Do Voice-Based Judgments of Socially Relevant Speaker Traits Differ Across Speech Types?

Agata Groyecka-Bernard, Katarzyna Pisanski, Tomasz Frąckowiak, Aleksander Kobylarek, Piotr Kupczyk, Anna Oleszkiewicz, Agnieszka Sabiniewicz, Monika Wróbel, Piotr Sorokowski

ACCEPTED VERSION

Do voice-based judgments of socially relevant speaker traits differ across speech types?

Agata Groyecka-Bernard[1,2], Katarzyna Pisanski[1,3,4], Tomasz Frąckowiak[1], Aleksander Kobylarek[5], Piotr Kupczyk[1], Anna Oleszkiewicz[1,5], Agnieszka Sabiniewicz[1,5], Monika Wróbel[1], Piotr Sorokowski[1]*

[1] University of Wrocław, Institute of Psychology, Wrocław, Poland
[2]Johannes Gutenberg University Mainz, Mainz, Germany
[3] ENES Bioacoustics Research Laboratory, University of Saint-Etienne, CRNL, CNRS UMR 5292, Inserm UMR_S 1028, Saint-Etienne, France
[4]CNRS Centre National de la Recherche Scientifique, Laboratoire Dynamique du Langage, Université Lyon 2, Lyon, France
[5]University of Wrocław, Institute of Pedagogy, Wrocław Poland

*Correspondence: piotr.sorokowski@uwr.edu.pl, University of Wrocław, Institute of Psychology, ul. Dawida 1, 50-529, Wrocław, Poland

Words count: 8503

Conflict of interest statement: The authors declare no conflict of interest.

Funding Statement: The study was funded by the Polish National Science Center grant OPUS (2016/23/B/HS6/00771) awarded to PS.

ABSTRACT

Purpose: The human voice is a powerful and evolved social tool, with hundreds of studies showing that nonverbal vocal parameters robustly influence listeners' perceptions of socially meaningful speaker traits, ranging from perceived gender and age to attractiveness and trustworthiness. Yet these studies have utilized a wide variety of voice stimuli to measure listeners' voice-based judgments of these traits. Here, in the largest scale study known to date, we test whether listeners judge the same unseen speakers differently depending on the complexity of the neutral speech stimulus, from single vowel sounds to a full paragraph.
Method: In a playback experiment testing 2618 listeners, we examine whether commonly studied voice-based judgments of attractiveness, trustworthiness, dominance, likability, femininity/masculinity and health differ if listeners hear isolated vowels, a series of vowels, single words, single sentences (greeting), counting from 1 to 10, or a full paragraph recited aloud (Rainbow Passage), recorded from the same 208 men and women. Data were collected using a custom designed interface in which vocalizers and traits were randomly assigned to raters.
Results: Linear mixed models show that the type of voice stimulus does indeed consistently affect listeners' judgments. Overall, ratings of attractiveness, trustworthiness, dominance, likability, health, masculinity among men and femininity among women increase as speech duration increases. At the same time, speaker-level regression analyses show that inter-individual differences in perceived speaker traits are largely preserved across voice stimuli, especially among those of a similar duration.
Conclusions: Socially relevant perceptions of speakers are not wholly changed but rather moderated by the length of their speech. Indeed, the same vocalizer is perceived in a similar way regardless of which neutral statements they speak, with the caveat that longer utterances explain the most shared variance in listeners' judgments and elicit the highest ratings on all traits, possibly by providing additional nonverbal information to listeners.

Key-words: voice, stimulus type, stimulus duration, voice perception, playback experiment

Introduction

The human voice is a source of abundant information about the vocalizer, which listeners can use to make socially relevant decisions. Hundreds if not thousands of experimental studies have shown that the human voice plays a central role in predicting listeners' perceptions of the social and biological qualities of vocalizers, including psychological and physical traits (reviewed in: Aung & Puts, 2020; Kamiloğlu & Sauter, 2021; Kreiman & Sidtis, 2011; Pisanski & Bryant, 2019). For example, the nonverbal properties of a person's voice, particularly fundamental frequency (f_o) perceived as voice pitch, and formant frequencies affecting voice timbre, can predict listeners' judgments of a vocalizer's personality traits (Stern et al., 2021), social dominance (Aung & Puts, 2020; Borkowska & Pawlowski, 2011; David A. Puts et al., 2016), attractiveness (reviewed in Pisanski & Feinberg, 2019), and health (Vukovic et al., 2010), to name only a few. In turn, these same voice features can predict who people vote for in an election (Klofstad et al., 2012; Mileva et al., 2020; Tigue et al., 2012), choose to hire in a job interview (Anderson et al., 2014), or choose as a romantic partner (Pisanski et al., 2018; Rosenfield et al., 2020).

Considering the increasing prevalence of voice research in the human behavioral sciences, and mounting empirical evidence that nonverbal voice parameters predict perceptions of the vocalizer in the 'ears' of the beholder, remarkably few studies have examined whether such voice-based perceptions depend on the type of voice stimulus being judged. Indeed, the abovementioned studies on human nonverbal voice production and perception utilized a variety of speech stimuli to test listeners' perceptions of vocalizer traits. Traditionally, these have often included affectively neutral speech, such as vowel sounds or scripted sentences and paragraphs, designed to control for linguistic content, but nevertheless varying in duration and complexity. The potential effects driven by these differences in duration or complexity across speech stimuli, and thus the amount of nonverbal information available to the listener,

remain largely unknown. There is hence a compelling need to gain knowledge about the validity and comparability of the diverse methods used to study vocal communication in humans. Testing the extent to which listeners' socially relevant judgments of speakers remain stable across different utterances produced by the same person also carries theoretical and social implications, namely regarding the underlying mechanisms driving voice-based judgments that are in turn known to predict important real-world outcomes for speakers. Here, to this aim, we directly test whether differences in the type of neutral speech stimulus used can affect listeners' judgements of the same vocalizers on a range of socially relevant traits.

Studies testing human voice perception often follow a similar basic protocol and are typically referred to as playback, psychoacoustic, or perception experiments. At the most basic level, the first step is to record vocalizers' voices, and the second is to subsequently play these speech stimuli to a sample of listeners who then evaluate them on various scales (e.g., pertaining to personality, attractiveness, physical traits). Some studies utilize unchanged, natural voice samples (Cartei et al., 2014; McAleer et al., 2014; Pisanski et al., 2014; Sorokowski et al., 2019) while others manipulate the nonverbal acoustic properties of speech to causally test how changes in specific acoustic parameters (for instance voice pitch) affect listeners' judgements (e.g., Albert et al., 2021; Belin et al., 2017; Feinberg et al., 2005; Krahé et al., 2021; Pisanski et al., 2012).

The type of voice stimuli used in playback experiments vary greatly. Some researchers use standard phrases like vowels (for example a e i o u, Feinberg et al., 2008), counting from 1 to 10 (e.g., Hughes et al., 2014), single words (e.g., greetings, see Apicella & Feinberg, 2009), single sentences (e.g., Jones et al., 2008) or a phonetically balanced paragraph such as the Rainbow Passage that is recited aloud (Fairbanks, 1960) – a passage that contains a broad representation of vowel sounds (e.g., Pisanski et al., 2016; Pisanski, Anikin, et al., 2021; Puts et al., 2006; Schild et al., 2020). These stimulus types differ in numerous ways, most notably

in their duration and complexity. Preliminary evidence points to a negative relationship between the duration of a voice stimulus and within-listener stability in voice perception (Ohno et al., 2014). One possible explanation for this is that longer stimuli provide more nonverbal information than do shorter stimuli, particularly regarding stable or dynamic acoustic parameters of the speaker's voice such as fundamental and formant frequencies. Longer stimuli also offer more time for listeners to consider their responses. On the other hand, by increasing the amount of nonverbal information available to listeners, longer speech may introduce more variability in listener's judgments of the vocalizer owing to more opportunities for listeners to express their individual differences in voice preferences and perceptions, and in turn, may reduce inter-rater agreement.

A recent study has shown that individual differences in voice fundamental frequency (f_o, perceived as voice pitch) are preserved across speech types (Pisanski, Groyecka-bernard, et al., 2021). Pisanski and colleagues (2021) analyzed f_o in six different types of neutral speech utterances (from vowels to longer bouts of spontaneous speech) and showed that inter-individual differences in this salient voice property are highly robust, such that a person's voice pitch when speaking a vowel sound correlates strongly with their voice pitch when speaking a full paragraph of free speech. These results thus demonstrate the methodological validity of comparing f_o measures across different neutral speech stimulus types. Nevertheless, these results, based only on acoustic measures, do not necessitate that listeners' perceptions will likewise be robust across speech types, wherein stimuli can differ on a range of other spectrotemporal parameters that may affect listeners' voice-based judgments. Some voice stimuli also have higher ecological validity than do others (as in the case of greetings versus vowels), which may differentially influence listeners' judgements. For example, single vowel sounds, in addition to providing limited information regarding articulation compared to longer phrases (Kreiman, 1997), are rarely used in isolation in everyday real-life

conversations. Some researchers have thus opted to use standardized greetings resembling statements one might use in everyday life (e.g., "Hi, I'm a student at UCLA": Bryant & Haselton, 2009; "Get out and be quiet": Sell et al., 2010), or entirely unscripted speech produced in response to a given context (e.g., discussing one's admirable traits in a competitive context: Puts et al., 2006), to increase the ecological validity of speech stimuli collected in the lab.

