Relationship between phoneme-level spectral acoustics and speech intelligibility in healthy speech: a systematic review

Timothy Pommée, Mathieu Balaguer, Julien Pinquier, Julie Mauclair, Virginie Woisard, Renée Speyer

Relationship between phoneme-level spectral acoustics and speech intelligibility in healthy speech: A systematic review

Timothy POMMÉE[a*], Mathieu BALAGUER[a,b], Julien PINQUIER[a], Julie MAUCLAIR[a], Virginie WOISARD[b,c,d], Renée SPEYER[e]

[a]Institut de Recherche en Informatique de Toulouse, CNRS, Université de Toulouse - Paul Sabatier, 118 Route de Narbonne, F-31062 Toulouse CEDEX 9, France; [b]Centre Hospitalier Universitaire Larrey, 24 Chemin de Pouvourville, 31400 Toulouse, France; [c]Oncopole, 1 Avenue Irène Joliot-Curie, 31100 Toulouse, France; [d]Unité de Recherche Interdisciplinaire Octogone Lordat, Maison de la Recherche, Université de Toulouse - Jean-Jaurès, 5, allée Antonio Machado, 31058 Toulouse CEDEX 9, France; [e]Faculty of Educational Sciences, University of Oslo, P.O Box 1161, Blindern 0318 Oslo, Norway

Correspondence to:
*Timothy Pommée
timothy.pommee@irit.fr
+33 6 03 03 51 48
ORCID : 0000-0001-7846-7282
IRIT Institut de Recherche en Informatique de Toulouse
118 Route de Narbonne
31062 TOULOUSE CEDEX 9

Biographical notes

Timothy Pommée received a master's degree in Speech and Language Pathology, specialized in Voice Therapy, at the University of Liège (Belgium). He is since 2018 a PhD student in computer sciences and telecommunications at the Institut de Recherche en Informatique de Toulouse (Université Toulouse III Paul Sabatier) in Toulouse (France). His research interests focus on the clinical relevance of speech intelligibility measures.

Mathieu Balaguer graduated as a Speech-Language Pathologist in 2007 at the Université Toulouse III Paul Sabatier in Toulouse (France). He has then worked as a private practitioner for two years, after what he worked as an employed SLP in hospital settings until 2019. He then received a master's degree in Clinical Epidemiology in 2018. He is since 2018 a PhD student in computer sciences and telecommunications at the Institut de Recherche en Informatique de Toulouse (Université Toulouse III

Paul Sabatier) in Toulouse (France). His research interests focus on the automatic assessment of the functional impact of speech disorders on daily communication acts in patients treated for cancer of the oral cavity or oropharynx. He is also involved in education and training of SLP students as lecturer and internships coordinator in the Faculty of Medicine Toulouse-Rangueil since 2010.

Julien Pinquier received a PhD (computer science specialty) in 2004, related to audio indexing and structuring by search of primary components: speech, music and keysounds. He received the HDR diploma of the University of Toulouse in 2014: this work was based on audio segmentation (speech, music and environmental sounds) and audiovisual segmentation.

Since 2005, he is an assistant professor at the Université Toulouse III Paul Sabatier where he works in the Institut de Recherche en Informatique de Toulouse. His objectives relate to the combination of the audio and the video, the multimedia indexing for automatic structuring of audiovisual documents. He is the author of more than 120 scientific publications. He is the team leader of the SAMoVA team of IRIT. He is now focusing on speech intelligibility measurements.

Julie Mauclair received a PhD at the Laboratoire d'Informatique de l'Université du Mans in 2006, entitled "Confidence measures in speech processing and applications". From 2007 to 2009, she was a postdoc in Ireland (UCD) working within the CNGL project addressing the adaptation of digital content to culture, locale and linguistic environment at high volume, speed and quality. Since 2009 she is an assistant professor, first at the University Paris Descartes within the LIPADE laboratory, then in the SAMoVA team at the Institut de Recherche en Informatique de Toulouse. Her research interests focus on characterizing the speech and voice of people presenting with various disorders, using automatic technologies.

Virginie Woisard is a phoniatrician and Associate Professor in ENT and Phoniatrics. She is working in the Voice and Swallowing Unit in the ENT department of the Rangueil-Larrey University Hospital in Toulouse, a unit she founded in 1992. She created the Oncorehabilitation Unit at the Cancer Institute of Toulouse in 2014 and is also the head of the Rehabilitaion center for laryngectomized patients at the University Hospital of Toulouse.

Author and co-author of numerous publications regarding the assessment and rehabilitation of oropharyngeal dysphagia and speech disorders, she actively participates in the promotion of research on this topic. Involved in several scientific societies (ENT, Phoniatrics, laryngology and dysphagia), she is also a devoted teacher and leader of the University Logopedic Training Center of Toulouse. Her research, as a member of Octogone-Lordat, Jean Jaurès University Toulouse II, mainly focuses on the fields of speech disorders. She is the General secretary of the French Phoniatricians Society since 2015.

