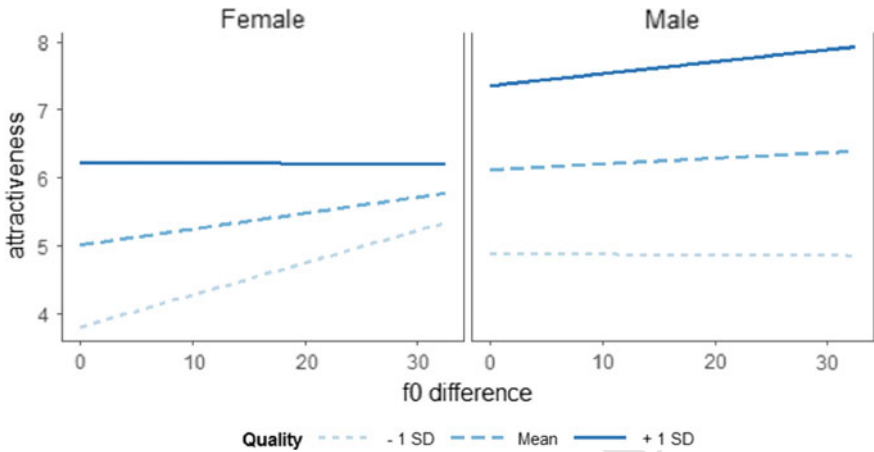


Fig. 12.4 Interaction plot for the effects of F0 DIFFERENCE, QUALITY, and SEX on perceived attractiveness

Table 12.8 Post hoc significant main effects and interactions of PROXIMITY on perceived attractiveness for the female speakers

Fixed factors	<i>b</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
F0 DIFFERENCE	0.10	0.01	7272.00	9.63	<0.001
QUALITY	0.58	0.02	7274.00	35.64	<0.001
F0 DIFFERENCE * QUALITY	-0.01	0.00	7271.00	-7.15	<0.001

Table 12.9 Post hoc significant main effects and interactions of PROXIMITY on perceived attractiveness for the male speakers

Fixed factors	<i>b</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
QUALITY	0.59	0.02	7306.00	32.14	<0.001
F0 DIFFERENCE * QUALITY	0.00	0.00	7302.00	2.43	<0.05

Table 12.10 Significant main effects and interactions of PROXIMITY on perceived conversational quality

Fixed factors	<i>b</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
ATTRACTIVENESS	0.93	0.02	14580.00	40.55	<0.001
SEX	1.14	0.38	38.24	3.01	<0.01
F0 DIFFERENCE * ATTRACTIVENESS	-0.01	0.00	14570.00	-4.48	<.001
ATTRACTIVENESS * SEX	-0.27	0.03	14580.00	-8.20	<0.001
F0 DIFFERENCE * ATTRACTIVENESS * SEX	0.01	0.00	14570.00	3.19	<0.01

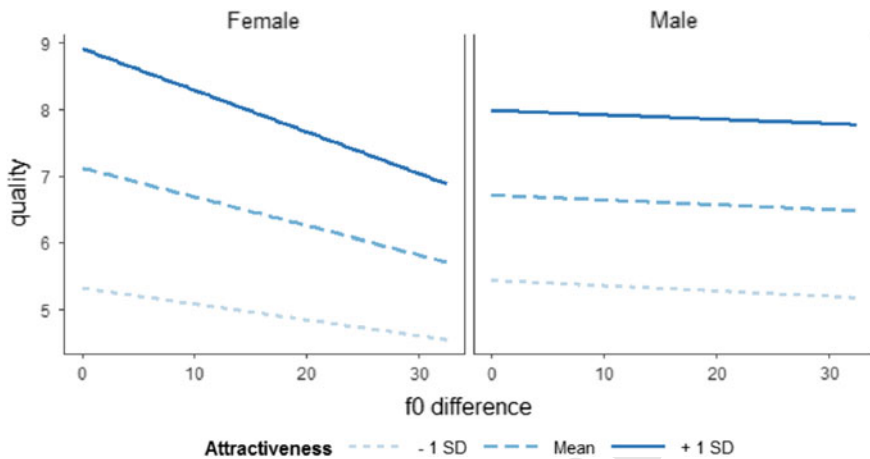


Fig. 12.5 Interaction plot for the effects of F0 DIFFERENCE, ATTRACTIVENESS, and SEX on perceived *quality*

Table 12.11 Post hoc significant main effects and interactions of PROXIMITY on perceived *conversational quality* for the female speakers

Fixed factors	<i>b</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
ATTRACTIVENESS	0.93	0.03	7279.00	36.81	<0.001
F0 DIFFERENCE * ATTRACTIVENESS	-0.01	0.00	7273.00	-4.07	<0.001

Table 12.12 Post hoc significant main effects and interactions of PROXIMITY on perceived *conversational quality* for the male speakers

Fixed factors	<i>b</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
ATTRACTIVENESS	0.66	0.01	7304.00	72.57	<0.001

560 way-interaction while Tables 12.11 and 12.12 report the post hoc results separated
 561 by SEX. Again, we see a common general tendency for both sexes but in contrast
 562 to perceived *attractiveness*, both sexes judge the conversation as better if the inter-
 563 locutor shows smaller F0 DIFFERENCE and hence greater degrees of entrainment.
 564 Again, these effects interact with the other social variable, in this case perceived
 565 ATTRACTIVENESS. For the female speakers, Fig. 12.5 shows that the effects of
 566 F0 DIFFERENCE on *quality* increase with perceived ATTRACTIVENESS, which
 567 is statistically significant in the post hoc test reported in Table 12.11. Male speakers
 568 show the same tendency but as shown in Fig. 12.5, the effects are much smaller and
 569 do not reach statistical significance in the post hoc test (s. Table 12.12).

570 The post hoc investigation of the conversational parts shows that female speakers
 571 show the same effects for *attractiveness* as for the conversations as a whole within
 572 both the first and the last five minutes of the conversation. For the male speakers,

Table 12.13 Post hoc significant main effects and interactions of PROXIMITY and ATTRACTIVENESS on perceived *conversational quality* for the male speakers for the last five minutes of a conversation

Fixed factors	<i>b</i>	<i>SE</i>	<i>df</i>	<i>t</i>	<i>p</i>
F0 DIFFERENCE	-0.02	0.01	2615.00	-2.81	<0.01
ATTRACTIVENESS	0.65	0.02	2615.00	43.00	<0.001

573 however, we only find significant effects for the conversation as a whole but not for
 574 the conversational parts. With respect to Fig. 12.4, we suspect that the effects for the
 575 conversation as a whole are already too small and are hence lost when splitting the
 576 data set.

577 Comparable to perceived *attractiveness*, the effects for *quality* are robust for the
 578 female speakers within the conversational parts. Female speakers show the same
 579 positive effects of entrainment on *quality* for both the start and the end of the con-
 580 versation. The male speakers, however, show deviating effects from the conversation
 581 as a whole, which also point in the opposite direction. While we find effects on
 582 *attractiveness* for the whole conversation but not for the parts, we find the opposite
 583 for *quality*. While entrainment does not significantly correlate with *quality* in the
 584 conversation as whole, we find a significant effect of F0 DIFFERENCE on *quality*
 585 in the part subsets (s. Table 12.13). Furthermore, these effects only occur in the last
 586 five minutes of the conversation but not in the first five minutes. Lastly, F0 DIFFER-
 587 ENCE does not interact with perceived ATTRACTIVENESS in contrast to any other
 588 entrainment effects reported in this chapter.

589 12.3.3 Convergence and Synchrony

590 Comparable to the effects of perceived attractiveness and conversational quality on
 591 *convergence* (s. chapter 12.3.1.2) we find no significant effects for convergence on
 592 either variable. However, in contrast to the effects found for perceived attractive-
 593 ness and conversational quality on *synchrony* (s. chapter 12.3.1.3) we also find no
 594 significant effects for synchrony on either variable.

595 12.4 Discussion

596 The results show that there is a strong connection between prosodic entrainment and
 597 both perceived visual attractiveness and conversational quality. We find that prosodic
 598 entrainment reflects the social relationship by showing effects for the perceived visual
 599 attractiveness of an interlocutor as well as effects for the perceived quality of the
 600 conversation. Furthermore, the degree of prosodic entrainment correlates with how

601 pleasant a speaker perceives a conversation as well as how visually attractive s/he
602 perceives his/her interlocutor. However, while there are core effects that suggest
603 a direct interpretation and are in accordance with previous studies as well as the
604 expectations given in chapter one, there are several findings that pose a challenge for
605 future research. Especially the synchrony effects, the reciprocity of the connection
606 between entrainment and social variables, as well as the interaction of perceived
607 attractiveness and conversational quality leave many open questions as discussed in
608 the following.

609 *12.4.1 Effects of Perceived Attractiveness and Conversational* 610 *Quality on Prosodic Entrainment*

611 Perceived attractiveness and conversational quality both significantly correlate with
612 the degree to which a speaker entrains to his/her interlocutor. However, both variables
613 correlate with entrainment differently and with notable differences depending on
614 speaker sex. In general, both female and male speakers show greater degrees of
615 entrainment in terms of proximity if they perceive the conversation as better. This
616 is compatible with our expectations from the link between prosodic entrainment
617 and conversational quality as well as social distance in general (cf. Nenkova et al.,
618 2008; Gonzales et al., 2009; Levitan et al., 2012). In contrast, both sexes show greater
619 degrees of disenitainment in conversations with more visually attractive interlocutors.
620 This is also in accordance with our expectations from previous research on the effects
621 of visual attractiveness on prosody in general (cf. Hughes et al., 2010; Fraccaro et al.,
622 2011). However, this also means that the effects are indeed diametrically opposing
623 each other.

624 For the female speakers, this results in a significant interaction between attrac-
625 tiveness and conversational quality with respect to entrainment. The effects of attrac-
626 tiveness decrease with higher degrees of conversational quality and are even absent
627 in conversations that are perceived as very pleasant. Accordingly, female speakers
628 emphasize conversational quality over attractiveness in terms of entrainment. This is
629 consistent across the entire conversation. The opposite is true for the male speakers.
630 Here we also find a significant interaction but male speakers show decreasing effects
631 of conversational quality as attractiveness increases. Accordingly, male speakers
632 emphasize visual attractiveness over conversational quality. However, this is only
633 true for the first five minutes of the conversation and neither for the last five minutes
634 nor the conversation as a whole. Hence, male speakers emphasize visual attractive-
635 ness when first engaging in a conversation but show more balanced prosodic effects
636 for both variables as the conversation emerges.

637 The picture is less clear for the other types of prosodic entrainment. Perceived
638 visual attractiveness and conversational quality both correlate with synchrony. How-
639 ever, the effects are difficult to interpret. Both variables show a positive correlation
640 with the degree of synchrony if the respective other variable is low, negatively if the

641 other variable is high, and marginally or not at all if the other one is average. We
 642 suggest that synchrony as measured in this study reflects the complex relationship
 643 between attractiveness and conversational quality and cannot be interpreted in its
 644 own right. Furthermore, we find a general convergence effect, i.e., a general trend
 645 for speakers to become more similar over time. However, this trend is independent
 646 from either social variable.

647 ***12.4.2 Effects of Prosodic Entrainment on Perceived*** 648 ***Attractiveness and Conversational Quality***

649 The effects of prosodic entrainment on perceived attractiveness and conversational
 650 quality show a nearly reciprocal relationship with the effects reported above. Both
 651 sexes judge conversations as better where the interlocutor shows a greater degree
 652 of prosodic entrainment in the form of proximity. Although the literature on the
 653 effects of entrainment on perception is scarce, these effects are in line with our
 654 expectations (cf. Nenkova et al., 2008; Gonzales et al., 2009; Levitan et al., 2012).
 655 Furthermore, both male and female speakers perceive interlocutors who show greater
 656 degrees of disentrainment as more visually attractive. This is also in line with our
 657 expectations since disentrainment generally leads male speakers to lower their voices
 658 and female speakers to raise their voices, which was found to increase perceived
 659 attractiveness (cf. Collins, 2000; Collins and Missing, 2003; Feinberg et al., 2005,
 660 2008; Hodges-Simeon et al., 2010; Jones et al. 2010; Xu et al., 2013). However, it is
 661 not the mere distinction between low and high which is connected to attractiveness
 662 but specifically the distance caused by disentrainment. An interlocutor's pitch is thus
 663 evaluated within his/her own natural register and not in absolute terms as comparable
 664 across speakers. Again, the effects of perceived attractiveness and conversational
 665 quality are contradicting.

666 For the female speakers, the effects of perceived attractiveness and conversational
 667 quality interact significantly. The effects of entrainment on conversational quality
 668 become stronger with an increased perceived visual attractiveness of the interlocutor.
 669 Simultaneously, the effects of attractiveness become weaker the better the conversa-
 670 tion. Both interactions are consistent across the entire conversation. Accordingly, we
 671 find the same dominance of conversational quality over visual attractiveness reported
 672 above. Again, the picture is vastly different for the male speakers. While the effects of
 673 entrainment on perceived attractiveness are statistically independent from conversa-
 674 tional quality, the effects of entrainment on conversational quality become weaker the
 675 more attractive the interlocutor. Accordingly, the male speakers again show a domi-
 676 nance of attractiveness over conversational quality. Furthermore, we find another
 677 effect compatible with the results reported above. While the dominance of visual
 678 attractiveness over conversational quality on entrainment disappears after the first 5
 679 min of the conversation, the effects of entrainment on conversational quality only
 680 appear after the first 5 min. Hence, although the male speakers show a general domi-

681 nance of visual attractiveness over conversational quality, these effects shift over time
682 with attractiveness being emphasized when first engaging in a conversation while
683 the perception of conversational quality manifests its effects in the later parts of the
684 conversation.

685 In contrast to the effects of the social variables on entrainment, we find no effects
686 of synchrony on either visual attractiveness or conversational quality. Again, we
687 suggest that this may be related to the complex interaction between the two social
688 factors. Furthermore, the data set may be too small for the reliable application of
689 Pearson correlation coefficients as measurements for synchrony (cf. Edlund et al.,
690 2009; Levitan, 2014), since we found effects of synchrony using other although
691 cruder measurements. The size of the data set could also explain the absence of
692 convergence effects.

693 **12.4.3 The Dilemma: Good Conversations with Attractive** 694 **Interlocutors**

695 This chapter has shown that high degrees of conversational quality and visual attrac-
696 tiveness within the same conversation do indeed lead to contradicting effects as
697 expected from the introduction. However, with respect to our initial expectations, we
698 do not find one factor completely canceling out the other. Instead, the results suggest
699 a weighting of the two variables. While female speakers emphasize conversational
700 quality, male speakers generally emphasize perceived attractiveness. Accordingly, if
701 both perceived attractiveness and conversational quality are high, female speakers
702 tend to show stronger entrainment while male speakers show stronger disentrain-
703 ment. This observation is complemented by the finding that male speakers also show
704 a shift in weighting. While female speakers consistently emphasized conversational
705 quality over attractiveness, male speakers show a tendency to emphasize perceived
706 attractiveness when first engaging in a conversation and then shifting the focus to
707 conversational quality as the conversation progresses. Accordingly, with respect to
708 our initial expectation we find both one factor overruling the other as well as differ-
709 ences in distribution across a conversation. However, we did not find an association
710 of different types of entrainment with different social variables.

711 A factor not considered within this study concerns differences in the weighting of
712 these social variables not only in their distribution by time but also by topic. Accord-
713 ingly, there may be conversational topics that are thematically closer to mating and
714 hence show a higher demand for signaling attractiveness versus topics closer related
715 to forming stronger bonds and hence related to signaling conversational quality com-
716 parable to the effects found for positive versus negative topics by Lee et al. (2010).
717 Such a topic related shift in signaling social variables would also mirror and thus add
718 to the interpretation of the effects observed for the male speakers as a higher density
719 of mating related topics in the first half of the conversation compared to the more
720 bonding related topics in the last half seems likely.

721 Lastly, the weighting of perceived attractiveness may also be related to person-
 722 ality. Emphasizing perceived conversational quality over perceived visual attrac-
 723 tiveness could be related to factors such as emotional empathy or agreeableness.
 724 Consequently, the differences we find for speaker sex may actually not be related to
 725 speaker sex itself but to gender-related personality attributes.

726 **12.4.4 Additional Thoughts and Further Implications**

727 Before discussing some further implications of the results, we would like to address
 728 two observations regarding the experiment itself to clarify the possible generaliz-
 729 ability of our results. Prior to the experiment, we assessed personal data from all
 730 participants including their intent to participate in the study. As reported above, all
 731 participants were informed that the experiment was designed as a dating study. How-
 732 ever, only three participants stated that they were actually interested in dating and
 733 eventually finding a partner. Furthermore, all of these three participants were male.
 734 The remaining participants all declared to be merely interested in having good conver-
 735 sations and meeting new people. Accordingly, the majority of the participants did not
 736 intend to date prior to the conversations or at least did not admit it. Inspecting the con-
 737 versations with respect to content leads to a mixed result. While most conversations
 738 confirm the assessment by a lack of flirting, several conversations suggest a strong
 739 intent for dating. One pair even requested to exchange contact information although
 740 both participants did not declare to be interested in finding a partner. Accordingly,
 741 engaging in a dating conversation is not necessarily something that happens inten-
 742 tionally. Furthermore, participants may just not be willing to admit their intent when
 743 participating in a scientific study. The fact, that we do find strong effects for perceived
 744 attractiveness may support this conclusion. However, as pointed out above, perceived
 745 attractiveness may play a strong role even in non-dating conversations which shifts
 746 the focus of the generalizability of this study.

747 The second observation regards the naturalness of the conversations. Although
 748 initially most participants were irritated by the setting and often commented on the
 749 recording situation, this issue quickly dissipated in most conversations. Overall, we
 750 perceive the majority of the conversations as resembling natural interactions. The
 751 participants engaged freely in spontaneous dialogues, choosing a wide variety of
 752 different topics and transitioning fluently between them. There are also several cases
 753 of participants talking about the researchers, other university staff, or even sharing
 754 personal information which they were explicitly instructed not to reveal, suggesting
 755 that participants quickly forgot about being recorded.

756 With respect to the ongoing debate about the function of entrainment, our study
 757 supports both categories of assumptions. The effects of perceived conversational
 758 quality strongly support the *communication accommodation theory* (Giles et al.,
 759 1991) and related models which link social closeness to greater entrainment. Greater
 760 conversational quality can be related to a stronger social bond between the interlocu-
 761 tors and hence a greater degree of social closeness. However, we also find an effect

762 of categorical convergence, which is not affected by either conversational quality or
763 attractiveness. Accordingly, speakers become categorically closer with respect to f0
764 over time. These effects support the assumption of entrainment as an automatism, for
765 example, to enhance intelligibility by matching speaking styles as suggested by the
766 *communication model* (Natale 1975) or the *perception behavior link* (Chartrand &
767 Bargh, 1999) among others. Lastly, the effects of perceived attractiveness allow for
768 two possible interpretations. On the one hand, perceived attractiveness may primarily
769 affect f0 lowering or raising with disentrainment just being a logical consequence
770 and not a feature in itself. On the other hand, the effect of social distance, which is
771 linked to disentrainment (cf. Giles et al., 1991), may actually be the primary effect.
772 Accordingly, there may be sociological/psychological reasons why a higher degree
773 of social distance is linked to greater perceived attractiveness.

774 We suggest that our findings should be generalizable to non-dating conversations
775 to some degree. As described above, the participants mostly stated that they did
776 not intend to flirt or date. Accordingly, we can characterize the conversations as a
777 hybrid of natural conversations in a dating setting leading to real dating conversa-
778 tions in some cases. Hence, we expect the effects of perceived attractiveness and
779 conversational quality to be slightly less pronounced in real non-dating mixed-sex
780 conversations and more pronounced in real intended dating conversations but present
781 in both.

782 Another follow-up question concerns the generalizability to same-sex dating con-
783 versations. The particular question regards the two possible interpretations of the
784 findings on perceived attractiveness as primarily leading to a raising or lowering in
785 f0 or to an effect of disentrainment with respect to the interlocutor. Accordingly,
786 for same-sex dating conversations we would either expect both female speakers to
787 raise and both male speakers to lower their f0 or both speakers to move away from
788 the interlocutor's f0. In the latter case, we would expect the speaker with the higher
789 register to raise his/her f0 and the other speaker to lower his/her f0. The fact that
790 female speakers consistently raised their f0, although both lowered and raised f0 is
791 perceived as attractive by male listeners (cf. Karpf, 2006), supports the assumption
792 that indeed disentrainment and not primarily f0 movement is linked to perceived
793 attractiveness.

794 With respect to other prosodic cues, the effects observed for f0 are not easily
795 generalizable. The effects found for f0 entrainment and conversational quality are in
796 line with studies on other prosodic parameters. For example, Schweitzer et al. (2017)
797 observe a link between social attractiveness and speaking rate. However, there are no
798 studies on the effects of visual attractiveness or any observations of disentrainment
799 regarding anything but f0. If the disentrainment in f0 is a secondary effect of raising or
800 lowering f0, then those effects are linked to the natural sex differences expected from
801 the frequency code (Ohala, 1983, 1984) and should not transfer to anything other
802 than f0. However, if disentrainment and hence signaling social distance is the primary
803 cue, we could expect other prosodic features to show similar effects. Accordingly,
804 taking other prosodic features into consideration could also further our understanding
805 concerning what to expect in same-sex conversations for reasons explained above.

12.5 Conclusion

This paper shows that the perceived quality of a conversation and the perceived visual attractiveness of an interlocutor are linked to f0 entrainment. This relationship is largely reciprocal with f0 entrainment both apparently affecting and reflecting the social variables. Regarding the different types of entrainment (cf. Edlund et al., 2009; Levitan, 2014), the effects are mainly restricted to f0 proximity with no systematic effects for synchrony or convergence. As expected from the literature, we find contradicting effects with conversational quality being linked to more entrainment and attractiveness being linked to more disentrainment. Additionally, both variables depend on as well as affect each other and the respective effects on and of entrainment. This contradiction is primarily resolved by emphasizing one over the other with female speakers emphasizing conversational quality over attractiveness and male speakers doing the opposite. However, male speakers also show a shift from emphasizing attractiveness to conversational quality over the course of the conversation. Future research needs to investigate how the connection of f0 entrainment and perceived attractiveness and conversational quality relates to conversational topics as well as personality profiles, as well as take other prosodic features such as speaking rate, intensity variation, or voice quality into consideration. Furthermore, the role of synchrony leaves several open questions for further investigation.

References

- Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48.
- Belz, M., Mooshammer, C., Fuchs, S., Jannedy, S., Rasskazova, O., & Žygis, M. (Eds.). (2018). *Proceedings of the Conference on Phonetics & Phonology in German-Speaking Countries*. Berlin: Humboldt Universität.
- BeBeňuš, S. (2014). Social aspects of entrainment in spoken interaction. *Cognitive Computation*, 6(4), 802–813.
- Beňuš, S., Trnka, M., Kuric, E., Matrák, L., Gravano, A., Hirschberg, J., & Levitan, R. (2018). Prosodic entrainment and trust in human-computer interaction. In *Proceedings of Speech Prosody 9, Poznań, Poland* (pp. 220–224).
- Boersma, P., & Weenink, D. (2016). Praat: Doing phonetics by computer. Retrieved from <http://www.fon.hum.uva.nl/praat/>
- Borkowska, B., & Pawlowski, B. (2011). Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour*, 82(1), 55–59.
- Brennen, S. E., & Clark, H. H. (1996). Conceptual pacts and lexical choice in conversation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 22(6), 1482–1493.
- Brooks, A. B., Huang, L., Kearney, S. W., & Murray, F. E. (2014). Investors prefer entrepreneurial ventures pitched by attractive men. *Proceedings of the National Academy of Sciences of the United States of America*, 111, 4427–4431.
- Chartrand, T. L., & Bargh, J. A. (1999). The chameleon effect: The perception-behavior link and social interaction. *Journal of Personality and Social Psychology*, 76(6), 893–910.
- Cialdini, R. B. (2009). *Influence: Science and Practice* (5th ed.). Boston: Allyn & Bacon.
- Core, R., & Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>.
- Collins, S. A. (2000). Men's voices and women's choices. *Animal Behaviour*, 60, 773–780.

- 851 Collins, S. A., & Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal*
852 *Behaviour*, 65, 997–1004.
- 853 Dunbar, R. I. M. Coevolution of neocortex size, group size and language in humans. *Behavioural*
854 *Brain Science*, 16, 681–735.
- 855 Edlund, J., Heldner, M., & Hirschberg, J. (2009). Pause and gap length in face-to-face interaction.
856 In *Proceedings of INTERSPEECH 2009*.
- 857 Feinberg, D. R., Debruine, L. M., Jones, B. C., & Perrett, D. I. (2005). Manipulations of fundamental
858 and formant frequencies influence the attractiveness of human male voices. *Animal Behaviour*,
859 69, 561–568.
- 860 Feinberg, D. R., Debruine, L. M., Jones, B. C., & Perrett, D. I. (2008). The role of femininity and
861 averageness of voice pitch in aesthetic judgements of women's voices. *Perception*, 37, 615–623.
- 862 Fraccaro, P. J., Jones, B. C., Vukovic, J., Smith, F. G., Watkins, C. D., Feinberg, D. R., et al. (2011).
863 Experimental evidence that women speak in higher voice pitch to men they find attractive. *Journal*
864 *of Evolutionary Psychology*, 9(1), 57–67.
- 865 Friedberg, H., Litman, D., & Paletz, S. (2012). Lexical entrainment and success in student engi-
866 neering groups. In *Spoken Language Technology Workshop (SLT) 2012* (pp. 404–409). IEEE.
- 867 Gessinger, I., Schweitzer, A., Andreeva, B., Raveh, E., Möbius, B., & Steiner, I. (2018). Convergence
868 of pitch accents in a shadowing task. In *Proceedings of Speech Prosody, Poznań, Poland* (vol. 9,
869 pp. 225–229).
- 870 Giles, H., Coupland, N., & Coupland, J. (1991). Accomodation theory: Communication, context,
871 and consequence. *Contexts of Accomodation. Developments in Applied Sociolinguistics*, 1.
- 872 Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2009). Language style matching as a predictor
873 of social dynamics in small groups. *Communication Research*.
- 874 Gregory, S. W. (1996). A nonverbal signal in voices of interview partners effectively predicts
875 communication accommodation and social status perceptions. *Journal of Personality and Social*
876 *Psychology*, 70, 1231–1240.
- 877 Hewstone, M., Stroebe, W., & Jonas, K. (2012). *An Introduction to Social Psychology* (5th ed.).
878 Hoboken, New Jersey: BPS Blackwell.
- 879 Hodges-Simeon, C. R., Gaulin, S. J. C., & Puts, D. A. (2010). Different vocal parameters predict
880 perceptions of dominance and attractiveness. *Human Nature*, 21, 406–427.
- 881 Hughes, S. M., Farley, S. D., & Rhodes, B. C. (2010). Vocal and physiological changes in response
882 to the physical attractiveness of conversational partners. *Journal of Nonverbal Behavior*, 34, 1–13.
- 883 Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J.
884 W. (2011). Language style matching predicts relationship initiation and stability. *Psychological*
885 *Science*, 22, 39–44.
- 886 Jones, B. C., Feinberg, D. R., Debruine, L. M., Little, A. C., & Vukovic, J. (2010). A domain-
887 specific opposite-sex bias in human preferences for manipulated voice pitch. *Animal Behaviour*,
888 79(57–62)
- 889 Karpf, A. (2006). *The Human Voice*. New York, NY: Bloomsbury Publishing.
- 890 Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2016). lmerTest: Tests in linear
891 mixed effects models. R package version 2.0-30. Retrieved from [https://CRAN.R-project.org/
892 package=lmerTest](https://CRAN.R-project.org/package=lmerTest)
- 893 Ladd, R. D., Silverman, K., Tolkmitt, F., Bergmann, G., & Scherer, K. (1985). Evidence for the
894 independent function of intonation contour type, voice quality, and f0 range in signaling speaker
895 affect. *Journal of the Acoustical Society of America*, 78, 435–444.
- 896 Leaderbrand, K., Dekam, J., Morey, A., & Tuma, L. (2008). The effects of voice pitch on perceptions
897 of attractiveness: Do you sound hot or not. *Winona State University Psychology Student Journal*.
- 898 Lee, C. C., Black, M. P., Katsamanis, A., Lammert, A. C., Baucom, B. R., Christensen, A., et al.
899 (2010). Quantification of prosodic entrainment in affective spontaneous spoken interactions of
900 married couples. *Proceedings of Interspeech*, 793–796.
- 901 Levitan, R. (2014). Acoustic-prosodic entrainment in human-human and human-computer dialogue.
902 Columbia University. Ph.D. thesis.

- 903 Levitan, R., Gravano, A., Willson, L., Beňuš, S., Hirschberg, J., & Nenkova, A. (2012). Acoustic-
 904 prosodic entrainment and social behavior. In *Proceedings of the 2012 Conference of the North*
 905 *American Chapter of the Association for Computational Linguistics: Human Language Tech-*
 906 *nologies* (pp. 11–19).
- 907 Linville, S. E. (1996). The sound of senescence. *Journal of Voice*, 10(2), 190–200.
- 908 Lubold, N., & Pon-Barry, H. (2014). Acoustic-Prosodic Entrainment and Rapport in Collabora-
 909 tive Learning Dialogues. Proceedings of the. (2014). *ACM Workshop on Multimodal Learning*
 910 *Analytics Workshop and Grand Challenge, November 12–12, 2014, Turkey, Istanbul.*
- 911 Michalsky, J. (2017). Pitch synchrony as an effect of perceived attractiveness and likability. In:
 912 *Proceedings of DAGA 2017.*
- 913 Michalsky, J., & Schoormann, H. (2016). Effects of perceived attractiveness and likability on global
 914 aspects of fundamental frequency. In *Proceedings of P&P12* (120–124).
- 915 Michalsky, J., & Schoormann, H. (2017). Pitch convergence as an effect of perceived attractiveness
 916 and likability. In *Proceedings of INTERSPEECH, 2017* (pp. 2253–2256).
- 917 Michalsky, J., Schoormann, H., & Niebuhr, O. (2018a). Conversational quality is affected by and
 918 reflected in prosodic entrainment. In: *Proceedings of Speech Prosody, Poznań, Poland* (vol. 9).
- 919 Michalsky, J., Schoormann, H., & Niebuhr, O. (2018b). Turn transitions as salient places for social
 920 signals—Local prosodic entrainment as a cue to perceived attractiveness and likability. In M.
 921 Belz, C. Mooshammer, S. Fuchs, S. Jannedy, O. Rasskazova, & M. Žgis (Eds.), *Proceedings of*
 922 *the Conference on Phonetics & Phonology in German-Speaking Countries* (pp. 169–172). Berlin:
 923 Humboldt Universität.
- 924 Natale, M. (1975). Convergence of mean vocal intensity in dyadic communication as a function of
 925 social desirability. *Journal of Personality and Social Psychology*, 32(5), 790–804.
- 926 Nenkova, A., Gravano, A., & Hirschberg, J. (2008). High frequency word entrainment in spoken
 927 dialogue. In *Proceedings of the 46th Annual Meeting of the Association of Computational*
 928 *Linguistics on Human Language Technologies: Short Papers* (pp. 169–172).
- 929 Oguchi, T., & Kikuchi, H. (1997). Voice and interpersonal attraction. *Japanese Psychological*
 930 *Research*, 39, 56–61.
- 931 Ohala, J. (1983). Cross-language use of pitch. An ethnological view. *Phonetica*, 40, 1–18.
- 932 Ohala, J. (1984). An ethnological perspective on common cross-language utilization of f0 in voice.
 933 *Phonetica*, 41, 1–16.
- 934 Pickering, M. J., & Garrod, S. (2006). Alignment as the basis for successful communication.
 935 *Research on Language and Computation*, 4, 203–228.
- 936 Puts, D. A., Gaulin, S. J. C., & Verdolini, J. (2006). Dominance and the evolution of sexual dimor-
 937 phism in human voice pitch. *Evolution and Human Behavior*, 27, 283–296.
- 938 Reitter, D., & Moore, J. D. (2007). Predicting success in dialogue. *Annual Meeting - Association*
 939 *for Computational Linguistics*, 45, 808.
- 940 Scherer, K., Ladd, R. D., & Silverman, K. (1984). Vocal cues to speaker affect: testing two models.
 941 *Journal of the Acoustical Society of America*, 76, 1346–1356.
- 942 Schweitzer, A., Lewandowski, N., & Duran, D. (2017). Social attractiveness in dialogs. In *Proceed-*
 943 *ings of INTERSPEECH 2017* (pp. 2243–2247).
- 944 Street, R. L. (1984). Speech convergence and speech evaluation in fact-finding interviews. *Human*
 945 *Communication Research*, 11(2), 139–169.
- 946 Taylor, J. G. (2009). Cognitive computation. *Cognitive Computation*, 1, 4–16.
- 947 Thomason, J., Nguyen, H. V., & Litman, D. (2013). Prosodic entrainment and tutoring dialogue
 948 success. *Artificial Intelligence in Education*, 750–753.
- 949 Vukovic, J., Jones, B. C., Debruine, L. M., Feinberg, D. R., Smith, F. G., Little, A. C., et al. (2010).
 950 Women's own voice pitch predicts their preferences for masculinity in men's voices. *Behavioral*
 951 *Ecology*, 21(4), 767–772.
- 952 Xu, Y., Lee, A., Wu, W.-L., Liu, X., & Birkholz, P. (2013). Human vocal attractiveness as signaled
 953 by body size projection. *PLoS ONE*, 8(4),
- 954 Zimmer-Gembeck, M. J., Hughes, N., Kelly, M., & Connolly, J. (2011). Intimacy, identity and status:
 955 Measuring dating goals in late adolescence and emerging adulthood. *Motivation and Emotion*,
 956 36(3), 311–322.