Very few studies have attempted to address how the use of different voice stimuli might affect the results of playback experiments. One study tested for differences in attractiveness ratings based on voice duration and type, but the researchers considered only very short (i.e., single vowel, three vowels, and 'bonjour') stimuli and artificial manipulation thereof (Ferdenzi et al., 2013). The authors showed that artificially shortening the voice samples decreased the perceived attractiveness of the same vocalizers and that words elicited higher attractiveness ratings compared to vowels. While Ferdenzi et al. focused solely on attractiveness ratings, Mahrholz and colleagues (2018) tested whether listeners' judgements of attractiveness, dominance, and trustworthiness vary for one word versus one sentence. In the one-sentence stimulus, the authors also manipulated content (i.e., social relevance). In that study, listeners' judgments were highly correlated for words and sentences produced by the same set of vocalizers, which could potentially mean that speech content did not influence judgments in that study. The findings thus suggest that, regardless of speech stimulus type, vocalizers are perceived similarly. However, these studies were limited to perceptions of only three vocalizer traits: attractiveness, dominance, and/or trustworthiness. They did not explore how the type of speech stimulus affects perceptions of other traits known to be highly relevant in human interpersonal relationships, and known to be perceptually linked to nonverbal parameters of the voice, including masculinity/femininity, likeability, and health (for review see Kreiman & Sidtis, 2011). Moreover, Mahrholz et al. (2018) used only two types of stimuli

(word, sentence), which, similar to Ferdenzi et al. (2013), cover only a small share of the methodological diversity observed in the literature, with both studies focusing on short stimuli.

The current study was designed to complement the scarce literature regarding the potential effects of the type of speech stimulus on listeners' perceptions of unseen vocalizers. In a large-scale playback experiment, we test whether a broad variety of utterances (consisting of isolated vowels, vowels pronounced in a series, single words, counting 1-10, greeting sentences, and reading aloud a phonetically balanced passage) elicit different assessments of the same vocalizers' perceived attractiveness, trustworthiness, dominance, likability, femininity/masculinity and health. The study examines the effects of specific stimulus types as well as the general effect of stimulus duration on perceptions of these traits as judged by a large sample of over 2000 male and female listeners. Because voice perception can heavily rely on inter-vocalizer or inter-listener factors, we controlled for variance due to individual differences in multilevel models.

The perceived vocalizer traits tested here were selected on the basis of countless studies that have shown that listeners' perceptions of these specific traits, while not exhaustive, are robustly influenced by nonverbal parameters of the voice, and are thus among the most intensively studied voice-based traits in the human voice sciences (for reviews see Kamiloğlu & Sauter, 2021; Kreiman & Sidtis, 2011; Pisanski & Bryant, 2019). Importantly, voice-based perceptions of these same traits are also known to predict a range of social decisions, such as mate preferences (Rosenfield et al., 2020) and election outcomes (Mileva et al., 2020).

The question of whether the type of speech uttered affects how a person is perceived is important from several perspectives. From a methodological perspective it can verify whether

results of past studies using myriad types of voice stimuli to test listeners' perceptions of the same trait are comparable. From a social perspective, it may provide novel insight into whether such perceptions are robustly preserved across different speech types. For example, would a speaker uttering a series of sentences be perceived as similarly attractive, trustworthy, or likeable if uttering only a single word? Likewise, are voice-based perceptions of masculinity or femininity, dominance, and health dependent on the complexity of a speech utterance? Given the importance of these perceived traits for interpersonal relationships, mate choice, and/or broader societal outcomes, the extent to which such judgments are made based on snippets of nonverbal vocal cues is of high public relevance. From a practical perspective, the research question is also relevant for voice-based technologies such as automated voice recognition and detection devices or voice-based clinical diagnostic tools, in which algorithms relying on artificial intelligence may be refined depending on the robustness of data obtained from various durations and types of speech.

Method

The study was conducted in accordance with the Declaration of Helsinki. Study protocols were accepted by the institutional ethics committee at the University of Wrocław. All participants (vocalizers and raters) provided informed consent prior to participation. Vocalizers were informed that their voice recordings will be further used for the purposes of the present study and that they will be played to other participants.

Participants

The number of raters was based on the number of vocalizers with the aim that each voice stimulus be rated approximately 20 times, based on evidence that high inter-rated agreement (alphas >0.80, $p < 0.001$) among listeners is typically achieved with relatively small sample

sizes (e.g., less than 15 listeners per sex for voice-based judgements of dominance or attractiveness: Kordsmeyer et al., 2018; Schild et al., 2020). The number of speakers (ca. 200) was determined as a trade-off between a sample that translates into a reasonable number of raters and one that allows us to detect inter-individual differences in traits of interest and nonverbal aspects of voice communication.

Vocalizers

208 vocalizers representing a broad age range and balanced sex ratio (M_{age} = 32.83, SD_{age} = 12.32, 48% women, 52% men) provided voice recordings for use as stimuli in playback experiments. Their native language was Polish and they were recruited through snowball sampling by researchers and research assistants who posted recruitment ads on their social media profiles, and around their city of residence, both inside and outside of the university. Vocalizers were not compensated for providing speech samples.

Voice raters

2618[1] voice raters representing a broad age range and balance sex ratio (M_{age} = 32.51, SD_{age} = 13.01, 54% women, 46% men) judged the speech samples. All raters reported normal hearing. To increase our sample size and the diversity of our rater sample, raters were recruited via a combination of the snowball sampling method and through a dedicated research recruitment firm. Snowball sampling followed the same procedures as for vocalizers, with additional recruitment efforts targeting older and elderly individuals. All raters (lab and online) were residents of Poland and understood written Polish as verified by the recruitment firm and/or in the attention and hearing tests preceding the experiment. Participants were reimbursed in cash (for the cohort recruited via the recruitment firm) or through a lottery draw of small prizes such as pen drives (for all remaining participants).

[1] Data from more participants were collected, however, they were asked different questions for the purposes of a different study.

Materials

Voice recording

Participants were recorded at the Institute of Psychology at the University of [covered for blind review]. First, participants were audio recorded in private sessions in a small quiet room with a low level of external noises. We used a Zoom H4n professional digital recorder with X/Y stereo microphone array positioned 10 cm from the mouth. Participants familiarized themselves with a script and were then instructed to say aloud five items in their native language (Polish), listed below in English translation:

I. Vowels a-e-i-o-u (/a/, /ə/, /i/, /ɔ/, /u/, International Phonetic Alphabet)[2]
II. One-syllable word "lat" (containing the vowel /a/)
III. Counting from 1 to 10
IV. Greeting sentence "Dzień dobry, jestem z Polski" [Good morning, I am from Poland]
V. First paragraph of the Rainbow Passage (Fairbanks, 1960)

After recording the voice samples, vocalizers filled in a short demographic survey. Participants were then thanked and debriefed. Voice recordings were saved as WAV files at 96 kHz sampling frequency and 16-bit resolution, and then further divided into short fragments each containing one specific stimulus type. Vowels were saved together within a single segment and also saved separately[2]. Therefore, 6 different stimulus types from each vocalizer (1248 speech stimuli in total) were prepared for the playback experiments. We coded not only the type of speech but also its length, i.e., number of syllables (1, 5, 5, 16, 8, 150 for isolated vowels, vowel series, word, counting, greeting and paragraph read aloud, respectively).

Stimulus preparation

[2] Vowels were recorded as a sequence. For playback experiments, they were then presented either in isolation, or as a sequence.

Voice samples were incorporated into a custom designed online web app. The survey included display of voice samples, assessment (rating) scales and demographic questions.

Procedure

The voice rating phase took place in person at the University lab, or online. Participants who took part at the lab conducted the playback experiment either separately, or in small groups (2 to 5 individuals) in rooms prepared accordingly. To ensure privacy and independence in responses, participants who conducted the study in a small group were placed in partitioned computer booths to reduce visual contact, and always wore headphones. All lab participants used high quality professional headphones (Sennheiser HD 210). Participants who completed the playback online were instructed to use good quality headphones and to complete the study in a quiet environment without distractions. This was verified with hearing and attention tests. Before beginning the experiment, participants were exposed to a test voice sample to ensure they can hear the stimuli properly, and to set and standardize playback volume. Eighty participants (3%) failed this hearing test and their data were thus omitted from further analyses. At a random time during the playback, participants were also presented with an additional attention check item that read, "This is an attention checking question – please, mark 1". Eleven additional participants (<1%) were excluded due to an incorrect answer. For each participant, a sample of 10 recorded individuals (5 men and 5 women) was randomly drawn. For each of these vocalizers, a set of 6 voice samples was presented (resulting in 60 voice stimuli per listener). These stimuli were presented in a random order, each time followed by a request to evaluate the vocalizer on one given dimension. Thus, each speech stimulus was judged independently on a single trial and for a single trait. Each listener evaluated 60 voice samples in total on the same randomly assigned trait, i.e., one of the following questions appeared after each speech recording:

1. How attractive is this person? (1 = very unattractive, 7 = very attractive)
2. How dominant is this person? (1 = not dominant at all, 7 = very dominant)
3. How likeable is this person? (1 = not likable at all, 7 = very likable)
4. How trustworthy is this person? (1 = not trustworthy at all, 7 = very trustworthy)
5. How feminine/masculine is this person? (1 = very feminine, 7 = very masculine)
6. How healthy is this person? (1 = very unhealthy, 7 = very healthy)

To align closely with most studies in this domain (e.g., Cartei et al., 2014; Feinberg et al., 2012; Hughes et al., 2014; McAleer et al., 2014; Pisanski & Rendall, 2011), participants were not provided with any definitions of these concepts. Following the experiment, they reported their age and sex and were debriefed. Rating sessions lasted an average of 30 minutes including instructions, hearing and attention checks, playback (rating speech stimuli), brief survey, and debriefing.

Statistical Analysis

In order to test the effects of stimulus type on perceived attractiveness, dominance, likability, trustworthiness, masculinity/femininity, and health, we proceeded with a series of Linear Mixed Models (LMMs). Due to sizeable sexual dimorphism in the nonverbal properties of the human voice (Titze, 1989) and known differences in perceived attractiveness, masculinity and dominance for male versus female voices (Borkowska & Pawlowski, 2011; Cartei et al., 2014; Pisanski & Rendall, 2011), we conducted separate LMMs for male and female vocalizers. The data fulfilled the assumption of residual normality which was identified by visual exploration of Q-Q plots of residuals. The only exception where the distribution was slightly different from normal was femininity-masculinity, however, this violation was not extreme.