Renée Speyer was appointed as Professor at the University of Oslo (UiO, Norway) in 2017 in recognition of her expertise in the illness trajectory of dysphagia (swallowing problems). After graduating as a speech and language pathologist in the Netherlands, she earned master's degrees in Speech and Language Pathology, Epidemiology and Health Professions Education and completed a

PhD in 2004 at Utrecht University. Dr Speyer's career included working as a speech pathologist in different health care settings and in academia. Over the last two decades Dr Speyer has developed an international research track record in the field of oropharyngeal dysphagia. She is currently undertaking research projects both nationally and internationally involving clinimetrics, instrument development, and developing evidenced based interventions in allied health. As an epidemiologist, she has a strong interest in evaluating the validity and reliability of assessments and testing the efficacy of interventions. Renée Speyer is a frequently invited speaker at international conferences and published over 90 scientific internationally peer-reviewed articles including many systematic reviews using PRISMA methodology.

Authors' CRediT role(s)

Timothy Pommée: Conceptualization; Formal analysis; Investigation; Writing - original draft; Visualization. ***Mathieu Balaguer***: Conceptualization; Formal analysis; Investigation; Writing - review & editing. **Julien Pinquier**: Conceptualization; Validation; Writing - review & editing; Supervision. ***Julie Mauclair***: Conceptualization; Validation; Writing - review & editing; Funding acquisition. ***Virginie Woisard***: Conceptualization; Validation; Writing - review & editing; Supervision. ***Renée Speyer***: Methodology; Investigation; Resources

Abstract

Purpose: To review papers investigating the link between spectral acoustic measures in healthy talkers and perceived speech intelligibility.

Study selection: This systematic review was carried out according to the PRISMA guidelines. Two independent raters selected articles from the Embase and PubMed databases. Only original articles written in English, addressing both notions of intelligibility and speech-related spectral acoustics in natural speech of at least five adult healthy speakers were retained. Papers with a methodological quality lower than 50% as rated by the QualSyst tool and research reports of level IV according to the NHMRC hierarchy as well as expert opinions were excluded from further analyses.

Data extraction: Study population characteristics, speech sample(s) used for the acoustic measure(s), acoustic parameter(s), perceptual intelligibility measure(s), main conclusion regarding the link between acoustics and intelligibility, as well as descriptive data if available.

Results: Twenty-two studies were retained. Eighteen papers investigated vowel acoustics, one studied glides and eight articles investigated consonants, mostly sibilants. Various spectral measures and intelligibility estimates were used. The following measures were shown to be linked to sub-lexical perceived speech intelligibility ratings: for vowels, steady-state F1 and F2 measures, the F1 range, the [i]-[u] F2 difference, F0-F1 and F1-F2 differences in [ɛ-æ] and [ɪ-ɛ], the vowel space area, the mean amount of formant movement, the vector length and the spectral change measure; for consonants, the

centroid energy and the spectral peak in the [s]-sound, as well as the steady-state F1 offset frequency in vowels preceding [t] and [d].
Limitations: There might be studies from other sources that address the topic but that are not referenced in the two considered databases. Regarding the acoustic measures considered in this review, time-domain measures were not taken into account in order to limit the noise in the initial database search. Only frequency-domain measures were included.
Conclusion: Speech is highly variable even in healthy adult speakers. A better understanding of the imprecisions in healthy spontaneous speech will provide a more realistic baseline for the investigation of disordered speech. To date, no acoustic measure is able to predict speech intelligibility to a large extent. There is still extensive research to be carried out to identify relevant acoustic combinations that could account for perceived speech variations (e.g. vowel and consonant reductions) and to gather normative data from a large number of healthy speakers. To that end, speech-related terms (e.g. intelligibility, comprehensibility, severity) need to be clearly defined and methodologies described in sufficient details to allow for replication, cross-comparisons/meta-analyses and pooling of data.

PROSPERO registration number: CRD42019129597
Keywords: *Acoustics, Speech, Intelligibility, PRISMA, Systematic Review*

Introduction

Speech is an essential function in everyday life that requires complex interactions between the generation of air pressure, the vibration of the vocal folds, and the modulation by the resonating cavities of the phonatory tract (Fitch, 2000; Honda, 2008). Not being correctly understood, for example in dysarthria (Stipancic, Tjaden, & Wilding, 2016), can limit educational, occupational and social participation, hence reducing the quality of life (Hustad, 2008). Therefore, when speech production is impaired, assessing and quantifying the deficit is essential to determine the overall degree of impairment as well as to provide a follow-up measure (Raymond D. Kent, 1992; Miller, 2013; Stipancic et al., 2016; Sussman & Tjaden, 2012).

However, speech is not only variable in a pathological context (Benzeguiba et al., 2007; Miller, 2013). Some ***healthy talkers*** are indeed more intelligible than others, which was shown to be linked to the speaker's acoustic-phonetic production rather than to the listener's

perception (Bond & Moore, 1994; Cox, Alexander, & Gilmore, 1987; Hazan & Markham, 2004; Hood & Poole, 1980). The analysis of speech 'errors' often leads to the well-documented speed-accuracy trade-off (Guenther, 1995; Meunier, 2007; Tremblay, Sato, & Deschamps, 2017). To understand this trade-off, Lindblom proposed the 'hyper/hypo-speech' ('H&H') theory (Bond & Moore, 1994; Lindblom, 1990), which posits that high intelligibility can be reached through different acoustic-phonetic strategies (Cox et al., 1987; Guenther, 1995; Hazan & Markham, 2004; Lavoie, 2002). It is therefore of paramount importance to get a good understanding of these strategies, to distinguish which variations can be attributed to the constraints of spontaneous speech in a natural communication context, and which deviations indicate disordered speech, to allow for a more accurate assessment.