1

2

Part IV Databases

Editor Proof

UNCORRECTED PROOF

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Voice Attractiveness
Series Title	
Chapter Title	Acoustic Correlates of Likable Speakers in the NSC Database
Copyright Year	2020
Copyright HolderName	Springer Nature Singapore Pte Ltd.
Corresponding Author	Family Name Weiss Particle Given Name Benjamin Prefix Suffix Role Division Organization Technische Universität Berlin Address Ernst-Reuter-Platz 7, 10405, Berlin, Germany Email benjamin.weiss@tu-berlin.de
Author	Family Name Trouvain Particle Given Name Jürgen Prefix Suffix Role Division Organization Saarland University Address Campus C7.2, 66123, Saarbrücken, Germany Email trouvain@coli.uni-saarland.de
Author	Family Name Burkhardt Particle Given Name Felix Prefix Suffix Role Division Organization audeERING GmbH Address Friedrichstraße 68, 10117, Berlin, Germany Email fburkhardt@audeering.com
Abstract	Speech stimuli from scenario-based conversations were analyzed regarding acoustic correlates of likability. Utterances from the pizza ordering scenario of the NSC corpus were selected, and the confederate's turns were excluded. These stimuli were recorded in high quality and were subjected to third-party listeners' ratings. Six promising acoustic parameters from related work are tested applying methods of correlation, regression, and regression trees. These parameters are average fundamental frequency, articulation rate, standard deviation of both and of intensity, as well as spectral center of gravity. The amount of variance

explained remains below 50%. Results confirm variability of the fundamental frequency as dominating correlate of likable voices in male and female speakers. It is concluded that the promising acoustic parameters are not robust to stimulus duration and scenario. Therefore, it is argued to explore the applicability of locally defined and linguistically motivated parameters.

Keywords

Voice - Acoustic parameters - Likability - Rating test - Database - Analysis - Modelling

Chapter 13

Acoustic Correlates of Likable Speakers in the NSC Database



Benjamin Weiss, Jürgen Trouvain, and Felix Burkhardt

Abstract Speech stimuli from scenario-based conversations were analyzed regarding acoustic correlates of likability. Utterances from the pizza ordering scenario of the NSC corpus were selected, and the confederate's turns were excluded. These stimuli were recorded in high quality and were subjected to third-party listeners' ratings. Six promising acoustic parameters from related work are tested applying methods of correlation, regression, and regression trees. These parameters are average fundamental frequency, articulation rate, standard deviation of both and of intensity, as well as spectral center of gravity. The amount of variance explained remains below 50%. Results confirm variability of the fundamental frequency as dominating correlate of likable voices in male and female speakers. It is concluded that the promising acoustic parameters are not robust to stimulus duration and scenario. Therefore, it is argued to explore the applicability of locally defined and linguistically motivated parameters.

Keywords Voice · Acoustic parameters · Likability · Rating test · Database · Analysis · Modelling

13.1 Introduction: Likability of Speakers

The aim of this chapter is twofold: First, acoustic correlates of likability ratings for the common stimulus length of a single utterance are presented as brief literature survey with a focus on re-occurring results. The second aim is to check whether such

B. Weiss (✉)
Technische Universität Berlin, Ernst-Reuter-Platz 7, 10405 Berlin, Germany
e-mail: benjamin.weiss@tu-berlin.de

J. Trouvain
Saarland University, Campus C7.2, 66123 Saarbrücken, Germany
e-mail: trouvain@coli.uni-saarland.de

F. Burkhardt
audEERING GmbH, Friedrichstraße 68, 10117 Berlin, Germany
e-mail: fburkhardt@audearing.com

© Springer Nature Singapore Pte Ltd. 2020
B. Weiss et al. (eds.), *Voice Attractiveness*, Prosody, Phonology and Phonetics,
https://doi.org/10.1007/978-981-15-6627-1_13

251

confirmed correlates can be applied for longer stimuli provided with a database for studying attributions and preference of speakers. The domain hereby is limited to evaluations of unacquainted speakers in order to maximize the impact of the first impression obtained from voice and speaking style of speakers.

The question of whether a person likes another person or not represents one of the most crucial social attitudes in humans, as it lays the basis for own social behavior (Chap. 1). In the most extremes of cases, liking somebody or not can determine not only the kind of social interaction, but whether there is avoidance or approach in the first place. In the research topic of first impressions, studies investigated the potential impact of surface signals, like clothing and facial expressions, but also voice and speaking style, on the formation of likability (Ambady & Skowronski, 2008). While people might not be inclined to immediately judge whether they truly like a person from a few seconds of interaction of recorded voice samples, listeners can express their gradual preference of the voice of a speaker, and thus his or her likability.

Such explicit ratings already show significant consistency between raters. For example, a standard measure of consistency between multiple raters is the intra-class correlation (ICC) with values between 0 and 1. It ranges for likability ratings from $ICC = 0.76$ (Burkhardt, Schuller, Weiss, & Wening, 2011) to $ICC = 0.93$ (Weiss & Burkhardt, 2010). This strong consistency documents that not only visual, but also acoustic information has a systematic relationship with a first impression. As the first impression is persistent over time and has predictive power (Ambady, Bernieri, & Richeson, 2000, 2006; Peterson, Cannito, & Brown, 1995; Hecht & LaFrance, 1995), it also potentially affects relationship building. As such acoustic or visual data in first encounters are sparse and superficial, attributions and even stereotypes play an important role in the formation of likability judgments. Salient attributions of regional or social background of speakers or disfluencies are relevant for likability (Scherer & Giles, (Scherer and Giles, 1979); Giles, 1980, Weiss & Burkhardt, 2012; McCroskey & Mehrley, 1969).

When studying acoustic correlates of likable voices, the effect of such attributions should therefore be minimized by providing homogeneous groups of speakers in terms of social and regional background, age, speech pathology, physical attractiveness, or gender (Murry, Singh, & Sargent, 1977; Murry & Singh, 1980; Linville, 2001; Brückl, 2011; Kreiman & Gerratt, 1996). Ideally, homogeneous groups of raters/listeners should be selected as well, or a diverse group that is balanced for these influencing factors can be recruited (Deal & Oyer, 1991). Of course, this statement holds not for the case of explicitly aiming to test for the effects of those attributions. Physical attractiveness that is inferred from voice and speaking style is such an example. Attractiveness is, like aesthetics for many other domains, a well-known and important factor for preference and liking, regardless of the sexual preference. The aim of this chapter is to present work that identifies which acoustic characteristics affect likability of unacquainted speakers, apart from the aforementioned attributions of age, gender, regional, and social background or speech-related pathologies.

13.2 A Review on Acoustic Correlates

In search for acoustic correlates of likability ratings, the dominant perspective is a global one, spanning the entire stimulus: The chosen stimuli are acoustically analyzed as a whole, and parameter values are aggregated, for example, to obtain the average fundamental frequency (F_0) for a complete utterance. As a consequence, vowel formants are typically only analyzed if the stimulus consists only of a single vowel (e.g., Bruckert et al., 2006).

Early research has brought a body of results on the so-called suprasegmentals, which are F_0 , intensity, and duration (Lehiste, 1970). For these suprasegmentals, analysis and re-synthesis studies have been conducted to identify acoustic parameters and test their impact on listeners' ratings. As outlined in Chap. 1, likability can be used as a synonym for social attractiveness. It can be related to unacquainted speakers to the attribution of warmth or benevolence—although there are social situations, in which competence might play a bigger role. Therefore, some of the results mentioned here stem not directly from ratings of likability but were elicited by questionnaires with related items or scales that also contribute to social attractiveness. Examples are friendliness, sympathy, or pleasantness (Weiss & Möller, 2011). If available in the studies, results for competence are also mentioned.

At least for male speakers and sometimes for females as well, the **fundamental frequency** (F_0) correlates negatively with ratings of benevolence, trust, likability, or pleasantness (Brown, Strong, & Rencher 1974; Apple, Streeter, & Krauss, 1979; Bruckert et al., 2006; Gravano et al., 2011; Weirich, 2010; Chattopadhyay, Dahl, Ritchie, & Shahin, 2003; Weiss & Burkhardt, 2012; Weiss, 2013). Noteworthy, however, are contradicting results of a positive correlation reported for German male speakers (Scherer, 1979). A similar opposing effect was found for the brief greeting “hello” in Scottish English (McAlear, 2014). Both results can be interpreted in a different communicative context, in which a raised average pitch is more appropriate, maybe to signal arousal.

The observed general tendency of a lower F_0 being evaluated more positively concurs with a positive association of perceived “darkness” for male speakers and the attribution of “being relaxed” for voices of the two genders considered (Weiss et al., 2018b). Variability or range of F_0 shows a positive effect on likability-related concepts of benevolence or warmth (Brown et al., 1973, 1974; Ray, 1986) but also on competence (Ray, 1986). There is also evidence for a positive effect of a rising F_0 contour (Bruckert et al., 2006; Weiss & Burkhardt, 2012; McAlear, 2014). However, this seems to be a more problematic acoustic correlate due to its dependency on the linguistic material.

Intensity as the second aspect reveals a negative correlation with benevolence but a positive one with competence (Ray, 1986), although this relationship might be more complex (Scherer, 1979). Effects caused by speech manipulation also have shown to add up or cancel each other out, dependent on the sign of correlation, which means that they can be considered as being independent of each other (Ray, 1986).

105 With respect to **articulation rate**, an ideal-point relation was found for likability-
 106 related concepts, especially for benevolence. This relation can be visualized as an
 107 inverted U-shaped line. This separates its effect from the more linear positive corre-
 108 lation with competence (Brown et al., 1974, 1975; Smith, Brown, Strong, & Rencher,
 109 1975; Apple et al., 1979; Street, Brady, & Putnam, 1983). For articulation rate, an
 110 additional effect could be found. Apparently, the raters' own intrinsic rates affect the
 111 evaluation of speakers' rates (Street et al., 1983; Feldstein, Dohm, & Crown, 2001).
 112 One interpretation is that listeners perceive articulation rate according to their own
 113 reference as high or low. A second interpretation would be that there is an effect of
 114 similarity preference, which may interfere with a linear relation between rate and lik-
 115 ability. In all cases, the result shows a positive correlation with moderate or slightly
 116 increased rates, confirming the rough inverted U shape of relationship, and thus a
 117 saturation for very fast conditions (Street & Brady, 1982). This is why Table 13.1
 118 gives results on both, positive correlations and an ideal-point relation with moderate
 119 or similar rates.

120 All these suprasegmentals are relatively easy to manipulate, e.g., Trouvain et al.
 121 (2006) could convincingly model personality dimensions such as sincerity, compe-
 122 tence, and excitement with speech synthesis. These suprasegmentals are also easy to
 123 measure automatically (maybe apart from articulation rate). This may be the reason
 124 that they have been studied extensively. More recently, spectral measures have been
 125 moved into focus with the aim to study voice quality, but also other, yet understud-
 126 ied, anatomical and articulatory sources of spectral aspects of speech. For example,
 127 shimmer, i.e., the local variability in amplitude, correlates positively with likability
 128 (Gravano et al., 2011), while measures of energy distribution, such as spectral tilt
 129 or center of gravity, show positive evaluation with less energy in higher frequencies
 130 (Weiss et al., 2017; Weiss, 2015). One reason could be a co-variation with the aver-
 131 age F_0 , i.e., the perception of "dark" or "relaxed" voices (Weiss et al., 2016; Weiss,
 132 2018b). However, a summary of many studies on this topic reveals non-significant
 133 results for spectral parameters (cf. Table 13.1).

134 There is some kind of tendency to be found in this summary. First of all, there
 135 are studies showing no effects, which are mostly analysis studies and thus might
 136 represent a non-sufficient variability in parameter values to show an effect. But also,
 137 F_0 mean, F_0 variability, and articulation rate seem to form a kind of majority vote
 138 to have a systematic effect, despite some contradicting results. For other parameters,
 139 such as variability in intensity or articulation rate, but also for spectral measures,
 140 such a systematic pattern is not obvious.

141 A particular issue is the status of the stimuli used in listening-and-rating experi-
 142 ments that are typically applied in this line of research. For example, the very short
 143 stimulus "hello" was rated and acoustically analyzed (McAlear, 2014). In this study,
 144 step-wise regression models of likability for male voices include average F_0 and,
 145 negatively, the harmonic-to-noise ratio (HNR). For female voices, a similar model
 146 is made up of HNR (negative sign), a rising F_0 contour, and the F_0 range. The posi-
 147 tive, and thus contradicting, result for pitch might have not appeared in the case of
 148 presenting the full utterances the "hello" was cut out from. While other studies used
 149 even shorter stimuli, i.e., vowels that have been excluded in this chapter, likability

Table 13.1 Summary of results from literature on acoustic correlates of likability and similar concepts. Positive, negative, and non-significant relations are depicted by +, -, and o, respectively. Gender of the speakers is indicated by *m, f*, respectively, and if the stimuli were re-synthesized. Reports on non-significant results may be incomplete

Reference	Gender	Language	Re-synthesized	F_0 mean	F_0 variation	F_0 raise	Intensity mean	Intensity variation	Articulation rate mean	Articulation rate variation	Harmonic-to-noise ratio	Jitter	Voiced-unvoiced ratio	Formant frequencies	Formant dispersion	Spectral tilt	Spectral center of gravity	Spectral skewness
Brown et al. (1973)	m	en	R	-	+													
Brown et al. (1974)	m	en	R	-	+													
Bruckert et al. (2006)	m	f*		-		+								o				
Duran (2017)	mf	de		o	+													
Feldstein et al. (2001)	mf	en		o	o		o	o	-									
Fernández Gallardo and Weiss (2016)	m	de		o	o		o	o								+	-	
Gravano et al. (2011)	f	de		o	o		o	+	o							o	o	
McAker (2014)	mf	en		-		o	+		o				o					
	m	en		+	o	o								o		o		
	f	en		o	+	+								o		o		
Ray (1986)	m	en	R	o	+													
Smith et al. (1975)	m	en	R				-											
Street and Brady (1982)	m	en	R															
Street et al. (1983)	m	en																
Weiss (2015)	m	de		o	o				+								o	o
	f	de		o	o				o								o	+
Weiss & Burkhardt (2010)	m	de		-					+								-	o
Weiss et al. (2010)	f	de		o					+								-	+
Weiss (2013)	m	de																
	f	de		-	o	o		-	o	o								
Weiss and Burkhardt (2012)	m	de		-	o	o		o	+	-								
	mf	de		-	o	+		o	+								o	o
Weiss et al. (2017)	f	de	R															+

*vowels only, female raters only

150 is a concept that emerges in social situations. It should therefore be studied not only
 151 concerning the voice quality but also the speaking style. In order to use more realistic
 152 data and to identify or verify correlates that emerge only with longer stimuli, such as
 153 variability of F_0 , the Nautilus Speaker Characteristics (NSC) database was recorded
 154 and used. It is described in the next section.

155 13.3 Material

156 The aim of this new analysis is to extend the insight into acoustic correlates of
 157 likable voices by avoiding several limitations of earlier research. First of all, the
 158 *number of speakers analyzed has often been very small*, about 20–30, for example.
 159 Secondly, the *social situation has been unclear*. Examples are the aforementioned
 160 utterance “hello” or reading aloud single sentences. And thirdly, the *stimuli have*
 161 *been very short*. Therefore, the Nautilus¹ database was created. It features 300 Ger-
 162 man speakers (aged 18–35, of which are 126 males) and was recorded with the
 163 aim to study speaker characteristics (Fernández Gallardo & Weiss, 2018). During
 164 recruitment, the speakers were subjectively checked for neither exhibiting a strong
 165 regional or social accent, nor displaying signs of a voice-related sickness or speech
 166 disorder. Although all speakers display Standard German, some speakers do exhibit
 167 some regional features and suprasegmental non-modal voice qualities. Hearing issues
 168 were not reported during collecting speakers’ details, which is important for properly
 169 conducting the interactive scenarios and understanding the instructions of the exper-
 170 imenter. The database and documentation of the Nautilus Speaker Characteristics
 171 (NSC) have been compiled by Fernández Gallardo (2018). NSC includes recordings
 172 from simulated telephone conversations, read passages, and read sentences in high
 173 signal quality.² From this database, telephone scenarios were chosen as appropriate
 174 material, as it contained a typical and well-defined social situation of unacquainted
 175 dyads that can be judged by third-party listeners. The scenario used for analysis here
 176 is ordering something to eat from a pizza service with a phone call. It stems from a
 177 list of pre-defined scenarios used for evaluating audio network transmission quality
 178 (Rec & P.805, 2007). The invited and recorded speakers all took over the role of
 179 the caller, while a student confederate played the pizza service. The caller obtained
 180 the following task information: a fake surname, address, and phone number. The
 181 instruction was to order a single pizza for two people, preferably a vegetarian option.
 182 During the conversation, the caller is asked to note down the exact final toppings,
 183 price, and duration until delivery. Such a conversation typically took about 60 s to
 184 complete.

¹Nautilus is the recording booth name used in the laboratory.

²The ISLRN of this corpus is 157-037-166-491-1. Is has been made available at the CLARIN repository: hdl.handle.net/11022/1009-0000-0007-C05F-6 under the CLARIN ACA+BY+NC+NORED license (freely available for scientific research).

185 In preparation of the stimuli for the listening-and-rating test on likability, all parts
 186 of the confederate in the pizza scenario were removed from the recordings. The
 187 resulting stimuli have an average duration of 23 s (SD = 3.3 s). Based on a question-
 188 naire with 34 items that was developed to assess voice-based personality attributions
 189 (Weiss & Möller, 2011; Fernández Gallardo & Weiss, 2017b), a final version was
 190 created with only minimal changes (Fernández Gallardo & Weiss, 2018).³ For the
 191 evaluation, each stimulus was rated by 15.1 listeners on average (sd = 1.17, due
 192 to splitting the students into groups). Altogether, 114 students, in the frame of a
 193 lecture's exercise, took part in this test (44 females, 70 males, aged on average 24.5
 194 years with an SD of 3.4). 93 of these were native German speakers, and the remaining
 195 participants were fluent in German. Each listener rated male and female stimuli in
 196 separate blocks with sliders on continuous scales. On average, each rater listened to
 197 about 16.9 males (sd = 0.49) and 23.2 females (sd = 2.33, due to splitting the data
 198 into sets). A single session took about 50 min.

199 The questionnaire itself includes items to cover major concepts of personal-
 200 ity attributions. It is based on existing instruments for the personality circumflex
 201 (Wiggins, Trapnell, & Phillips, 1988) for the first impression of warmth and agency,
 202 the OCEAN personality taxonomy (Rammstedt & John, 2007), the three-dimensional
 203 model of emotional states with valence, activity, and potency (Osgood, Suci, & Tan-
 204 nenbaum, 1957), and estimation of physical attractiveness that is affecting person-
 205 ality attributions and frequent attributions observed empirically for unacquainted
 206 voices (Weiss et al., 2018b). The questionnaire also includes the item of likability.
 207 A screenshot shows all scales with sliders on one page Fig. 13.1.

208 13.4 Analysis

209 Data analysis is presented in four sections. First, the comprehensive questionnaire
 210 responses are reduced in dimensionality to obtain values for the concept of likability.
 211 The subsequent correlation analysis aims at testing promising acoustic parameters
 212 from Sect. 13.2 on the new stimuli. Two simple modeling approaches are presented
 213 with different aims, mainly to find out how much variance the acoustic correlates
 214 of likability can explain. In order to inspect potentially non-linear relationships, a
 215 regression tree is applied.

³likable/non-likable, insecure/secure, unattractive/attractive, sympathetic/unsympathetic, decided/
 indecisive, obtrusive/unobtrusive, close/distant, interested/bored, unemotional/emotional, irri-
 tated/not irritated, passive/active, unpleasant/pleasant, characterful/characterless, reserved/sociable,
 nervous/relaxed, distant/affectionate, conformable/dominant, affected/unaffected, cold/hearty,
 young/old, factual/not factual, excited/calm, competent/incompetent, beautiful/ugly, unfriendly/
 friendly, feminine/masculine, offensive/submissive, committed/indifferent, boring/interesting,
 compliant/cynical, genuine/artificial, stupid/intelligent, adult/childish, bold/modest.

Inwieweit treffen die folgenden Attribute auf den Sprecher zu?

sympathisch _____	unsympathisch _____	affektiert _____	unaffektiert _____
unsicher _____	sicher _____	gefühllos _____	herzlich _____
unattraktiv _____	attraktiv _____	jung _____	alt _____
verständnisvoll _____	verständnislos _____	sachlich _____	unsachlich _____
entschieden _____	unentschieden _____	aufgeregt _____	ruhig _____
aufdringlich _____	unaufdringlich _____	kompetent _____	inkompetent _____
nah _____	distanziert _____	schön _____	hässlich _____
interessiert _____	gelangweilt _____	unfreundlich _____	freundlich _____
emotionslos _____	emotional _____	weiblich _____	männlich _____
genervt _____	nicht genervt _____	provokativ _____	gehorsam _____
passiv _____	aktiv _____	engagiert _____	gleichgültig _____
unangenehm _____	angenehm _____	langweilig _____	interessant _____
charaktervoll _____	charakterlos _____	folgsam _____	zynisch _____
reserviert _____	gesellig _____	unaufgesetzt _____	aufgesetzt _____
nervös _____	entspannt _____	dumm _____	intelligent _____
distanziert _____	mitfühlend _____	erwachsen _____	kindlich _____
unterwürfig _____	dominant _____	frech _____	bescheiden _____

Fig. 13.1 Screenshot of the first page of the rating interface. After pressing “Start” a new playback and continue button appears, while the scales remain

216 13.4.1 Factor Analysis

217 A factor analysis of the personality questionnaire was conducted to identify the most
 218 relevant basic dimension that explains the ratings. With this method, co-variabilities
 219 are represented by a smaller number of underlying factors each representing multiple
 220 questionnaire items for subsequent analysis. As human social evaluation concepts
 221 can be expected to be correlated to some degree, a non-orthogonal method was
 222 applied. The result of the factor analysis reveals five factors. These are named after
 223 inspecting the items that contribute to each one as warmth, attractiveness, confidence,
 224 compliance, and maturity (Fernández Gallardo & Weiss, 2017a, 2018). The first
 225 two show a strong correlation with each other ($r = 0.77$). Not only because of this
 226 correlation, but also due to the single questionnaire item “likability” correlating with
 227 these two dimensions (with warmth: $r = 0.87$, with attractiveness: $r = 0.83$), these

228 two dimensions are apparently related to the attitude toward speakers. Considering
 229 the small number and inconsistent groups of raters, the first principal component
 230 of warmth and attractiveness is used to represent the concept of likability more
 231 robust than the single item “likability”. This principal component is used as target
 232 for identifying acoustic correlates and represents likability on values from -3 to $+3$.

233 13.4.2 Correlation Analysis

234 For the first analysis, we tested the most important and promising acoustic param-
 235 eters that can be derived from Table 13.1.⁴ The chosen candidates are F_0 mean,
 236 F_0 SD, intensity SD, articulation rate mean, articulation rate SD, and Center of
 237 Gravity (CoG). Although variability in intensity and rate are not very promising candi-
 238 dates according to Table 13.1, they were chosen nevertheless. This was done to test
 239 whether the claim of Ketrow (1990) can be supported that variability in supraseg-
 240 mentals generally is signaling benevolence and positively affects likability. Except
 241 articulation rate, all acoustic parameters were measured with Praat (Boersma, 2001).
 242 Average articulation rate and its SD were estimated by an acoustic model (Weiss
 243 et al., 2018a) that was trained on the perceptually motivated “perceived local speak-
 244 ing rate” (PLSR) (Pfitzinger, 1990). The reason for applying this method is that
 245 stimulus duration would not be appropriate because of the varying linguistic mate-
 246 rial of the spontaneous utterances and that other established methods (De Jong &
 247 Wempe, 2009) sometimes have issues with the detection of silence and of unstressed
 248 syllables.

249 The results of linear bivariate correlations are presented in Table 13.2, separately
 250 for females and males. This separation reflects different value ranges of acoustic
 251 parameters but also the potentially different references and relations in likability
 252 formation. Gender of the raters was not analyzed due to the small number of listeners
 253 for each stimulus. False discovery rate approach is used to adjust for multiple testing
 254 (Benjamini & Hochberg, 1995). It is not as conservative as Bonferroni correction.
 255 The false discover rate sorts all p -values from lowest ($i = 1$) to highest ($i = \max(i)$)
 256 and adjusts the α -level by $i / \max(i) \cdot \alpha$. There are only two significant results for
 257 male and female speakers, respectively (see Sect. 13.5), indicated by bold p -values.
 258 Using the divergence from a global mean in articulation rate in order to represent an
 259 ideal-point relation is not significant in either gender. Before discussing these results,
 260 simple modeling of the data is conducted.

⁴While articulation rate is not an acoustic parameter in a narrow sense, the estimates used here are a prediction result based on spectral data, and it is also called acoustic parameter for convenience.

Table 13.2 Pearson's correlation between selected acoustic parameters and likability

Parameter	Female speakers		Male speakers	
	Pearson's r	p-value	Pearson's r	p-value
F_0 mean	0.25	0.0008	0.16	0.0688
F_0 SD	0.44	<0.0001	0.52	<0.0001
Intensity SD	-0.04	0.5975	-0.07	0.4427
Artic. Rate mean	0.05	0.4937	0.30	0.0006
Artic. Rate SD	0.02	0.7845	0.15	0.0913
CoG	0.05	0.5521	0.18	0.0445

Table 13.3 Linear models for Likability: Females ($p < 0.0001$, $R^2 = 0.206$); males ($p < 0.0001$, $R^2 = 0.345$). Parameters not included into a model are represented by “—”. (significance levels of $< .05^*$, $< .01^*$, and $< .001^{***}$ are applied)

Parameters	Females	Males
Intercept	-0.56**	-0.15
F_0 mean	0.38	-0.67*
F_0 SD	0.49***	0.83***
Artic. rate mean	—	0.22**
Artic. rate SD	—	—
Intensity SD	—	—
CoG	—	—

13.4.3 Linear Regression Analysis

As a second step, describing likability ratings with these selected acoustic parameters can shed a light on the amount of variance explained. Due to the relatively large number of stimuli, acoustic modeling can furthermore help to identify additional candidates of acoustic correlates that have non-linear relationships or meaningful interaction effects with other parameters, as attempted in the next subsection. As linear baseline, linear regression with step-wise inclusion of parameters was performed.⁵ Overall, the resulting models are significant but explain only about 1/5 of the variance for female and about 1/3 for male speakers (Table 13.3). While, for males, articulation rate mean is included in addition to the two pitch-related parameters, F_0 mean does contribute significantly to the model with F_0 SD, most likely due to a cross-correlation between them ($r = 0.34^{***}$). The resulting estimates are depicted in Fig. 13.2.

⁵Based on AIC, and single inclusion and exclusion of variables; only main effects.

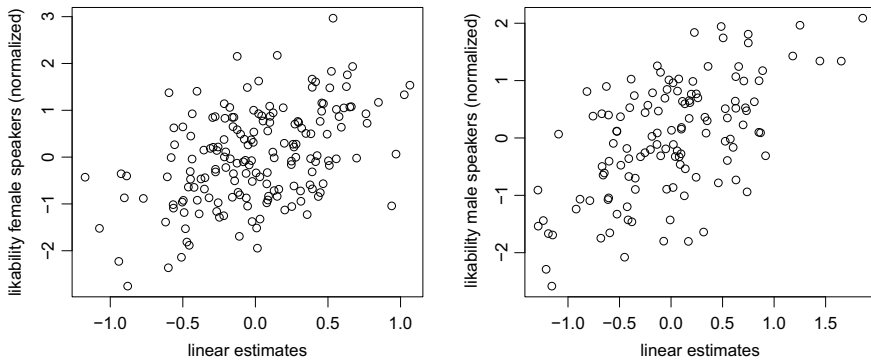


Fig. 13.2 Results of the two linear models for likability of female (left) and male (right) speakers. Average likability values versus model estimates from acoustics

274 13.4.4 Non-linear Modeling

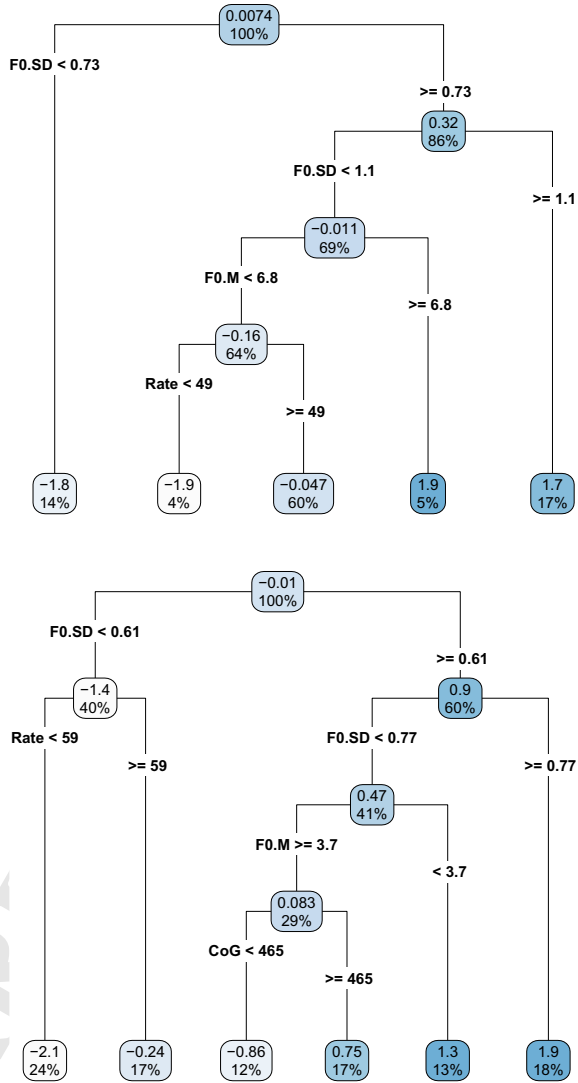
275 A second simple approach to modeling is using regression trees to better take non-
 276 linearities into account. The main aim is not a better fit, but a better view for identify-
 277 ing acoustic correlates of likability in voices. Such trees, pruned on cross-correlation
 278 errors to avoid overfitting, improve the amount of variance a little, but without suc-
 279 ceeding 50% of variance. Figure 13.3 shows the two regression trees, with the target
 280 likability value and the percentage of data points given in the boxes, while the joints
 281 are labeled with the conditional value of the acoustic parameter used for splitting the
 282 data.

283 13.5 Discussion

284 The positive correlation of likability with F_0 SD for both speaker sets and articulation
 285 rate for males is in line with results of other studies presented in the literature survey in
 286 Sect. 13.2. However, other promising parameters are not significantly correlated for
 287 the stimuli from our pizza order scenario, in particular, CoG and, for males, F_0 mean.
 288 Actually, average F_0 is even correlated negatively for female speakers, which is in
 289 contrast to the state of the art. One reason for these results may be the small number
 290 and inconsistently selected raters (Fernández Gallardo & Weiss, 2018). However, a
 291 more probable cause can be suspected in the much longer stimulus durations and
 292 the different genres compared to other experiments. In particular, the strong effect
 293 of variability in the fundamental frequency (F_0 SD) may mask smaller effects, for
 294 example, CoG as timbre-related parameter. The two modeling approaches indicated
 295 that F_0 mean is relevant for male speakers.

296 The unexpected positive correlation with average F_0 for females is still surprising,
 297 as other work with shorter German stimuli repeatedly showed an opposite effect. This

Fig. 13.3 Results of the two regression trees for likability of female (top; $R^2 = 0.312$) and male (bottom; $R^2 = 0.437$) speakers. **F0.M** refers to F_0 mean in ERB (as normalization attempt from Hz); values of **Rate** are in the units of PLSR (Perceived Local Speaking Rate, usually between 50 and 150) instead of syllables/s



298 even holds in combination with a positive impact of higher F_0 SD, which seems to
 299 exclude the possibility of articulatory grounding of raising F_0 mean by increasing F_0
 300 variability. This contradicting result may indicate a situational difference between
 301 reading single utterances in a rather factual tone, where low articulatory tension or
 302 even larger vocal folds in males may be rewarded, whereas in a truly conversational
 303 situation a social signal of interest (a raised voice due to higher tension), benevolence,
 304 or even a biological signal of female attractiveness might be positively perceived.
 305 This kind of speculation has of course to be tested, for example, with re-synthesis

306 experiments for both kinds of situations. At least, for the non-significant correlation
 307 in males, the two models solve this issue by including F_0 both times with a negative
 308 relationship.

309 The attempt to describe likability ratings with simple models reveals only low
 310 performing results. Not even half of the variance is explained, despite applying
 311 approaches that use all data for training. More interesting are the systematics incor-
 312 porated in the models. First of all, parameters, which are non-significant in the corre-
 313 lation analysis, are included in order to explain more variance, i.e., male F_0 mean in
 314 males for the linear and the tree model, and CoG in the regression tree. For females,
 315 rate is included in the regression tree. For one split in the female data, the positive
 316 impact of higher F_0 mean is confirmed. For male speakers, a lower F_0 has a posi-
 317 tive impact, just as expected from literature. Additionally, still very simple regression
 318 trees perform better than the linear baseline, indicating non-linearities observed else-
 319 where (Weiss & Burkhardt, 2012).

320 13.6 Conclusion

321 Despite some agreement in English and German studies, the attempt to confirm a
 322 set of potential acoustic correlates of likable voices was not overall successful. The
 323 analysis of the Nautilus data confirms only F_0 variation and articulation rate as rele-
 324 vant parameters. Especially, F_0 variation seems to be a very salient parameter in this
 325 conversational data. The role of F_0 , or pitch level in general, has to be re-examined.
 326 Currently, a stereotype of low-pitched male voices and high-pitched female ones
 327 seem to be too simple for German. In light of other studies, there seems to be a pool
 328 of potential correlates that not necessarily show a relation in each analysis. However,
 329 generalization seems not to be possible from the given results.