The dependent variables (ratings based on a single voice stimulus) were nested within vocalizers and within listeners. The models differed only by the outcome variable (each time a different perceived trait as an outcome) and were estimated using a Restricted Maximum Likelihood (REML) estimator. We included type of voice stimulus as a factor, vocalizer's age and listener's sex (0-F, 1-M) and method of data collection during playback experiments (0-online raters, 1-lab raters) as covariates. These variables were treated as fixed effects. Random effects of vocalizer numeric ID and listener numeric ID were included to control for noise in the models owing to potential individual differences in speech stimuli across vocalizers or systematic biases in different listeners' ratings, i.e., random effects refer to speaker- and vocalizer-level variance. All continuous predictors were grand-mean centered. Each model was followed up by Bonferroni corrected pairwise comparisons. As a measure of effect size, we compared the amount of explained variance in trait ratings in models that included fixed effects of speech stimulus type against a null model (without predictors, accounted for clustering).

Results

Our results show that listeners' ratings on almost every trait generally increased with the duration of the voice stimulus. In general, single vowels and single words elicited the lowest ratings, longer utterances (counting, greetings and paragraph read aloud) elicited the highest ratings, and vowels pronounced in a series often obtained intermediate ratings (see dominance for an exception). The absolute differences between the lowest and highest rated stimulus types were significant, for all six dimensions, and ranged between 0.30 for masculinity among male vocalizers and 1.08 for trustworthiness in female vocalizers (on a 1-7 scale). The vast majority of differences in ratings across stimuli were statistically significant (for exceptions see Figure 1a-f).

Attractiveness

Overall average perceived voice attractiveness of female vocalizers, controlling for other predictors, was 3.88 ± 0.06 (mean ± standard error of the mean). The effect of voice speech stimulus type was significant $F(5, 12515) = 219.14$, $p < .001$. For the estimates of fixed and random effects see Table 1. The highest attractiveness ratings overall were those assessed after listening to the recited paragraph (longest utterance, 4.24 ± 0.07), followed by counting (4.17 ± 0.07) and greetings (4.10 ± 0.07) (intermediate duration stimuli), and then by the series of vowels (4.04 ± 0.07). These same vocalizers were judged as least attractive when speaking only a single word (3.48 ± 0.07) or single vowel (3.34 ± 0.07), the shortest speech utterances. Estimated marginal means controlling for covariates are presented in Table S1 in Supplementary Materials. Pairwise comparisons with Bonferroni correction are presented in Figure 1a and Table S2. Perceptions of attractiveness were positively related to the age of the vocalizer (b = -0.03, $p < .001$) but not to the rater's sex (b = 0.01, $p = 0.99$). There were no significant differences in ratings between raters who completed the playback experiment in the lab versus online (b = -0.17, $p = 0.06$).

Overall perceived attractiveness of male vocalizers, controlling for other predictors, was 3.53 ± 0.06). The effect of stimulus type was significant $F(5, 12260) = 168.30$, $p < .001$. The pattern of results was almost identical to that observed for female vocalizers with a slight difference in the most highly evaluated stimuli: greetings were evaluated as similarly attractive (3.84 ± 0.07) as counting (3.79 ± 0.07), and there was no significant difference between greetings and recited paragraphs (3.91 ± 0.07). Thus, for both sexes of vocalizers, attractiveness ratings tended to increase with the duration of the speech stimulus (See Figure 1a, and Tables S3 and S4 for post-hoc comparisons and estimated marginal means). Younger vocalizers were evaluated as more attractive than were older vocalizers (b = -0.01, $p = .018$) and female listeners rated men's voices as more attractive than did male listeners (b = -0.48, $p < .001$).

Comparing groups of raters who completed the playback experiment in the lab versus online, those who completed the study in the lab rated men's voices as less attractive overall (b = -0.22, $p = 0.02$) compared to those who completed it online (3.68 ± 0.02 vs. 3.47 ± 0.02, respectively). Adding speech stimulus type as a predictor did not substantially increase the amount of variance in attractiveness ratings explained by the model (0.08% and 0.06% for female and male voices, respectively). Due to a significant effect of rater group, we conducted additional analogous analyses separately for each group. The key pattern of results did not change, that is, the effect of speech stimulus type on attractiveness ratings was the same for both groups of raters. The results of these analyses are presented in Tables S5-S9 in the Supplementary Materials.

Dominance

Overall average perceived dominance of female vocalizers, controlling for other predictors, was 3.60 ± 0.07. Speech stimulus type had a significant effect $F(5, 11931.9) = 138.20$, $p < .001$. Listeners rated the series of vowels (3.96 ± 0.07), followed by the greeting (3.81 ± 0.07), counting (3.80 ± 0.07), the recited paragraph (3.70 ± 0.07), single words (3.40 ± 0.07) and vowels (3.04 ± 0.07) as most dominant, respectively (Figure 1b, Table S10 and S11 for exact differences and their significance).

For male vocalizers, average perceived dominance, controlling for other predictors, was 3.68 ± 0.07. The type of voice stimulus also affected perceived dominance $F(5, 11754) = 80.33$, $p < .001$. Listeners rated vocalizer's producing a greeting (3.89 ± 0.07), followed by the series of vowels (3.80 ± 0.07), counting (3.90 ± 0.07), recited paragraph (3.68 ± 0.07), single words (3.58 ± 0.07) and vowels (3.20 ± 0.07), as most dominant, respectively (see Tables S12 and S13 and Figure 1b). None of the covariates yielded a significant influence for dominance ratings of female nor male vocalizers (see Table 2).

There were no significant differences in ratings between raters who completed the playback experiment in the lab versus online for either vocalizer sex (b = -0.09, $p = .36$ for female and b = -0.12, $p = .21$ for male vocalizers). Adding speech stimulus type as a predictor did not significantly increase the amount of variance in dominance ratings explained by the model (0.05% and 0.03% for female and male voices, respectively).

Likability

Overall average perceived likability of female vocalizers, controlling for other predictors, was 4.15 ± 0.06. Speech stimulus type was a significant predictor $F(5, 12363.7) = 174.11$, $p < .001$. The order of likability ratings from highest to lowest was as follows: greetings (4.52 ± 0.06), recited paragraph (4.45 ± 0.06), counting (4.32 ± 0.06), series of vowels (4.17 ± 0.06), single word (3.77 ± 0.06), and vowels pronounced individually (3.73 ± 0.06; see Figure 1c and corresponding Tables S14 and S15 in SOM). Older vocalizers were evaluated as relatively slightly less likeable (b = -0.01, $p<.001$, see Table 3 for estimates of all effects).

For male vocalizers, average perceived likability, controlling for other predictors, was 3.90 ± 0.05. The effect of speech stimulus type was significant $F(5, 12158) = 157.53$, $p<.001$, and almost identical for male and female vocalizers, with the only difference in male vocalizers indicating slightly higher ratings based on a series of vowels (3.80 ± 0.07) compared to a single word (3.67 ± 0.07; Figure 2c). See Tables S16 and S17 for estimated marginal means and post-hoc tests. Male listeners judged voices as significantly less likeable than did female listeners (b = -0.27, $p = .003$).

There were no significant differences in ratings between raters who completed the playback experiment in the lab versus online for either vocalizer sex (b = 0.04, $p = .64$ for female and b = -0.01, $p = .87$ for male vocalizers). Adding fixed effects of speech stimulus type into the models did not significantly increase the amount of variance in likability ratings explained by the model (0.07% and 0.06% for female and male voices, respectively).

Trustworthiness

Overall average perceived trustworthiness of female vocalizers, controlling for other predictors, was 3.79 ± 0.06. Among female vocalizers, speech stimulus type also had a significant effect on perceived trustworthiness $F(5, 11466.8) = 256.44$, $p < .001$. The highest trustworthiness ratings were assigned to greetings (4.20 ± 0.07), then counting (4.12 ± 0.07), recited paragraph (4.10 ± 0.07), series of vowels (3.97 ± 0.07), single word (3.26 ± 0.07) and individual vowels (3.12 ± 0.07), respectively (Figure 1d, see also Tables S18 and S19 in SOM for estimated marginal means and pairwise comparisons). Covariates were not significant (see Table 4).

In male vocalizers, average perceived trustworthiness, controlling for other predictors, was 3.68 ± 0.06. Speech stimulus type had a significant effect on trustworthiness ratings $F(5, 11156) = 216.67$, $p < .001$). The highest trustworthiness ratings were assigned to vocalizers when reciting a paragraph (4.05 ± 0.07), followed by the greeting (4.01 ± 0.07), counting (4.00 ± 0.07), series of vowels (3.73 ± 0.07), single word (3.23 ± 0.07) and individual vowels (3.05 ± 0.07), respectively (see Table S20 and S21 for estimated marginal means and pairwise comparisons). None of the covariates yielded significant effects (see Table 4).

There were no significant differences in ratings between raters who completed the playback experiment in the lab versus online for either vocalizer sex (b = 0.01, $p = .95$ for female and $b = 0.04$, $p = .66$ for male vocalizers). Adding fixed effect of speech stimulus type into the model increased the amount of variance in trustworthiness ratings explained by the model only slightly (0.1% and 0.09% for female and male voices, respectively).

Femininity-Masculinity

The scale was coded such that higher scores indicate masculinity, and lower scores indicate femininity. In female vocalizers, overall average perceived femininity-masculinity, controlling for other predictors, was 2.35 ± 0.05 (thus on the 'feminine' side of the scale). Speech stimulus type had a significant effect on perceived masculinity-femininity $F(5, 11704.9) = 75.31$, $p < .001$. Single vowels (2.62 ± 0.06) followed by a word (2.52 ± 0.06) and vowels in a series (2.32 ± 0.06) elicited the highest masculinity ratings (least feminine). Longer utterances, including greetings (2.23 ± 0.06), counting (2.20 ± 0.06), and the recited paragraph (2.19 ± 0.06), respectively, were rated as least masculine (most feminine) (see Figure 1e and Tables S22 and S23 for means and pairwise comparisons). Higher ratings were, on average, assigned by slightly older raters (b = 0.01, $p < .001$, see Table 5).