330 This reveals a more general issue with the material. Most data referred to as
 331 related work are single short sentences, which are sometimes difficult to discern as
 332 read or not, but for which simple aggregated values are intuitive parameter choices.
 333 With longer durations, as in the Nautilus database, not only more material, including
 334 several sentences and utterances, are available, but also a specific social situation
 335 is evident. Apart from obvious differences due to this kind of styles, aggregating
 336 simple acoustic parameter values over time could result in unreliable correlates. As
 337 almost all parameters are globally defined, they seem to be fragile for changes in
 338 material. In order to better compare acoustics between for example brief greetings
 339 (“hello”) (McAlear, 2014) to longer utterances or even short conversations, the value
 340 of locally defined or dynamic parameters has to be tested.

341 In order to define more robust parameters and even more automatic measurement,
 342 segment-based and articulatorily defined candidates should be defined to better rep-
 343 resent perceptually salient aspects that are relevant for likability, especially when
 344 studying timbre. One example is the so-called speakers’ or actors’ formant to assess
 345 a potentially positive effect of trained voices. It manifests as a peak in the acoustic
 346 spectrum: 3–4 kHz for males (Nawka, Anders, Cebulla, & Zurakowski, 1997), and

347 4–5 kHz for females (Tayal, Stone, & Biskholz, 2017). This resonance seems to be
 348 caused by an epi-laryngeal narrow and pharyngeal wide configuration that is evident
 349 in professional speakers, and it is considered as pleasant also for non-trained speak-
 350 ers (Leino, Laukkanen, & Radolf, 2011). The issue is to properly re-synthesize this
 351 phenomenon in a valid and salient way. A recent analysis shows a relation for males
 352 voices (Weiss, 2015) that is even stronger than the typical average F_0 . However, this
 353 effect was not confirmed by a first attempt of overall spectral manipulation (Karnop
 354 & Weiss, 2016), maybe due to missing representation in other acoustic features that
 355 are perceptually relevant for stimulating the acoustic effect of this configuration.
 356 Other, more phonetically or phonologically defined parameters such as vowel for-
 357 mant dispersion as a measure of articulatory precision or aspects of intonation, have
 358 not yet been studied in depths for likability, simply because they require manual or
 359 automatic phonetic analysis for segmental selection.

360 There are further factors in the research of likability of voices that remain under-
 361 explored or simply ignored. Among them is the question of how audible smiling
 362 in voices has an influence on whether somebody likes a formerly unknown person.
 363 Certain types of smiling are perceived as displays of happiness (Krys, 2016). For
 364 instance, in a Brazilian study smiling faces were considered as happier and even as
 365 more attractive than a neutral expression (Otta, Abrosio, & Hoshino, 1996). How-
 366 ever, there is evidence that in some cultures visually transmitted smiling faces of
 367 unknown persons may have a *negative* image on side of the viewers (Krys, 2016).
 368 Thus, it could be that similar patterns could occur for audibly transmitted smiling.

369 As mentioned in the introduction, the level of speech fluency can also have an effect
 370 on the perceived attractiveness of voices and thus might affect social attractiveness
 371 as well. For instance, Zuta (2007) showed that in retold narratives male voices were
 372 considered least attractive by female listeners when comparatively many disfluencies
 373 occurred, along with less varied F_0 and a high degree of nasality. Also, the number
 374 and the duration of pauses is a strong marker of fluency but also of the valence of
 375 speech (Tisljár-Szabó and Pléh, 2014). Too long pauses seem to have a tendency
 376 toward a negative and less likable image of the speaker, also in dialogs.

377 With regard to intonation contours, the impression of politeness and pleasant-
 378 ness obviously depend on the sentence mode. For instance, Uldall (1960) found
 379 that declarative sentences were perceived with a high degree of pleasantness when
 380 produced with either a falling or rising pitch at the end; however, questions and
 381 commands tend to be felt pleasantly only when they showed a final rise.

382 Audible smiling, fluency, pauses, sentence accents, and phrase tones can be con-
 383 sidered as local phenomena of spoken sentences and longer stretches of speech. In
 384 contrast, a regional or a foreign accent is always a global phenomenon. Regarding
 385 accents, people sometimes have more or less strong attitudes which can heavily
 386 influence the likability of the speakers in a negative and likewise in a positive way.

387 Lastly, acoustic correlates of sexual preference and physical attractiveness
 388 have been mostly neglected in this line of research. While there are some cross-
 389 correlations found for likability as social attractiveness and subjective estimates
 390 of physical attractiveness from voice or ratings of vocal attractiveness directly
 391 (McAlear, 2014), well-founded correlates, such as formant dispersion in males (Fitch
 392 & Giedd, 1999; Bruckert et al., 2006) might increase insight into the cause of a likable
 393 first impression in speech.

394 References

- 395 Ambady, N., & Skowronski, J. J. (Eds.). (2008). *First Impressions*. New York: Guilford Press.
- 396 Ambady, N., Bernieri, F. J., & Richeson, J. A. (2000). Toward a histology of social behavior:
 397 Judgmental accuracy from thin slices of the behavioral stream. In M. P. Zanna (Ed.), *Advances*
 398 *in experimental social psychology* (Vol. 32, pp. 201–272). San Diego: Academic Press.
- 399 Ambady, N., Krabbenhoft, M. A., & Hogan, D. (2006). The 30-sec sale: Using thin slice judgments
 400 to evaluate sales effectiveness. *Journal of Consumer Psychology*, 16, 4–13.
- 401 Apple, W., Streeter, L. A., & Krauss, R. M. (1979). Effects of pitch and speech rate on personal
 402 attributions. *Journal of Personality and Social Psychology*, 37(5), 715–727.
- 403 Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Pow-
 404 erful Approach to Multiple Testing. *Journal of the Royal Statistical Society - Series B (Method-*
 405 *ological)*, 57, 289–300.
- 406 Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glott International*, 5, 341–345.
- 407 Brown, B. L., Strong, W. J., & Rencher, A. C. (1973). Perceptions of personality from speech:
 408 Effects of manipulations of acoustical parameter. *Journal of the Acoustical Society of America*,
 409 54(1), 29–35.
- 410 Brown, B. L., Strong, W. J., & Rencher, A. C. (1974). Fifty-four voices from two: The effects of
 411 simultaneous manipulations of rate, mean fundamental frequency, and variance of fundamental
 412 frequency on ratings of personality from speech. *Journal of the Acoustical Society of America*,
 413 55(2), 313–318.
- 414 Brown, B. L., Strong, W. J., & Rencher, A. C. (1975). Acoustic determinants of perceptions of
 415 personality from speech. *Linguistics*, 13(166), 11–32.
- 416 Bruckert, L., Liénard, J.-S., Lacroix, A., Kreutzer, M., & Leboucher, G. (2006). Women use voice
 417 parameter to assess men's characteristics. *Proceedings of the Royal Society B: Biological Sciences*,
 418 237(1582), 83–89.
- 419 Brückl, M. (2011). Altersbedingte Veränderungen der Stimme und Sprechweise von Frauen. Dis-
 420 sertation. Berlin: Technische Universität Berlin.
- 421 Burkhardt, F.; Schuller, B.; Weiss, B. & Weninger, F. (2011). Would you buy a car from me?—On
 422 the likability of telephone voices. In *Proceedings of 12th Interspeech* (pp. 1557–1560), Florence.
- 423 Chattopadhyay, A., Dahl, D. W., Ritchie, R. J., & Shahin, K. N. (2003). Hearing voices: The impact
 424 of announcer speech characteristics on consumer response to broadcast advertising. *Journal of*
 425 *Consumer Psychology*, 13(3), 198–204.
- 426 De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate
 427 automatically. *Behavior Research Methods*, 41, 385–390.
- 428 Deal, L. V., & Oyer, H. J. (1991). Ratings of vocal pleasantness and the aging process. *Folia*
 429 *Phoniatica*, 43, 44–48.
- 430 Duran, D., Lewandowski, N., Bruni, J., & Schweitzer, A. (2017). Akustische Korrelate
 431 wahrgenommener Persönlichkeitsmerkmale und Stimmattraktivität. In *Proceedings of Elektron-*
 432 *ische Sprachsignalverarbeitung* (pp. 91–98). TUD Press.

- 433 Feldstein, S., Dohm, F.-A., & Crown, C. L. (2001). Gender and speech rate in the perception of
434 competence and social attractiveness. *Journal of Social Psychology, 141*, 785–806.
- 435 Fernández Gallardo, L. (2018). The Nautilus Speaker Characterization Corpus. ISLRN: 157-037-
436 166-491-1. hdl.handle.net/11022/1009-0000-0007-C05F-6.
- 437 Fernández Gallardo, L., & Weiss, B. (2016). Speech likability and personality-based social relations:
438 A round-robin analysis over communication channels. In *Proceedings of 17th Interspeech* (pp.
439 903–907), San Francisco.
- 440 Fernández Gallardo, L., & Weiss, B. (2017a). Perceived interpersonal speaker attributes and their
441 acoustic features. *I3*, 61–64. Berlin: Tagung Phonetik & Phonologie im deutschsprachigem Raum.
- 442 Fernández Gallardo, L., & Weiss, B. (2017b). Towards speaker characterization: Identifying and
443 predicting dimensions of person attribution. In *Proceedings of 18th Interspeech* (pp. 904–908),
444 Stockholm.
- 445 Fernández Gallardo, L., & Weiss, B. (2018). The Nautilus Speaker characterization corpus: Speech
446 recordings and labels of speaker characteristics and voice descriptions. In *Proceedings of 11th*
447 *Language Resources and Evaluation Conference (LREC)* (pp. 2837–2842), Miyazaki.
- 448 Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study
449 using magnetic resonance imaging. *Journal of the Acoustical Society of America, 106*, 1511–1522.
- 450 Giles, H. (1980). Accommodation theory: Some new directions. *York Papers in Linguistics, 9*,
451 105–136.
- 452 Gravano, A., Levitan, R., Willson, L., Beňuš, Š., Hirschberg, J., & Nenkova, A. (2011). Acoustic
453 and prosodic correlates of social behavior. In *Proceedings of Interspeech* (pp. 97–100).
- 454 Hecht, M. A., & LaFrance, M. (1995). How (Fast) can i help you? Tone of voice and telephone
455 operator efficiency in interactions. *Journal of Applied Social Psychology, 25*(23), 2086–2098.
- 456 Rec, I. T. U.-T., & P.805. (2007). *Subjective evaluation of conversational quality*. Geneva: Interna-
457 tional Telecommunication Union.
- 458 Karnop, C., & Weiss, B. (2016). Zum Effekt von Tempo, Tonhöhe und Sprecherformant auf Sym-
459 pathiebewertungen: Ein Resyntheseexperiment. *27, 206–213*. Leipzig: Konferenz Elektronische
460 Sprachsignalverarbeitung.
- 461 Ketrow, S. M. (1990). Attributes of a telemarketer’s voice and persuasiveness: A review and synthesis
462 of the literature. *Journal of Direct Marketing, 4*, 7–21.
- 463 Kreiman, J., & Gerratt, B. R. (1996). The perceptual structure of pathologic voice quality. *Journal*
464 *of the Acoustical Society of America, 100*, 1787–1795.
- 465 Kryś, et al. (2016). Be careful where you smile: Culture shapes judgments of intelligence and
466 honesty of smiling individuals. *Journal of Nonverbal Behavior, 40*, 101–116.
- 467 Lehiste, I. (1970). *Suprasegmentals*. Cambridge, Massachusetts: MIT Press.
- 468 Leino, T., Laukkanen, A.-M., & Radolf, V. (2011). Formation of the actor’s/speakers’ formant: A
469 study applying spectrum analysis and computer modeling. *Journal of Voice, 25*(2), 150–158.
- 470 Linville, S. E. (2001). *Vocal aging*. San Diego: Singular Thomson Learning.
- 471 McAleer, P., Todorov, A., & Berlin, P. (2014). How do you say ‘Hello’? Personality impressions
472 from brief novel voices. *PLOS ONE, 9*(3).
- 473 McCroskey, J. C., & Mehrley, R. S. (1969). The effects of disorganization and nonfluency on attitude
474 change and source credibility. *Speech Monographs, 36*, 13–21.
- 475 Murry, T., & Singh, S. (1980). Multidimensional analysis of male and female voices. *Journal of the*
476 *Acoustical Society of America, 68*, 1294–1300.
- 477 Murry, T., Singh, S., & Sargent, M. (1977). Multidimensional classification of abnormal voice
478 qualities. *Journal of the Acoustic Society of America, 61*, 1630–1635.
- 479 Nawka, T., Anders, L. C., Cebulla, M., & Zurakowski, D. (1997). The speaker’s formant in male
480 voices. *Journal of Voice, 11*(4), 422–428.
- 481 Osgood, C. E., Suci, G., & Tannenbaum, P. (1957). *The measurement of meaning*. Urbana, IL:
482 University of Illinois Press.
- 483 Otta, E., Abrosio, F. F. E., & Hoshino, R. L. (1996). Reading a smiling face: Messages conveyed
484 by various forms of smiling. *Perceptual and Motor Skills, 82*, 1111–1121.

- 485 Peterson, R., Cannito, M., & Brown, S. (1995). An exploratory investigation of voice characteristics
486 and selling effectiveness. *Journal of Personal Selling & Sales Management*, 15(1), 1–15.
- 487 Pfitzinger, H. R. (1990). Local speech rate perception in German speech. In *Proceedings of ICPhS*
488 (pp. 893–896).
- 489 Rammstedt, B., & John, O. P. (2007). Measuring personality in one minute or less: A 10-item short
490 version of the big five inventory in English and German. *Journal of Research in Personality*, 41,
491 203–212.
- 492 Ray, G. B. (1986). Vocally cued personality prototypes: An implicit personality theory approach.
493 *Journal of Communication Monographs*, 53(3), 266–176.
- 494 Scherer, K. R. (1979). Personality markers in speech. In K. Scherer & H. Giles (Eds.), *Social*
495 *markers in speech* (pp. 147–209). Cambridge: Cambridge University Press.
- 496 Scherer, K. R., & Giles, H. (1979). *Social markers in speech*. Cambridge: Cambridge University
497 Press.
- 498 Smith, B. L., Brown, B. L., Strong, W. J., & Rencher, A. C. (1975). Effects of speech rate on
499 personality perception. *Language and Speech*, 18, 145–152.
- 500 Street, R. L, Jr., & Brady, R. M. (1982). Speech rate acceptance ranges as a function of evaluative
501 domain, listener speech rate and communication context. *Communication Monographs*, 49(4),
502 290–308.
- 503 Street, R. L, Jr., Brady, R. M., & Putnam, W. B. (1983). The influence of speech rate stereotypes and
504 rate similarity or listeners' evaluation of speakers. *Journal of Language and Social Psychology*,
505 2(1), 37–56.
- 506 Tayal, S., Stone, S., & Birkholz, P. (2017). Towards the measurement of the actor's formant in
507 female voices. In *Proceedings of Elektronische Sprachsignalverarbeitung* (pp. 286–293). TUD
508 Press.
- 509 Tisljár-Szabó, E., & Pléh, C. (2014). Ascribing emotions depending on pause length in native and
510 foreign language speech. *Speech Communication*, 56, 35–48.
- 511 Trouvain, J., Schmidt, S., Schröder, M., Schmitz, M., & Barry, W. J. (2006). Modelling personality
512 features by changing prosody in synthetic speech. In *Proceedings of Speech Prosody*.
- 513 Uldall, E. (1960). Attitudinal meanings conveyed by intonation contours. *Language and Speech*, 3,
514 223–234.
- 515 Weirich, M. (2010). *Die attraktive Stimme: Vocal stereotypes. Eine phonetische Analyse anhand*
516 *akustischer und auditiver Parameter*. Saarbrücken: Verlag Dr. Müller.
- 517 Weiss, B. (2013). Prosodische Elemente vokaler Sympathie. In *Konferenz Elektronische Sprachsig-*
518 *nalverarbeitung* (Vol. 24, pp. 212–217), Bielefeld.
- 519 Weiss, B. (2015). Akustische Korrelate von Sympathieurteilen bei Hörern gleichen Geschlechts.
520 26, 165–171. Eichstädt: Konferenz Elektronische Sprachsignalverarbeitung.
- 521 Weiss, B. (2016). Voice descriptions by non-experts: Validation of a questionnaire. In *Tagung*
522 *Phonetik & Phonologie im deutschsprachigem Raum* (Vol. 12, pp. 228–231), München.
- 523 Weiss, B., & Burkhardt, F. (2010). Voice attributes affecting likability perception. In *11th Interspeech*
524 (pp. 1934–1937), Makuhari.
- 525 Weiss, B., & Burkhardt, F. (2012). Is 'not bad' good enough? Aspects of unknown voices' likability.
526 In *13th Interspeech* (pp. 1–4), Portland.
- 527 Weiss, B., & Möller, S. (2011). Wahrnehmungsdimensionen von Stimme und Sprechweise. In
528 *Konferenz Elektronische Sprachsignalverarbeitung* (Vol. 22, pp. 261–268), Aachen.
- 529 Weiss, B., Möller, S., & Polzehl, T. (2010). Zur Wirkung menschlicher Stimme auf
530 diewahrgenommene Sympathie—Einfluss der Stimmanregung. In *Konferenz Elektronische*
531 *Sprachsignalverarbeitung* (Vol. 21, pp. 56–63), Berlin.
- 532 Weiss, B., Hacker, A., Moshona, C., Rudawski, F., & Ruhland, M. (2017). Studying vocal social
533 attractiveness by re-synthesis—Results from two student projects applying acoustic morphing
534 with Tandem-Straight. 28, 316–323. Saarbrücken: Konferenz Elektronische Sprachsignalver-
535 arbeitung.

- 536 Weiss, B., Hillmann, S., & Michael, T. (2018a). Kontinuierliche Schätzung von
537 Sprechgeschwindigkeit mit einem Rekurrenten Neuronalen Netzwerk. 29, 186–191. Ulm:
538 Konferenz Elektronische Sprachsignalverarbeitung.
- 539 Weiss, B., Estival, D., & Stiefelhagen, U. (2018b). Studying vocal perceptual dimension of non-
540 experts by assigning overall speaker (dis-)similarities. *Acta Acustica united with Acustica*, 104,
541 174–184.
- 542 Wiggins, J. S., Trapnell, P., & Phillips, N. (1988). Psychometric and geometric characteristics
543 of the revised interpersonal adjective scales (IAS-R). *Multivariate Behavioral Research*, 23(4),
544 517–530.
- 545 Zuta, V. (2007). Phonetic criteria of attractive male voices. In *Proceedings of 16th ICPhS, Saar-*
546 *brücken* (pp 1837–1840).

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Voice Attractiveness
Series Title	
Chapter Title	Ranking and Comparing Speakers Based on Crowdsourced Pairwise Listener Ratings
Copyright Year	2020
Copyright HolderName	Springer Nature Singapore Pte Ltd.
Corresponding Author	Family Name Baumann Particle Given Name Timo Prefix Suffix Role Division Organization Universität Hamburg, Language Technology Group Address Hamburg, Germany Email mail@timobaumann.de
Abstract	Speech quality and likability is a multi-faceted phenomenon consisting of a combination of perceptory features that cannot easily be computed nor weighed automatically. Yet, it is often easy to decide which of two voices one likes better, even though it would be hard to describe why, or to name the underlying basic perceptory features. Although likability is inherently subjective and individual preferences differ, generalizations are useful and there is often a broad intersubjective consensus about whether one speaker is more likeable than another. We present a methodology to efficiently create a likability ranking for many speakers from crowdsourced pairwise likability ratings which focuses manual rating effort on pairs of similar quality using an active sampling technique. Using this methodology, we collected pairwise likability ratings for many speakers (>220) from many raters (>160). We analyze listener preferences by correlating the resulting ranking with various acoustic and prosodic features. We also present a neural network that is able to model the complexity of listener preferences and the underlying temporal evolution of features. The recurrent neural network achieves remarkably high performance in estimating the pairwise decisions and an ablation study points toward the criticality of modeling temporal aspects in speech quality assessment.
Keywords	Ranking - Speech quality - Likability ratings - Found data - Crowdsourcing - Sequence modelling

Chapter 14

Ranking and Comparing Speakers Based on Crowdsourced Pairwise Listener Ratings



Timo Baumann

Abstract Speech quality and likability is a multi-faceted phenomenon consisting of a combination of perceptory features that cannot easily be computed nor weighed automatically. Yet, it is often easy to decide which of two voices one likes better, even though it would be hard to describe why, or to name the underlying basic perceptory features. Although likability is inherently subjective and individual preferences differ, generalizations are useful and there is often a broad intersubjective consensus about whether one speaker is more likeable than another. We present a methodology to efficiently create a likability ranking for many speakers from crowdsourced pairwise likability ratings which focuses manual rating effort on pairs of similar quality using an active sampling technique. Using this methodology, we collected pairwise likability ratings for many speakers (>220) from many raters (>160). We analyze listener preferences by correlating the resulting ranking with various acoustic and prosodic features. We also present a neural network that is able to model the complexity of listener preferences and the underlying temporal evolution of features. The recurrent neural network achieves remarkably high performance in estimating the pairwise decisions and an ablation study points toward the criticality of modeling temporal aspects in speech quality assessment.

Keywords Ranking · Speech quality · Likability ratings · Found data
Crowdsourcing · Sequence modelling

14.1 Introduction

Speaker traits (such as age or gender), emotional coloring (such as anger or distress), socio-cultural aspects (such as accent or dialects), conscious or subconscious coloring toward the addressee (such as friendliness or positivity), and other paralinguistic aspects (such as clarity and comprehensibility) are expressed through various prosodic, suprasegmental, segmental, and non-segmental aspects of one's speech

T. Baumann (✉)
Universität Hamburg, Language Technology Group, Hamburg, Germany
e-mail: mail@timobaumann.de

© Springer Nature Singapore Pte Ltd. 2020
B. Weiss et al. (eds.), *Voice Attractiveness*, Prosody, Phonology and Phonetics,
https://doi.org/10.1007/978-981-15-6627-1_14

269

and voice, where the combination of features and their temporal evolution are far from trivial. Intermittent deficiencies (e.g., a lisp) or deviations limited to a few features (e.g., nasalisation) can already have strong influences on the perceived quality. Together, they form the ‘quality’ of speech. It is important to note that no one ‘best’ combination of all features exists that would constitute ‘ideal’ speech.

Voice is a highly personal and subjective matter such that a multitude of combinations of these features result in a ‘good’ voice. This often makes likability comparisons hard and inherently subjective. Despite subjective preferences, *intersubjective* agreement on the preferences can often be found by-and-large, making generalizations useful. Generalizations are also necessary, for example, to cast news speakers, readers, or other speaking roles that need to approximate an intersubjective consensus. Such castings are typically performed by small expert juries (potentially limiting the universality of decisions) and for small numbers of speaker candidates (for practical reasons).

In our work, we use rankings to analyze the influencing factors of speaker likability for broad speaker populations, or to eventually ‘score’ a voice sample along a range of speakers. Hence, we are interested in full rankings rather than in who is the best speaker for a task. Our aim is to create rankings for large speaker populations, by large and diverse juries, and while keeping the effort as low as possible.

To simplify the human effort involved in creating the ranking, we have participants take many pairwise decisions on which of two stimuli is better. We then create a ranking from the pairwise comparisons (see below). The number of possible pairs grows quadratically with the number of the stimuli compared. Thus, while full comparisons for each rater are possible for small speaker groups (10 speakers → 45 rating pairs), these are infeasible for large speaker groups (225 speakers → 25200 rating pairs), in particular when relying on volunteer raters. Thus, our method must be able to build rankings from incomplete comparisons. Note, however, that many of the ratings will have predictable outcomes if one known-strong and one known-weak speaker are paired. It will be helpful to not waste too much human effort on such pairs; in contrast, human input on speakers of similar (or unknown) quality is most informative.

The main idea is to start from an initial ranking (based on some initial ratings) which is iteratively revised as more evidence becomes available with more ratings. Once the initial ranking is available, rating outcomes can be predicted and human effort can be directed away from comparisons with clear outcomes and toward the most informative pairs; this will be described in detail in Sect. 14.2.

Section 14.3 describes the corpus developed via crowdsourcing and based on the iterative method, both in terms of the stimuli used, as well as the resultant preference ratings. Section 14.4 examines the overall preference ranking derived from all pairwise ratings and finds some explaining factors in terms of high-level properties of the speech stimuli (and their speakers) via linear correlations.

As outlined above, however, prosody is a highly non-linear phenomenon and we hence build a recurrent neural network-based model that successfully identifies listener preferences using non-linear (but opaque) aggregation functions. Via an ablation study we find that the tunes in to phone-specific prosodic aspects given

71 phonetic identity as additional features. Section 14.5 describes model for estimating
72 the preferences of raters and analyzes the importance of features for modeling speaker
73 preference. We conclude that modeling the *temporal aspects* of speech is critical for
74 preference estimation.

75 14.2 Rankings from Pairwise Comparisons

76 Rankings have a long history in competitive sports, where individuals or teams play
77 against each other in order to determine who's best. Two common forms, elimination
78 and round-robin tournaments, both require a high degree of control over who plays
79 who, which is not always possible. In addition, they may lead to only partial rank-
80 ings. In chess, Elo's system (Elo, 1978) was designed to overcome these issues: a
81 player's skill is estimated based on prior match outcomes, and skills are updated after
82 each match. Skill changes correspond to the surprisal of the system by the match
83 outcome. A ranking can be derived by ordering players by their skill. Microsoft
84 TrueSkill™ (Herbrich, 2007) uses a Bayesian estimation of rankings from pairwise
85 comparisons originally developed for ranking players of online games (based on
86 their win/loss performance). TrueSkill models skill as a normal distribution, i.e., it
87 makes the system's uncertainty about skill explicit, which enables smoother updates
88 and more robust results when few match outcomes are available.

89 Most work in speech quality estimation has used direct scalar ratings of individual
90 stimuli (Burkhardt, Schuller, Weiss, & Weninger, 2011) or required each subject to
91 assign a complete ranking for all stimuli. Gallardo (2016) feeds paired comparisons
92 into a Bradley–Terry–Luce model (Bradley & Terry, 1952) and finds similar results
93 to direct scaling. Both of these methods have been limited to few raters and/or
94 few stimuli. We extend the methodology introduced by Sakaguchi, Post, and Van
95 Durme (2014) who created rankings for machine translation systems from pairwise
96 comparisons using Microsoft TrueSkill™. In our metaphor, we view each rating as
97 a 'match' in which the preferred stimulus wins against the dispreferred stimulus.
98 We then compute the 'skill' of stimuli and their ranking. TrueSkill also provides
99 *match making* capabilities that, given one player, select an opponent that has the most
100 similar skill and where uncertainty of the skill difference is low (technically, TrueSkill
101 estimates the probability of a draw and prefers matches with high draw probability).
102 This is meant to lead to interesting matches with similarly skillful opponents. We use
103 match making to select stimulus pairs for human rating in an iterative fashion which
104 uses the ratings collected so far to steer our *active sampling* approach to select among
105 the possible stimulus pairs to be compared. We actively select stimulus pairs that are
106 expected to be informative for the full ranking based on a preliminary ranking of all
107 ratings performed so far.

108 In our application, we found the abovementioned strategy for match making to be
109 flawed: as scores tend to get more certain with more data, stimuli are preferred that
110 already participated in many comparisons. As a result, the number of comparisons

111 is not balanced on all stimuli but accumulates on few, well-known anchor points.¹
 112 We use an approach that better balances the number of ratings per stimulus: We
 113 (1) pick a first stimulus based on the system’s uncertainty about its ranking and (2)
 114 compute the match quality for all opponents and pick the opponent based on the
 115 predicted match quality with a dampening factor for the number of comparisons
 116 that the opponent has played so far. As a result, we (a) favor little-tested stimuli
 117 over well-tested ones and (b) select informative games over predictable ones. We
 118 randomly select pairs weighted by the criteria mentioned above which enables us to
 119 sample multiple ‘interesting’ pairs at once.

120 In comparison to Sakaguchi et al. (2014), which ranked 13 translation systems
 121 for which complete evaluation data had already been collected, we rank a total of
 122 223 speakers, thus well over an order of magnitude more, in a live setting without
 123 external reference ranking.

124 14.3 Stimuli and Rating Collection via Crowdsourcing

125 We limit our likability judgements to one specific reading genre: the reading of
 126 encyclopaedic entries in Wikipedia. We use recordings from the Spoken Wikipedia.²
 127 as a broad sample of read *speech in the wild*. The Spoken Wikipedia project unites
 128 volunteer readers who devote significant amounts of time and effort into producing
 129 read versions of Wikipedia articles as an alternate form of access to encyclopaedic
 130 content. It can thus be considered a valid source of speech produced by ambitious
 131 but not always perfect readers. The data has been prepared as a corpus (Baumann,
 132 Köhn, & Hennig, 2018) and the German subset of the corpus, which we use here,
 133 contains ~300 h of speech read by ~300 speakers.

134 To avoid rating preferences based on *what* is spoken rather than how, we choose
 135 as stimuli the opening that is read for every article in the Spoken Wikipedia, which
 136 is (supposed to be) identical for all articles except for the article lemma.³ We extract
 137 that stimulus for every speaker in the German subset of the Spoken Wikipedia Corpus
 138 using the alignment information given in Baumann et al. (2018). As some alignment
 139 information was missing or clearly wrong, our stimulus pool is reduced to 227 speak-
 140 ers. We then masked the article lemma with noise in a length that matches the average
 141 reading speed of the stimulus. The mean/median duration per stimulus is 4.7/4.57 s
 142 with 5/95% quantiles at 3.74/6.03 s.

143 For every rating pair, participants were asked to rate which of the two voices they
 144 would prefer for having a Wikipedia article read out to them. We realized a web-
 145 based rating experiment on the basis of BeagleJS Kraft and Zölzer (2014) which we

¹This may not be a problem when using TrueSkill for match assignment, as participation in games is limited by the players’ availability.

²https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Spoken_Wikipedia.

³Expected reading: ‘Sie hören den Artikel *article lemma* aus Wikipedia, der freien Enzyklopädie.’ (You are listening to the article *article lemma* from Wikipedia, the free encyclopedia.)

146 extended to allow for an open number of pairwise ratings for each participant. The
147 experiment operated with a mini-batch cache of 1000 rating pairs from which clients
148 sampled randomly. The cache was updated manually whenever more than 200 ratings
149 had been submitted by re-creating a new best ranking and selecting stimulus pairs
150 as outlined above. We opted against an active backend with immediate update and
151 selection of the next most relevant rating pair to ensure availability in times of high
152 system usage (e.g., during the minutes after a mailing list advertised our experiment).

153 We solicited participants to our experiment via the German Wikipedia ‘off-topic
154 bulletin board’ and various open mailing lists of student organizations (particularly
155 CS students), as well as the Chaos Computer Club in Germany, Austria, and Switzer-
156 land in order to reach a wide variety of dialect and age groups. We deliberately did
157 not explicitly invite the Spoken Wikipedia community to participate, as they could
158 have been particularly biased.

159 Statistics of the participants’ self-reported meta data are shown in Table 14.1. As
160 can be seen, Northern Germans, males, and 20–30 years olds are over-represented in
161 our data (presumably computer science students at Universität Hamburg). However,
162 almost all other demographic groups are included as well, at least to some extent.
163 In total, we collected 5440 ratings from 168 participants. Participation was strictly
164 voluntary and without compensation and hence the resulting ratings are unlikely to
165 be prone to vandalistic behavior.

166 Although participants could perform as many ratings as they liked, they were
167 instructed that 10 ratings are sufficient, 30–50 preferable, and that they should take a
168 break after 100 ratings (and possibly return the next day). We excluded participants
169 who submitted a single rating only. The median ratings per participant were 26 with
170 half the participants between 11 and 43 ratings and 5/95% quantiles at 4 and 101
171 ratings, respectively.

172 Participants were asked to always state a preference, even if unsure, and did not
173 explicitly have the option to state that they could not decide. It is more informative for
174 our setup to get contradicting preferences than to explicitly invite the participants to
175 omit a decision. As our method steers toward ‘difficult’ comparisons, many omitted
176 decisions could otherwise have been expected. Our software, however, did allow to
177 skip ahead without making a decision and sometimes participants did not provide
178 a decision (accidentally or on purpose). These instances were ignored in further
179 processing, as no rating has been recorded.

180 We also measured the time taken for each rating. The median time per rating is
181 14.3 s with half the ratings between 11.3 and 21.3 s and 5/95% quantiles at 6.3 and
182 39.7 s, respectively. 6.3 seconds can still be considered a reasonable lower bound for
183 listening to both stimuli and then taking the decision quickly. In total, participants
184 spent ~26 h on rating stimulus pairs.⁴

⁴We substitute the median for the slowest 2.5% of ratings, as participants were obviously side-tracked who spent more than 55 s for a single rating.

Table 14.1 Breakdown of self-reported meta information of participants and their rating counts

	Total	Participants	Ratings
		168	5440
Gender	Female	41	1665
	Male	109	3221
	<i>Unreported</i>	18	554
Age	<20	18	358
	20–30	78	2593
	30–40	34	1030
	40–60	24	886
	>60	6	418
	<i>Unreported</i>	8	155
	Dialectal origin	Northern Germany	83
Berlin/Brandenburg		8	128
Northrhine-Westphalia		11	464
Middle Germany		9	443
Rhine-/Saarland		3	82
Baden-Wuerttemberg		15	432
Bavaria		8	405
Austria		5	179
Switzerland		0	0
Unsure/other		26	651

185 The stimulus ordering was randomized. Participants have a slight tendency for
 186 stimulus B over A (2784 versus 2656, n.s.: sign test, $p = .09$), which could be
 187 interpreted as a recency effect.

188 We measure the degree of disagreement by constructing a directed acyclic graph
 189 of the preference relation expressed through all ratings (i.e., the stimuli are nodes
 190 and one edge is introduced per rating). If ratings were consistent, there would not
 191 be any rating circles ($a < b$, $b < c$ but $c < a$) and the proportion of feedback arcs
 192 can be taken as a measure of consistency. We heuristically compute the minimum
 193 feedback arc set of all ratings (Eades, Lin, & Smyth, 1993) and find the proportion
 194 to be 29%. In a preliminary experiment using only 10 stimuli and all 45 possible
 195 comparisons, only one rater was ‘perfect’ in not producing any circles. Hence, we
 196 know that both within-rater and across-rater inconsistencies occur. In addition, our
 197 stimulus selection process is tailored towards choosing pairs that are expected to be
 198 hard to rate (and the disagreeing proportion grew over the runtime of the experiment).

14.4 Ranking Analyses

We feed all pairwise ratings into TrueSkill™ to derive rankings. In TrueSkill, more recent ratings are more influential for the final ranking due to the iterative update mechanism.⁵ As proposed by Sakaguchi et al. (2014), we use the fact that rankings depend on the rating order to validate our method: we permute the ratings and create many rankings for the same set of ratings (below: $N = 300$). We then take the median ranking as the final decision. Thus, we are also able to report ranking confidence levels.⁶

Rankings can be compared using correlation coefficients like Kendall's Tau (Langville & Meyer, 2012, Chap. 16). We find that pairwise correlations of the 300 rankings result in $\tau > 0.92$ and that each ranking against the median ranking gives $\tau > 0.95$. Thus, we conclude that TrueSkill leads to consistent rankings (within bounds) and that the median ranking is a meaningful middle ground for all rankings.

The final median ranking with quartile and 5/95% confidence ranks is shown in Fig. 14.1. As can be seen in the figure, there is no one clear ranking of all speakers. While there is a best and worst stimulus shared among all rankings, variability is larger in the middle. Overall, the average rank variability is 6.7 ranks within the 25–75% confidence interval and 16.4 ranks within the 90% confidence interval. Interestingly, some clusters of similarly 'good' stimuli emerge, e.g., as highlighted in the green circled area where 11 stimuli share similar ranks with a high variability that are delimited with high confidence to higher ranks (upper right of circled area) and slightly less to lower ranks.

Finally, we use rankings to predict the outcome of ratings as another way of testing the ranking validity. We assume that a rating will be 'won' by the better ranked stimulus (although similarly ranked stimuli could easily have any outcome). We use 100-fold cross-validation and find that on average, the prediction performance is 68%. Given that 29% of ratings can be expected to be mis-predicted due to the rating inconsistencies, the rankings have a high level of predictive value. As described above, TrueSkill can compute match quality, effectively describing how likely a rating will lead to disagreement among raters. We find that prediction performance highly correlates with that score (Kendall's $\tau = -0.81$, $p < .001$).

We investigate which stimulus pairs have been selected for comparison to find out whether the method proposed in Sect. 14.2 works effectively. The rated pairs are presented in Fig. 14.2. We find that pairs along the diagonal (i.e., with similar ranks) have been tested more densely than pairs further apart. Furthermore, the plot shows that 'better' stimuli (as per the ranking) win more often against inferior stimuli (green/blue division of the plot) and multiple controversial ratings (red) mostly occur along the diagonal. Overall, our 5440 ratings spread over 4000 different pairs, that is,

⁵This is a feature when ranking human players, as their true performance may change over time – but this is not the case in our experiment.

⁶The confidence is about TrueSkill producing a preference ordering given another permutation of ratings. We cannot make any guarantee with respect to some 'gold' ranking, which does not exist for our data.

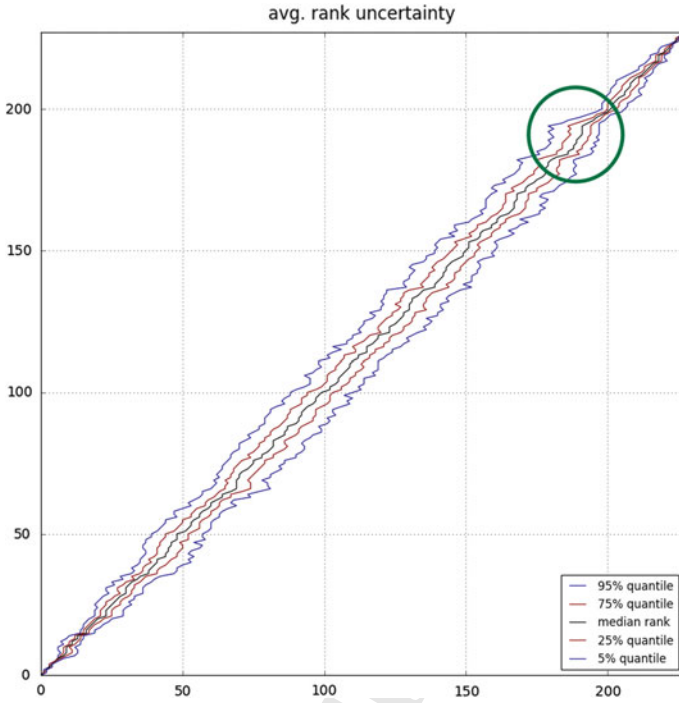


Fig. 14.1 Ranking results (both axes ordered by median ranking) including rank confidence on the x-axis. The circled area is further discussed in the text

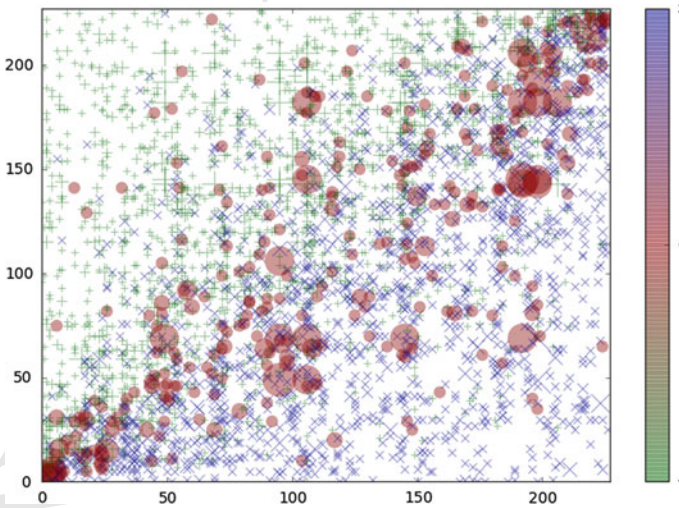


Fig. 14.2 Scatter plot of pairs compared (axes ordered by median ranking, color-coding indicates the avg. outcome of comparisons). The plot is more dense along the diagonal, as stimuli are compared more often when they are of comparable rank

Editor Proof



Fig. 14.3 Line graph comparison of median rankings for female (top) and male (bottom) raters. Stimuli spoken by females are shown in red

237 7,7% of all possible comparisons. 3057 pairs have been tested once, 666 pairs twice,
 238 and the remaining pairs up to 9 times (which seem to be artefacts of older versions of
 239 pair selection). Overall, the average stimulus has been rated 46 times with the 5/95%
 240 quantiles at 39 and 56 ratings. Thus, our rating pair selection strategy successfully
 241 balances stimulus selection and opponent assignment.

242 **14.4.1 The Influence of Rater Population on Ranking** 243 **Outcome**

244 Finally, we analyze the rankings wrt. to gender. We produce one median ranking
 245 each for ratings from female and male listeners (randomly subsampling the male
 246 ratings to the number of female ratings; see Table 14.1). We find only a moderate
 247 correlation ($\tau = 0.44$, $p \ll .001$) between female and male listener rankings, which
 248 indicates different preferences between these listener groups. We further analyze the
 249 ranking wrt. to speaker gender of the stimuli.⁷ The rank assigned to a female speaker
 250 is on average 12.7 ranks better for female than for male listeners (half of the stimuli
 251 between -32 and $+60$ ranks), indicating that one major difference between female
 252 and male listeners is their preference toward female voices.

253 Figure 14.3 compares the gender-dependent rankings (each line corresponds to
 254 a stimulus, female stimuli in red). The less inclined a line, the more similar the
 255 rank for female/male listeners. As can be seen, preferences differ both in ranking
 256 female speakers as for male speakers. It is interesting to note that Dykema, Diloreto,
 257 Price, White, and Schaeffer (2012) find that male speakers respond more truthfully
 258 to questions posed by female voices, yet they seem to disprefer them in our data. The
 259 results highlight the importance of gender-appropriate voice selection for reading
 260 encyclopaedic, and possibly other factual information.

261 We also divide our data by age (<30 versus >30) and dialect (Northern German
 262 versus all other dialects as there is insufficient data to further differentiate among
 263 dialects). In both cases, correlation between the groups is stronger (age: $\tau = 0.50$,
 264 dialect: $\tau = 0.54$) than in the gender partition. No age or dialect information is
 265 available for the speakers, hence we cannot compare within/across-group effects
 266 (e.g., we would expect matched dialects of speaker and listener being preferred).

⁷Unfortunately, just 20 of 227 stimuli (9%) were spoken by females.

14.4.2 Acoustic Correlates of Ranking Quality

We finally experiment with acoustic factors that could explain the speaker likability expressed by the median ranking shown in Fig. 14.1. First, we compute the perceptual quality of audio stimuli as standardized by ITU-T P.563 (Malfait, Berger, & Kastner, 2006). We find a low (but significant) correlation ($\tau = 0.14$, $p < .002$) of achieved median ranking and estimated MOS for the audio transmission quality.⁸ We conclude that carefully arranged recording conditions could coincide with better speech quality, or that listener judgements are influenced by encoding quality—in contrast to Burkhardt et al. (2011) where no such influence was found in a similar task.

We estimate the liveliness of the speaker's prosody as it might be a relevant factor of likability. We compute the pitch range in semi-tones and take the 50% (25–75%) and 90% (5–95%) ranges of the measured pitch. On average, the 50/90% ranges are 4.3/12.8 semi-tones for all speakers. We find very slight but non-significant correlations between either liveliness measure and the ranking. As this could be due to very little data from each short stimulus, we also extract pitch from the full articles. This allows us to estimate each speaker's liveliness *in general*, not just in the opening of the article. Here we find that the inter-quartile (50%) pitch range correlates somewhat ($\tau = 0.10$, $p < .03$) with the ranking.

14.5 Listener Preference Classification

In previous work (Eyben, Wöllmer, & Schuller, (Burkhardt et al., 2011)), speaker likability has been modeled using OpenSmile (Eyben et al., 2010) features based on linear and non-linear aggregation functions (such as means and medians) to aggregate over the duration of the stimulus. Features were used to train classifiers such as SVMs which resulted in moderately high (better than chance) performance in classifying speakers as being above or below median likability (Burkhardt et al., 2011). Like the analyses in Sect. 14.4.2, the abovementioned aggregation functions cannot take into account the context of feature characteristics in the stimulus, and are unlikely to accurately express more fine-grained details relevant for speech quality (such as where and how a pitch accent is realized, beyond mean pitch). In this section, we experiment with neural sequence-learning methods (RNNs) to encode the complex temporal evolution of features of speech quality into a latent feature space and use the difference in these for pairs of speech stimuli to train our classifier.

⁸We must mention that all speech in the Spoken Wikipedia is distributed as OGG/Vorbis, with varying bit rates and under diverse recording conditions.

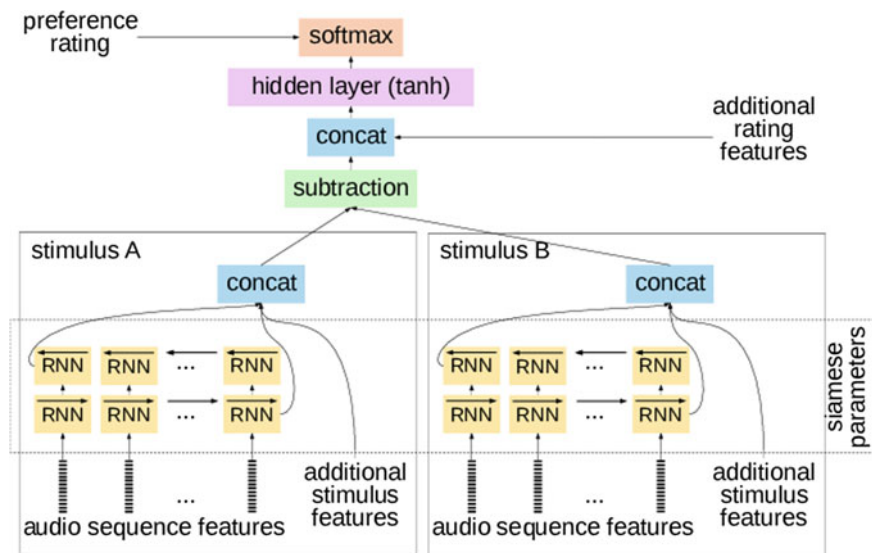


Fig. 14.4 Diagram of the neural architecture for speech likability preference. The task is symmetric (whether a stimulus is A or B is irrelevant) and hence the parameters for the RNNs can be shared (siamese network). Additional features about stimuli and the rating can be concatenated in

14.5.1 Model Architecture

The task of preference ranking is *asymmetric* in the sense that if the two stimuli to be compared are swapped, then the comparison result is the opposite. This has two consequences: (a) parameters for sequence analysis of both stimuli can be shared which is called a *siamese model* (Bromley et al., 1994) and reduces the degrees of freedom of the model, making learning more efficient, and (b) the outputs from sequence analysis of each stimulus can simply be subtracted and the difference be subjected to a final decision layer.

In our model and as shown in Fig. 14.4, we use two layers of bidirectional RNN (LSTMs Hochreiter & Schmidhuber, 1997 or GRUs (Cho et al., 2014)) to model the feature sequence of each stimulus and concatenate the outputs of the forward and backward pass. We can also concatenate additional stimulus-level features into the representation at this time, e.g., measures of signal quality such as ITU-T P.563 (Malfait et al., 2006) (cmp. Sect. 14.4.2), or meta information about the speaker or the audio recording (such as gender or bitrate, cmp. Sect. 14.4.1).

Given that our final decision is based on the quality *difference* alone, not the overall quality, we subtract both stimuli's vectors.⁹ We then pass the difference to one hidden layer and a final binary softmax layer that models the preference decision. We can

⁹We found, in initial experimentation, that this performs much better than concatenating the outputs of each speaker.

318 also optionally include additional meta features of the rating (such as identity, age, or
 319 dialectal region of the rater). These can easily be concatenated in before the hidden
 320 layer, in order to model the relative preferences of individual raters or rater groups.
 321 As we found that preferences differ, this could be useful information.

322 14.5.2 Data and Evaluation

323 The original purpose of the rating collection reported in Sect. 14.3 was to create a
 324 ranking and effort was put into maximizing the efficiency of human annotation by
 325 focusing the human effort on ‘difficult’ pairs using *active sampling* of stimulus pairs
 326 that focus human annotation effort on ‘similarly good’ speakers. As a result, the
 327 stimulus pairs that were rated by participants are much more similar in their quality
 328 than randomly selected stimulus pairs would be.

329 In addition, inconsistency in the data set is high, as are pairs of stimuli that have
 330 been rated multiple times. Above, we have computed the minimum feedback arc set,
 331 i.e., the subset of ratings that lead to a fully consistent ranking (Eades et al., 1993).
 332 We found the proportion of conflicting arcs to be 29%. This can act as an indicator of
 333 the proportion of ratings that are inconsistent (where potentially different raters have
 334 different preferences, or simply cannot reliably tell the difference). In addition, we
 335 here compute an oracle correctness for all pairs that have been rated more than once,
 336 by checking for each rating, if it is the majority rating for this pair (deciding randomly
 337 to resolve draws). We find that such an *oracle classifier* reaches a correctness of only
 338 65% for those pairs that have been rated more than once. Pairs that were rated just
 339 once *potentially* are easier to classify, which makes it possible to beat this oracle.

340 For evaluations, we report multiple settings below. The settings are meant to
 341 counterbalance the difficulties introduced by the data elicitation technique and to
 342 test different aspects of listener preference classification:

- 343 **naïve** we sample randomly among the evaluation instances from the corpus of
 344 human-rated pairs; as the corpus focuses on difficult pairs, we cannot expect
 345 a spectacular performance;
- 346 **easy** based on the median ranking derived in Sect. 14.4, we sample instances with
 347 ‘large’ ranking differences (distance on the ranking scale >0.25 or >0.5), in
 348 order to test if our classifiers fare better with stronger preference differences
 349 (and hence easier to identify differences in speech quality).

350 Given that stimuli were presented in random order, the data set is balanced in
 351 terms of which stimulus outperforms the other. Thus, we focus on accuracy as the
 352 only evaluation metric.

14.5.3 Features and Conditions

Using a sliding window, we derive a multitude of local features from the audio stream that might capture aspects of speech quality. All features use a frame shift of 10 ms. In particular, we measure Mel-frequency cepstral coefficients (MFCCs, 12 + 1 energy) to capture voice and recording characteristics, f_0 (measured using `Snack`'s `esps` implementation) as a first measure of speech melody, and fundamental frequency variation (FFV) features (Laskowski, Heldner, & Edlund, 2008) as these are more robust (and might contain more valuable information) than single f_0 . Using Praat (Boersma, 2002), we compute jitter (PPQ5), shimmer (APQ5), and harmonics-to-noise ratio (Boersma, 1993). We do not perform z-scale normalization on the feature streams.

The Spoken Wikipedia Corpus also contains phonetic alignments that were computed using the MAUS tool (Schiel, 2004). The alignments allow us to assign phone annotations to every frame. With this information, the model is informed explicitly that different phones have different phonetic characteristics (as expressed in the MFCCs) and can condition its learning of speech quality on these characteristics. In other words: the model can learn to focus on a phone's quality aspects (e.g., nasalization) without needing to learn to differentiate phones.

One frame of features for every 10 ms may overwhelm the model with very large amounts of parameters, reducing training efficiency as well as effectiveness. In order to keep training tractable, we subsample the feature frames with various values (see *seq. step size* in Table 14.2). When we do so, we use mean aggregation for numeric values (ignoring missing values for pitch and HNR).

14.5.4 Experiments and Results

We separate out about 1/10th of the 5440 ratings as the test data: the **naïve** test set contains 400 ratings, and we sample among ratings with 'large' differences 100 ratings each for the >0.25 and >0.5 **easy** test sets.

We implemented our network in `dynet` (Neubig, 2017). In the experiments reported below, we train for 50 epochs using `AdamTrainer` and no dropout. We concatenate the various audio features that are computed for every frame. We use embeddings to characterize the phonetic labels.

14.5.4.1 Meta Parameter Optimization

As originally reported in Baumann (2018), we have performed an optimization to find good sizes for various meta parameters of the model:

- To reduce the length of the sequence that need to be learned by the LSTMs (and to avoid the problem of vanishing gradients through long sequences), we subsample

Table 14.2 Meta parameters considered in grid search. Best values are shown in boldface

Meta parameter	Values
Sequence step size	5 , 10, 15
Phone embed size	8, 16 , 24
Sequence state size	24, 32, 48 , 64
Hidden layer size	2, 3 , 4 × sequence state size

389 the audio features by mean-aggregating values over a number of frames (5, 10, or
390 15).

- 391 • To represent the discrete phonetic labels, we use embeddings of varying sizes (8,
392 16, or 24), in order to allow the model to cluster similar phones.
- 393 • The sequential LSTM state size determines how many dimensions can be consid-
394 ered during the sequence analysis and we experiment with various sizes (24, 32,
395 48, or 64).
- 396 • The output from concatenation of both forward and backward LSTMs doubles the
397 size of the next layer’s input. For the hidden layer size, we hence consider scaling
398 factors (2, 3, or 4) over the size of the sequential state size.

399 We performed a grid search over the possible meta parameter values as summa-
400 rized in Table 14.2 and focusing on the naïve data set. We found an optimum for
401 sequence step size of 5 (i.e., one feature frame for every 50 ms of speech), phone
402 embedding with 16 dimensions, sequence state size of 48, and hidden layer size of
403 $3 \times 48 = 144$ (sequence state size of 32 and $4 \times 32 = 128$ was a close contender).

404 At these settings, our model yields an accuracy of 67.25% on the naïve test set,
405 93% on the **easy-0.25** test set and 97% on the **easy-0.5** test set. The accuracy on the
406 naïve test set is close to what we estimated as the upper limit for the harder part of
407 our training data.

408 14.5.4.2 Ablation Study on Phonemic Alignments

409 We hypothesized above that our performance gain over previous work may be largely
410 due to the model being able to perform prosodically meaningful aggregations and
411 could, for example, relate prosodic parameters to the phones spoken. To test this
412 hypothesis, we perform an ablation study and remove the phoneme embeddings
413 from the input features. We perform this experiment with the other meta parameters
414 set to their optima as found in the previous subsection. As shown in Table 14.3, we
415 find performance to drop substantially when the phone identity feature is removed.
416 We believe this is because the model is unable to make maximum sense of features
417 such as MFCCs given speech quality is obviously just a secondary feature, far behind
418 phone identities. If the model is not informed about the phonetic identities, it needs to
419 resolve whether input has good quality, whereas the full model only needs to resolve
420 the quality of a feature given the particular speech sound.

Table 14.3 Accuracy (in percent) of full and reduced feature set (without phone alignment)

Setting	Accuracy		
	naïve	easy-0.25	easy-0.5
Full mode	67.25	93	97
Without phone alignment	58.75	73	80

14.6 Conclusions and Future Work

We have presented a method for creating crowdsourced speaker likability rankings from pairwise comparisons. The material that we base our ratings on is freely available and we likewise publish the ratings and the software to derive rankings from those ratings under the same terms. Unlike Gallardo (2016) which uses Bradley–Terry–Luce models, our method does not require a complete comparison of all pairs, and works on a small subset (in our case: 7% of possible comparisons) jointly provided by many participants.

One advantage of the Spoken Wikipedia corpus is the availability of much more data from each speaker beyond the short stimuli that are used in the ranking experiment. Thus, more complex characteristics of a speaker, such as accentuation and other prosodic idiosyncrasies (which listeners presumably would be able to judge in one sentence), can be derived from up to an hour of (closely transcribed and aligned) speech. In fact, we found in Sect. 14.4.2 that extracting the 50% pitch range as an estimate of liveliness significantly correlates with likability, at least if liveliness is extracted from the full speech, rather than just the one sentence used in human ratings, potentially because this circumvents effects from faulty fundamental frequency extraction.

We have also presented a neural architecture for determining which of two speech stimuli is rated as the better of the two in noisy human annotations. Our model yields good performance most likely because the RNN provides for complex aggregations of the (conventional) feature sequences. Our model’s aggregations are able account for sequential information, in particular it is able to relate acoustic features to the phones spoken, unlike more coarse-grained aggregation functions as have been used before.

In Burkhardt et al. (2011), the authors train classifiers to differentiate whether a stimulus is better/worse than average and reach a classification accuracy of 67.6%. Their setup is comparable to our decisions for stimuli that are relatively far apart on the rating scale, in which case the neural aggregation and classification yields a classification accuracy of 93–97%. We believe this to be caused by the better temporal modeling of our approach and the use of phonetic identities during aggregation.

Despite the relatively good results, our method is still basic in terms of the neural architecture employed. In particular, our method does not yet employ an attention mechanism that could help to better weigh the speech quality encoding. Given that all speakers in our corpus speak (more or less) the same content, we envision that

our model would profit greatly if the comparison between both stimuli could attend to particular differences rather than only the comparison of the final BiLSTM output vectors. An attention model would also help the analysis of *why* a speaker is rated as better than another, as it would indicate the relative importance of parts of the stimuli in the comparison. Another venue, at least for comparisons on shared text would be connectionist temporal classification to temporally relate the feature streams before comparison for a better notion of timing differences between the stimuli. Finally, it might be worthwhile to pre-train the intermediate representations of the model.

In the end, our model could weigh slight mis-pronunciations against voice quality or prosodic phrasing, and we intend to use analysis techniques to ultimately understand the relative weights of these aspects in comparisons.

We have limited our study to one identical stimulus sentence in order to exclude contextual differences, and to one stimulus per speaker. We plan to extend the study to other stimulus pairs where the sentences (or sentence fragments) are spoken by different speakers across the Spoken Wikipedia. In this way, we hope to get a better judgement of the speakers, based on more than (on average) 4.7 s of their speech.

Acknowledgments We thank our listeners/raters as well as the volunteers of the Spoken Wikipedia for donating their time and voice. This work has been partially supported by a Daimler and Benz Foundation PostDoc grant.

References

- Baumann, T. (2018). Learning to determine who is the better speaker. In *Proceedings of Speech Prosody*.
- Baumann, T., Köhn, A., & Hennig, F. (2018). The Spoken Wikipedia corpus collection: Harvesting, alignment and an application to hyper listening. In *Language resources and evaluation. Special Issue representing significant contributions of LREC 2016*.
- Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences*, 17, 97–110.
- Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glott International*, 5(9/10), 341–345.
- Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika*, 39(3/4), 324–345.
- Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1994). Signature verification using a "siamese" time delay neural network. In *Advances in neural information processing systems* (pp. 737–744).
- Burkhardt, F., Schuller, B., Weiss, B., & Wening, F. (2011). Would you buy a car from me? On the likability of telephone voices. In *Proceedings of Interspeech. ISCA*.
- Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1724–1734). Doha, Qatar: Association for Computational Linguistics. <http://www.aclweb.org/anthology/D14-1179>.
- Dykema, J., DiIoreto, K., Price, J. L., White, E., & Schaeffer, N. C. (2012). ACASI gender-of-interviewer voice effects on reports to questions about sensitive behaviors among young adults. *Public Opinion Quarterly*, 76(2), 311–325.

- 500 Eades, P., Lin, X., & Smyth, W. F. (1993). A fast and effective heuristic for the feedback arc set
501 problem. *Information Processing Letters*, 47(6), 319–323.
- 502 Elo, A. E. (1978). The rating of chessplayers, past and present. Arco Pub.
- 503 Eyben, F., Wöllmer, M., & Schuller, B. (2010). Opensmile: The Munich versatile and fast open-
504 source audio feature extractor. In *Proceedings of the 18th ACM International Conference on*
505 *Multimedia* (pp. 1459–1462). ACM.
- 506 Gallardo, L. F. (2016). A paired-comparison listening test for collecting voice likability scores. In
507 *Speech Communication; 12. ITG Symposium; Proceedings of VDE* (pp. 1–5).
- 508 Herbrich, R., Minka, T., & Graepel, T. (2007). TrueSkill™: A Bayesian skill rating system. In
509 *Advances in neural information processing systems* (vol. 20, pp. 569–576). MIT Press.
- 510 Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8),
511 1735–1780.
- 512 Kraft, S., & Zölzer, U. (2014). BeagleJS: HTML5 and JavaScript based framework for the subjective
513 evaluation of audio quality. In *Linux Audio Conference*.
- 514 Langville, A. N., & Meyer, C. D. (2012). *Who's #1? The Science of Rating and Ranking*. Princeton
515 University Press.
- 516 Laskowski, K., Heldner, M., & Edlund, J. (2008). The fundamental frequency variation spectrum.
517 In *Proceedings of FONETIK 2008*.
- 518 Malfait, L., Berger, J., & Kastner, M. (2006). P.563—The ITU-T standard for single-ended speech
519 quality assessment. *IEEE Transactions on Audio, Speech and Language Processing*, 14(6), 1924–
520 1934.
- 521 Neubig, G. et al. (2017). DyNet: The dynamic neural network toolkit. In arXiv preprint
522 [arXiv:1701.03980](https://arxiv.org/abs/1701.03980).
- 523 Sakaguchi, K., Post, M., & Van Durme, B. (2014). Efficient elicitation of annotations for human
524 evaluation of machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine*
525 *Translation* (pp. 1–11). Baltimore, Maryland, USA: ACL.
- 526 Schiel, F. (2004). MAUS goes iterative. In *Proceedings of the LREC*.

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Voice Attractiveness
Series Title	
Chapter Title	Multidimensional Mapping of Voice Attractiveness and Listener's Preference: Optimization and Estimation from Audio Signal
Copyright Year	2020
Copyright HolderName	Springer Nature Singapore Pte Ltd.
Corresponding Author	Family Name Obuchi Particle Given Name Yasunari Prefix Suffix Role Division Organization School of Media Science, Tokyo University of Technology Address 1404-1 Katakura, Hachioji, Tokyo, 192-0982, Japan Email obuchiysnr@stf.teu.ac.jp
Abstract	In this chapter, a new framework of listener-dependent quantification of voice attractiveness is introduced. The probabilistic model of paired comparison results is extended to the multidimensional merit space, in which the intrinsic attractiveness of voices and the preference of listeners are both expressed as vectors. The attractiveness for a specific listener is then obtained by calculating the inner product of two vectors. The mapping from the paired comparison results to the multidimensional merit space is formulated as the maximization problem of the likelihood function. After the optimal mapping is obtained, we also discuss the possibility of predicting the attractiveness from the acoustic features. Machine learning approach is introduced to analyze the real data of Japanese greeting phrase "irasshaimase," and the effectiveness is confirmed by the higher prediction accuracy.
Keywords	Paired comparison - Mapping - Optimization - Listener's preference - Multidimensional - Acoustic feature - Machine learning

Chapter 15

Multidimensional Mapping of Voice Attractiveness and Listener's Preference: Optimization and Estimation from Audio Signal



Yasunari Obuchi

Abstract In this chapter, a new framework of listener-dependent quantification of voice attractiveness is introduced. The probabilistic model of paired comparison results is extended to the multidimensional merit space, in which the intrinsic attractiveness of voices and the preference of listeners are both expressed as vectors. The attractiveness for a specific listener is then obtained by calculating the inner product of two vectors. The mapping from the paired comparison results to the multidimensional merit space is formulated as the maximization problem of the likelihood function. After the optimal mapping is obtained, we also discuss the possibility of predicting the attractiveness from the acoustic features. Machine learning approach is introduced to analyze the real data of Japanese greeting phrase “irasshaimase,” and the effectiveness is confirmed by the higher prediction accuracy.

Keywords Paired comparison · Mapping · Optimization · Listener's preference
Multidimensional · Acoustic feature · Machine learning

15.1 Introduction

Most people believe that there are attractive voices and unattractive voices. However, they also believe that there are voices that are attractive and unattractive depending on who is listening. This chapter deals with such objectivity and subjectivity of voice attractiveness. We start the discussion by establishing a framework of voice attractiveness quantification based on the probabilistic analysis of experimental results. Once the quantification framework is given, we then try to predict the attractiveness of a new voice from its acoustic characteristics.

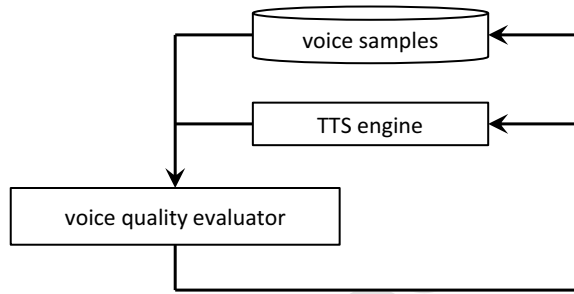
In this chapter, we focus on the social attractiveness, especially in a commercial context. For example, if you make a commercial video for your product with some narration, its attractiveness has strong influence on your business. In the pre-Internet

Y. Obuchi (✉)
School of Media Science, Tokyo University of Technology, 1404-1 Katakura, Hachioji, Tokyo
192-0982, Japan
e-mail: obuchiysnr@stf.teu.ac.jp

© Springer Nature Singapore Pte Ltd. 2020
B. Weiss et al. (eds.), *Voice Attractiveness*, Prosody, Phonology and Phonetics,
https://doi.org/10.1007/978-981-15-6627-1_15

287

Fig. 15.1 Schematic diagram of voice selection process



era, voice attractiveness was evaluated as a scalar, typically the average of evaluation by a mass audience. However, in the Internet era, in which the contents can be customized on delivery, it is important to evaluate voice attractiveness for each customer.

The voice selection process in commercial applications is illustrated in Fig. 15.1. If the system uses recorded voice samples, the quality of each sample is evaluated, and the sample with the highest score is selected. If the system uses text-to-speech (TTS) software, the evaluation score is fed back to the TTS engine, and the system parameters are adjusted. In both cases, the voice quality evaluator plays the central role.

Voice quality evaluation can be realized by collecting a mass of human judgments. Typical evaluation methods include the mean opinion score (MOS) test and preference test. The MOS test is designed to give an absolute score for each voice sample, whereas the preference test focuses on relative quality of two or more voice samples. Both tests are based on subjective judgment of human listeners, and it requires days or weeks of evaluation process in the development cycle. If we can replace the human-based evaluator with the computer-based automatic evaluator, the development cycle would be accelerated dramatically.

Although we have limited insight into the physical features representing voice attractiveness, an automatic evaluator can be developed using the machine learning framework. If we have plenty of data with correct attractiveness label, machine learning algorithms such as support vector machine can provide a model which connects voice signals and their attractiveness.

Figure 15.2 shows the way to train the model from a large database. Before starting the training process by a machine learning tool, we have to prepare appropriate labels of voice attractiveness. We know from our daily experiences that the definition of attractiveness is ambiguous, and the evaluation results of human listeners are often inconsistent. Therefore, the first important problem is how to prepare correct attractiveness labels. For this problem, we start with the paired comparison test (Shah et al., 2014). Since it is difficult to give a concrete definition of the attractiveness scale, it is easier for a typical listener to answer the question “which voice is more attractive, A or B?” than the question “how attractive is this voice in the scale of one to five?”.

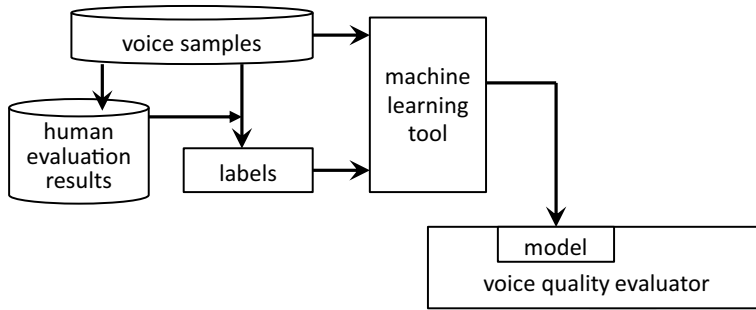


Fig. 15.2 Schematic diagram of voice quality evaluator training

58 The drawback of paired comparison test is that it includes more inconsistency than
 59 the absolute evaluation such as MOS test. In addition to the inconsistency between
 60 listeners, sometimes a listener gives inconsistent evaluation results within his/her
 61 own comparisons. A typical example is that A is better than B, B is better than C,
 62 and C is better than A. Such kind of discussion leads us to the probabilistic model of
 63 paired comparison test, in which the intrinsic attractiveness of voices works as the
 64 parameter of winning probability.