The average femininity-masculinity rating in males, controlling for other predictors, was 5.64 ± 0.06 (thus on the 'masculine' side of the scale). Speech stimulus type was significant $F(5, 11459) = 35.35$, $p < .001$. The pattern of results was opposite to that observed in female vocalizers. Specifically, longer utterances including the recited paragraph (5.68 ± 0.06), counting (5.76 ± 0.06), and greeting (5.75 ± 0.06) yielded higher masculinity ratings whereas shorter speech samples yielded the lowest, including vowels in a series (5.66 ± 0.06), single word (5.56 ± 0.06) and individual vowels (5.45 ± 0.06). For both sexes, the series of vowels elicited intermediate ratings (see Figure 1e). The opposing results show that longer utterances generally elicited higher sex-typical ratings (i.e., higher masculinity ratings for men and higher femininity ratings for women) (see Figure 1e; Tables S24 and S25). The effects of covariates were nonsignificant.

There were no significant differences in ratings between raters who completed the playback experiment in the lab versus online for either vocalizer sex (b = 0.18, $p = .07$ for female and b = -0.08, $p = .28$ for male vocalizers). Adding the fixed effect of speech stimulus

type into the model did not significantly increase the amount of variance in femininity/masculinity ratings explained by the model (0.03% and 0.01% for female and male voices, respectively).

Health

Overall average perceived health of female vocalizers, controlling for other predictors, was 5.03 ± 0.06. Among female vocalizers, stimulus type also predicted differences in voice-based health assessments $F(5, 12691.1) = 143.40$, $p < .001$. The perceived healthiness of vocalizers based on listening to their utterances increased in the following order: single vowels (4.53 ± 0.06), single word (4.74 ± 0.06), series of vowels (5.23 ± 0.06), counting (5.25 ± 0.06), greetings (5.36 ± 0.06), and recited paragraph (5.19 ± 0.06; see Figure 1f and Tables S26 and S27 in Supplementary materials). The younger the vocalizers were, the higher their evaluated health (b = -0.02, $p < .001$). The effect of the rater's group was not significant (see Table 6).

For male vocalizers, average perceived health, controlling for other predictors, was 4.80 ± 0.06. Like female vocalizers, stimulus type significantly predicted judgements of men's health $F(5, 12355) = 94.66$, $p < .001$. The perceived healthiness of male vocalizers based on listening to their utterances increased in a following order: single vowels (4.36 ± 0.07), single word (4.71 ± 0.07), series of vowels (4.78 ± 0.07), recited paragraph (4.95 ± 0.07), counting (5.04 ± 0.07) and greetings (5.11 ± 0.07; see Tables S28 and S29 and Figure 1f for means and comparisons). Model summaries are given in Table 6. Younger male vocalizers were also evaluated as slightly healthier than were older male vocalizers (b = -0.01, $p = .045$).

The effect of rater group was significant for male (b = -0.22, $p = .023$) but not female vocalizers (b = -0.17, $p = .06$), with online raters assessing male vocalizers' health as slightly higher than did lab raters (M/SEM = 4.93/0.02 vs. 4.72/0.02). Adding fixed effects of speech stimulus type into the model did not substantially increase the amount of variance in health ratings explained by the model (0.05% and 0.04% for female and male voices, respectively). We present separate analyses of health judgments of male vocalizers made by online versus lab raters in the Supplementary Materials (Tables S30-S34). The key pattern of results did not

change, that is, the effect of speech stimulus type on health ratings was the same for both groups of raters.

Additional analyses

To corroborate our findings and to test whether inter-vocalizer differences in perceived speaker traits are preserved across stimuli, we proceeded with a series of Pearson's correlation analyses conducted on the vocalizer level, i.e., for each speaker, we averaged their ratings for each stimulus type separately and regressed these averaged scores on one another. The results are presented in Figures 2-8. All ratings were significantly and positively correlated at the inter-individual level, suggesting that individual differences in social judgments were preserved across stimulus types. For example, the average attractiveness rating that a given vocalizer received when producing one type of speech utterance explained 10%-74% of the variance in the attractiveness ratings given to that same vocalizer when producing any other utterance (Fig 2). The relationships were moderate to strong for all traits, wherein 9.6% to 73% of variance in each rating dimension was explained across speech types. However, the fact that the correlations in many cases did not exceed $r = 0.5$ implies that the ratings, although meaningfully correlated, are not identical.

In all cases, the weakest (r =0.31-0.6) correlations were observed between pairs consisting of one short (one syllable) utterance and one longer utterance (greeting, counting or recited paragraph; Table S25). In contrast, for almost all traits, the strongest correlations (r = 0.64-0.86) were observed between two longer utterances. In only two cases (dominance judgments of females and trustworthiness judgments of males), the strongest coefficient was found between a series of vowels (mid-length utterance) and a longer speech stimulus.

Discussion

Studies on perceptions of speakers based on the nonverbal properties of their voices vary methodologically owing to the use of a wide range of speech stimulus types, from single vowels to longer paragraphs of speech. This raises concerns regarding comparability of results across studies, namely whether different methodological choices are equally valid and whether listeners judge the exact same unseen speaker differently based on different types of neutral speech utterances, with a range of social and practical implications. The goal of the current study was to compare listeners' judgements of various traits of male and female vocalizers based on six content-neutral speech utterances, differing in duration and complexity, to test the degree to which variance in listeners' voice-based judgments is shared across speech types produced by the same person. In a large-scale playback experiment, we presented over 2000 raters with recordings of various utterances produced by the same approximately 200 vocalizers, and asked them to rate the unseen vocalizers on six different socially relevant dimensions. These included traits relevant to human mate selection and to other non-sexual social contexts (Pisanski & Feinberg, 2019). All of the considered traits have been of high interest to researchers studying the evolution and social outcomes of human vocal communication for decades (for reviews of the literature see: Aung & Puts, 2020; Kamiloğlu & Sauter, 2021; Kreiman & Sidtis, 2011; Oleszkiewicz et al., 2017; Pisanski & Bryant, 2019; Pisanski & Feinberg, 2019; Stern et al., 2021).

Our results show that while listeners' assessments of socially relevant traits are highly correlated across different neutral speech utterances produced by the same vocalizers, listeners' ratings generally increase for utterances of longer duration. Generally, vocalizers of both sexes are likely to be evaluated as more attractive, likeable, trustworthy, healthy, and dominant when producing longer than shorter utterances, particularly compared to single vowels or single words each comprised of only one syllable. In the case of perceived

femininity-masculinity, longer utterances elicited higher femininity ratings for women and higher masculinity ratings for men. At the same time, ratings among different types of longer utterances (reciting a paragraph aloud, counting, greeting, and in some cases, series of vowels; 150, 16, 8 and 5 syllables, respectively) were not remarkably different from one another. For instance, although the multilevel models show significant differences in ratings of dominance among men when comparing counting/greeting versus reciting a full paragraph, the differences in mean scores between these speech stimulus types are very small (0.22/0.21) and a high degree of variance is still shared between ratings based on these stimuli (37-42%).

Indeed, listeners' judgments shared a high degree of variance across stimulus types, within vocalizers. For example, a vocalizer who received high attractiveness ratings based on her or his vowel series was likely to also be judged as highly attractive when producing a full paragraph. However, based on vocalizer-level analyses, relatively more variance was shared between pairs of two long utterances than between two short utterances, or than between a short and long utterance. Importantly, then, while we show that longer utterances elicit relatively higher ratings of attractiveness, trustworthiness, likability, health, dominance, masculinity in males, and femininity in females, we also show moderate to strong relationships between ratings of the same individuals across speech types, indicating stability intra-individual stability in speaker perception. The strongest relationships were observed between stimulus types of a similar length, wherein a given speech type could explain upwards of 73% of the variance in ratings of the same vocalizers based on another speech type.

Although the relationships between pairs of short-long utterances were weaker than between pairs of longer utterances, they were still significant and never lower than r =.26. While acoustic parameters were not measured in the present study, this finding raises the possibility that listeners' social judgments are tapping into the same underlying acoustic

properties in vocalizer's voices, whether those vocalizers are producing vowels or a paragraph, however that longer utterances provide additional information that listeners readily utilize. These results are in line with previous research reporting that listeners can judge a range of personality traits from nothing more than a single utterance (e.g., "hello", McAleer et al., 2014) and that ratings of trustworthiness, dominance, and attractiveness are highly correlated when based on a word vs. sentence (Mahrholz et al., 2018). At the same time, corroborating a previous study showing higher attractiveness ratings for words compared to single vowels (Ferdenzi et al., 2013), our results provide novel insight into the importance of stimulus duration on listeners' perceptions.

There is growing evidence from acoustic analyses of nonverbal vocal parameters of the human voice that certain vocal parameters are stable across speech types. For example, Pisanski and colleagues (2021) found that fundamental frequency (f_o), perceived as voice pitch and one of the most extensively studied and socially meaningful acoustic parameters in the human voice (Aung & Puts, 2020), is highly stable across neutral utterances of different lengths produced by the same vocalizers. At least half and up to 80% of the variance in f_o measured from one utterance was explained by the f_o of any other utterance within speakers. In another study, Pisanski and colleagues show that these inter-individual differences in f_o also extent to emotional speech and nonverbal vocalisations such as screams, roars and cries produced by the same men and women (Pisanski et al., 2020). This suggests that individual differences in voice pitch (often measured by researchers to test for relationships between f_o and vocalizer traits such as body size, testosterone levels, attractiveness, or dominance, to name a few) are robust regardless of speech type or duration, within the same group of vocalizers. However, long and short speech recordings can differ in a number of other characteristics like formant patterns, prosody, articulation, or speed of speech (Leung et al., 2018). Longer utterances are likely to convey more information than shorter utterances

612 regarding not only the nonverbal parameters of the vocalizer's voice but also their prosody or
613 the vocalizer's way of speaking, which might be perceptually related to personality (Zellner
614 Keller, 2005). Indeed, voice pitch only partly explains the personality judgements of unseen
615 vocalizers (Stern et al., 2021).