65 In this chapter, a new probabilistic model of voice attractiveness is introduced, in
 66 which the intrinsic attractiveness of voices and the intrinsic preference of listeners are
 67 handled in a unified form. In this model, the voice attractiveness is represented as a
 68 multidimensional vector in the “merit space.” The preference of listener is described
 69 as a direction in the merit space, and the evaluation of a voice by a listener is expressed
 70 as the inner product of those vectors. The process to obtain the optimal mapping from
 71 the paired comparison results will also be discussed.

72 After the optimal mapping is obtained, we move on to the discussion of merit
 73 vector estimation from the acoustic features of a new voice. If the estimation scheme
 74 is established, we can predict the comparison results for the new voice, and the effec-
 75 tiveness of the whole framework can be confirmed by the correctness of prediction.

76 15.2 Analysis of Paired Comparison

77 In this section, we discuss how to model the paired comparison test of voice attrac-
 78 tiveness. In the development process of text-to-speech (TTS) systems, voice quality
 79 assessment by MOS (Ribeiro, Florêncio, Zhang, & Seltzer, 2011) and paired compar-
 80 ison (Zen, Tokuda, Masuko, Kobayashi, & Kitamura, 2004; Junichi, Onishi, Masuko,
 81 & Kobayashi, 2005) test are both used. Although the MOS test has the advantage
 82 that the results can be used directly as the absolute attractiveness value, it imposes a
 83 heavy burden on listeners and the results are often unreliable. That is the reason why
 84 we decided to use the paired comparison test. Below we introduce various models
 85 of paired comparison result interpretation.

15.2.1 Universal Attractiveness Model

A straightforward interpretation of paired comparison test is that each voice has own attractiveness and the listener compares two attractiveness values to make a decision. In this paper, such interpretation is referred to as the *universal attractiveness model* because each voice sample is assumed to have one attractiveness value which is applicable to everyone.

The easiest case is that all comparison results are completely consistent. If so, we can obtain at least the order of the voices. Any mapping rule can be acceptable if it satisfies the order. However, such results are rarely obtained, and we need a mapping rule to handle inconsistent results that are unavoidable. From that viewpoint, various analysis methods were applied in various fields, including an example in sports competition (Cattelan, Varin, & Firth, 2013).

A typical approach of probabilistic modeling is based on minimization of the log likelihood function. Assuming that the probability of the voice i 's winning against the voice j only depends on the difference of their attractiveness values, the total log likelihood function L can be written as follows.

$$L = \sum_i \sum_j w_{ij} \log f(d_{ij}) \quad (15.1)$$

$$d_{ij} = a_i - a_j \quad (15.2)$$

where a_i and a_j are the attractiveness of the voice i and j , $f(d_{ij})$ is the probability that the voice i wins against the voice j , and w_{ij} is the number of voice i 's winning against the voice j .

Historically, there have been two major models of $f(d_{ij})$. In the Bradley–Terry model (Bradley & Terry, 1952), two voices behave as though competing for the shared resource and the probability represents the ratio of one's resource over the other.

$$f(d_{ij}) = \frac{e^{a_i}}{e^{a_i} + e^{a_j}} = \frac{1}{1 + e^{-d_{ij}}} \quad (15.3)$$

In the Thurstone–Mosteller Case V model (Mosteller, 1951), the observation probability of each voice is assumed to be a Gaussian whose mean corresponds to its own attractiveness. The probability that the voice i wins against the voice j is equal to the probability that the observation of voice i is larger than the observation of voice j , which is calculated by

$$f(d_{ij}) = \frac{1}{2}(1 + \operatorname{erf}(d_{ij})) \quad (15.4)$$

where erf represents the error function.

In both cases, a scaling factor can be multiplied to d_{ij} . A larger scaling factor induces less frequent “upset,” in which a less attractive voice wins against a more

attractive voice. However, it was omitted in the definition of $f(d_{ij})$ because the attractiveness a_i itself has the freedom of scale.

If the definition of $f(d_{ij})$ is fixed, it is easy to obtain the optimal set of $\{a_i\}$ for a given result of paired comparison. We start with randomly selected initial values of $\{a_i\}$, and update them on the direction of gradient ascent of L .

$$\frac{\partial L}{\partial a_i} = \sum_{j \neq i} f'(d_{ij}) \left(\frac{w_{ij}}{f(d_{ij})} - \frac{w_{ji}}{1-f(d_{ij})} \right) \quad (15.5)$$

15.2.2 Personalized Attractiveness Model

In paired comparison test, two listeners may have opposing opinions for a given pair. In the universal attractiveness model, such event is interpreted simply as an occurrence of less likely event. However, in our social experience, we would react to such situations by saying “tastes differ.” However, we also feel that there are some voices which many people tend to like. These two aspects of voice attractiveness—listener’s preference and voice likability—can be modeled by defining voice attractiveness as a function of voice-originated character and listener-originated character. Such relationship can be visualized by mapping voices onto a multidimensional merit space, in which the voices are given as points and the preferences are given as directions.

There have been some studies proposing multidimensional extension of Bradley–Terry model. Fujimoto, Hino, and Murata (2009) proposed a mixture model and applied it as a visualization method to the movie rating task. The idea of calculating the inner product between the voice-originated vector and listener-originated vector is an extension of Fujimoto’s model. Another example is the work of Causeur and Husson (2005), in which a two-dimensional model representing ranking and relevance axes is proposed and applied to the consumer’s preference of cornflakes.

In the proposed personalized attractiveness model, the likelihood function has the same form as the universal attractiveness model, but an additional parameter of listener’s index k is introduced.

$$L = \sum_i \sum_j \sum_k w_{ijk} \log f(d_{ijk}) \quad (15.6)$$

where w_{ijk} is the number of times the listener k prefers the voice i to the voice j . The fact that d_{ijk} includes the listener index k means that the attractiveness a_{ik} depends on the listener k . A simple model to define the listener-dependent attractiveness is

$$d_{ijk} = a_{ik} - a_{jk} \quad (15.7)$$

$$a_{ik} = \mathbf{p}_k \cdot \mathbf{m}_i \quad (15.8)$$

153 where \mathbf{p}_k ($|\mathbf{p}_k| = 1$) is the preference vector for the listener k and \mathbf{m}_i is the merit
154 vector intrinsic for the voice i .

155 The process to obtain the optimal set of preference vectors and merit vectors is
156 similar to the case of universal attractiveness model. We start with randomly selected
157 initial values of $\{\mathbf{p}_k\}$ and $\{\mathbf{m}_i\}$, and update them on the direction of gradient ascent
158 of L .

159 In the 2-dimensional case, we can describe the preference and merit vectors as
160 $\mathbf{p}_k = (\cos \theta_k, \sin \theta_k)^T$ and $\mathbf{m}_i = (\xi_i, \eta_i)^T$, where θ_k , ξ_i , and η_i are the parameters to
161 be adjusted. The parameter θ represents the preference angle. Two parameters ξ and η
162 are interchangeable, and represent the elements of merit vector. The differentiation
163 of L in terms of those parameters are calculated as follows.

$$164 \quad \frac{\partial L}{\partial \xi_i} = \sum_{j \neq i} \sum_k f'(d_{ijk}) \left(\frac{w_{ijk}}{f(d_{ijk})} - \frac{w_{jik}}{1 - f(d_{ijk})} \right) \cos \theta_k \quad (15.9)$$

$$165 \quad \frac{\partial L}{\partial \eta_i} = \sum_{j \neq i} \sum_k f'(d_{ijk}) \left(\frac{w_{ijk}}{f(d_{ijk})} - \frac{w_{jik}}{1 - f(d_{ijk})} \right) \sin \theta_k \quad (15.10)$$

$$166 \quad \frac{\partial L}{\partial \theta_k} = \sum_i \sum_{j \neq i} f'(d_{ijk}) \left(\frac{w_{ijk}}{f(d_{ijk})} - \frac{w_{jik}}{1 - f(d_{ijk})} \right) r_{ijk} \quad (15.11)$$

$$167 \quad d_{ijk} = (\xi_i - \xi_j) \cos \theta_k + (\eta_i - \eta_j) \sin \theta_k \quad (15.12)$$

$$168 \quad r_{ijk} = (\eta_i - \eta_j) \cos \theta_k - (\xi_i - \xi_j) \sin \theta_k \quad (15.13)$$

169 Additional restriction is applied to constrain the vectors in the unit square.

$$170 \quad 0 \leq \xi_i \leq 1 \quad (15.14)$$

$$171 \quad 0 \leq \eta_i \leq 1 \quad (15.15)$$

$$172 \quad 0 \leq \theta_k \leq \pi/2 \quad (15.16)$$

173 Using above equations, the optimization procedure can be described by the pseudo-
174 code shown in Fig. 15.3. Since the quality of solution strongly depends on the
175 initial values, the procedure is repeated using various initial values, and the best
176 combination is selected as the final solution.

177 The process described above can be extended to the higher dimensional cases
178 easily. In the N -dimensional space, we assume

$$179 \quad \mathbf{m}_i = [\xi_{1i}, \xi_{2i}, \dots, \xi_{Ni}] \quad (15.17)$$

$$180 \quad \mathbf{p}_k = [\sin \theta_{1k} \sin \theta_{2k} \cdots \sin \theta_{N-1,k} \cos \theta_{Nk},$$

$$181 \quad \sin \theta_{1k} \sin \theta_{2k} \cdots \sin \theta_{Nk},$$

$$182 \quad \sin \theta_{1k} \sin \theta_{2k} \cdots \sin \theta_{N-2,k} \cos \theta_{N-1,k},$$

$$183 \quad \vdots$$

$$184 \quad \sin \theta_{1k} \cos \theta_{2k},$$

Fig. 15.3 Pseudocode for log likelihood maximization in the two-dimensional merit space

```

1: set small step value of  $s$ 
2: repeat
3:   initialize  $\{\xi_i\}, \{\eta_i\}, \{\theta_k\}$  randomly
4:   repeat
5:     for all  $i, k$  do
6:       calculate  $\frac{\partial L}{\partial \xi_i}, \frac{\partial L}{\partial \eta_i}, \frac{\partial L}{\partial \theta_k}$  using (15.9),(15.10),(15.11)
7:     end for
8:     for all  $i, k$  do
9:        $\xi_i \leftarrow \xi_i + \frac{\partial L}{\partial \xi_i} s$ 
10:       $\xi_i \leftarrow \max(\min(\xi_i, 1), 0)$ 
11:       $\eta_i \leftarrow \eta_i + \frac{\partial L}{\partial \eta_i} s$ 
12:       $\eta_i \leftarrow \max(\min(\eta_i, 1), 0)$ 
13:       $\theta_k \leftarrow \theta_k + \frac{\partial L}{\partial \theta_k} s$ 
14:       $\theta_k \leftarrow \max(\min(\theta_k, \frac{\pi}{2}), 0)$ 
15:    end for
16:  until converge
17:  calculate  $L$  using (15.6) and store  $\{\xi_i\}, \{\eta_i\}, \{\theta_k\}, L$ 
18: until  $N$  times
19: return  $\{\xi_i\}, \{\eta_i\}, \{\theta_k\}$  that yielded the largest  $L$ 

```

$$\cos \theta_{1k}] \quad (15.18)$$

and obtain the update rule, which is a straightforward extension of Eqs. (15.9)–(15.13).

Finally, we selected the Thurstone-Mosteller Case V model (15.4) for f , and the derivative is given by

$$f'(d_{ijk}) = \frac{1}{\sqrt{\pi}} e^{-d_{ijk}^2} \quad (15.19)$$

In fact, the factor $1/\sqrt{\pi}$ can be omitted because the step s has the freedom of arbitrary scale.

15.3 Estimating Merits from Acoustic Features

There have been many studies that tried to connect the subjective nature of voices and their physical characteristics. The largest field is emotion recognition from speech. Various acoustic features and machine learning techniques were applied to predict the emotional state of speaking person, and the achievements were compared in challenges (Ringeval et al., 2017; Schuller et al., 2017). Early researches focused on the prosodic features such as F0 and loudness (Tato, Santos, Kompe, & Pardo, 2002), but the cepstral features were also found to be effective (Sato & Obuchi, 2007). In recent years, it is common to prepare many features and apply machine learning algorithms to find the best feature set. The success of those studies encouraged us to connect the voice attractiveness in the multidimensional merit space and the acoustic features using the machine learning framework.

We start the analysis by preparing a redundant set of acoustic features. Those features are extracted using **OpenSMILE** (Eyben, Wenginger, Groß, & Schuller, 2013) and **Julius** (Lee & Kawahara, 2009).

OpenSMILE is a multi-purpose feature extractor from audio signal. It divides the input voice into 25 ms overlapping frames with 10 ms frame interval, and extract various low-level descriptors (LLDs) including energy, pitch, and spectral centroid. Those LLDs are accumulated from all frames, and then various interframe features (functionals) are extracted from each type of LLD. As shown in Tables 15.1, we prepared 14 LLDs related to energy, pitch, and spectral features, and 23 functionals related to extremes, regression, and segment. The total number of features extracted by OpenSMILE is 322.

Julius is an open-source speech recognition engine. We assume that the transcription of voice is given, and Julius is used as the forced-alignment tool. The features provided by Julius include the acoustic model score, total utterance length, and the length of the final phoneme (mostly vowels in Japanese). The first feature indicates how typically the utterance was pronounced, because it represents the distance between the utterance and the standard acoustic model. The second feature indicates how fast the utterance was pronounced. The third feature represents the hesitation, which is frequently observed in Japanese conversation.

Starting with 325 baseline features (322 from OpenSMILE and 3 from Julius), we try to reduce the number of features using the backward stepwise selection (BSS) framework. For any set of features, candidate subsets are made by removing single feature, and each subset is evaluated by cross validation. After evaluating all subsets, the subset with the highest score survives as the set for next step. The same procedure is repeated until only one feature remains. We also tried forward stepwise selection (FSS) in which a null set is prepared as the baseline, and candidate features are added

Table 15.1 List of LLDs and functionals. Linreg stands for linear regression, qreg for quadratic regression, and seglen for segment length

LLDs		Functionals		
Energy/Pitch	Spectral	Extremes	Regression	Segment
RMS energy	Max position	Max	Linreg slope	Number of seg
Log energy	Min position	Min	Linreg offset	Seglen mean
F0	Centroid	Range	Linreg linear error	Seglen max
Voicing prob.	Entropy	Max position	Linreg quadratic error	Seglen min
	Variance	Min position	Qreg coef 1	Seglen std. dev.
	Skewness	Mean	Qreg coef 2	
	Kurtosis	Max–mean	Qreg coef 3	
	Slope	Mean–min	Qreg linear error	
	Harmonicity		Qreg quadratic error	
	Sharpness		Qreg contour centroid	

231 step by step. However, FSS achieves much worse results than BSS, so the detailed
 232 investigation was done with BSS only.

233 Prediction of the multidimensional merit values are realized by the regression
 234 algorithm called SMOREg (Shevade, Keerthi, Bhattacharyya, & Murthy, 2000, which
 235 is an extension of support vector machine algorithm. We use WEKA (Hall et al.,
 236 2009), which provides various machine learning algorithms including SMOREg.

237 15.4 Experimental Results

238 In this section, two important issues are examined by the experiments using a real
 239 database. The first issue is how efficient mappings of voices can be obtained by the
 240 optimization process of personalized attractiveness model. The second issue is how
 241 accurately those mappings can be reproduced from the unknown voice using acoustic
 242 features.

243 15.4.1 Recordings and Comparisons

244 For the experiments, we recorded voices of Japanese greeting “irasshaimase (wel-
 245 come)” uttered by 115 university students. They were recorded using Panasonic
 246 RR-XS355 digital voice recorders with 44.1 kHz sampling rate, stereo recording and
 247 16-bit quantization condition. Since “irasshaimase” is the phrase given by the shop
 248 clerk every time a customer comes in, it is uttered very frequently in commercial
 249 situations and its impression is very important for the business. The recording was
 250 done in a typical classroom situation, in which voluntary students with no payment
 251 were asked to say “irasshaimase” one by one. No instruction was given as for the
 252 speaking style. Silence was not kept during the recording and the recorded voices
 253 include some environmental noises.

254 In the feature extraction process, OpenSMILE version 2.3.0 rc1 and Julius version
 255 4.2 grammar kit were used. OpenSMILE used the recorded data as their original
 256 format, and Julius used the converted version to 16 kHz monaural sampling. The
 257 original Japanese acoustic model delivered with the Julius main program was used.

258 Eighteen listeners participated in comparison experiments. Since we used a
 259 browser-based comparison system equipped with anonymous login function, gender
 260 and age distribution of the listeners are not available. However, we assume that the
 261 majority are in their twenties and there are more male listeners than female listeners.
 262 Each listener was given 38 or 39 sets of triplet voices, and asked to choose the most
 263 attractive one. The sets were made randomly. We used triplet comparison instead
 264 of paired comparison simply because we can obtain more comparison results with
 265 smaller number of trials, although we understand that it is controversial whether
 266 triplet comparison provides as reliable results as paired comparison. A single triplet
 267 comparison result was interpreted as two paired comparison results. If voice A was

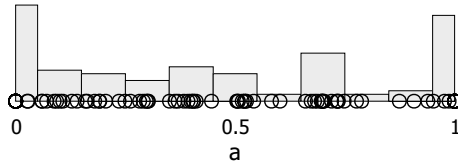


Fig. 15.4 Optimal mapping of 1-dimensional merit space. The variable a is the 1-dimensional merit value. Since there are many circles that are completely overlapped, bars were added to show the histogram. There are 26 voice mapped to $a = 0$ and 22 voices mapped to $a = 1$

268 chosen from the triplet $\{A, B, C\}$, it was interpreted that A won in the comparisons
 269 $\{A, B\}$ and $\{A, C\}$. Finally, we collected 1,388 paired comparison results (76 or
 270 78 randomly chosen comparisons for each listener) over 115 voice samples.

271 15.4.2 Mapping to Multidimensional Merit Space

272 First, we confirmed the effectiveness of mapping to 1-dimensional merit space, which
 273 corresponds to the universal attractiveness model. The likelihood function L of Eq.
 274 (15.1) was minimized in terms of 115 scaler values $\{a_1, a_2, \dots, a_{115}\}$ using Eq.
 275 (15.5).

276 Figure 15.4 shows the obtained mapping. Using the attractiveness values shown in
 277 Fig. 15.4, we can calculate which voice deserves a win for each comparison.
 278 Accordingly, the human judgments are categorized into anticipated or surprising.
 279 The mapping efficiency is defined by the ratio of anticipated judgments.¹

280 A common metric for such efficiency is called “Kendall rank correlation coefficient,”
 281 or “Kendall’s τ ” in short. To deal with the incomplete comparison with ties,
 282 we modify “Kendall’s τ_b ” as

$$283 \tau_b = \frac{N_C - N_D}{\sqrt{N_C + N_D + N_T} \sqrt{N_C + N_D}} \quad (15.20)$$

284 where N_C is the number of concordant (anticipated) pairs, N_D is the number of
 285 discordant (surprising) pairs, and N_T is the number of tied pairs in which two voices
 286 have the same attractiveness. τ_b becomes 1 if all comparisons are concordant and -1
 287 if all comparisons are discordant. In the case of Fig. 15.4, τ_b was 0.622.

288 Next, the same procedure was applied to the higher dimensional cases. The pro-
 289 cedure in 2-d was described in Fig. 15.3. In the cases with higher dimension, it was
 290 extended in a natural manner. In each case, we repeated the update 80 times with
 291 random initialization, but they converged to several mappings only.

¹It is similar to the athletes’ ranking. If the high-ranked player always wins, the ranking is efficient. If there are many upsets in which the low-ranked player wins, the ranking is not efficient.

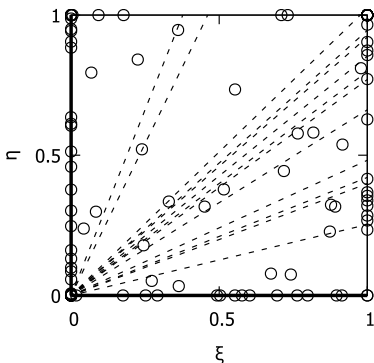


Fig. 15.5 Optimal mapping of 2-dimensional merit space. The voices are represented by circles. The listeners are represented by dashed lines. There are two listeners with $\theta_k = 0$ (x-axis) and three listeners with $\theta_k = \pi/2$ (y-axis)

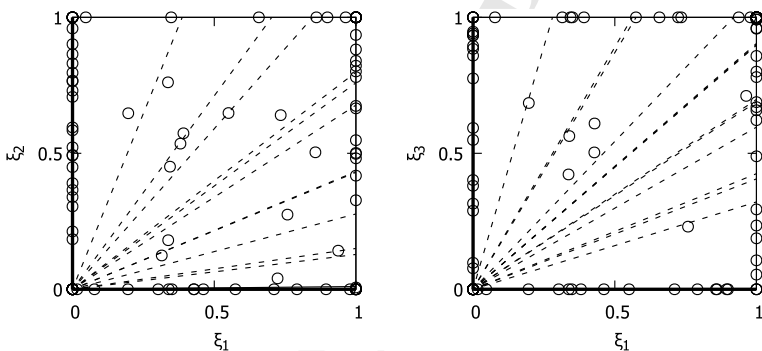


Fig. 15.6 Optimal mapping of 3-dimensional merit space. The left plot shows the first and second dimension, and the right plot shows the first and third dimension. There is one listener with $\theta_{1k} = 0$, three listeners with $\theta_{1k} = \pi/2$, four listeners with $\theta_{2k} = 0$, and two listeners with $\theta_{2k} = \pi/2$

292 Figures 15.5 and 15.6 are the optimal mapping in 2-d and 3-d cases. It can be seen
 293 that many voices have either 0 or 1 as an element of \mathbf{m} , meaning that the goodness
 294 or badness in terms of specific viewpoint is judged unanimously. The voices located
 295 on the right-top corner are perfectly attractive voices for everyone. The voices on the
 296 left-bottom are perfectly unattractive for everyone. There are 16 perfectly attractive
 297 and 15 perfectly unattractive voices in 2-d mapping, and 8 perfectly attractive and 7
 298 unattractive voices in 3-d mapping.

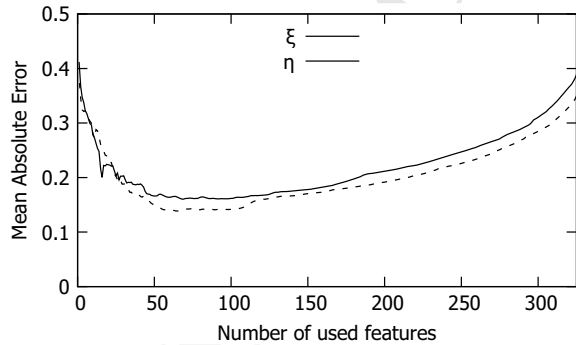
299 Table 15.2 shows τ_b values in the mapping in various dimensions up to eight.
 300 Since the larger number of free parameters have more power to solve inconsistency of
 301 comparison results, it is natural that τ_b increases as the larger dimension is introduced.
 302 However, it can be noticed that τ_b seems to saturate at around $D = 5$.

Editor Proof

Table 15.2 Values of τ_b calculated from the optimal map

Dim	1	2	3	4	5	6	7	8
N_C	1068	1143	1200	1237	1306	1262	1309	1310
N_D	232	186	159	132	79	121	71	76
N_T	88	59	29	19	3	5	8	2
τ_b	0.622	0.705	0.758	0.802	0.885	0.824	0.895	0.890

Fig. 15.7 Results of backward stepwise selection (BSS). The experiments started with 325 features (far right), and proceeded to the left



15.4.3 Merit Estimation from Acoustic Features

After confirming the effectiveness of multidimensional mapping, our concern shifted to the relationship between the merit space and the acoustic features. In particular, the most interesting question would be whether we can predict \mathbf{m}_i from the voice itself. If \mathbf{m}_i is predictable, we can predict the voice attractiveness at least for the known listeners whose preference vector \mathbf{p}_k is given.

Since the size of our database is not large enough for two-stage (optimal mapping and merit estimation) fully open condition experiments, we conducted evaluation experiments under a semi-open condition. The optimal mapping in multidimensional merit space was obtained using all data. However, after the merit values for all voices are fixed, the predictability of those merit values from the acoustic features was evaluated using WEKA version 3.6.13 under an open condition using tenfold cross validation. As mentioned before, we started the experiments using all of the 325 features. SMOreg estimator with the second-order polynomial kernel was used. BSS was carried out using the criteria of mean absolute error between real and predicted values.

Figure 15.7 shows how BSS reduced the mean square error when the number of used features changed in the 2-dimensional case. In the case of ξ , the error drops from 0.39 with all features to 0.16 with 68 carefully selected features. The error of η drops from 0.35 with all features to 0.14 with 66 features. Using the predicted values of ξ and η in cross-validation for all 115 voices, we obtained the estimation map shown in Fig. 15.8.

Fig. 15.8 Optimal mapping of 2-dimensional merit space. The voices are represented by circles. The listeners are represented by dashed lines. There are two listeners with $\theta_k = 0$ (x-axis) and three listeners with $\theta_k = \pi/2$ (y-axis)

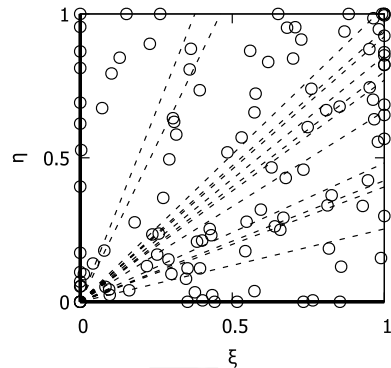


Table 15.3 Values of τ_b calculated from the estimated mapping

Dim	1	2	3	4	5	6	7	8
N_C	1057	1127	1175	1206	1221	1194	1243	1252
N_D	320	260	211	182	167	194	144	136
N_T	11	1	2	0	0	0	1	0
τ_b	0.533	0.625	0.695	0.738	0.759	0.720	0.792	0.804

325 It is straightforward to predict paired comparison results from the mapping of
 326 Fig. 15.8. The efficiency of prediction is quantified by τ_b . Among 1,388 comparisons,
 327 we obtained 1,127 concordant pairs, 260 discordant pairs, and 1 tie prediction. The
 328 value of τ_b was calculated as 0.625, which is slightly better than τ_b obtained with
 329 1-dimensional optimal mapping.

330 After all, we carried out experiments in various dimensions, and obtained τ_b values
 331 of estimated attractiveness as shown in Table 15.3. Since the mapping itself becomes
 332 more powerful as the dimension increases, the value of τ_b also increases as the higher
 333 dimension is introduced. The tendency that the efficiency saturates at around $D = 5$
 334 does not change.

335 15.5 Conclusions

336 In this chapter, a multidimensional mapping scheme of voice attractiveness was
 337 proposed. Intrinsic attractiveness of voice samples are represented as vectors in the
 338 merit space, and listener-dependent preferences are represented as directions in the
 339 same space. The attractiveness of a voice for a listener is calculated as the inner
 340 product of these two vectors. This mapping scheme provides a better-fit model for
 341 the comparison results to which the universal attractiveness model assigned small
 342 likelihood values.

343 The effectiveness of the proposed model was confirmed by the experiments
 344 using real voices and their attractiveness judgments. The multidimensional map-
 345 ping scheme achieves the higher likelihood for the Thurstone-Mosteller model, and
 346 better prediction of comparisons.

347 We also tried to predict the merit values of a new voice sample from its acoustic
 348 features. If we can do so, we can predict the comparison result at least for the known
 349 listener. A set of machine learning-based experiments confirmed the feasibility of
 350 such prediction. If we use four or higher dimension merit space, more than 1,200 of
 351 1,388 paired comparisons can be predicted correctly.

352 The proposed scheme can be applied to select attractive voice samples for com-
 353 mercial applications. Moreover, it can also be applied to the development process of
 354 TTS systems. Since it is easier to synthesize various voices than to prepare a large
 355 set of recorded voices, a TTS-based system can speak with the tailor-made voice for
 356 the customer.

357 Although the results presented in this chapter are promising, there are three impor-
 358 tant problems to solve. First, the experiments presented in this chapter are not fully
 359 open. In a sense, the optimization process of multidimensional mapping and feature
 360 selection are optimized using the evaluation data. If such optimization tends to over-
 361 fit the training data, we would have less accurate results with completely new data,
 362 especially in the higher dimension cases. To avoid that problem, we need more data
 363 and more experiments under the fully open condition.

364 The second problem is that the paired comparison data were collected with only
 365 small number of listeners. Due to the limited data size, it remains an open question
 366 whether the model trained in a certain listener group is transferable to another listener
 367 group. In addition to the data size problem, anonymousness of the listeners made it
 368 impossible to analyze the age and gender dependency of the preference.

369 The third problem is that all results in this chapter were obtained for just one
 370 phrase “irasshaimase.” Although it is a very important phrase in the commercial
 371 context, we may have something different if we use different phrases. However, the
 372 methodology to deal with the merit space and acoustic features is applicable to any
 373 phrase, and that is the most important achievement of the work described in this
 374 chapter.

375 References

- 376 Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method
 377 of paired comparisons. *Biometrika*, 39(3–4), 324–345.
- 378 Cattelan, M., Varin, C., & Firth, D. (2013). Dynamic Bradley-Terry modelling of sports tournaments.
 379 *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 62(1), 135–150.
- 380 Causeur, D., & Husson, F. (2005). A 2-dimensional extension of the Bradley-Terry model for paired
 381 comparisons. *Journal of Statistical Planning and Inference*, 135, 245–259.
- 382 Eyben, F., Weninger, F., Groß, F., & Schuller, B. (2013). Recent developments in openSMILE, the
 383 Munich open-source multimedia feature extractor. In *Proceedings of ACM Multimedia (MM)*,
 384 *Barcelona, Spain* (pp. 835–838).

- 385 Fujimoto, Y., Hino, H., & Murata, N. (2009). Item-user preference mapping with mixture models—
 386 Data visualization for item preference. In *Proceedings of International Conference on Knowledge*
 387 *Discovery and Information Retrieval* (pp. 105–111).
- 388 Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA
 389 data mining software: an update. *SIGKDD Explorations* 11(1), 10–18.
- 390 Junichi, Y., Onishi, K., Masuko, T., & Kobayashi, T. (2005). Acoustic modeling of speaking styles
 391 and emotional expressions in HMM-based speech synthesis. *IEICE Transaction on Information*
 392 *and Systems* 88(3), 502–509.
- 393 Lee, A., & Kawahara, T. (2009). Recent development of open-source speech recognition engine
 394 Julius. In *Proceedings of APSIPA Annual Summit and Conference, Sapporo, Japan* (pp. 1–7).
- 395 Mosteller, F. (1951). Remarks on the method of paired comparisons: I. The least squares solution
 396 assuming equal standard deviations and equal correlations. *Psychometrika*, 16(1), 3–9.
- 397 Ribeiro, F., Florêncio, D., Zhang, C., & Seltzer, M. (2011). CROWDMOS: An approach for crowd-
 398 sourcing mean opinion score studies. In *Proceedings of IEEE International Conference on Acous-*
 399 *tics, Speech, and Signal Processing, Prague, Czech Republic* (pp. 1–7).
- 400 Ringeval, F. et al. (2017). AVEC2017 Real-life depression, and affect recognition workshop and
 401 challenge. In *Proceedings of 7th Annual Workshop on Audio/Visual Emotion Challenge, Mountain*
 402 *View, CA, USA* (pp. 3–9).
- 403 Schuller, B. et al. (2017). The interspeech 2017 computational para linguistics challenge: addressee,
 404 cold & snoring. In *Proceedings of INTERSPEECH 2017, Stockholm, Sweden* (pp. 3442–3446).
- 405 Sato, N., & Obuchi, Y. (2007). Emotion recognition using mel-frequency cepstral coefficients.
 406 *Journal of Natural Language Processing* 14(4), 83–96.
- 407 Shah, N. B., Balakrishnan, S., Bradley, J., Parekh, A., Ramchandran, K. & Wainwright, M. (2014).
 408 When is it better to compare than to score? CoRR abs/1406.6618.
- 409 Shevade, S. K., Keerthi, S. S., Bhattacharyya, C., & Murthy, K. (2000). Improvements to the SMO
 410 algorithm for SVM regression. *IEEE Transaction on Neural Networks* 11(5), 1188–1193.
- 411 Tato, R., Santos, R., Kompe, R., & Pardo, J. (2002). Emotional space improves emotion recognition.
 412 In *Proceedings of 7th International Conference on Spoken Language Processing (ICSPL2002),*
 413 *Denver, USA* (pp. 2029–2032).
- 414 Zen, H., Tokuda, K., Masuko, T., Kobayashi, T., & Kitamura, T. (2004). Hidden semi-Markov model
 415 based speech synthesis. In *Proceedings of Interspeech 2004, Jeju Island, Korea* (pp. 1393–1396).

1

2

Part V Technological Applications

Editor Proof

UNCORRECTED PROOF

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Voice Attractiveness
Series Title	
Chapter Title	Trust in Vocal Human–Robot Interaction: Implications for Robot Voice Design
Copyright Year	2020
Copyright HolderName	Springer Nature Singapore Pte Ltd.
Corresponding Author	Family Name Torre Particle Given Name Ilaria Prefix Suffix Role Division Department of Electronic and Electrical Engineering Organization Trinity College Dublin Address Dublin, Ireland Email torrei@tcd.ie ilariat@kth.se
Author	Family Name White Particle Given Name Laurence Prefix Suffix Role Division School of Education, Communication and Language Sciences Organization Newcastle University Address Newcastle, UK Email laurence.white@ncl.ac.uk
Abstract	Trust is fundamental for successful human interactions. As robots become increasingly active in human society, it is essential to determine what characteristics of robots influence trust in human–robot interaction, in order to design robots with which people feel comfortable interacting. Many interactions are vocal by nature, yet the vocal correlates of trust behaviours have received relatively little attention to date. Here we examine the existing evidence about dimensions of vocal variation that influence trust: voice naturalness, gender, accent, prosody and interaction context. Furthermore, we argue that robot voices should be designed with specific robot appearance, function and task performance in mind, to avoid inducing unrealistic expectations of robot performance in human users.
Keywords	Speech - Robots - Voice design - Human-robot interaction - Trustworthiness - Speech prosody

Chapter 16

Trust in Vocal Human–Robot Interaction: Implications for Robot Voice Design



Ilaria Torre and Laurence White

Abstract Trust is fundamental for successful human interactions. As robots become increasingly active in human society, it is essential to determine what characteristics of robots influence trust in human–robot interaction, in order to design robots with which people feel comfortable interacting. Many interactions are vocal by nature, yet the vocal correlates of trust behaviours have received relatively little attention to date. Here we examine the existing evidence about dimensions of vocal variation that influence trust: voice naturalness, gender, accent, prosody and interaction context. Furthermore, we argue that robot voices should be designed with specific robot appearance, function and task performance in mind, to avoid inducing unrealistic expectations of robot performance in human users.

Keywords Speech · Robots · Voice design · Human-robot interaction · Trustworthiness · Speech prosody

16.1 Introduction

Trust is an essential foundation for human societies. Numerous approaches have been taken towards understanding the means by which it is negotiated. For background, the reader is referred to texts in biology (Bateson, 2000), evolutionary theory (Harcourt, 1991), sociology (Luhmann, 1979), economics (Berg, Dickhaut, & McCabe, 1995) and neuroscience (Bzdok et al., 2011). Here, it will suffice to say that trust relates both to attribution—when someone makes a decision to trust someone else—and to states and traits, when someone acts, in the short term or over the long term, in a trustworthy manner.

I. Torre (✉)

Department of Electronic and Electrical Engineering, Trinity College Dublin, Dublin, Ireland
e-mail: torrei@tcd.ie; ilariat@kth.se

L. White

School of Education, Communication and Language Sciences, Newcastle University,
Newcastle, UK
e-mail: laurence.white@ncl.ac.uk

© Springer Nature Singapore Pte Ltd. 2020

B. Weiss et al. (eds.), *Voice Attractiveness*, Prosody, Phonology and Phonetics,
https://doi.org/10.1007/978-981-15-6627-1_16

305



Human social evolution has made us very sensitive to cues that may provide information about state or trait trustworthiness in others (e.g. Jones & George, 1998), to the point that a short extract of someone’s speech (McAlear, Todorov, & Belin, 2014), or a short exposure to someone’s face (Willis, 2006) are enough to make us form a consistent impression of that person’s trustworthiness. As robots increasingly become part of our daily lives, it is important to understand what makes people trust robots and, conversely, how we can design robots to appear trustworthy, in order to facilitate human–robot interaction (HRI) and collaboration. While much effort is put into designing robots to look trustworthy and appropriate for their task (e.g. Saldien, Goris, Yilmazyildiz, Werner, & Lefebvre, 2008; DiSalvo, Gemperle, Forlizzi, & Kiesler, 2002; Lütkebohle et al., 2010; Oh et al., 2006), less consideration is given to designing voices for these robots (e.g. Sandygulova & O’Hare, 2015). As we argue in this chapter, voice is a very powerful cue used in judgements of trustworthiness, and it should be carefully considered—in conjunction with the robot’s appearance/function and with the users’ expectations—when designing a robot.

With regard to vocal attractiveness more generally, this chapter considers how characteristics of a robot’s voice that contribute to an impression of trustworthiness reinforce, and are reinforced by, features of vocal attractiveness. Dion, Berscheid, & Walster (1972) used the phrase ‘what is beautiful is good’ to refer to the fact that attractiveness is strongly perceived as related to other positive traits, including that of trustworthiness. Indeed, attractiveness and trustworthiness loaded on the same factor in McAlear et al. (2014)’s study of vocal features. Additional evidence of the close link between attractiveness and trustworthiness comes from neuroscience (Bzdok et al., 2011) and neurobiology Theodoridou, Rowe, Penton-Voak and Rogers (2009), Bzdok et al. (2011), for example, concluded that specific brain regions, such as the amygdala, might selectively reinforce sensory information with high social importance, such as information concerning potential relationships (e.g.: ‘Is this person attractive? I might date them in the future.’; ‘Is this person trustworthy? I might collaborate with them in the future’.). Here, we examine what characteristics of a robot’s voice, by analogy with human voices, contribute to an impression of trustworthiness as a socially relevant cue for human–robot interaction.

16.2 Trust in Voices

Most human communication is predicated on some degree of mutual trust between interlocutors. When we ask a stranger for directions, we trust that they will give us the correct information, to the best of their knowledge (cf, Cooperative Principle, Grice, 1975). Moreover, like the fabled ‘boy who cried wolf’, untrustworthy communicators tend to be downgraded as interlocutors once their deceitfulness has been exposed.

As the spoken channel is typically our main mode of communication, we have naturally developed vocal means to signal our trustworthiness and to detect it in others. Indeed, the natural tendency to trust speech is mediated by heuristics that give us indicators about when the speaker might not be trustworthy. Not being able

63 to determine a speaker’s background can contribute to this impression, as can vocal
64 identifiers of social affiliations disfavoured by the perceiver, or prosodic indicators
65 of aggression or dominance. Conversely, positive evidence for trustworthiness can
66 be inferred from many vocal features, such as accent (e.g. LevpsAri & Keysar,
67 2010), prosodic cues (e.g. Miller, Maruyama, Beaver, & Valone, 1976), or emotional
68 expressions (e.g. Schug et al., 2010).

69 Regarding accents, the literature suggests that foreign accents tend to be trusted
70 less than native accents (LevpsAri & Keysar, 2010), and that, within a language,
71 ‘prestigious’ and ‘standard’ accents are trusted more than regional accents (Giles,
72 1970). For example, in the context of the UK, Standard Southern British English
73 (SSBE) is generally evaluated as more trustworthy than, for example, typical London
74 or Birmingham accents (Bishop, Coupland, & Garrett, 2005). Furthermore, experi-
75 mental evidence suggests that such first impressions of trustworthiness might persist
76 over time, despite being mediated by experience of a speaker’s actual behaviour
77 (Torre, White, & Gosli, 2016).

78 Results are less conclusive, indeed sometimes contradictory, regarding the direc-
79 tion of influence of various prosodic features on trust attributions. For example,
80 OConnor and Barclay (2017) found that people had greater trust in higher pitched
81 male and female voices (based on average fundamental frequency, f_0). By contrast,
82 Villar, Arciuli, and Paterson (2013), amongst other studies, have found that partic-
83 ipants raise their vocal pitch when lying, and Apple (1979) showed that speakers
84 with a high f_0 and slow speech rate were rated as ‘less truthful’. A fast speaking
85 rate was found to be a feature of charismatic and persuasive speakers (Jiang & Pell,
86 2017; Chaiken, 1979), but has also been found to detract from charisma in speech
87 (Niebuhr, Brem, Novák-Tóth, & Voße, 2016). Finally, higher pitch and slower speech
88 rate predicted greater trusting behaviour in an economic game (Torre et al., 2016).
89 Such variable results might be due to the fact that the studies employed different
90 methods, such as questionnaires or behavioural measures, and looked at different
91 aspects of trust, such as deception, economic trust, voting behaviour, charisma, and
92 so on. They might also reflect quantitative variation in the prosodic features examined
93 in the different studies: it is unlikely that the relationships between trust attributions
94 and, for example, speech rate or pitch range are strictly linear. Additionally, rather
95 than individual vocal characteristics, it is more likely to be a combination of features
96 that determine the perceiver’s assessment of trustworthiness, along with how vocal
97 features interact with physical appearance, interaction context and the perceiver’s
98 emotional state.

99 Voice is a powerful medium through which a diversity of speaker-specific index-
100 ical information is transmitted and interpreted, and robot voices are likely to be
101 subjected to similar appraisals. Thus, the design of robot voices should be influenced
102 by the nature of the attributions appropriate to the purposes of particular human–robot
103 interactions.

16.3 Trust in Robot Voices

People tend to attribute personality traits to computers and robots as if they were human agents (Nass & Lee, 2001; Nass, Moon, Fogg, Reeves, & Dryer, 1995; Walters, Syrdal, Dautenhahn, Te Boekhorst, & Koay, 2008), and to respond to robots as if they had a personality (Lee, Peng, Jin, & Yan, 2006). Given also that people attribute traits to human speakers based on subtle speech characteristics (e.g. Torre et al., 2016; McAleer et al., 2014), there is reason to assume that voice information will be used to attribute traits—e.g. of trustworthiness—to robots as well. Thus, voice selection should be an integral part of the overall robot’s design. Issues to take into consideration are numerous and diverse, the following being just a selection. Should large robots have lower pitched voices than small robots, congruent with anthropomorphic expectations about larger larynxes? Should human-like robots have natural human voices? Should robot voices have regional accents? If so, should these be chosen to reflect the accent of the person with whom they are interacting or, for example, to reflect a stereotyped association between particular voice styles and the specific functions that the robot will perform? The latter approach risks reinforcing stereotypes, but ignoring any considerations of voice-function congruency could be problematic for the naturalness of the interaction.

It seems, however, that relatively little attention is currently paid to how the selection of robot voices in HRI research might affect our interaction with robots. For example, McGinn and Torre (2019) conducted an informal survey of researchers whose paper at the Human–Robot Interaction 2018 conference featured a speaking robot. Specifically, they asked if they chose the voice of their robot and, if so, why. Of the 18 responses received, six had used the Nao robot built-in voice, seven had used a voice generated with a Text-To-Speech system, either because it was freely available or because it was the voice that the robot came with, three pre-recorded the voice using actors, and two simply described what the voice sounded like (e.g. ‘androgynous, child-like voice’). In addition, six of these authors specified that they had adjusted the robot voice in terms of pitch or speech rate to increase intelligibility or to elicit the perception of a particular voice age. Only one author mentioned the accent that the voice had, and only one author mentioned looking for a voice that would specifically suit the task that the robot had to carry out in the experiment. About the reasons for their choice, two authors specified that ‘it was the only good one’ and ‘because it was open source’. While 11 mentioned the gender of the robot’s voice, only a minority considered other voice characteristics such as prosody or accent, or the context in which the interaction would take place. However, as we show in this chapter, all of these features influence human perception of robots, and should not be neglected.

Studies experimentally manipulating a robot voice in order to measure its effect on users’ perceptions and behaviours are relatively scarce, but here we review work in which a robot’s voice was manipulated, or where vocal characteristics were considered in the analyses. As trust is related to other positive traits—a typical ‘halo effect’ Dion et al. (1972)—and as studies examining the effect of robot voices on

147 trust are limited, we evaluate voice-based research in human–robot interaction in
148 general, considering the implications for trustworthiness in particular.

149 **16.3.1 Voice Naturalness**

150 One key aspect of voice that is often taken into account when designing robots is
151 naturalness. While current efforts in the speech technology community are dedicated
152 to creating the most natural-sounding artificial voices, it might not be the case that
153 people actually prefer interacting with a robot or other artificial agent with a perfect
154 natural-sounding voice (Hinds, 2004). For example, Sims et al. (2009) showed that
155 being able to speak with either a synthetic or a natural voice was enough for a robot
156 to be treated as a competent agent: people gave more commands to a robot that
157 had a voice, whether synthetic or natural, and fewer to a robot that communicated
158 with beeps. They hypothesised that people assumed that speechless robots would
159 not understand language, and thus they did not speak either. Within the speaking
160 robot condition, however, participants gave more commands to the synthetic-voiced
161 robot than the natural one: (Sims et al., 2009) suggested that participants might have
162 thought that a robot with a human voice was more competent and therefore needed
163 fewer commands. Taking a different perspective, Mitchell et al. (2011) argued that
164 incongruence in the human likeness of a character’s face and voice can elicit feel-
165 ings of eeriness. In contrast, Tamagawa, Watson, Kuo, MacDonald, and Broadbent
166 (2011) argued that, for the sake of clarity and familiarity, people would prefer such
167 an ‘incongruent’ robot. In Eyssel, Kuchenbrandt, Hegel, and de Ruiter (2012), par-
168 ticipants were shown a video of a Flobi robot saying: ‘it’s quarter past three’ and
169 were asked to rate the robot in terms of anthropomorphism, likeability, psycholog-
170 ical closeness and intentions. The robot had either a natural or a synthetic voice.
171 Interestingly, voice had an effect only on participants’ ratings of likeability, with
172 people rating the natural voice higher. On the other hand, in Theodoridou et al.
173 (2009), people implicitly trusted robots with synthetic voices more than those with
174 natural voices when they were behaving trustworthily, but found the opposite effect
175 when the robots were behaving untrustworthily. This also points to the importance
176 of interaction context for robot voice design (Sect. 16.3.5).

177 More generally, Hegel (2012) did not find strong evidence that the human like-
178 ness of a robot’s appearance influenced the perception of its social capabilities. If the
179 same were true for the human likeness of robot *voices*, this would argue that voice
180 naturalness might not be critical for creating feelings of trust. However, another
181 factor to take into account when considering naturalness is listening effort: thus,
182 listening to synthetic voices can increase cognitive load relative to natural voices
183 (Simantiraki, Cooke, & King, 2018; Francis & Nusbaum, 2009). In turn, high cogni-
184 tive load hinders strategic thinking and can lead to trust misplacement, for example,
185 to trusting untrustworthy individuals (Duffy & Smith, 2014; Samson & Kostyszyn,
186 2015). This suggests that—especially if the robot is meant to sustain an extended

187 vocal communication with a person—it should be given a natural—or high-quality
 188 synthetic,—voice, notwithstanding any contradictions with the robot’s mechanical
 189 looks.

190 **16.3.2 Voice Gender**

191 Talking specifically about trust, research on human–human interaction has not found
 192 consistent differences in trust judgments towards men or women (e.g. Nass & Brave,
 193 2005; Chaudhuri, 2007; Boenin & Serra, 2009; Slonim & Guillen, 2010). Given that
 194 people’s mental models of humanoid social robots are generally similar to human
 195 models (Lee et al., 2005; Kiesler & Goetz, 2002), it would be reasonable to expect a
 196 lack of overall difference, when it comes to trusting a ‘female’ or ‘male’ robot. Indeed,
 197 Crowell, Scheutz, Schermerhorn, and Villano (2009) failed to find any difference
 198 in how people reacted to a mechanical robot that had either a female or a male
 199 voice. Thus, in terms of voice design, the straightforward expectation would be that
 200 robots that are designed to look more feminine or masculine should have a voice
 201 corresponding to their apparent gender.

202 The problem of voice gender selection may be further simplified by the fact that
 203 many robots are not perceived as having a clearly defined gender. For example, in a
 204 recent study (partially described in Theodoridou et al. (2009)), we used a Nao robot
 205 with two different natural female voices, with participants interacting with both. At
 206 the end of the experiment, a random sample of the 120 participants was asked what
 207 gender they thought the robots had. Of the 66 randomly sampled participants, 23 said
 208 they thought the robot was always female, 17 always male, 20 did not associate any
 209 gender, and 6 associated a different gender to the two robots they played with. This
 210 suggests that even a natural female voice does not consistently convey information
 211 about the gender of the robot with that voice. Similarly, the majority of participants
 212 in Walters et al. (2008) who interacted with a robot that had either a pre-recorded
 213 male voice, a pre-recorded female voice, or a synthesised voice, gave either a male
 214 or a neutral name to the robot, even when the robot had a female voice.

215 Thus, it seems that the gender of a robot voice does not necessarily influence
 216 whether people will perceive the robot to have the same gender. However, describing
 217 a study involving 9–11-year-old children, Sandygulova and O’Hare (2015) suggested
 218 that children assigned a gender to a Nao robot on the basis of the voice alone. This was
 219 a synthetic male or female voice. However, participants heard all the possible voices
 220 in succession with the same robot, and so a contrast effect may have contributed to
 221 the gender attribution being based on voice in this case.

222 While there is no evidence that voice gender influences a positive human–robot
 223 interaction, it is possible that it could interact with presumed gender-specific knowl-
 224 edge (e.g. Powers et al., 2005). As discussed later (Sect. 16.3.5), the context in which
 225 the interaction takes place might be more important for trustworthy voice design than
 226 voice gender as an isolated feature.

227 **16.3.3 Voice Accent**

228 Everyone has an accent. The term ‘accent’ refers to systematic patterns of realisation
229 of the sounds of a language—phonetic and phonological—that people belonging
230 to certain geographically or socially defined groups tend to have in common (Lip-
231 pispisGreen, 1997). Accents thus provide immediate information about whether or
232 not two interlocutors belong to similar social and/or regional groups, information
233 that we tend to implicitly use in judgements of trustworthiness (e.g. Kinzler et al.,
234 2009). Specifically, in-group membership elicits favourable first impressions, includ-
235 ing with robots (Kuchenbrandt, Eyssel, Bobinger, & Neufeld, 2013). Given that every
236 speaker has an accent, and that these accents affect the way we interpret interpersonal
237 communication, should speaking machines have purpose-specific human accents?

238 There is, unsurprisingly, evidence of straightforward accent preferences in inter-
239 actions with robots. For example, children based in Ireland showed a preference
240 towards male and female UK English over US English in a Nao robot (Sandygulova
241 & O’Hare, 2015). We can also contribute some survey data regarding overall prefer-
242 ences for robot accents. All the participants of various UK-based studies run over
243 3 years were asked what accent they would like a robot to have. The question was
244 open-ended, so we re-coded the answers to fit in broad categories (e.g. ‘West Coun-
245 try’ and ‘South West’ would both be coded as ‘South West’; labels such as ‘English’,
246 ‘British’, ‘RP’ would be coded as ‘SSBE’). Figure 16.1 shows these standardised
247 answers from all 503 participants who answered this question. As the figure shows,
248 the majority of respondents answered with ‘SSBE’, followed by ‘Neutral’ accent
249 (which in the UK is also likely to mean the non-regional SSBE), followed by ‘Irish’.
250 All of the respondents were native British English speakers, with the following self-
251 reported regional identities: southwest England (58%), southeast England (22%),
252 Midlands (8%), Wales (5%), East Anglia (3%), with participants from northeast
253 England, northwest England and Scotland comprising almost all of the remaining
254 3–4%. As shown in Fig. 16.1, very few people reported a preference for a robot to
255 have a machine-like voice. There were also relatively few preferences for a regional
256 accent reflecting one’s own origins: 58% of respondents were from the southwest but
257 only about 5% of all respondents said they would like the robot to have a southw-
258 ern accent (which here we use to encompass Bristol, Cornwall, Devon, Plymouth or
259 general South-West).

260 Preferences for robot accents may well also be influenced by the nature of the inter-
261 action, however. For example, research from Andrist (2015) on the Arabic language
262 showed an interaction between accent and behaviour in human–robot interaction
263 (see Sect. 16.3.5): participants believed that robots with the same regional accent as
264 theirs were more credible—when the robots were knowledgeable—than those with
265 a standard accent, whereas robots with standard accents were perceived to be the
266 more credible when the robots had little knowledge. Similar interactions between
267 accents and behaviour are, of course, likely with other languages. For example, Tam-
268 agawa et al. (2011) ran two experiments comparing synthesised British, American,
269 and New Zealand English accents. In the first experiment, participants from New

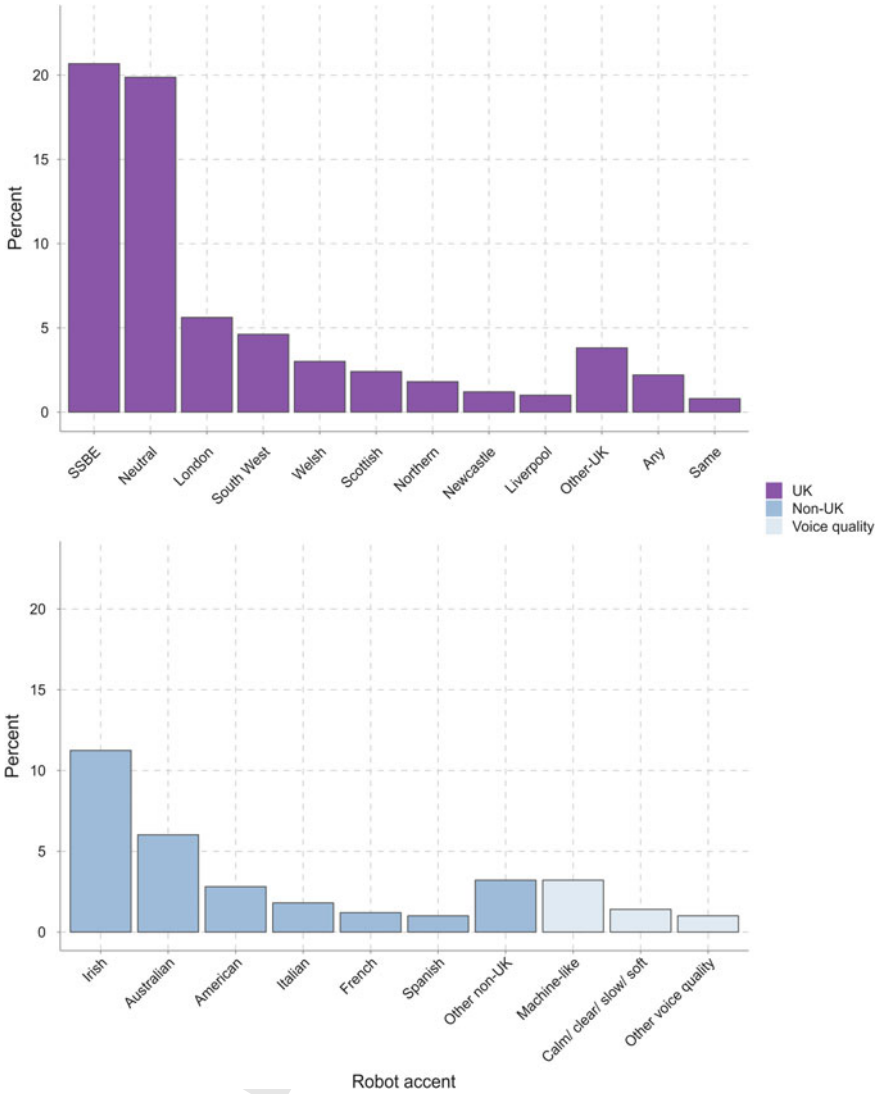


Fig. 16.1 Preference for a robot’s accent from a survey of 503 native British English speakers. The question allowed a free response and the bars indicate the overall proportion of all responses that fitted within each accent category (see text for how accent responses were categorised). ‘Any’ means ‘any accent’; ‘Same’ means ‘the same accent as me’

270 Zealand explicitly rated the disembodied UK accent more positively than the US
 271 one, while their own New Zealand accent was not rated significantly differently
 272 to either of the other accents. In the second experiment, participants were told by
 273 a healthcare robot, in one of the three accents, how to take blood pressure mea-

Editor Proof

274 surements on themselves. After the interaction, they completed a questionnaire, and
275 reported more positive emotions towards the New Zealand-accented robot than the
276 US-accented robot, and thought that the New Zealand robot performed better (no
277 other pairwise comparison was statistically significant). These results also point to
278 the differential effect that accents might have in different interaction contexts (e.g.
279 disembodied voice versus speaking robot).

280 **16.3.4 Voice Prosody**

281 Pitch gives information about a speaker’s size, being inversely proportional to the
282 body mass (Ohala, 1983). Thus, it might be straightforward to think that bigger
283 robots should have lower pitched voices than smaller robots. However, as we saw
284 in the discussion on voice gender (Sect. 16.3.2), intuitive assumptions regarding
285 appropriate voices do not necessarily apply in practice, and experimental work is
286 required. In Niculescu, van Dijk, Nijholt, and See (2011), after interacting with a robot
287 with either a high-pitched or a low-pitched female voice, participants’ questionnaire
288 responses indicated an overall preference for the higher pitched voice. In another
289 study, Yilmazyildiz et al. (2012) asked participants which of two voices, with lower
290 or higher pitch, was more suitable for a NAO—a child-like humanoid robot—or a
291 Probo—a green furry elephant-like robot: participants preferred the higher pitched
292 voice for NAO and the lower pitched voice for Probo.

293 Vocal prosody is also a feature that often manifests convergence. Linguistic
294 convergence—sometimes also called adaptation, entrainment or synchrony, although
295 we prefer the specificity of ‘convergence’—is a phenomenon by which two speak-
296 ers tend to unconsciously imitate each other’s speech characteristics as interac-
297 tions proceed (Benuš, 2014). According to Communication Accommodation Theory
298 (Giles, Coupland, & Coupland, 1991), convergence is a signal of openness and
299 positive attitude—including trust—towards the interlocutor. For example, Manson,
300 Bryant, Gervais, and Kline (2013) showed that people who converged in terms of
301 their speech rate trusted each other more in a Prisoner’s Dilemma task. Looking at
302 linguistic convergence more generally, Scissors, Gill, Geraghty, and Gergle (2009)
303 found that some types of linguistic similarity positively correlate with behavioural
304 trust in text-based interaction, while others negatively correlate with it: for exam-
305 ple, trusting individuals exhibited convergence in the use of words linked to positive
306 emotions, while deceiving individuals exhibited convergence in the use of negative
307 emotions words.

308 The well-documented occurrence of convergence phenomenon in human–human
309 interaction led researchers to examine it in human-agent interaction as well. In a
310 computer game where participants followed the advice of an owl-shaped avatar that
311 was either converging or diverging from the participants’ own prosody, Benuš et al.
312 (2018) found that female participants followed the advice of the diverging avatar
313 more often than the converging one, while there was no effect for male participants.
314 Also contrary to some expectations, Strupka, Niebuhr, and Fischer (2016) found that

315 participants tended to diverge prosodically from Keepon robots whose prosody was
 316 manipulated. On the other hand, Sadoughi et al. (2017) found that children who
 317 played a game with a converging social robot had higher levels of engagement at the
 318 end of the interactions than children who played the game with a non-converging
 319 robot. The apparent differences in convergence behaviour might be due to several
 320 factors, notably whether one is concerned with factors promoting convergence by
 321 human speakers towards the vocal features of agents or with the impact of conver-
 322 gence by agents on human behaviours and attitudes. Additionally, there may be an
 323 influence of age differences in participants, adults in Benušet al. (2018) and Strupka
 324 et al. (2016), compared with children in Sadoughi et al. (2017): potentially, for exam-
 325 ple, children may have fewer implicit socio-cognitive biases towards artificial agents.
 326 More generally, discrepancies between studies may arise because of intrinsic differ-
 327 ences in the artificial voices used. Human speech has been shown to converge with
 328 artificial voices in terms of phonetics and prosody when the artificial voice is of high
 329 quality, but less so when it is of low quality (Gessinger, Raveh, Le Maguer, Möbius,
 330 & Steiner, 2017; Gessinger et al., 2018). Differences could also be due to appearance
 331 contrasts between artificial agents: for example, in the studies reported above, there
 332 was an owl-shaped avatar in Benušet al. (2018), a small, rudimentarily humanoid
 333 robot in Strupka et al. (2016) and a life-size humanoid robot in Sadoughi et al. (2017).
 334 Interactions between robot appearance and convergence behaviours cannot be ruled
 335 out.

336 Prosody conveys important information on the emotional state of the speaker (e.g.
 337 Bänziger, & Scherer, 2005; Auberge & Cathiard, 2003). In this regard, it is known that
 338 displaying a positive emotion generally leads to attributions of other positive traits—
 339 including trustworthiness—a typical ‘halo’ effect (Lau, 1982; Penton-Voak, Pound,
 340 Little, & Perrett, 2006; Schug et al., 2010). Indeed, voice-based Embodied Conversa-
 341 tional Agents that were smiling were trusted more than those with a neutral facial
 342 expression (Elkins, 2013). Smiling in the face also led to trusting avatars and robots
 343 more (Krumhuber et al., 2007; Mathur & Reichling, 2016). Thus, a robot express-
 344 ing positive affect in its prosody could similarly increase the human user’s feeling
 345 that it can be trusted. The situation-congruent expression of affect might increase
 346 trust even when it is not displaying a positive emotion. For example, portraying
 347 stress and urgency through the voice increased performance in a joint human–robot
 348 collaborative task (Scheutz et al., 2006).

349 Apart from signalling a speaker’s mood or emotional state, prosodic cues also con-
 350 tribute to an individual’s vocal profile, that is, what makes a voice unique. Arguably,
 351 distinct-looking robots should have different-sounding voices, in order to: (a) con-
 352 tribute to the impression that they are individual agents; (b) be congruent with their
 353 physical appearance; (c) elicit personality attributions congruent with the primary
 354 functions. In a recent study (partially described in Theodoridou, Rowe, Penton-Voak,
 355 and Rogers, (2009), people played a trust game with robots having different voices.
 356 We obtained a natural recording of two female SSBE speakers, which we then resyn-
 357 thesised to sound robotic, thus generating four voices altogether: Speaker 1 natural,
 358 Speaker 1 synthetic, Speaker 2 natural, Speaker 2 synthetic. As mentioned earlier
 359 (Sect. 16.3.1), much of the variance in trust was explained by the voice naturalness

360 variable: specifically, people trusted robots with synthetic voices more than those
361 with natural voices when they were behaving trustworthily, but the opposite when
362 the robots were behaving untrustworthily (Theodoridou et al., 2009). However, peo-
363 ple also demonstrated greater implicit trust to one of the two speaker voices over the
364 other, both in natural–natural and synthetic–synthetic comparisons. This is consis-
365 tent with previous studies showing that very fine speech characteristics, which are
366 independent from higher level features such as accent, affect impression formation
367 (e.g. Gobl & Chasaide, 2003; Trouvain, 2006). It also suggests that people’s prefer-
368 ence for certain individual voices might apply when these voices are embodied in a
369 robot. Thus, idiolectal characteristics, such as those conveyed by prosody, seem to
370 contribute to trusting behaviours as well.

371 Overall, it seems simplistic to relate trustworthiness judgments purely to isolated
372 vocal features—such as gender, naturalness or pitch—and a holistic view of voice
373 might be better suited for promoting positive interactions, rather than only consider-
374 ing specific individual vocal features.

375 *16.3.5 Voice Context and Expectations*

376 As discussed earlier, some studies have shown that people perceive robots differ-
377 ently depending on the context in which the interaction takes place (Sims et al.,
378 2009; Andrist, 2015). Thus, the nature of the specific human–robot interaction may
379 affect the optimal characteristics of the robot (see also Theodoridou et al., 2009). For
380 example, Wang, Arndt, Singh, Biernat, and Liu (2013) found that, in a favourable
381 context, such as a satisfactory customer/employee call centre interaction, customers
382 with an American English background tended to suppress their negative prejudices
383 towards employees with an Indian English accent. On the other hand, when the
384 interaction was not satisfactory, customers tended not to suppress their accent preju-
385 dice (Wang et al., 2013). Similarly, Bresnahan, Ohashi, Nebashi, Liu, and Shearman
386 (2002) examined accent perception as a function of the message that the accented
387 speaker was delivering. They recorded two non-native speakers of American English,
388 one very intelligible and one not very intelligible, and one native speaker, reading
389 passages in a ‘friend’ and ‘teaching assistant’ condition. Participants were under-
390 graduate students of various ethnic origins, but mostly white Americans. They found
391 that the ‘friend’ context was judged as more attractive and dynamic than the ‘teach-
392 ing assistant’ context, in all accent conditions. Also, participants with and a strong
393 ethnic identity regarded the native accent as higher in status, dynamism and attrac-
394 tiveness, while the opposite was found for participants with a weak ethnic identity,
395 who attributed higher status and attractiveness to the not very intelligible foreign
396 accent, as compared to the native one. Thus, not only the interaction context, but also
397 the specific background context of the human interlocutor is likely to influence the
398 interaction success.

399 In HRI, Salem, Ziadee, and Sakr (2013) found that participants’ perception—in
400 terms of politeness, competency, extraversion, perceived warmth and shared reality—

401 of a receptionist robot differed according to the context of the interaction, which was
 402 either goal-oriented or open-ended. By contrast, the variation in the robot's politeness
 403 level did not influence participants' perception. Additionally, in the aforementioned
 404 study by Sims et al. (2009), participants watched videos of a robot in different scen-
 405 arios (robot damaged, robot in danger, robot requiring more information, robot has
 406 located target, robot has completed task). They found that, for example, participants
 407 gave more commands to the robot in the videos where the robot needed assistance,
 408 and concluded that a robot's voice should be chosen based on task context. In par-
 409 ticular, this would allow for the transmission of pragmatic information which may
 410 increase the operation success. For example, in a search and rescue operation, a syn-
 411 thetic voice for a robot might be the appropriate choice, because—while it conveys
 412 to the person being rescued that the robot is able to help and may be capable of under-
 413 standing human speech—the fact that the robot voice is not fully human-like could
 414 suggest to its human teammates that their input in the operation is still necessary.