616 Short utterances such as vowels and counting have traditionally been used in voice
617 perception studies due to their contextual neutrality, standardized nature, and thus high
618 experimental control. Despite the brevity and neutrality of these voice stimuli, studies have
619 generally shown that listeners can gauge various social traits from short utterances such as
620 vowels or a single word (Apicella & Feinberg, 2009; McAleer et al., 2014; Pisanski et al.,
621 2014). Nevertheless, there exists an important difference between very short utterances (like
622 single vowels) and longer utterances: the former are typically less ecologically valid. In real
623 life conversations, people are unlikely to base judgements solely on single vowels, as they
624 rarely hear them in separation from the rest of the statement. Therefore, another possibility is
625 that voice recordings consisting of one vowel, or a neutral one-syllable word, may be
626 relatively unfamiliar to listeners and, in turn, may thus be evaluated less positively.

627

628 *Limitations and future research recommendations*

629 Our research design does not allow to draw direct conclusions regarding the
630 'accuracy' of listeners' judgments, as no objective measures of speaker traits were obtained.
631 As such, the results cannot speak to the question of superiority of one methodological
632 stimulus choice over the other. However, a close look at the result patterns shows that short
633 utterances are in almost all cases closer to the scale mean (a rating of 4; for an exception see
634 attractiveness among male vocalizers). This suggests that listeners might have been less
635 certain about their ratings of shorter versus longer stimuli, resulting in judgments closer to
636 what they might have perceived as "average". The geographic homogeneity of our vocalizer

and listener samples, paired with a lack of data about various socially relevant participant variables including their gender identity, sexual orientation and socioeconomic status, represent limitations of our study that, if known, could broaden its generalizability. Regarding the study design, we observed a tendency for raters to make judgements closer to the scale mean for shorter than for longer utterances, possibly reflecting a level of uncertainty for short speech utterances. Providing raters with the option to omit the question, to mark "I don't know", or to rate the confidence of their ratings, could help to clarify the mechanisms driving the small observed differences in ratings across speech stimulus types. Moreover, although online raters were instructed to use high quality headphones, and headphone use was verified with hearing tests, we cannot be certain that the quality of their headphones was comparable to that of those used by lab participants. Our results suggest, however, that the online sample of raters produced qualitatively similar ratings to the lab sample that used professional headphones.

Our results may not necessarily generalize to judgments of other traits not tested here, for example, the Big Five personality traits (McCrae & Costa, 1989). Traits like extraversion or neuroticism may be especially closely perceptually interrelated with the prosody of speech (Feldstein & Sloan, 1984). Future studies could therefore extend the list of evaluated traits. We also limited our analyses to judgments of traits described by a single word, such as 'trustworthiness', wherein people may differ in the way they understand such terms. For masculinity and femininity, for example, some listeners may associate these terms with biological sex while others may judge masculinity and femininity independently of the perceived gender of the vocalizer. While using single undefined terms on rating scales is common practice in this research domain (Cartei et al., 2014; Feinberg et al., 2012; Hughes et al., 2014; McAleer et al., 2014), researchers may provide a working definition of these constructs or assess them through behavioral measures (e.g., probability of certain behaviors

or implicit association tasks). Moreover, femininity and masculinity were represented on a single rating scale, rather than representing two qualitatively independent constructs. We acknowledge that femininity and masculinity do not necessarily represent two ends of one continuum (Bem, 1974; Donnelly & Twenge, 2017) and that in future research may be treated separately.

It may also be of practical interest to test the 'accuracy' of social judgements. This is methodologically challenging because of the imperfection of measurement scales and the social approval factor that biases self-ratings toward more desired profiles. However, researchers may be able to commit to reliable proxies (behavioral or self-assessed) to test how accurately social traits can be inferred from voice samples depending on speech stimulus type. Another important avenue for future studies will be to test the effect of stimulus type in more ecologically valid or multi-modal conditions, i.e., where a speaker is seen (or smelled). Different modalities (e.g., voice, physical appearance or body odor) interact with one other to jointly affect impressions we form of people (for reviews and discussion see Feinberg, 2008; Groyecka et al., 2017; Krumpholz et al., 2021). Multi-modal studies can help to clarify whether speech complexity affects ratings of person dimensions to the same extent when voice stimuli are perceived alongside other modalities as when they are perceived in isolation.

Conclusion: Methodological, social and practical implications

We show that inter-individual differences in voice-based judgments are relatively stable across neutral speech stimuli, particularly relatively longer utterances, for the same vocalizers. Interestingly, while single-syllable voice stimuli elicit significantly lower ratings than do multi-syllable stimuli, a stimulus with 16 syllables of length appears to be comparable with a 150-syllable one. This suggests that in terms of playback experiments, using lengthy excerpts from passages like the Rainbow Passage, although well established and commonly used

(Cartei et al., 2014; Pisanski et al., 2016; Schild et al., 2020; Tigue et al., 2012), can unnecessarily extend the time of the experimental procedure while being similarly informative to listeners as one short sentence or counting from one to ten. Judgments based on five vowels are also quite comparable to full sentences and paragraphs, at least in their effects on listeners' social judgments of vocalizers.

Our results also corroborate earlier acoustic analyses and playback studies showing that a great deal of information about a person can be encoded in very short speech segments. Vowel sounds, for example, encode enough acoustic information, namely in voice fundamental and formant frequencies, for listeners to reliably judge static speaker traits such as body size, age, and biological sex, owing largely to anatomical or physiological constraints on vocal production (for reviews see Aung & Puts, 2020; Charlton et al., 2020; Pisanski & Bryant, 2019). Listeners also show high agreement on judgments of social or personality traits from a single word (McAleer et al., 2014). Our results suggest that listeners' judgments do not change drastically if they are presented with longer bouts of neutral speech from the same speaker, indicating that voice-based person impressions, which can meaningfully impact social and societal outcomes, may be formed early during first interactions and/or may rely heavily on low-level and relatively static acoustic features that do not vary greatly across speech utterances.

Finally, paired with emerging evidence from acoustic analyses that show remarkable intra-individual stability in people's voices across time and context (Fouquet et al., 2016; Levrero et al., 2018; Pisanski et al., 2020; 2021), voice-based human perception studies can inform practical technologies, including automated voice recognition devices in mobile phones and computers that increasingly rely on large amounts of human user response data to build high-performance predictive algorithms using artificial intelligence and deep learning (Deng, 2018).

Data Availability Statement

Dataset and measurement instruments are available at https://osf.io/cevpd/?view_only=e3ecf2aad04d4e01be5314e13c3446b1.

References

Albert, G., Arnocky, S., Puts, D. A., & Hodges-Simeon, C. R. (2021). Can listeners assess men's self-reported health from their voice? *Evolution and Human Behavior*, *42*(2), 91–103. https://doi.org/10.1016/j.evolhumbehav.2020.08.001

Anderson, R. C., Klofstad, C. A., Mayew, W. J., & Venkatachalam, M. (2014). Vocal fry may undermine the success of young women in the labor market. *PLoS ONE*, *9*(5), 1–8. https://doi.org/10.1371/journal.pone.0097506

Apicella, C. L., & Feinberg, D. R. (2009). Voice pitch alters mate-choice-relevant perception in hunter–gatherers. *Proceedings of the Royal Society B: Biological Sciences*, *276*(1659), 1077–1082.

Aung, T., & Puts, D. (2020). Voice pitch: a window into the communication of social power. *Current Opinion in Psychology*, *33*, 154–161.

Belin, P., Boehme, B., & McAleer, P. (2017). The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PLoS ONE*, *12*(10), 4–12. https://doi.org/10.1371/journal.pone.0185651

Bem, S. L. (1974). The measurement of psychological androgyny. *Journal of Consulting and Clinical Psychology*, *42*(2), 155.

Borkowska, B., & Pawlowski, B. (2011). Female voice frequency in the context of dominance and attractiveness perception. *Animal Behaviour*, *82*(1), 55–59. https://doi.org/10.1016/j.anbehav.2011.03.024

Bryant, G. A., & Haselton, M. G. (2009). Vocal cues of ovulation in human females. *Biology Letters*, *5*(1), 12–15.

Cartei, V., Bond, R., & Reby, D. (2014). What makes a voice masculine: Physiological and acoustical correlates of women's ratings of men's vocal masculinity. *Hormones and Behavior*, *66*(4), 569–576. https://doi.org/10.1016/j.yhbeh.2014.08.006

Charlton, B. D., Pisanski, K., Raine, J., & Reby, D. (2020). Coding of static information in terrestrial mammal vocal signals. In *Coding strategies in vertebrate acoustic communication* (pp. 115–136). Springer.

Deng, L. (2018). Artificial intelligence in the rising wave of deep learning: The historical path and future outlook [perspectives]. *IEEE Signal Processing Magazine*, *35*(1), 177–180.

Donnelly, K., & Twenge, J. M. (2017). Masculine and feminine traits on the Bem Sex-Role Inventory, 1993–2012: A cross-temporal meta-analysis. *Sex Roles*, *76*(9), 556–565.

Fairbanks, G. (1960). *Voice and articulation drillbook*. Addison-Wesley Educational Publishers.

Feinberg, D. R., Jones, B. C., Little, A. C., Burt, D. M., & Perrett, D. I. (2005). Manipulations of fundamental and formant frequencies influence the attractiveness of human male voices. *Animal Behaviour*, *69*(3), 561–568. https://doi.org/10.1016/j.anbehav.2004.06.012

Feinberg, D. R. (2008). Are human faces and voices ornaments signaling common underlying cues to mate value? *Evolutionary Anthropology: Issues, News, and Reviews: Issues, News, and Reviews*, *17*(2), 112–118.

Feinberg, D. R, DeBruine, L. M., Jones, B. C., Little, A. C., O'Connor, J. J. M., & Tigue, C. C. (2012). Women's self-perceived health and attractiveness predict their male vocal masculinity preferences in different directions across short-and long-term relationship contexts. *Behavioral Ecology and Sociobiology*, *66*(3), 413–418.