415 As reviewed above, a robot's voice, along with its appearance, will have an influ-
 416 ence on the first impressions of that robot's trustworthiness. Given the role of inter-
 417 action context, however, these first impressions should be validated over long-term
 418 interactions with that robot. In fact, several experiments on trusting behaviour in
 419 human-machine interaction showed that incongruency between first impressions of
 420 trustworthiness and experience of a speaker's actual trustworthiness can drastically
 421 reduce trust (Theodoridou et al., 2009). Thus, if a robot's voice gives the impression
 422 that the robot will function well, people might have more negative reactions in the
 423 case that the robot's performance does not live up to expectations. If it is expected
 424 that a robot will operate with some degree of error, perhaps its design (appearance,
 425 voice) should reflect the fact that its performance will not always be flawless, so as
 426 not to set the users' expectations too high from the beginning (Van den Brule, Dotsch,
 427 Bijlstra, Wigboldus, & Haselager, 2014). For example, Hegel (2012) found that peo-
 428 ple attributed higher social capabilities, including honesty, to robots that looked more
 429 sophisticated. Whether robots can deliver on their promise of sophisticated perfor-
 430 mance is a different matter, however, and over-reliance on a robot according to posi-
 431 tive first impressions could have major negative consequences (Robinette, Li, Allen,
 432 Howard, & Wagner, 2016; Hancock et al., 2011; Salem, Lakatos, Amirabdollahian,
 433 & Dautenhahn, 2015).

434 Emotional expression might also elicit different trusting behaviours depending on
 435 the interaction context. Van Kleef, De Dreu, and Manstead, (2010), in the 'Emotions
 436 as Social Information' (EASI) model, suggest that emotions are used to make sense
 437 of ambiguous situations, and that their effect depends on the situation in which the
 438 interaction takes place, being specifically mediated by its cooperative or competitive
 439 nature. Thus, displaying a positive emotion, such as happiness, in a cooperative con-
 440 text will reinforce the parties' belief that everyone is gaining, and will elicit more
 441 cooperative behaviours. By contrast, displaying a negative emotion, such as anger,
 442 in a cooperative context will hinder future cooperative behaviours. Accordingly,
 443 Antos (2011) found that, in a negotiation game, participants tended to select as part-
 444 ners those computer agents which displayed emotions congruent with their actions.
 Those agents were also perceived as more trustworthy than agents whose emotional

445 expression and action strategy did not match, even though the strategy itself was
446 the same. In summary, emotional expression is helpful only if it is congruent with
447 behaviour.

448 16.4 Conclusion

449 This chapter offers an overview of some of the aspects to consider when designing a
450 trustworthy voice to be used in human–robot interaction. Given that many studies in
451 HRI employing a speaking robot have not carefully considered their robot’s voice, the
452 present work aims to be a starting point for subsequent research involving speaking
453 robots.

454 In particular, we summarised work on the effect that voice naturalness, gender,
455 accent, and prosody can have on trust attributions in human–robot interactions, along
456 with the interactions of such vocal features with the characteristics and demands of
457 the specific human–robot encounter. Naturalness, accent, and prosody seem to be
458 the features with the highest likelihood of shaping trusting behaviour, while voice
459 gender appears secondary. Moreover, carefully controlling for context might be more
460 important than, for example, manipulations of naturalness in the voice: specifically,
461 successful interactions over time may be hindered by inaccurate user expectations
462 arising from mismatches between robot’s voice features and robot’s competence and
463 performance.

464 It is possible that voice has been a secondary concern in human–robot interaction
465 research so far because vocal interactions have often been scripted, or generated by
466 an imperfect dialogue system, meaning that other aspects of the interaction, such
467 as the robot’s movements or attention, might have been prioritised. However, recent
468 advances in the field of natural language and speech processing (such as WaveNet)
469 mean that fluent autonomous human–robot conversations are getting closer to being
470 commonplace. It is time to consider more carefully what the robot’s input into these
471 conversations should actually sound like.

472 **Acknowledgments** The first author is funded by the European Union’s Horizon 2020 research
473 and innovation programme under the Marie Skłodowska–Curie grant agreement No. 713567. She
474 is also funded by the ADAPT Centre for Digital Content Technology, which is funded under the
475 SFI Research Centres Programme (Grant 13/RC/2016) and is co-funded by the European Regional
476 Development Fund. We are grateful to all the HRI conference authors who generously took time to
477 reply to the survey questions.

478 References

479 Andrist, S., Ziadee, M., Boukaram, H., Mutlu, B., & Sakr, M. (2015). Effects of culture on the cred-
480 ibility of robot speech. In *Proceedings of the Tenth Annual ACM/IEEE International Conference*
481 *on Human-Robot Interaction—HRI’15* (pp. 157–164). ACM. ACM Press.

- 482 Antos, D., De Melo, C., Gratch, J., & Grosz, B. J. (2011). The influence of emotion expression on
 483 perceptions of trustworthiness in negotiation. In *Proceedings of the 25th AAAI Conference on*
 484 *Artificial Intelligence*.
- 485 Apple, W., Streeter, L. A., & Krauss, R. M. (1979). Effects of pitch and speech rate on personal
 486 attributions. *Journal of Personality and Social Psychology*, 37(5), 715–727.
- 487 Aubergé, V., & Cathiard, M. (2003). Can we hear the prosody of smile? *Speech Communication*,
 488 40(1–2), 87–97.
- 489 Bänziger, T., & Scherer, K. R. (2005). The role of intonation in emotional expressions. *Speech*
 490 *Communication*, 46(3–4), 252–267.
- 491 Bateson, P. (2000). The biological evolution of cooperation and trust. In D. Gambetta (Ed.), *Trust:*
 492 *Making and breaking cooperative relations* (pp. 14–30). Oxford: Department of Sociology, Uni-
 493 versity of.
- 494 Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and*
 495 *Economic Behavior*, 10(1), 122–142.
- 496 Benuš, V. (2014). Social aspects of entrainment in spoken interaction. *Cognitive Computation*, 6(4),
 497 802–813.
- 498 Benuš, V., Trnka, M., Kuric, E., Marták, L., Gravano, A., Hirschberg, J., & Levitan, R. (2018).
 499 Prosodic entrainment and trust in human-computer interaction. In *Proceedings of the Ninth Inter-*
 500 *national Conference on Speech Prosody 2018*. Poznan, Poland: ISCA.
- 501 Bishop, H., Coupland, N., & Garrett, P. (2005). Conceptual accent evaluation: Thirty years of accent
 502 prejudice in the UK. *Acta Linguistica Hafniensia*, 37(1), 131–154.
- 503 Boenin, A., & Serra, D. (2009). Gender pairing bias in trustworthiness. *The Journal of Socio-*
 504 *Economics*, 38, 779–789.
- 505 Bresnahan, M. J., Ohashi, R., Nebashi, R., Liu, W. Y., & Shearman, S. M. (2002). Attitudinal and
 506 affective response toward accented English. *Language & Communication*, 22(2), 171–185.
- 507 Bzdok, D., Langner, R., Caspers, S., Kurth, F., Habel, U., Zilles, K., et al. (2011). ALE meta-
 508 analysis on facial judgments of trustworthiness and attractiveness. *Brain Structure and Function*,
 509 215(3–4), 209–223.
- 510 Chaiken, S. (1979). Communicator physical attractiveness and persuasion. *Journal of Personality*
 511 *and Social Psychology*, 37(8), 1387.
- 512 Chaudhuri, A., & Gangadharan, L. (2007). An experimental analysis of trust and trustworthiness.
 513 *Southern Economic Journal*, 959–985.
- 514 Crowell, C. R., Scheutz, M., Schermerhorn, P., & Villano, M. (2009). Gendered voice and robot
 515 entities: Perceptions and reactions of male and female subjects. In *IEEE/RSJ International Con-*
 516 *ference on Intelligent Robots and Systems, 2009. IROS 2009* (pp. 3735–3741). IEEE.
- 517 DiSalvo, C. F., Gemperle, F., Forlizzi, J., & Kiesler, S. (2002). All robots are not created equal:
 518 the design and perception of humanoid robot heads. In *Proceedings of the 4th Conference on*
 519 *Designing Interactive Systems: Processes, Practices, Methods, and Techniques* (pp. 321–326).
 520 ACM.
- 521 Dion, K., Berscheid, E., & Walster, E. (1972). What is beautiful is good. *Journal of Personality and*
 522 *Social Psychology*, 24(3), 285.
- 523 Duffy, S., & Smith, J. (2014). Cognitive load in the multi-player prisoner's dilemma game: Are
 524 there brains in games? *Journal of Behavioral and Experimental Economics*, 51, 47–56.
- 525 Elkins, A. C., & Derrick, D. C. (2013). The sound of trust: Voice as a measurement of trust during
 526 interactions with embodied conversational agents. *Group Decision and Negotiation*, 22(5), 897–
 527 913.
- 528 Eyssel, F.; Kuchenbrandt, D.; Hegel, F. & de Ruitter, L. (2012). Activating elicited agent knowledge:
 529 How robot and user features shape the perception of social robots. In *IEEE International Workshop*
 530 *on Robot and Human Interactive Communication, 2012. ROMAN 2012* (pp. 851–857). IEEE.
- 531 Francis, A. L., & Nusbaum, H. C. (2009). Effects of intelligibility on working memory demand for
 532 speech perception. *Attention, Perception, & Psychophysics*, 71(6), 1360–1374.

- 533 Gessinger, I., Raveh, E., Le Maguer, S., Möbius, B., & Steiner, I. (2017). Shadowing synthesized
534 speech—Segmental analysis of phonetic convergence. *Proceedings of Interspeech, 2017*, 3797–
535 3801.
- 536 Gessinger, I., Schweitzer, A., Andreeva, B., Raveh, E., Möbius, B., & Steiner, I. (2018). Convergence
537 of pitch accents in a shadowing task. In *Proceedings of the 9th International Conference on Speech*
538 *Prosody 2018* (pp. 225–229).
- 539 Giles, H. (1970). Evaluative reactions to accents. *Educational Review*, 22(3), 211–227.
- 540 Giles, H., Coupland, N., & Coupland, J. (1991). Accommodation theory: Communication, context,
541 and consequence. In H. Giles, N. Coupland, & J. Coupland (Eds.), *Contexts of accommodation:*
542 *Developments in applied sociolinguistics* (pp. 1–68). Press: Cambridge University.
- 543 Gobl, C., & Ní Chasaide, A. (2003). The role of voice quality in communicating emotion, mood
544 and attitude. *Speech Communication*, 40, 189–212.
- 545 Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics:*
546 *Speech acts* (vol. 3, pp. 41–58). New York: Academic Press.
- 547 Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., & Parasuraman,
548 R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors:*
549 *The Journal of the Human Factors and Ergonomics Society*, 53(5), 517–527.
- 550 Harcourt, A. H. (1991). Help, cooperation and trust in animals. In R. A. Hinde & J. Groebel (Eds.),
551 *Cooperation and Prosocial Behaviour* (pp. 15–26). Cambridge University Press.
- 552 Hegel, F. (2012). Effects of a robot’s aesthetic design on the attribution of social capabilities.
553 In *IEEE RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive*
554 *Communication* (pp. 469–475). IEEE.
- 555 Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose job is it anyway? A study of human-robot
556 interaction in a collaborative task. *Human-Computer Interaction*, 19(1), 151–181.
- 557 Jiang, X., & Pell, M. D. (2017). The sound of confidence and doubt. *Speech Communication*, 88,
558 106–126.
- 559 Jones, G. R., & George, J. M. (1998). The experience and evolution of trust: Implications for
560 cooperation and teamwork. *Academy of Management Review*, 23(3), 531–546.
- 561 Kiesler, S., & Goetz, J. (2002). Mental models of robotic assistants. In *Proceedings of the CHI*
562 *(2002) Conference on Human Factors in Computer Systems*. New York: ACM.
- 563 Kinzler, K. D., Shutts, K., DeJesus, J. M., & Spelke, E. S. (2009). Accent trumps race in guiding
564 children’s social preferences. *Social Cognition*, 27(4), 623.
- 565 Krumhuber, E., Manstead, A. S. R., Cosker, D., Marshall, D., Rosin, P. L., & Kappas, A. (2007).
566 Facial dynamics as indicators of trustworthiness and cooperative behavior. *Emotion*, 7(4), 730–
567 735.
- 568 Kuchenbrandt, D., Eyssel, F., Bobinger, S., & Neufeld, M. (2013). When a robot’s group membership
569 matters. *International Journal of Social Robotics*, 5(3), 409–417.
- 570 Lau, S. (1982). The effect of smiling on person perception. *The Journal of Social Psychology*,
571 117(1), 63–67.
- 572 Lee, K. M., Peng, W., Jin, S., & Yan, C. (2006). Can robots manifest personality? An empirical
573 test of personality recognition, social responses, and social presence in human-robot interaction.
574 *Journal of Communication*, 56(4), 754–772.
- 575 Lee, S., Lau, I. Y., Kiesler, S., & Chiu, C. (2005). Human mental models of humanoid robots.
576 In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation* (pp.
577 2767–2772). IEEE.
- 578 Lev-Ari, S., & Keysar, B. (2010). Why don’t we believe non-native speakers? The influence of
579 accent on credibility. *Journal of Experimental Social Psychology*, 46(6), 1093–1096.
- 580 Lippi-Green, R. (1997). *English with an accent: Language, ideology, and discrimination in the*
581 *United States*. Psychology Press.
- 582 Luhmann, N. (1979). *Trust and power*. Wiley.
- 583 Lütkebohle, I., Hegel, F., Schulz, S., Hackel, M., Wrede, B., Wachsmuth, S., & Sagerer, G. (2010).
584 The bielefeld anthropomorphic robot head “Flobi”. In *2010 IEEE International Conference on*
585 *Robotics and Automation (ICRA)* (pp. 3384–3391). IEEE.

- 586 Manson, J. H., Bryant, G. A., Gervais, M. M., & Kline, M. A. (2013). Convergence of speech rate
587 in conversation predicts cooperation. *Evolution and Human Behavior*, 34(6), 419–426.
- 588 Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quan-
589 titative cartography of the Uncanny Valley. *Cognition*, 146, 22–32.
- 590 McAleer, P., Todorov, A., & Belin, P. (2014). How do you say Hello? Personality impressions from
591 brief novel voices. *PLoS ONE*, 9(3), e90779.
- 592 McGinn, C., & Torre, I. (2019). Can you tell the robot by the voice? An exploratory study on the role
593 of voice in the perception of robots. In *Proceedings of the 14th Annual ACM/IEEE International*
594 *Conference on Human-Robot Interaction—HRI'19*. ACM.
- 595 Miller, N., Maruyama, G., Beaber, R. J., & Valone, K. (1976). Speed of speech and persuasion.
596 *Journal of Personality and Social Psychology*, 34(4), 615.
- 597 Mitchell, W. J., Szerszen, K. A., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & MacDorman, K.
598 F. (2011). A mismatch in the human realism of face and voice produces an uncanny valley. In
599 *i-Perception 2.1* (pp. 10–12).
- 600 Nass, C. I., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-*
601 *computer relationship*. Cambridge, MA: MIT Press.
- 602 Nass, C. I., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Exper-
603 imental tests of recognition, similarity-attraction, and consistencyattraction. *Journal of Exper-*
604 *imental Psychology: Applied*, 7(3), 171–181.
- 605 Nass, C. I., Moon, Y., Fogg, B. J., Reeves, B., & Dryer, C. (1995). Can computer personalities be
606 human personalities? *International Journal of Human-Computer Studies*, 43(2), 223–239.
- 607 Niculescu, A., van Dijk, B., Nijholt, A., & See, S. L. (2011). The influence of voice pitch on the
608 evaluation of a social robot receptionist. In *2011 International Conference on User Science and*
609 *Engineering (i-USER)* (pp. 18–23). Shah Alam, Selangor, Malaysia: IEEE.
- 610 Niebuhr, O., Brem, A., Novák-Tóth, E., & Voße, J. (2016). Charisma in business speeches—A
611 contrastive acoustic-prosodic analysis of Steve Jobs and Mark Zuckerberg. In *Proceedings of the*
612 *8th International Conference on Speech Prosody*. Boston, MA, USA.
- 613 O'Connor, J. J. M. & Barclay, P. (2017). The influence of voice pitch on perceptions of trustworthi-
614 ness across social contexts. In *Evolution and human behavior*.
- 615 Oh, J. -H., Hanson, D., Kim, W. -S., Han, Y., Kim, J. -Y., & Park, I. -W. (2006). Design of android
616 type humanoid robot Albert HUBO. In *2006 IEEE/RSJ International Conference on Intelligent*
617 *Robots and Systems* (pp. 1428–1433). IEEE.
- 618 Ohala, J. J. (1983). Cross-language use of pitch: An ethological view. *Phonetica*, 40, 1–18.
- 619 Penton-Voak, I. S., Pound, N., Little, A. C., & Perrett, D. I. (2006). Personality judgments from
620 natural and composite facial images: More evidence for a "Kernel Of Truth" in social perception.
621 *Social Cognition*, 24(5), 607–640.
- 622 Powers, A., Kramer, A. D. I., Lim, S., Kuo, J., Lee, S.-I., & Kiesler, S. (2005). Eliciting information
623 from people with a gendered humanoid robot. In *2005 IEEE International Workshop on Robot*
624 *and Human Interactive Communication, ROMAN 2005* (pp. 158–163). IEEE.
- 625 Robinette, P., Li, W., Allen, R., Howard, A. M., & Wagner, A. R. (2016). Overtrust of robots in
626 emergency evacuation scenarios. In *Proceedings of the 11th Annual ACM/IEEE International*
627 *Conference on Human-Robot Interaction—HRI '16* (pp. 101–108). IEEE Press.
- 628 Sadoughi, N., Pereira, A., Jain, R., Leite, I., & Lehman, J. F. (2017). Creating prosodic synchrony
629 for a robot co-player in a speech-controlled game for children. In *Proceedings of the 12th Annual*
630 *ACM/IEEE International Conference on Human-Robot Interaction—HRI '17* (pp. 91–99). ACM.
- 631 Saldien, J., Goris, K., Yilmazyildiz, S., Werner, V., & Lefebvre, D. (2008). On the design of the
632 huggable robot probot. *Journal of Physical Agents*, 2(2), 3–11.
- 633 Salem, M., Ziadee, M., & Sakr, M. (2013). Effects of politeness and interaction context on perception
634 and experience of HRI. In *International Conference on Social Robotics* (pp. 531–541). Springer.
- 635 Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would you trust a (faulty)
636 robot?: Effects of error, task type and personality on human robot cooperation and trust. In
637 *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interac-*
638 *tion—HRI '15* (pp. 141–148). ACM.

- 639 Samson, K., & Kostyszyn, P. (2015). Effects of cognitive load on trusting behavior—An experiment
640 using the trust game. *PLoS ONE*, *10*(5), e0127680.
- 641 Sandygulova, A., & O'Hare, G. M. P. (2015). Children's perception of synthesized voice: Robot's
642 gender, age and accent. In A. Tapus, E. André, J.-C. Martin, F. Ferland & M. Ammi (Eds.), *Social*
643 *robotics* (pp. 594–602). Springer International Publishing.
- 644 Scheutz, M., Schermerhorn, P. W., & Kramer, J. (2006). The utility of affect expression in natural
645 language interactions in joint human-robot tasks. In *Proceedings of the First Annual ACM/IEEE*
646 *International Conference on Human-Robot Interaction—HRI '06* (pp. 226–233).
- 647 Schug, J., Matsumoto, D., Horita, Y., Yamagishi, T., & Bonnet, K. (2010). Emotional expressivity
648 as a signal of cooperation. *Evolution and Human Behavior*, *31*(2), 87–94.
- 649 Scissors, L. E., Gill, A. J., Geraghty, K., & Gergle, D. (2009). In CMC we trust: The role of
650 similarity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*
651 (pp. 527–536). ACM.
- 652 Simantiraki, O., Cooke, M., & King, S. (2018). Impact of different speech types on listening effort. In
653 *Proceedings of Interspeech 2018* (pp. 2267–2271). [https://doi.org/10.21437/Interspeech.2018-](https://doi.org/10.21437/Interspeech.2018-1358)
654 [1358](https://doi.org/10.21437/Interspeech.2018-1358).
- 655 Sims, V. K., Chin, M. G., Lum, H. C., Upham-Ellis, L., Ballion, T., & Lagattuta, N. C. (2009).
656 Robots' auditory cues are subject to anthropomorphism. In *Proceedings of the Human Factors*
657 *and Ergonomics Society Annual Meeting* (Vol. 53, pp. 1418–1421). San Antonio, Texas, USA:
658 SAGE Publications.
- 659 Slonim, R., & Guillen, P. (2010). Gender selection discrimination: Evidence from a trust game.
660 *Journal of Economic Behavior & Organization*, *76*(2), 385–405.
- 661 Strupka, E., Niebuhr, O., & Fischer, K. (2016). Influence of robot gender and speaker gender
662 on prosodic entrainment in HRI. In *2016 IEEE International Workshop on Robot and Human*
663 *Interactive Communication, ROMAN 2016*. IEEE.
- 664 Tamagawa, R., Watson, C. I., Kuo, I. H., MacDonald, B. A., & Broadbent, E. (2011). The effects of
665 synthesized voice accents on user perceptions of robots. *International Journal of Social Robotics*,
666 *3*(3), 253–262.
- 667 Theodoridou, A., Rowe, A. C., Penton-Voak, I. S., & Rogers, P. J. (2009). Oxytocin and social
668 perception: Oxytocin increases perceived facial trustworthiness and attractiveness. *Hormones*
669 *and Behavior*, *56*(1), 128–132.
- 670 Torre, I., White, L., & Goslin, J. (2016). Behavioural mediation of prosodic cues to implicit judge-
671 ments of trustworthiness. In *Proceedings of the eighth International Conference on Speech*
672 *Prosody 2016*. Boston, MA, USA: ISCA.
- 673 Torre, I., Goslin, J., White, L., & Zanatto, D. (2018). Trust in artificial voices: A "congruency effect"
674 of first impressions and behavioural experience." In *Proceedings of APAScience '18: Technology,*
675 *Mind, and Society (TechMindSociety'18)*. Washington, DC, USA.
- 676 Trouvain, J., Schmidt, S., Schröder, M., Schmitz, M., & Barry, W. J. (2006). Modelling personal-
677 ity features by changing prosody in synthetic speech. In *Proceedings of the 3rd International*
678 *Conference on Speech Prosody, Dresden, Germany*.
- 679 Van Kleef, G. A., De Dreu, C. K. W., & Manstead, A. S. R. (2010). An interpersonal approach
680 to emotion in social decision making: The emotions as social information model. *Advances in*
681 *Experimental Social Psychology*, *42*, 45–96.
- 682 Van den Brule, R., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., & Haselager, P. (2014). Do robot
683 performance and behavioral style affect human trust? *International Journal of Social Robotics*,
684 *6*(4), 519–531.
- 685 Villar, G., Arciuli, J., & Paterson, H. (2013). Vocal pitch production during lying: Beliefs about
686 deception matter. *Psychiatry, Psychology and Law*, *20*(1), 123–132.
- 687 Walters, M. L., Syrdal, D. S., Dautenhahn, K., Te Boekhorst, R., & Koay, K. L. (2008). Avoiding
688 the uncanny valley: Robot appearance, personality and consistency of behavior in an attention-
689 seeking home scenario for a robot companion. *Autonomous Robots*, *24*(2), 159–178.

- 690 Wang, Z., Arndt, A. D., Singh, S. N., Biernat, M., & Liu, F. (2013). "You Lost Me at Hello": How
691 and when accent-based biases are expressed and suppressed. *International Journal of Research*
692 *in Marketing*, 30, 185–196.
- 693 Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure
694 to a face. *Psychological Science*, 17(7), 592–598.
- 695 Yilmazyildiz, S., Patsis, G., Verhelst, W., Henderickx, D., Soetens, E., Athanasopoulos, G., Sahli, H.,
696 Vanderborght, B., & Lefebvre, D. (2012). Voice style study for human-friendly robots: Influence of
697 the physical appearance. In: *Proceedings of the 5th International Workshop on Human-Friendly*
698 *Robotics*.

Metadata of the chapter that will be visualized in SpringerLink

Book Title	Voice Attractiveness	
Series Title		
Chapter Title	Exploring Verbal Uncanny Valley Effects with Vague Language in Computer Speech	
Copyright Year	2020	
Copyright HolderName	Springer Nature Singapore Pte Ltd.	
Corresponding Author	Family Name	Clark
	Particle	
	Given Name	L.
	Prefix	
	Suffix	
	Role	
	Division	School of Information, & Communication Studies
	Organization	University College Dublin
	Address	Dublin, Ireland
	Division	Computational Foundry
	Organization	Swansea University
	Address	Swansea, UK
	Email	leigh.clark@ucd.ie l.m.h.clark@swansea.ac.uk
Author	Family Name	Ofemile
	Particle	
	Given Name	A.
	Prefix	
	Suffix	
	Role	
	Division	English Department
	Organization	FCT College of Education, Zuba
	Address	Abuja, Nigeria
	Email	abdulmalikkuka@gmail.com
Author	Family Name	Cowan
	Particle	
	Given Name	B. R.
	Prefix	
	Suffix	
	Role	
	Division	School of Information, & Communication Studies
	Organization	University College Dublin
	Address	Dublin, Ireland
	Email	benjamin.cowan@ucd.ie

Abstract

Interactions with speech interfaces are growing, helped by the advent of intelligent personal assistants like Amazon Alexa and Google Assistant. This software is utilised in hardware such as smart home devices (e.g. Amazon Echo and Google Home), smartphones and vehicles. Given the unprecedented level of spoken interactions with machines, it is important we understand what is considered appropriate, desirable and attractive computer speech. Previous research has suggested that the overuse of humanlike voices in limited-communication devices can induce uncanny valley effects—a perceptual tension arising from mismatched stimuli causing incongruence between users' expectations of a system and its actual capabilities. This chapter explores the possibility of verbal uncanny valley effects in computer speech by utilising the interpersonal linguistic strategies of politeness, relational work and vague language. This work highlights that using these strategies can create perceptual tension and negative experiences due to the conflicting stimuli of computer speech and 'humanlike' language. This tension can be somewhat moderated with more humanlike than robotic voices, though not alleviated completely. Considerations for the design of computer speech and subsequent future research directions are discussed.

Keywords

Speech interface - Voice interface - Intelligent personal assistant - Uncanny valley - Humanlike - Politeness - Vague language

Chapter 17

Exploring Verbal Uncanny Valley Effects with Vague Language in Computer Speech



L. Clark, A. Ofemile, and B. R. Cowan

Abstract Interactions with speech interfaces are growing, helped by the advent of intelligent personal assistants like Amazon Alexa and Google Assistant. This software is utilised in hardware such as smart home devices (e.g. Amazon Echo and Google Home), smartphones and vehicles. Given the unprecedented level of spoken interactions with machines, it is important we understand what is considered appropriate, desirable and attractive computer speech. Previous research has suggested that the overuse of humanlike voices in limited-communication devices can induce uncanny valley effects—a perceptual tension arising from mismatched stimuli causing incongruence between users’ expectations of a system and its actual capabilities. This chapter explores the possibility of verbal uncanny valley effects in computer speech by utilising the interpersonal linguistic strategies of politeness, relational work and vague language. This work highlights that using these strategies can create perceptual tension and negative experiences due to the conflicting stimuli of computer speech and ‘humanlike’ language. This tension can be somewhat moderated with more humanlike than robotic voices, though not alleviated completely. Considerations for the design of computer speech and subsequent future research directions are discussed.

Keywords Speech interface · Voice interface · Intelligent personal assistant · Uncanny valley · Humanlike · Politeness · Vague language

L. Clark (✉) · B. R. Cowan
School of Information, & Communication Studies, University College Dublin, Dublin, Ireland
e-mail: leigh.clark@ucd.ie; l.m.h.clark@swansea.ac.uk

B. R. Cowan
e-mail: benjamin.cowan@ucd.ie

L. Clark
Computational Foundry, Swansea University, Swansea, UK

A. Ofemile
English Department, FCT College of Education, Zuba, Abuja, Nigeria
e-mail: abdulmalikkuka@gmail.com

© Springer Nature Singapore Pte Ltd. 2020

B. Weiss et al. (eds.), *Voice Attractiveness, Prosody, Phonology and Phonetics*,
https://doi.org/10.1007/978-981-15-6627-1_17

323

20 17.1 Introduction

21 As a mode of interaction, speech can affect peoples' perceptions of others in terms
 22 of identity, personality, power and attractiveness (Cameron, 2001; Coulthard, 2013;
 23 Goffman, 2005; Zuckerman & Driver, 1988). Speech can impact these perceptions
 24 in both the language used and the voice quality used to produce it; the latter defined
 25 here as 'those characteristics which are present more or less all the time that a person
 26 is talking' (Abercrombie, 1967, p. 91 in Laver, 1980, p. 1). As with human–human
 27 interaction (HHI), this impact on perceptions can be seen in human–computer inter-
 28 action (HCI), where speech has become a more prominent mode of interaction. This
 29 prominence has been accelerated with the advent of intelligent personal assistants
 30 (IPAs) such as Amazon Alexa and Google Assistant featuring in home-based smart
 31 speakers like Amazon Echo and Google Home, as well as in mobile devices and
 32 vehicles. These are in addition to longer standing speech-based technologies like
 33 interactive voice response (IVR) and navigation systems. Although we are begin-
 34 ning to understand more about how people use and communicate with these types
 35 of devices (Cowan et al., 2017; Luger & Sellen, 2016; Porcheron, Fischer, Reeves,
 36 & Sharples, 2018; Porcheron, Fischer, & Sharples, 2017), less is known about the
 37 psychological and behavioural effects of speech interface design choices on users
 38 (Clark, Cabral, & Cowan, 2018).

39 While we are aware that design choices in speech-based HCI can affect user expe-
 40 rience (UX) and interaction behaviour, we are still lacking theoretical understandings
 41 and subsequent design considerations supporting them (Clark et al., 2019b). Conse-
 42 quently, it is not always clear what linguistic or voice styles may be appropriate,
 43 desirable or even attractive to users in HCI. Mimicking aspects of humanness in
 44 speech interfaces, for example, may not always be an appropriate design objective
 45 and can result in systems being perceived as creepy or even deceitful (Aylett, Cowan,
 46 & Clark, 2019). Recent research (Moore, 2017a) has argued that humanlike voices
 47 are not always appropriate for non-human artefacts, as they may heighten peoples'
 48 expectations of what artefacts are capable of, in contrast to more robotic voices. This
 49 heightened perception of humanness can result in a gap between users' perceptions
 50 of a system's abilities or *partner models* and the reality of its limitations observed
 51 through interaction (Cowan et al., 2017). As well as the quality of a system's voice,
 52 there are also less explored questions as to what are considered appropriate styles
 53 of language for computer speech, and how humanlike or 'machinelike' they are
 54 expected to be Clark (2018), Clark et al., (2019a).

55 This chapter explores the concepts of three interpersonal linguistic strategies—
 56 politeness, relational work and vague language (VL)—as a lens to examine the
 57 possibility of *verbal uncanny valley effects* that exist in users' perceptions towards
 58 both voice and language in computer speech. This may underpin some of the user
 59 behaviour and perceptions of appropriateness, desirability and attractiveness directed
 60 towards speech interfaces in previous research, as well as peoples' expectations and
 61 partner models of their computer interlocutors. It is hoped that these discussions may

62 drive theoretical understandings of our interactions with speech interfaces, which
63 may in turn encourage design considerations in the field.

64 17.2 Uncanny Valley

65 The *uncanny valley* hypothesis suggests that non-human artefacts approaching close
66 to human likeness, but retaining obvious differences from human norms, can induce
67 negative responses from people due to one or more obvious differences from expected
68 human appearance or behaviour (Mori, 1970; Mori, MacDorman, & Kageki, 2012).
69 These responses may be referred to as concepts like eeriness, revulsion, or a sense of
70 unease, signifying perceptions of undesirable or unattractive characteristics. Disflu-
71 encies between appearance and motion, for instance, may be more disliked than
72 entities displaying more congruent features—contrasting an android that is human-
73 like in appearance yet displaying robotic movements with an all human and all robot
74 alternative (Carr, Hofree, Sheldon, Saygin, & Winkielman, 2017).

75 While empirical evidence for the uncanny valley is somewhat scarce, a review of
76 uncanny valley research papers highlighted support for two perceptual mismatch
77 hypotheses (Kätsyri, Förger, Mäkäraäinen, & Takala, 2015). The first of these
78 hypotheses suggests that uncanny valley effects arise due to mismatches between
79 the human likeness of different sensory cues (e.g. obviously non-human eyes on a
80 fully humanlike face). The second hypothesis posits that the effects occur because of
81 a higher sensitivity towards exaggerated features on more humanlike characters that
82 differ from expected humanlike norms (e.g. ‘grossly enlarged eyes, Kätsyri et al.,
83 2015, p. 7). Similar explanations for uncanny valley effects are discussed by Moore
84 (2012). In developing a Bayesian explanation for the uncanny valley effect, Moore
85 points to conflicting cues creating a perceptual distortion and subsequent perceptual
86 tension at category boundaries. These categories refer to stimuli that are discrim-
87 inately perceived as being different from one another. Stimuli perceived to be at
88 the boundaries of these categories may incur more perceptual distortion than those
89 stimuli perceived to be prototypical examples of those categories.

90 Whereas most uncanny valley research has focused on the visual, there are an
91 increasing number of works that include audio as an additional modality of interest
92 in exploring perceptual mismatches. Grimshaw (2009) discusses the concept of an
93 audio uncanny valley, with the view that further theoretical understandings may be
94 useful for sound design in horror-based computer games in creating perceptions of
95 fear and apprehension. The author provides examples of features that may induce
96 uncanny valley effects, including uncertainty about the location of sound sources
97 and exaggerated articulation of the mouth whilst speaking. Mitchell et al. (2011)
98 and Meah and Moore (2014) explored the concepts of misaligned voice and face
99 cues (or mismatched stimuli) in robots and humans. Both experiments showed that
100 mismatches in voice and face (e.g. robotic voice and human face or human voice and
101 robotic face) result in higher ratings of perceived eeriness than matched stimuli.

102 These experiments give credence to the uncanny valley existing in audio as well
 103 as visual stimuli, although the focus in the above work is on multimodal cues and
 104 the audio is primarily centred on the voice quality. With the increasing number of
 105 speech interfaces, users are exposed to unprecedented levels of primarily speech-
 106 based interactions with machines. However, there remain important design consider-
 107 ations on what is considered appropriate speech output by speech interfaces. Moore
 108 (2017a), for example, highlights the proliferation of humanlike rather than more
 109 robotic sounding voices in computer speech is not always an appropriate design
 110 choice. Using humanlike voices can create mismatches between users' expectations
 111 of a machine's capabilities and the reality of what it can achieve through speech. This
 112 may result in unsuccessful engagement with speech-based, non-human artefacts.
 113 Less is understood as to what may be considered appropriate language in spoken
 114 interactions with machines—perceptual mismatches may also occur on a linguistic
 115 as well as a voice level, potentially resulting in unwanted negative effects to UX
 116 (Clark, 2018). The subsequent sections of this chapter reflect on recent research into
 117 the use of interpersonal linguistic strategies in spoken computer instructions and
 118 discuss the possible boundaries of appropriate language use (as opposed to solely
 119 the appropriate humanlike synthesis choices) in light of uncanny valley theories and
 120 mismatched stimuli (Clark, Bachour, Ofemile, Adolphs, & Rodden, 2014; Clark,
 121 Ofemile, Adolphs, & Rodden, 2016).

122 17.3 Politeness and Relational Work

123 The concept of politeness is often discussed in terms of Brown and Levinson's (1987)
 124 work that associates politeness with the concept of *face*—the social self-image that we
 125 present to others during interaction (Goffman, 1955). This self-image is dependent
 126 on sociocultural and contextual factors and dynamically progresses between and
 127 within interactions. Face theory discusses it being in speakers' own interests to avoid
 128 damaging the face of oneself or the face of others during interaction. Conducting
 129 this is known as *facework*.

130 In Brown and Levinson's (1987) research, facework can be accomplished using
 131 politeness strategies. *Positive face* refers to desires of being liked and approved.
 132 Positive politeness strategies include showing group membership between partners,
 133 paying attention to the wants and desires of others, and presenting approval. *Negative*
 134 *face* refers mainly to the desire not to be imposed upon by others. Negative politeness
 135 strategies often focus on minimising this potential imposition. This can be accom-
 136 plished by being indirect rather than direct, for example, when issuing instructions
 137 or making requests that may create an imbalance of power.

138 *Relational work* seeks to expand Brown and Levinson's (1987) politeness theory
 139 to include the whole polite–impolite spectrum (Locher, 2004, 2006; Locher & Watts,
 140 2005, 2008). This includes all work by individuals for the 'construction, maintenance,
 141 reproduction and transformation of interpersonal relationships among those engaged
 142 in social practice' (Locher & Watts, 2008, p. 96). As with facework and politeness

143 described above, relational work is similarly discursive and on-going (Locher &
144 Watts, 2005, 2008; Watts, 2003).

145 17.3.1 *Politeness in Machines*

146 While there are disagreements in politeness and relational work, the politeness strate-
147 gies discussed in this chapter focus on the polite end of the relational work spectrum
148 and discuss a combination of positive and negative politeness strategies discussed in
149 Brown and Levinson's (1987) theory. In some previous research, politeness strategies
150 have been explored in both the HCI and human–robot interaction (HRI) communi-
151 ties, although the visual modality and/or the use of embodiment was as prominent
152 as speech. For example, Wang et al. (2008) employed politeness strategies in a
153 Wizard-of-Oz experiment providing tutorial feedback to students. The tutorial inter-
154 face contained visual features—in the form of text and an animated robotic char-
155 acter that produces gestures—and text-to-speech (TTS) synthesis that would appear
156 to come from the robotic character. In comparing polite and direct feedback, the
157 authors note that students receiving the polite tutorial feedback learned better than
158 those receiving the direct feedback. Furthermore, politeness appeared to be espe-
159 cially effective for students who displayed a preference for indirect help or were
160 judged to have less ability to complete the task.

161 In an HRI-based experiment, positive attitudinal results were observed. Torrey,
162 Fussell and Kiesler (2013) conducted a study in which participants observed videos
163 of human and robot helpers giving advice to a person learning to make cupcakes.
164 In creating the communication conditions, the authors used combinations of hedges
165 and discourse markers. Hedges (e.g. *sort of, I guess*) are described by the authors as a
166 negative politeness strategy mitigating the force of messages and reducing threats to a
167 listener's autonomy. The authors acknowledge that descriptions of discourse markers
168 (e.g. *like, you know*) have no standard definition,¹ though for the purposes of their
169 study they are described in similar terms hedges in being used to 'soften commands'
170 (Torrey et al., 2013, p. 277). Four communication conditions were created: direct (no
171 hedges/discourse markers), hedges with discourse markers, hedges without discourse
172 markers and discourse markers without hedges. Results of the experiment showed
173 that hedges and discourse markers as individual strategies improved perceptions
174 towards helpers in terms of considerateness, likeability and the helper being control-
175 ling compared to the direct condition. However, the combination of the two strategies
176 did not show significant differences compared to the individual strategies. While
177 positive improvements in perceptions towards both human and robot helpers were

¹Discourse markers may also be referred to, amongst other terms, as *discourse particles*, *pragmatic particles* and *pragmatic expressions*. Their purposes can include switching topics, marking boundaries between segments of talk, helping to conduct linguistic repair and being used as hedging devices (Jucker & Ziv, 1998).

178 observed, participants only observed videos of interactions with helpers, rather than
 179 interact with any themselves.

180 In a similar study, Strait, Canning and Scheutz (2014) analysed both observations
 181 and actual interactions with robots providing advice in a drawing task. The authors
 182 created an experiment comparing three different interaction modalities: remote third-
 183 person (observations of interactions), remote first person (one-to-one with a robot
 184 via a laptop) and co-located first person (one-to-one with robot in the same room).
 185 As with the experiment by Torrey, Fussell, and Kiesler (2013), two communication
 186 conditions were presented. The indirect condition used a combination of positive
 187 politeness strategies (e.g. giving praise, being inclusive) and negative politeness
 188 strategies (e.g. being indirect, using discourse markers), whereas the direct condition
 189 referred to the absence of these strategies in the robot helper's speech. A further
 190 condition was included in the robot's appearance, which compared one robot with
 191 a more humanlike appearance and another with a more typical robotic appearance.
 192 The results of the experiment showed politeness strategies in the indirect speech
 193 condition improve ratings of likeability and reduced ratings of perceived aggression
 194 when compared to the direct speech condition. Improved ratings for considerateness
 195 were also observed in indirect speech, but only in the remote third-person interaction
 196 modality. The findings showed that previous results from observations of interaction
 197 of robots do not necessarily transfer to actual interaction.

198 ***17.3.2 Politeness in Non-embodied Computer Speech***

199 The above studies highlight the mixed user responses towards different types of
 200 machines and interaction modalities using politeness strategies, focusing in particular
 201 on interactions with partners who are embodied or are represented visually. Many
 202 modern speech interface technologies like Google Assistant can include a minimal
 203 amount of visual output, depending on the device being used but do not necessarily
 204 include embodied features.

205 With this in mind, two further studies explored the use of politeness strategies in
 206 HCI, in which participants were tasked with constructing models under the instruction
 207 of a speech interface (Clark et al., 2014, 2016). In both studies, VL was used to create
 208 indirectness as a form of overall negative politeness strategy.² VL refers to language
 209 that is deliberately imprecise and can achieve a wide range of functional and inter-
 210 personal goals (Channell, 1994). For example, lexical hedges like *just* and *partly* can
 211 be used as a tension-management device to play down the perceived significance of
 212 research during academic conferences (Trappes-Lomax, 2007). Furthermore, vague
 213 nouns such as *thing* and *whatsit* can be used to replace a typical noun if speakers

²These were adaptors, e.g. *more or less*, *somewhat* (reduce assertiveness, minimise imposition); discourse markers, e.g. *so*, *now* (structure talk, mitigate assertive impact of utterance); minimisers, e.g. *just*, *basically* (structure talk, reduce perceived difficulty, mitigate utterance impact) and vague nouns, e.g. *thing*, *bit* (improve language efficiency) (Clark et al., 2016).

214 and listeners have both established what the vague nouns are referring to (Channel,
215 1994). While not all VL has functions in being polite, this is the primary purpose
216 of which it used in the speech interface studies—the indirectness and imprecision
217 of VL can contribute to lessening the perception of speakers being too authoritative
218 (McCarthy & Carter, 2006) and help create an informal and less direct atmosphere
219 during interaction.

220 In the first speech interface study using VL, two communication conditions were
221 developed—a vague condition containing politeness strategies and a non-vague
222 condition excluding these politeness strategies (Clark et al., 2014). Participants were
223 tasked with building Lego models under the verbal instructions of a computer inter-
224 face, the speech of which was produced by the TTS voice Cepstral Lawrence.³ During
225 this study, participants interacted with an interface on a MacBook Pro 10.2. This was
226 a minimalistic interface using an HTML file linked to a library of pre-recorded speech
227 files. The interface allowed participants to proceed to the next instruction or repeat
228 a current instruction, with the pace being dictated by the participants. Results of
229 this study showed that the non-vague interface was rated as significantly more direct
230 and authoritative than the vague interface. However, post-task interviews revealed
231 participants perceived the vague interface to be inappropriate in terms of its language
232 choice. This was partly a result of the quality of the voice. People’s expectations of
233 a relatively robotic voice were matched more with the non-vague interface than the
234 vague interface, with the latter discussed as being insincere and its language more
235 appropriately suited to a more natural (i.e. humanlike) sounding voice.

236 A follow-up experiment explored vague communication conditions across three
237 different voices (Clark et al., 2016). Two of these were TTS-synthesised voices—
238 Cepstral Lawrence as per the previous experiment—and CereProc Giles.⁴ The third
239 voice was provided by a professional voice actor who was deemed to sound similar in
240 age and accent to the two synthesised voices. Participants followed verbal instructions
241 to build models using two of the three voices in two separate tasks. These tasks
242 used the same style of interface as the first experiment. Results showed that the
243 voice actor was perceived as significantly more likeable, more humanlike and less
244 annoying than the two synthesised voices. Furthermore, it was perceived as more
245 coherent than Giles, and both the voice actor and Lawrence were rated as allowing
246 more task completion than Giles. Analysis of post-task interview data also revealed
247 that VL in both synthesised voices was perceived negatively. Participants cited it as
248 inappropriate and often commented on the jarring nature between the quality of the
249 voice and the language being used. However, while the voice actor was seen as a
250 more appropriate fit for VL, results were not wholly convincing. Despite the increased
251 naturalness and humanlikeness, participants still highlighted the disparity between
252 the more machinelike nature of the voice and the humanlike nature of the language.
253 Even with a human voice, there were comments discussing it as ‘just a machine’
254 that is not capable of executing VL or politeness strategies, unlike other people, due
255 to their inherent interpersonal and social linguistics purposes. This suggests that the

³<https://www.cepstral.com>.

⁴<https://www.cereproc.com>.

256 medium of speech delivery, in this case a machine, can also impact on perceptions
 257 of appropriateness and attractiveness.

258 17.4 Implications for Verbal Uncanny Valley Effects

259 In terms of what may be considered appropriate computer and human speech, the
 260 experiments discussed above raise the possibility of category boundaries existing on
 261 a linguistic level—verbal uncanny valley effects. While participants could not always
 262 explicitly identify individual lexical items that caused negative reactions towards the
 263 interfaces, they were able to identify a general disparity between the language being
 264 used and the interface that provided the language. Although this was not the case for
 265 all participants, there was a general trend towards describing the vague conditions
 266 in both experiments as humanlike language, whereas in Clark et al. (2014), the
 267 non-vague condition was cited as being appropriately machinelike.

268 In the sense of the latter, the use of direct and non-vague language was seen to
 269 match people’s expectations of appropriate language use with a robotic synthesised
 270 voice. This is an example of matched speech-based stimuli, whereby categories of
 271 preconceived ‘machine likeness’ are aligned. Subsequently, there is little discussion
 272 about feelings of the uncanny arising, which are focused more on misaligned stimuli
 273 (Mitchell et al., 2011; Moore, 2012a). This also draws similarities with Moore’s
 274 (2012a) discussion of appropriate voices in non-human artefacts. With non-vague and
 275 direct instructions provided by a robotic voice, appropriateness is seemingly deter-
 276 mined as it matches people’s expectations of what their interaction partner is capable
 277 of. These expectations and beliefs of what a communicative partner can produce may
 278 be referred to as peoples’ partner models (e.g. Cowan, Branigan, Obregón, Bugis,
 279 & Beale, 2015). Previous research with infrequent users of IPAs has suggested that
 280 speech qualities such as regional accents can signal the communicative attributions
 281 people make towards artificial assistants (Cowan et al., 2017). Similarly, this may
 282 operate with the quality of a system’s voice, the language it uses, and how these
 283 two relate to one another. A robotic voice may relate more to signals of using direct
 284 than indirect language that is absent in relational work, vague language or politeness
 285 strategies. In terms of users’ expectations, these linguistic concepts may not be seen
 286 as residing in the category of appropriate computer speech.

287 This can be observed in the vague conditions of the two experiments (Clark
 288 et al., 2014, 2016). In the synthesised voices, in particular, the combination of a
 289 robotic sounding voice with language that is used to undertake social goals creates a
 290 mismatch in stimuli. Subsequently, uncanny valley effects can be observed, especially
 291 in participants’ descriptions of their interactions with the interfaces. In the second
 292 experiment (Clark et al., 2016), however, using a pre-recorded human voice appeared
 293 to cause less perceived stimuli mismatch in the vague conditions than the synthesised
 294 voices. This may indicate that perceived categories of appropriate computer and
 295 human speech can be blurred somewhat with the introduction of more humanlike
 296 voices—a human voice can signal a perceptual cue of being capable of producing

297 more humanlike language, even in a computer interface. However, the mismatch is not
298 alleviated completely. Other cues, such as the medium and/or context of interaction
299 (laptop interface providing task-based instructions), may alter what is perceived as
300 appropriate speech even with a human voice.

301 **17.4.1 Identifying Appropriateness in Computer Speech**

302 Indeed, the combination of socially driven linguistic cues and computer speech output
303 may create a *habitability gap* (Moore, 2017b), whereby there is a gap between a
304 users' model of a system and the reality of the actual system (Hone & Graham,
305 2000). Users' models of computer speech may not include the use of interpersonal
306 linguistic strategies and subsequently the presentation of actual computer speech that
307 includes these creates feelings of unease or *perceptual tension* (Moore, 2012).

308 The mismatching of cues and accompanying perceptual tension in spoken inter-
309 actions with computers and other machines appears strongly linked to perceptions of
310 what is considered appropriate communication. In addition to a possible habitability
311 gap, it may also be the case that perceived inappropriateness of politeness, rela-
312 tional work or vague language in computer speech is aligned with the socially driven
313 nature of these concepts. Relational work and politeness strategies, for example, are
314 primarily focused on establishing and maintaining interpersonal relationships with
315 other people (Locher & Watts, 2008; Brown & Levinson, 1987). It is debatable as
316 to what extent this can be accomplished in HCI, how achievable this is as a design
317 goal, and how much users would desire this feature in a speech-based device. The
318 social rules that underpin much HHI do not automatically transfer to HCI and the
319 latter may be markedly diminished in comparison. Moore (2017b, p. 8) highlights
320 a similar possible phenomenon—that there may be a 'fundamental limit' to the
321 linguistic interactions between humans and machines due to them being '*unequal*
322 partners'. The very nature of humans and machines means there are inherent differ-
323 ences in capabilities, and this is likely present in the partner models users create
324 in speech-based HCI. When these partner models clash with experiences, this may
325 lead to negative user experiences and perceptions of inappropriate, undesirable or
326 unattractive speech interface partners.

327 The social rules underpinning HCI and HHI also do not automatically align.
328 Relational work and politeness strategies are primarily focused on interpersonal
329 relationships. Brown and Levinson's (1987) theory on politeness in particular is
330 strongly associated with the process of facework during interaction. However, the
331 maintenance of face during interaction with machines is different than with other
332 people—machines do not have a face as such to protect and, in turn, users do not
333 have another self-image they have to consider during interaction. There may be
334 elements of corporate rather than individual self-images present during interaction,
335 and users can still be imposed upon by machines. However, this remains markedly
336 different from interaction with other people. Indeed, recent research observed that,
337 while descriptions of conversations with people often discuss social and interpersonal

338 wants and needs, interactions with machines are described in very functional and
 339 tool-like terms (Clark et al., 2019a). This may be due to a lack of familiarity and
 340 experience from which to draw upon. However, spoken interactions with machines
 341 lack many of the conversational complexities seen in human communication and are
 342 often limited to isolated question–answer pairs (Porcheron et al., 2018).

343 17.5 Future Work and Considerations for Computer 344 Speech

345 This chapter has presented the possible existence of verbal uncanny valley effects—
 346 that perceptual tension and negative user experiences and attitudes can emerge in
 347 spoken interactions with computers when using linguistic strategies that are inher-
 348 ently social and interpersonal. This effect appears to be intensified with more robotic
 349 voices and lessened, though not entirely, with more humanlike voices. This differs
 350 from previous discussions of an auditory uncanny valley (e.g. Grimshaw, 2009; Meah
 351 & Moore, 2014) in that it focuses on both language and voice quality, and the rela-
 352 tionship between them. Verbal uncanny valley effects suggest there may be category
 353 memberships that exist with styles of language that focus on relational work—i.e.
 354 that other people are members of this category, whereas computers do not become
 355 automatic members by virtue of employing the same strategies. Doing so may create
 356 an impression of machines encroaching upon the verbal space of people. This is
 357 similar to Moore’s (2017b) discussion of there being a fundamental limit to spoken
 358 interaction between humans and machines. Moore (2015) mentions that endowing
 359 machines with features like humanlike voices can create the mismatched stimuli that
 360 lead to perceptual tension, and this may also hold true for certain linguistic styles.
 361 With similar considerations, it appears that reducing perceptual tension with verbal
 362 uncanny valley effects may depend partly on the relationship between voice and
 363 language. If using a very robotic voice, interpersonal linguistic strategies may not
 364 be appropriate and may be subsequently undesirable and unattractive. Conversely, if
 365 wanting to employ these strategies, a more humanlike voice would be more appro-
 366 priate. However, there remains the possibility that no matter what voice is used,
 367 certain interpersonal language may be evaluated negatively regardless due to funda-
 368 mental and embedded differences in user expectation between humans and computers
 369 as interlocutors.

370 It is likely that this is not always the case—this argument stops short of saying all
 371 types of interpersonal linguistic strategies are off-limits. However, there are design
 372 choices around voice and language to consider for computers using speech. There
 373 are also other choices to consider. The discussions of politeness strategies and VL
 374 in this chapter tend to focus on task-based scenarios in HCI. While this is arguably
 375 where most speech-based HCI still currently remains at a linguistic level, it may be
 376 the case that instruction-giving or advice-giving computers in task-based scenarios
 377 are not appropriate vessels for interpersonal language. If the aim of an interaction

378 between speaking computers and humans is fundamentally an interpersonal one (e.g.
379 social talk Gilmartin, Cowan, Vogel, & Campbell, 2017) or in healthcare dialogues
380 (Bickmore et al., 2018), then these linguistic styles may be more appropriate. Simi-
381 larly, the role in which both computer and human play in any given interaction may
382 also influence evaluations of speech—an instruction-giver may be treated differently
383 to a machine that operates more on a peer-level or as a caregiver, due to varying levels
384 of power and exactly what linguistic possibilities these roles may afford. Similarly,
385 human-controlled speech synthesis output, such as the use of a vocal synthesiser to
386 create the ability to speak, may be evaluated differently to speech synthesis output
387 that is controlled by a machine. Furthermore, the direction of interaction may have
388 an effect. Previous experiments often focus on speech output only from a system,
389 whereas two-way dialogue may induce different evaluations. Previous research has
390 shown that politeness can be reciprocated back and forth in an interaction with an
391 in-car help system (Large, Clark, Quandt, Burnett, & Skrypchuk, 2017), though the
392 work does not provide insight into people’s actual evaluations of the system.

393 However, while these ideas are rooted in evidence from previous research, there
394 is still the need to test them further. As noted in Sect. 17.2, the evidence for the
395 uncanny valley alone is scarce, with Moore’s (2012) Bayesian approach offering a
396 rare quantitative verification of its existence. Future research endeavours can explore
397 the concept of a verbal uncanny valley and its effects further in both quantitative
398 and qualitative means, although any notions of a valley in terms of the shape are
399 arguably less important than the effects caused by underlying concepts of funda-
400 mental communicative limits. Comparisons with actual human stimuli as well as
401 computers may also prove beneficial. Indeed, quantifying what constitutes ‘human-
402 like’ or ‘machinelike’ communication is a complex process. Given the increasing
403 prevalence of computer speech, what is perceived as ‘machinelike’ may well change
404 over the years as familiarity with these devices increases. Longitudinal studies may
405 also uncover further evidence on the effects of prolonged interaction with devices
406 and the extent to which this may affect any verbal uncanny valley effects.

407 17.6 Summary and Conclusion

408 Determining what is considered appropriate speech in HCI remains a challenge.
409 Moore (2017a) offers examples of how to determine appropriateness in the voices of
410 non-human artefacts and avoid uncanny valley effects—robotic rather than human-
411 like in less sophisticated systems may be better at matching users’ expectations of a
412 system with reality. Language use, however, is arguably a more complex affair. This
413 chapter discusses three concepts of interpersonal linguistic strategies (politeness,
414 relational work and VL) to explore what may be considered appropriate language
415 use in speech-based HCI. In linking previous experiments on these strategies with
416 research on the uncanny valley, we find that the social rules that underpin human
417 interaction do not automatically transfer to HCI. The concept of face—the social
418 self-image presented to others—is mostly non-existent on the part of the system

419 during interaction. The need to conduct facework, i.e. protecting this self-image,
 420 is then diminished. While users can still be imposed upon by an interface, using
 421 strategies like politeness and VL may not always be appropriate and may be unde-
 422 sirable. The combination of computer speech and interpersonal language gives rise
 423 to perceptual mismatch at the category boundaries between human and computer
 424 speech, creating potential for negative user evaluations of systems. Consequently,
 425 this raises the potential of verbal uncanny valley effects, whereby the use of very
 426 ‘humanlike’ language creates feelings of perceptual tension in HCI. While a human-
 427 like voice can act as a moderator for these effects, it does not alleviate perceptual
 428 tension completely. Future research should explore the empirical testing of the verbal
 429 uncanny valley and its effects, identify what linguistic concepts are seen to reside
 430 in the category of appropriate and inappropriate computer speech, and understand
 431 what further phenomena (like voice) may influence its evaluation by users.

432 **Acknowledgments** This research was funded by a New Horizons grant from the Irish Research
 433 Council entitled “The COG-SIS Project: Cognitive effects of Speech Interface Synthesis” (Grant
 434 R17339).

435 References

- 436 Abercrombie, D. (1967). *Elements of general phonetics* (Vol. 203). Edinburgh: Edinburgh University
 437 Press.
- 438 Aylett, M. P., Cowan, B. R., & Clark, L. (2019). Siri, echo and performance: You have to suffer
 439 darling. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.
 440 ACM.
- 441 Bickmore, T. W., Trinh, H., Olafsson, S., O’Leary, T. K., Asadi, R., Rickles, N. M., & Cruz,
 442 R. (2018). Patient and consumer safety risks when using conversational assistants for medical
 443 information: An observational study of Siri, Alexa, and Google Assistant. *Journal of Medical*
 444 *Internet Research*, 20(9). <https://doi.org/10.2196/11510>.
- 445 Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage*. Cambridge
 446 University Press.
- 447 Cameron, D. (2001). *Working with spoken discourse*. SAGE.
- 448 Carr, E. W., Hofree, G., Sheldon, K., Saygin, A. P., & Winkielman, P. (2017). Is that a human?
 449 Categorization (dis)fluency drives evaluations of agents ambiguous on human-likeness. *Journal*
 450 *of Experimental Psychology: Human Perception and Performance*, 43(4), 651–666. <https://doi.org/10.1037/xhp0000304>.
- 451 Channell, J. (1994). *Vague language*. Oxford University Press.
- 452 Clark, L. (2018). Social boundaries of appropriate speech in HCI: A politeness perspective. In
 453 *Proceedings of British HCI*.
- 454 Clark, L., Cabral, J. & Cowan, B. R. (2018). The CogSIS project: Examining the cognitive effects
 455 of speech interface synthesis. In *Proceedings of British HCI*.
- 456 Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., ... & Cowan, B. R. (2019a).
 457 The state of speech in HCI: Trends, themes and challenges. *Interacting with Computers*, 31(4),
 458 349–371. <https://doi.org/10.1093/iwc/iwz016>.
- 459 Clark, L., Pantidi, N., Cooney, O., Doyle, P., Garaialde, D., Edwards, J., ... & Cowan, B.R. (2019b,
 460 May). What makes a good conversation? challenges in designing truly conversational agents.
 461 In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–12).
 462 <https://doi.org/10.1145/3290605.3300705>.

- 464 Clark, L. M. H., Bachour, K., Ofemile, A., Adolphs, S., & Rodden, T. (2014). Potential of imprecision: Exploring vague language in agent instructors (pp. 339–344). ACM Press. <https://doi.org/10.1145/2658861.2658895>
- 465
- 466 Clark, L., Ofemile, A., Adolphs, S., & Rodden, T. (2016). A multimodal approach to assessing user experiences with agent helpers. *ACM Transactions on Interactive Intelligent Systems*, 6(4), 29:1–29:31. <https://doi.org/10.1145/2983926>.
- 467
- 468
- 469 Coulthard, M. (2013). *Advances in spoken discourse analysis*. Routledge.
- 470
- 471 Cowan, B. R., Branigan, H. P., Obregón, M., Bugis, E., & Beale, R. (2015). Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human – computer dialogue. *International Journal of Human-Computer Studies*, 83, 27–42. <https://doi.org/10.1016/j.ijhcs.2015.05.008>.
- 472
- 473
- 474
- 475 Cowan, B. R., Pantidi, N., Coyle, D., Morrissey, K., Clarke, P., Al-Shehri, S., ... Bandeira, N. (2017). ‘What can I help you with?’: Infrequent users’ experiences of intelligent personal assistants (pp. 1–12). ACM Press. <https://doi.org/10.1145/3098279.3098539>.
- 476
- 477
- 478 Gilmartin, E., Cowan, B. R., Vogel, C., & Campbell, N. (2017). Exploring multiparty casual talk for social human-machine dialogue. In *International Conference on Speech and Computer* (pp. 370–378). Springer.
- 479
- 480
- 481 Goffman, E. (1955). On face-work. *Psychiatry*, 18(3), 213–231. <https://doi.org/10.1080/00332747.1955.11023008>.
- 482
- 483 Goffman, E. (2005). *Interaction ritual: Essays in face to face behavior*. AldineTransaction.
- 484
- 485 Grimshaw, M. (2009). The audio Uncanny Valley: Sound, fear and the horror game. *Audio Mostly*, 21–26.
- 486
- 487 Hone, K. S., & Graham, R. (2000). Towards a tool for the subjective assessment of speech system interfaces (SASSI). *Natural Language Engineering*, 6(3–4), 287–303.
- 488
- 489 Jucker, A. H., & Ziv, Y. (1998). *Discourse markers: Descriptions and theory*. John Benjamins Publishing.
- 490
- 491 Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.00390>.
- 492
- 493 Large, D. R., Clark, L., Quandt, A., Burnett, G., & Skrypchuk, L. (2017). Steering the conversation: A linguistic exploration of natural language interactions with a digital assistant during simulated driving. *Applied Ergonomics*, 63, 53–61. <https://doi.org/10.1016/j.apergo.2017.04.003>.
- 494
- 495
- 496 Laver, J. (1980). *The phonetic description of voice quality: Cambridge Studies in Linguistics*. Cambridge: Cambridge University Press.
- 497
- 498 Locher, M. A. (2004). *Power and politeness in action: Disagreements in oral communication*. Walter de Gruyter.
- 499
- 500 Locher, M. A. (2006). *Polite behavior within relational work: The discursive approach to politeness*. Walter de Gruyter.
- 501
- 502 Locher, M. A., & Watts, R. J. (2005). Politeness theory and relational work. *Journal of Politeness Research. Language, Behaviour, Culture*, 1(1). <https://doi.org/10.1515/jplr.2005.1.1.9>
- 503
- 504 Locher, M. A., & Watts, R. J. (2008). *Relational work and impoliteness: Negotiating norms of linguistic behaviour*. Mouton de Gruyter.
- 505
- 506 Luger, E., & Sellen, A. (2016). ‘Like having a really bad PA’: The gulf between user expectation and experience of conversational agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 5286–5297). New York, NY, USA: ACM. <https://doi.org/10.1145/2858036.2858288>.
- 507
- 508
- 509
- 510 McCarthy, M., & Carter, R. (2006). This that and the other: Multi-word clusters in spoken English as visible patterns of interaction. *Explorations in Corpus Linguistics*, 7.
- 511
- 512 Meah, L. F. S., & Moore, R. K. (2014). The Uncanny Valley: A focus on misaligned cues. In M. Beetz, B. Johnston, & M.-A. Williams (Eds.), *Social robotics* (pp. 256–265). Springer International Publishing.
- 513
- 514

- 515 Mitchell, W. J., Szerszen, K. A., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & MacDorman,
 516 K. F. (2011). A mismatch in the human realism of face and voice produces an Uncanny Valley.
 517 *I-Perception*, 2(1), 10–12. <https://doi.org/10.1068/i0415>.
- 518 Moore, R. K. (2012). A Bayesian explanation of the ‘Uncanny Valley’ effect and related
 519 psychological phenomena. *Scientific Reports*, 2(1). <https://doi.org/10.1038/srep00864>.
- 520 Moore, R. K. (2015). *From talking and listening robots to intelligent communicative machines*. In
 521 *Robots that talk and listen: de Gruyter*.
- 522 Moore, R. K. (2017a). Appropriate voices for artefacts: Some key insights. In *1st International*
 523 *Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*.
- 524 Moore, R. K. (2017b). Is spoken language all-or-nothing? Implications for future speech-based
 525 human-machine interaction. In *Dialogues with Social Robots* (pp. 281–291). Springer, Singapore.
 526 https://doi.org/10.1007/978-981-10-2585-3_22.
- 527 Mori, M. (1970). The Uncanny Valley. *Energy*, 7(4), 33–35.
- 528 Mori, M., MacDorman, K. F., & Kageki, N. (2012). The Uncanny Valley [from the field]. *IEEE*
 529 *Robotics and Automation Magazine*, 19(2), 98–100.
- 530 Porcheron, M., Fischer, J. E., Reeves, S., & Sharples, S. (2018). Voice interfaces in everyday life.
 531 In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (p. 640).
 532 ACM.
- 533 Porcheron, M., Fischer, J. E., & Sharples, S. (2017). ‘Do animals have accents?’: Talking with
 534 agents in multi-party conversation (pp. 207–219). ACM Press. <https://doi.org/10.1145/2998181.2998298>.
- 535
- 536 Strait, M., Canning, C., & Scheutz, M. (2014). *Let me tell you! Investigating the effects of robot*
 537 *communication strategies in advice-giving situations based on robot appearance, interaction*
 538 *modality and distance* (pp. 479–486). ACM Press. <https://doi.org/10.1145/2559636.2559670>.
- 539 Torrey, C., Fussell, S. R., & Kiesler, S. (2013). *How a robot should give advice* (pp. 275–282).
 540 IEEE. <https://doi.org/10.1109/HRI.2013.6483599>
- 541 Trappes-Lomax, H. (2007). Vague language as a means of self-protective avoidance: Tension
 542 management in conference talks. In *Vague language explored* (pp. 117–137). Springer.
- 543 Wang, N., Johnson, W. L., Mayer, R. E., Rizzo, P., Shaw, E., & Collins, H. (2008). The politeness
 544 effect: Pedagogical agents and learning outcomes. *International Journal of Human-Computer*
 545 *Studies*, 66(2), 98–112. <https://doi.org/10.1016/j.ijhcs.2007.09.003>.
- 546 Watts, R. J. (2003). *Politeness*. Cambridge University Press.
- 547 Zuckerman, M., & Driver, R. E. (1988). What sounds beautiful is good: The vocal attractiveness
 548 stereotype. *Journal of Nonverbal Behavior*, 13(2), 67–82. <https://doi.org/10.1007/BF00990791>.

MARKED PROOF

Please correct and return this set

Please use the proof correction marks shown below for all alterations and corrections. If you wish to return your proof by fax you should ensure that all amendments are written clearly in dark ink and are made well within the page margins.

<i>Instruction to printer</i>	<i>Textual mark</i>	<i>Marginal mark</i>
Leave unchanged	... under matter to remain	Ⓟ
Insert in text the matter indicated in the margin	∧	New matter followed by ∧ or ∧ [Ⓢ]
Delete	/ through single character, rule or underline or ┌───┐ through all characters to be deleted	Ⓞ or Ⓞ [Ⓢ]
Substitute character or substitute part of one or more word(s)	/ through letter or ┌───┐ through characters	new character / or new characters /
Change to italics	— under matter to be changed	↵
Change to capitals	≡ under matter to be changed	≡
Change to small capitals	≡ under matter to be changed	≡
Change to bold type	~ under matter to be changed	~
Change to bold italic	⌘ under matter to be changed	⌘
Change to lower case	Encircle matter to be changed	⊖
Change italic to upright type	(As above)	⊕
Change bold to non-bold type	(As above)	⊖
Insert 'superior' character	/ through character or ∧ where required	Υ or Υ under character e.g. Υ or Υ
Insert 'inferior' character	(As above)	∧ over character e.g. ∧
Insert full stop	(As above)	⊙
Insert comma	(As above)	,
Insert single quotation marks	(As above)	ʹ or ʸ and/or ʹ or ʸ
Insert double quotation marks	(As above)	“ or ” and/or ” or ”
Insert hyphen	(As above)	⊖
Start new paragraph	┌	┌
No new paragraph	┐	┐
Transpose	└┐	└┐
Close up	linking ○ characters	Ⓞ
Insert or substitute space between characters or words	/ through character or ∧ where required	Υ
Reduce space between characters or words		↑