762 Feinberg, D. R, DeBruine, L. M., Jones, B. C., & Perrett, D. I. (2008). The role of femininity
763 and averageness of voice pitch in aesthetic judgments of women's voices. *Perception*,
764 *37*(4), 615–623.
765 Feldstein, S., & Sloan, B. (1984). Actual and stereotyped speech tempos of extraverts and
766 introverts. *Journal of Personality*, *52*(2), 188–204.
767 Ferdenzi, C., Patel, S., Mehu-Blantar, I., Khidasheli, M., Sander, D., & Delplanque, S. (2013).
768 Voice attractiveness: Influence of stimulus duration and type. *Behavior Research*
769 *Methods*, *45*(2), 405–413. https://doi.org/10.3758/s13428-012-0275-0
770 Fouquet, M., Pisanski, K., Mathevon, N., & Reby, D. (2016). Seven and up: individual
771 differences in male voice fundamental frequency emerge before puberty and remain
772 stable throughout adulthood. *Royal Society Open Science*, *3*(10), 160395.
773 Groyecka, A., Pisanski, K., Sorokowska, A., Havlícek, J., Karwowski, M., Puts, D., Craig
774 Roberts, S., & Sorokowski, P. (2017). Attractiveness is multimodal: Beauty is also in the
775 nose and ear of the beholder. *Frontiers in Psychology*, *8*(MAY).
776 https://doi.org/10.3389/fpsyg.2017.00778
777 Hughes, S. M., Mogilski, J. K., & Harrison, M. A. (2014). The perception and parameters of
778 intentional voice manipulation. *Journal of Nonverbal Behavior*, *38*(1), 107–127.
779 Jones, B. C., Feinberg, D. R., DeBruine, L. M., Little, A. C., & Vukovic, J. (2008).
780 Integrating cues of social interest and voice pitch in men's preferences for women's
781 voices. *Biology Letters*, *4*(2), 192–194. https://doi.org/10.1098/rsbl.2007.0626
782 Kamiloğlu, R. G., & Sauter, D. A. (2021). Voice Production and Perception. In *Oxford*
783 *Research Encyclopedia of Psychology*.
784 Klofstad, C. A., Anderson, R. C., & Peters, S. (2012). Sounds like a winner: Voice pitch
785 influences perception of leadership capacity in both men and women. *Proceedings of the*
786 *Royal Society B: Biological Sciences*, *279*(1738), 2698–2704.

https://doi.org/10.1098/rspb.2012.0311

Kordsmeyer, T. L., Hunt, J., Puts, D. A., Ostner, J., & Penke, L. (2018). The relative importance of intra-and intersexual selection on human male sexually dimorphic traits. *Evolution and Human Behavior*, *39*(4), 424–436.

Krahé, B., Uhlmann, A., & Herzberg, M. (2021). The voice gives it away: Male and female pitch as a cue for gender stereotyping. *Social Psychology*, *52*(2), 101–113. https://doi.org/10.1027/1864-9335/a000441

Kreiman, J. (1997). Listening to voices: theory and practice in voice perception research. *Talker variability in speech processing*, 85-108.

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons.

Krumpholz, C., Quigley, C., Little, A. C., Zäske, R., & Riebel, K. (2021). Multimodal signalling of attractiveness. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *43*(43).

Leung, Y., Oates, J., & Chan, S. P. (2018). Voice, articulation, and prosody contribute to listener perceptions of speaker gender: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research*, *61*(2), 266–297.

Levrero, F., Mathevon, N., Pisanski, K., Gustafsson, E., & Reby, D. (2018). The pitch of babies' cries predicts their voice pitch at age 5. *Biology Letters*, *14*(7), 20180065.

Mahrholz, G., Belin, P., & McAleer, P. (2018). Judgements of a speaker's personality are correlated across differing content and stimulus type. *PLoS ONE*, *13*(10), 1–22. https://doi.org/10.1371/journal.pone.0204991

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say "hello"? Personality impressions from brief novel voices. *PLoS ONE*, *9*(3), 1–9. https://doi.org/10.1371/journal.pone.0090779

McCrae, R. R., & Costa, P. T. (1989). The structure of interpersonal traits: Wiggins's circumplex and the five-factor model. *Journal of Personality and Social Psychology*, *56*(4), 586.

Mileva, M., Tompkinson, J., Watt, D., & Burton, A. M. (2020). The role of face and voice cues in predicting the outcome of student representative elections. *Personality and Social Psychology Bulletin*, *46*(4), 617–625. https://doi.org/10.1177/0146167219867965

Ohno, R., Masanori, K., & Tetsuro, M. (2014). Relationship between perception of cuteness in female voices and their durations. *International Conference on Speech and Computer*, *8773*. https://doi.org/10.1007/978-3-319-11581-8

Oleszkiewicz, A., Pisanski, K., Lachowicz-Tabaczek, K., & Sorokowska, A. (2017). Voice-based assessments of trustworthiness, competence, and warmth in blind and sighted adults. *Psychonomic Bulletin & Review*, *24*(3), 856–862.

Pisanski, K., Anikin, A., & Reby, D. (2021). Static and dynamic formant scaling conveys body size and aggression. *Royal Society Open Science, The Royal Society*. https://doi.org/10.1098/rsos.211496 . hal-03501112

Pisanski, K., & Bryant, G. A. (2019). The evolution of voice perception. *The oxford handbook of voice studies*, 269-300.

Pisanski, K., & Feinberg, D. R. (2019). Vocal attractiveness. *Oxford Handbooks. The Oxford Handbook of Voice Perception*, 606–626.

Pisanski, K., Fraccaro, P. J., Tigue, C. C., Connor, J. J. M. O., David, R., Pisanski, K., Fraccaro, P. J., Tigue, C. C., Connor, J. J. M. O., & Feinberg, D. R. (2014). Return to Oz: Voice pitch facilitates assessments of men's body size. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4), 1316.

Pisanski, K., Groyecka-Bernard, A., & Sorokowski, P. (2021). Human voice pitch measures are robust across a variety of speech recordings: methodological and theoretical

implications. *Biology Letters*, *17*, 20210356. https://doi.org/https://doi.org/10.1098/rsbl.2021.0356

Pisanski, K., Mishra, S., & Rendall, D. (2012). The evolved psychology of voice: Evaluating interrelationships in listeners' assessments of the size, masculinity, and attractiveness of unseen speakers. *Evolution and Human Behavior*, *33*(5), 509–519. https://doi.org/10.1016/j.evolhumbehav.2012.01.004

Pisanski, K., Nowak, J., & Sorokowski, P. (2016). Individual differences in cortisol stress response predict increases in voice pitch during exam stress. *Physiology and Behavior*, *163*, 234–238. https://doi.org/10.1016/j.physbeh.2016.05.018

Pisanski, K., Oleszkiewicz, A., Plachetka, J., Gmiterek, M., & Reby, D. (2018). Voice pitch modulation in human mate choice. *Proceedings of the Royal Society B: Biological Sciences*, *285*(1893). https://doi.org/10.1098/rspb.2018.1634

Pisanski, K., Raine, J., & Reby, D. (2020). Individual differences in human voice pitch are preserved from speech to screams, roars and pain cries. *Royal Society Open Science*, *7*(2), 191642.

Pisanski, K., & Rendall, D. (2011). The prioritization of voice fundamental frequency or formants in listeners' assessments of speaker size, masculinity, and attractiveness. *The Journal of the Acoustical Society of America*, *129*(4), 2201–2212. https://doi.org/10.1121/1.3552866

Puts, D. A., Hill, A. K., Bailey, D. H., Walker, R. S., Rendall, D., Wheatley, J. R., Welling, L. L. M., Dawood, K., Cárdenas, R., Burriss, R. P., Jablonski, N. G., Shriver, M. D., Weiss, D., Lameira, A. R., Apicella, C. L., Owren, M. J., Barelli, C., Glenn, M. E., & Ramos-Fernandez, G. (2016). Sexual selection on male vocal fundamental frequency in humans and other anthropoids. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1829), 0–7. https://doi.org/10.1098/rspb.2015.2830

Puts, D. Andrew, Gaulin, S. J. C., & Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and Human Behavior*, *27*(4), 283–296. https://doi.org/10.1016/j.evolhumbehav.2005.11.003

Rosenfield, K. A., Sorokowska, A., Sorokowski, P., & Puts, D. A. (2020). Sexual selection for low male voice pitch among Amazonian forager-horticulturists. *Evolution and Human Behavior*, *41*(1), 3–11.

Schild, C., Stern, J., & Zettler, I. (2020). Linking men's voice pitch to actual and perceived trustworthiness across domains. *Behavioral Ecology*, *31*(1), 164–175. https://doi.org/10.1093/beheco/arz173

Sell, A., Bryant, G. A., Cosmides, L., Tooby, J., Sznycer, D., Von Rueden, C., Krauss, A., & Gurven, M. (2010). Adaptations in humans for assessing physical strength from the voice. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1699), 3509–3518.

Sorokowski, P., Puts, D., Johnson, J., Żółkiewicz, O., Oleszkiewicz, A., Sorokowska, A., Kowal, M., Borkowska, B., & Pisanski, K. (2019). Voice of authority: Professionals lower their vocal frequencies when giving expert advice. *Journal of Nonverbal Behavior*, *43*(2), 257–269. https://doi.org/10.1007/s10919-019-00307-0

Stern, J., Schild, C., Jones, B. C., DeBruine, L. M., Hahn, A., Puts, D. A., Zettler, I., Kordsmeyer, T. L., Feinberg, D., Zamfir, D., Penke, L., & Arslan, R. C. (2021). Do voices carry valid information about a speaker's personality? *Journal of Research in Personality*, 104092. https://doi.org/10.1016/j.jrp.2021.104092

Tigue, C. C., Borak, D. J., O'Connor, J. J. M., Schandl, C., & Feinberg, D. R. (2012). Voice pitch influences voting behavior. *Evolution and Human Behavior*, *33*(3), 210–216. https://doi.org/10.1016/j.evolhumbehav.2011.09.004

Titze, I. R. (1989). Physiologic and acoustic differences between male and female voices. *Journal of the Acoustical Society of America*, *85*(4), 1699–1707.

https://doi.org/10.1121/1.397959

Vukovic, J., Feinberg, D., Debruine, L., Smith, F., & Jones, B. (2010). Women's voice pitch is negatively correlated with health risk factors. *Journal of Evolutionary Psychology*, *8*(3), 217–225. https://doi.org/10.1556/JEP.8.2010.3.2

Zellner Keller, B. (2005). Speech prosody, voice quality and personality. *Logopedics Phoniatrics Vocology*, *30*(2), 72–78. https://doi.org/10.1080/14015430500256543

Figure 1. Listeners' judgments generally increased with speech stimulus duration. *Differences in perceived vocalizer traits across speech stimulus types for listeners' judgments of (a) Attractiveness; (b) Dominance; (c) Likability; (d) Trustworthiness; (e) Femininity-masculinity; (f) Health. For exact differences see Tables S2, S4, S11, S13, S15, S17, S19, S21, S23, S25, S29 and S29 in SOM. Error bars depict 95% CI. Dashed lines depict nonsignificant differences based on pairwise comparisons following Bonferroni correction for multiple comparisons (see supplementary tables listed above). All remaining pairwise comparisons (those without dashed lines joining them) are statistically significant at $p<.05$.*

Figure 2. *Distributions and correlations (Pearson's r coefficient) between ratings of* ***attractiveness*** *based on six different stimulus types. Red – female vocalizers, green – male vocalizers.*

Figure 3. *Distributions and correlations (Pearson's r coefficient) between ratings of* ***dominance*** *based on six different stimulus types. Red – female vocalizers, green – male vocalizers.*

Figure 4. *Distributions and correlations (Pearson's r coefficient) between ratings of* ***likability*** *based on six different stimulus types. Red – female vocalizers, green – male vocalizers.*

Figure 5. *Distributions and correlations (Pearson's r coefficient) between ratings of* ***trustworthiness*** *based on six different stimulus types. Red – female vocalizers, green – male vocalizers.*

Figure 6. *Distributions and correlations (Pearson's r coefficient) between ratings of* ***femininity-masculinity*** *based on six different stimulus types. Higher scores denote higher masculinity (lower femininity). Red – female vocalizers, green – male vocalizers.*

Figure 7. *Distributions and correlations (Pearson's r coefficient) between ratings of* ***health*** *based on six different stimulus types. Red – female vocalizers, green – male vocalizers.*

Table 1. Estimated fixed and random effects of the model with perceived **attractiveness** as an outcome variable

FEMALE VOCALIZERS

Fixed effects				95% CI				
		Estimate	SE	Lower	Upper	Df	t	*p*
Intercept		3.8816	0.06044	3.7630	4.0002	313	64.142	< .001
Stimulus type	2 – 1	0.6995	0.03659	0.6278	0.7712	12515	19.118	< .001
	3 – 1	0.1418	0.03659	0.0701	0.2135	12515	3.875	< .001
	4 – 1	0.8296	0.03659	0.7579	0.9013	12515	22.675	< .001
	5 – 1	0.7620	0.03659	0.6903	0.8337	12515	20.827	< .001
	6 – 1	0.8949	0.03659	0.8232	0.9666	12515	24.459	< .001
Vocalizer's age		-0.02237	0.00340	-0.0290	-0.0157	100	-6.592	< .001
Listener's sex (0 – F, 1 – M)		-0.0129	0.08954	-0.1770	0.1744	432	-0.0144	0.988
Rater group (0 – online, 1 – lab)		-0.17260	0.09041	-0.3498	0.00460	432	-1.9090	0.057

Random effects	Variance	ICC
Listener's ID	0.808	0.355
Vocalizer's ID	0.170	0.105
Residuals	1.457	

$N_{observations}$ = 13055, $N_{listeners}$ = 435, $N_{vocalizers}$ = 101

MALE VOCALIZERS

Fixed effects				95% CI				
		Estimate	SE	Lower	Upper	Df	t	*p*
Intercept		3.52695	0.06037	3.4086	3.64527	353	58.43	< .001
Stimulus type	2 – 1	0.41500	0.03543	0.3456	0.48446	12260	11.71	< .001
	3 – 1	0.19720	0.03543	0.1278	0.26664	12260	5.57	< .001
	4 – 1	0.68385	0.03543	0.6144	0.75329	12260	19.30	< .001
	5 – 1	0.73959	0.03543	0.6701	0.80903	12260	20.87	< .001
	6 – 1	0.80375	0.03543	0.7343	0.87320	12260	22.69	< .001
Vocalizer's age		-0.00860	0.00359	-0.0156	-0.00157	104	-2.40	0.018
Listener's sex (0 – F, 1 – M)		-0.50296	0.09230	-0.6639	-0.30206	432	-5.45	< .001
Rater group (0 – online, 1 – lab)		-0.21543	0.09311	-0.3979	-0.03295	432	-2.31	0.021

Random effects	Variance	ICC
Listener's ID	0.857	0.390
Vocalizer's ID	0.165	0.110
Residuals	1.338	

$N_{observations}$ = 12805 $N_{listeners}$ = 435, $N_{vocalizers}$ = 106

Note. 1 = single vowels, 2 = series of vowels, 3 = single word, 4 = counting, 5 = greeting, 6 = recited paragraph (Rainbow Passage)

Table 2. Estimated fixed and random effects of the model with perceived **dominance** as an outcome variable

FEMALE VOCALIZERS

Fixed effects				95% CI				
		Estimate	SE	Lower	Upper	Df	t	*p*
Intercept		3.60406	0.06661	3.4735	3.73463	257.8	54.101	< .001
Stimulus type	2 – 1	0.91370	0.04053	0.8343	0.99315	11932.0	22.542	< .001
	3 – 1	0.36286	0.04053	0.2834	0.44230	11932.0	8.952	< .001
	4 – 1	0.76045	0.04053	0.6810	0.83989	11932.0	18.761	< .001
	5 – 1	0.76623	0.04053	0.6868	0.84568	11932.0	18.904	< .001
	6 – 1	0.65394	0.04053	0.5745	0.73339	11932.0	16.134	< .001
Vocalizer's age		-0.00255	0.00400	-0.0104	0.00528	98.4	-0.638	0.525
Listener's sex (0 – F, 1 – M)		-0.09743	0.09306	-0.2688	0.09732	411.6	-0.918	0.359
Rater group (0 – online, 1 – lab)		-0.08572	0.09339	-0.2688	0.09732	411.6	-0.918	0.359

Random effects	Variance	ICC
Listener's ID	0.798	0.319
Vocalizer's ID	0.239	0.123
Residuals	1.706	

$N_{observations}$ = 12455, $N_{listeners}$ = 415, $N_{vocalizers}$ = 101

MALE VOCALIZERS

Fixed effects				95% CI				
		Estimate	SE	Lower	Upper	Df	t	*p*
Intercept		3.68345	0.06855	3.54908	3.8178	241	53.730	< .001
Stimulus type	2 – 1	0.59922	0.04153	0.51783	0.6806	11755	14.429	< .001
	3 – 1	0.38194	0.04153	0.30054	0.4633	11755	9.197	< .001
	4 – 1	0.70030	0.04153	0.61890	0.7817	11755	16.863	< .001
	5 – 1	0.68955	0.04153	0.60816	0.7709	11755	16.604	< .001
	6 – 1	0.47569	0.04153	0.39429	0.5571	11755	11.454	< .001
Vocalizer's age		0.00154	0.00471	-0.00769	0.0108	104	0.327	0.745
Listener's sex (0 – F, 1 – M)		-0.03377	0.09083	-0.21180	0.1443	412	-0.372	0.710
Rater group (0 – online, 1 – lab)		0.11532	0.09112	-0.06328	0.2939	412	1.266	0.206

Random effects	Variance	ICC
Listener's ID	0.755	0.300
Vocalizer's ID	0.289	0.141
Residuals	1.763	

$N_{observations}$ = 12283 $N_{listeners}$ = 415, $N_{vocalizers}$ = 106

Note. 1 = single vowels, 2 = series of vowels, 3 = single word, 4 = counting, 5 = greeting, 6 = recited paragraph (Rainbow Passage)

924

Table 3. Estimated fixed and random effects of the model with perceived **likability** as an outcome variable

FEMALE VOCALIZERS

Fixed effects				95% CI				
		Estimate	SE	Lower	Upper	Df	t	*p*
Intercept		4.1529	0.05688	4.0414	4.26435	323.8	73.01	< .001
Stimulus type	2 – 1	0.4423	0.03652	0.3707	0.51391	12363.7	12.11	< .001
	3 – 1	0.0372	0.03652	-0.0344	0.10880	12363.7	1.02	0.308
	4 – 1	0.5898	0.03652	0.5182	0.66135	12363.7	16.15	< .001
	5 – 1	0.7944	0.03652	0.7228	0.86601	12363.7	21.75	< .001
	6 – 1	0.7181	0.03652	0.6466	0.78973	12363.7	19.66	< .001
Vocalizer's age		-0.0116	0.00313	-0.0177	-0.00546	99.0	-3.71	< .001
Listener's sex (0 – F, 1 – M)		-0.1563	0.08708	-0.3270	0.01437	426.7	-1.795	0.073
Rater group (0 – online, 1 – lab)		0.0414	0.08764	-0.1304	0.21320	426.7	0.473	0.637

Random effects	Variance	ICC
Listener's ID	0.737	0.3395
Vocalizer's ID	0.142	0.0901
Residuals	1.434	

$N_{observations}$ = 12900, $N_{listeners}$ = 430, $N_{vocalizers}$ = 101

MALE VOCALIZERS

Fixed effects				95% CI				
		Estimate	SE	Lower	Upper	Df	t	*p*
Intercept		3.89720	0.05298	3.79337	4.00104	412	73.563	< .001
Stimulus type	2 – 1	0.30043	0.03547	0.23091	0.36994	12158	8.470	< .001
	3 – 1	0.17241	0.03547	0.10290	0.24193	12158	4.861	< .001
	4 – 1	0.55598	0.03547	0.48646	0.62549	12158	15.676	< .001
	5 – 1	0.76996	0.03547	0.70044	0.83947	12158	21.709	< .001
	6 – 1	0.74020	0.03547	0.67068	0.80971	12158	20.870	< .001
Vocalizer's age		0.00120	0.00283	-0.00434	0.00674	105	0.424	0.673
Listener's sex (0 – F, 1 – M)		-0.26678	0.08859	-0.44041	-0.09315	426	-3.011	0.003
Rater group (0 – online, 1 – lab)		-0.01412	0.08923	-0.18901	-0.16077	427	-0.158	0.874

Random effects	Variance	ICC
Listener's ID	0.7676	0.3657
Vocalizer's ID	0.0966	0.0676
Residuals	1.3315	

$N_{observations}$ = 12702 $N_{listeners}$ = 430, $N_{vocalizers}$ = 106

Note. 1 = single vowels, 2 = series of vowels, 3 = single word, 4 = counting, 5 = greeting, 6 = recited paragraph (Rainbow Passage)

Table 4. Estimated fixed and random effects of the model with perceived **trustworthiness** as an outcome variable

FEMALE VOCALIZERS

Fixed effects				95% CI				
		Estimate	SE	Lower	Upper	Df	t	*p*
Intercept		3.7917	0.05966	3.67465	3.90871	399.0	63.556	< .001
Stimulus type	2 – 1	0.8481	0.04185	0.76609	0.93015	11466.8	20.265	< .001
	3 – 1	0.1383	0.04185	0.05632	0.22037	11466.8	3.306	< .001
	4 – 1	0.9935	0.04185	0.91145	1.07551	11466.8	23.738	< .001
	5 – 1	1.0722	0.04185	0.99015	1.15421	11466.8	25.618	< .001
	6 – 1	0.9754	0.04185	0.89341	1.05747	11466.8	23.307	< .001
Vocalizer's age		-3.32e−4	0.00283	-0.00588	0.00521	98.7	-0.118	0.906
Listener's sex (0 – F, 1 – M)		-0.07119	0.10060	-0.26836	0.12598	396.9	-0.7077	0.480
Rater group (0 – online, 1 – lab)		0.00690	0.10256	-0.19412	0.20792	396.3	0.0673	0.946

Random effects	Variance	ICC
Listener's ID	0.924	0.3466
Vocalizer's ID	0.110	0.0592
Residuals	1.747	

$N_{observations}$ = 11970, $N_{listeners}$ = 399, $N_{vocalizers}$ = 101

MALE VOCALIZERS

Fixed effects				95% CI				
		Estimate	SE	Lower	Upper	Df	t	*p*
Intercept		3.67739	0.05999	3.55974	3.79490	397	61.299	< .001
Stimulus type	2 – 1	0.67541	0.04183	0.59342	0.75740	11156	16.145	< .001
	3 – 1	0.17490	0.04183	0.09291	0.25689	11156	4.181	< .001
	4 – 1	0.94444	0.04183	0.86245	1.02644	11156	22.577	< .001
	5 – 1	0.95936	0.04183	0.87737	1.04135	11156	22.933	< .001
	6 – 1	0.99691	0.04183	0.91492	1.07890	11156	23.831	< .001
Vocalizer's age		0.00208	0.00322	-0.00424	0.00840	105	0.646	0.520
Listener's sex (0 – F, 1 – M)		-0.07742	0.04970	-0.17482	0.01999	396	-1.558	0.120
Rater group (0 – online, 1 – lab)		0.04457	0.10159	-0.15455	0.24369	396	0.439	0.661

Random effects	Variance	ICC
Listener's ID	0.907	0.3482
Vocalizer's ID	0.124	0.0679
Residuals	1.701	

$N_{observations}$ = 11664 $N_{listeners}$ = 399, $N_{vocalizers}$ = 106

Note. 1 = single vowels, 2 = series of vowels, 3 = single word, 4 = counting, 5 = greeting, 6 = recited paragraph (Rainbow Passage)

Table 5. Estimated fixed and random effects of the model with perceived **femininity-masculinity** as an outcome variable

FEMALE VOCALIZERS

Fixed effects				95% CI				
		Estimate	SE	Lower	Upper	Df	t	*p*
Intercept		2.3539	0.05384	2.24838	2.4594	459.3	43.72	< .001
Stimulus type	2 – 1	-0.2972	0.02931	-0.35461	-0.2397	11705.0	-10.14	< .001
	3 – 1	-0.1036	0.02931	-0.16100	-0.0461	11705.0	-3.53	< .001
	4 – 1	-0.4181	0.02931	-0.47550	-0.3606	11705.0	-14.27	< .001
	5 – 1	-0.3915	0.02931	-0.44896	-0.3341	11705.0	-13.36	< .001
	6 – 1	-0.4240	0.02931	-0.48139	-0.3665	11705.0	-14.47	< .001
Vocalizer's age		0.0108	0.00217	0.00653	0.0151	97.6	4.96	< .001
Listener's sex (0 – F, 1 – M)		-0.1601	0.09488	-0.34608	0.0258	403.8	-1.69	0.090
Rater group (0 – online, 1 – lab)		0.1755	0.09547	-0.01162	0.3626	403.9	1.84	0.067

Random effects	Variance	ICC
Listener's ID	0.8856	0.5028
Vocalizer's ID	0.0664	0.0705
Residuals	0.8757	

$N_{observations}$ = 12220, $N_{listeners}$ = 407, $N_{vocalizers}$ = 101

MALE VOCALIZERS

Fixed effects				95% CI				
		Estimate	SE	Lower	Upper	Df	t	*p*
Intercept		5.63938	0.05638	5.5289	5.7499	279	100.03	< .001
Stimulus type	2 – 1	0.20407	0.02837	0.1485	0.2597	11459	7.19	< .001
	3 – 1	0.10107	0.02838	0.0455	0.1567	11460	3.56	< .001
	4 – 1	0.30717	0.02837	0.2516	0.3628	11459	10.83	< .001
	5 – 1	0.29966	0.02837	0.2441	0.3553	11459	10.56	< .001
	6 – 1	0.22459	0.02837	0.1690	0.2802	11459	7.92	< .001
Vocalizer's age		0.00719	0.00369	-5.85e−5	0.0144	104	1.94	0.055
Listener's sex (0 – F, 1 – M)		-0.10305	0.07717	-0.2543	0.0482	404	-1.34	0.182
Rater group (0 – online, 1 – lab)		-0.08457	0.07768	-0.2368	0.0677	405	-1.09	0.277

Random effects	Variance	ICC
Listener's ID	0.574	0.417
Vocalizer's ID	0.180	0.183
Residuals	0.802	

$N_{observations}$ = 11977 $N_{listeners}$ = 407, $N_{vocalizers}$ = 106

Note. Higher values indicate higher masculinity. 1 = single vowels, 2 = series of vowels, 3 = single word, 4 = counting, 5 = greeting, 6 = recited paragraph (Rainbow Passage)

Table 6. Estimated fixed and random effects of the model with perceived **health** as an outcome variable

FEMALE VOCALIZERS

Fixed effects				95% CI				
		Estimate	SE	Lower	Upper	Df	t	*p*
Intercept		5.0332	0.05752	4.9205	5.14605	312.5	87.357	< .001
Stimulus type	2 – 1	0.7026	0.03937	0.6254	0.77976	12691.4	17.845	< .001
	3 – 1	0.2096	0.03937	0.1325	0.28679	12691.4	5.324	< .001
	4 – 1	0.7203	0.03937	0.6431	0.79745	12691.4	18.295	< .001
	5 – 1	0.8305	0.03937	0.7533	0.90765	12691.4	21.094	< .001
	6 – 1	0.6582	0.03937	0.5810	0.73532	12691.4	16.717	< .001
Vocalizer's age		-0.0155	0.00321	-0.0218	-0.00924	98.6	-4.842	< .001
Listener's sex (0 – F, 1 – M)		-0.1010	0.08721	-0.2720	0.06989	437.4	-1.16	0.247
Rater group (0 – online, 1 – lab)		-0.1667	0.08885	-0.3408	0.00743	437.1	-1.88	0.061

Random effects	Variance	ICC
Listener's ID	0.746	0.3036
Vocalizer's ID	0.149	0.0802
Residuals	1.712	

$N_{observations}$ = 13240, $N_{listeners}$ = 441, $N_{vocalizers}$ = 101

MALE VOCALIZERS

Fixed effects				95% CI				
		Estimate	SE	Lower	Upper	Df	t	*p*
Intercept		4.80384	0.06304	4.6803	4.9274	325	76.209	< .001
Stimulus type	2 – 1	0.41164	0.03963	0.3340	0.4893	12355	10.387	< .001
	3 – 1	0.34336	0.03963	0.2657	0.4210	12355	8.664	< .001
	4 – 1	0.67963	0.03963	0.6020	0.7573	12355	17.149	< .001
	5 – 1	0.74280	0.03963	0.6651	0.8205	12355	18.743	< .001
	6 – 1	0.58674	0.03963	0.5091	0.6644	12355	14.805	< .001
Vocalizer's age		-0.00788	0.00388	-0.0155	-2.87e−4	103	-2.034	0.045
Listener's sex (0 – F, 1 – M)		-0.05583	0.09500	-0.2420	0.1304	438	-0.1588	0.557
Rater group (0 – online, 1 – lab)		-0.21749	0.09678	-0.4072	-0.0278	438	-2.247	0.025

Random effects	Variance	ICC
Listener's ID	0.895	0.347
Vocalizer's ID	0.191	0.102
Residuals	1.686	

$N_{observations}$ = 12908 $N_{listeners}$ = 441, $N_{vocalizers}$ = 106

Note. 1 = single vowels, 2 = series of vowels, 3 = single word, 4 = counting, 5 = greeting, 6 = recited paragraph (Rainbow Passage)

926