The variability in healthy speech is addressed under various angles and referred to through different concepts, such as speech clarity, precision, comprehensibility and, as already mentioned, intelligibility. In this study, we will use the psycholinguistic model of Levelt (Levelt, 1995; Levelt, Roelofs, & Meyer, 1999) as the reference model of speech production. In this model (1995, 1999), the constituent segments (phonemes) as well as the metrical frame (syllable number and lexical stress position) are retrieved for each word. The phonemes are then associated with the frame, and the resulting phonological syllable is confronted with the 'syllabary' (Schiller, 2006). The syllabary contains the articulatory gesture plans of frequent phonological syllables; for infrequent syllables, sub-syllabic units must be retrieved (Aichert & Ziegler, 2004; Levelt, 1995; Levelt et al., 1999). Other speech production models, such as Guenther's DIVA-model (Bohland & Guenther, 2006; Guenther, 1995; Guenther, Ghosh, & Tourville, 2006), also consider phonemes and syllables as the basic units. Level's model leads us to the term '***intelligibility***'. While it is used in various contexts determining the colour of its definition, in this work and in accordance with Levelt's model, intelligibility is defined as the accuracy with which the acoustic signal is decoded by

the listener at the segmental (phoneme and syllable) levels (Ghio et al., 2018; Hustad, 2008; Lalain et al., 2020; Yorkston, Strand, & Kennedy, 1996). Both the chosen speech production model and definition of intelligibility thus led us to focus on ***phoneme-level*** measures in this review, keeping in mind that syllable-level measures also contribute to speech intelligibility in running speech.

As per the above definition, the most appropriate way to perceptually assess intelligibility would be the minimization of signal-independent (lexical, syntactic and semantic) cues (Ghio et al., 2018; Lindblom, 1990), in order to focus on the speech production processes of sub-lexical units. This can be done using vowel, consonant, syllable or word identification scores, or pseudowords (Ghio et al., 2018; Lalain et al., 2020; Tremblay et al., 2017). The Frenchay Dysarthria Assessment (Enderby & Palmer, 2008), for example, makes use of orthographic transcriptions to compute a percentage of correctly identified items. Speech intelligibility can also be assessed with an identification task using minimal pairs, as in the Diagnostic Rhyme Test – DRT (Voiers, 1983) or in the Single Word Intelligibility Test (Ray D. Kent, Weismer, Kent, & Rosenbek, 1989). Other tasks exist, such as overall ratings of speech intelligibility on visual analog scales, using sentences. Although not in line with the above definition, they are a substantial part of the measures commonly used in clinical practice under the umbrella term "intelligibility". They must therefore be considered but differentiated from measures that fit the more specific definition.

While these various perceptual tasks are very informative, they rely on subjective ratings, biased, among other factors, by the familiarity of the rater with the subject's speech and with the test stimuli (Middag, Martens, Van Nuffelen, & De Bodt, 2009); they are also usually time-consuming (Fontan et al., 2014). While perceptual measures still remain the gold standard in clinical settings (Kent & Kim, 2003; Stipancic et al., 2016; Van Nuffelen et al., 2009), the ***acoustic analysis*** of speech provides a more objective assessment method that

helps alleviate the various inherent biases of perceptual methods. Therefore, these objective measures are increasingly gaining interest for speech assessment purposes (Carmichael & Green, 2004; T. Lee et al., 2016; Maniwa, Jongman, & Wade, 2009). An important question that arises is whether the imprecisions in healthy speech can be captured by acoustic-phonetic measures. Several studies, such as in healthy ageing (Hazan, 2017; Kuruvilla-Dugdale, Dietrich, McKinley, & Deroche, 2020), indicate that a large part of the variability in healthy speech is indeed 'traceable to specific acoustic-phonetic characteristics of the talker' (Bradlow, Torretta, & Pisoni, 1996; Metz, Schiavetti, Samar, & Sitler, 1990). The field of study of acoustic measurements in speech is vast, and we have chosen to conduct this systematic review on one of its aspects, the frequency-domain measures. Indeed, it has been shown that spectral cues have a greater contribution than temporal features in stimuli identification by normal-hearing listeners (Souza, Wright, Blackburn, Tatman, & Gallun, 2015). Furthermore, hearing-impaired listeners have difficulties in the identification of consonants (Dubno, Dirks, & Schaefer, 1989; Preminger & Wiley, 1985) and vowels (Li, Ning, Brashears, & Rife, 2008; Molis & Leek, 2011) due to a loss in the frequency content, which highlights the importance of spectral cues in phoneme-level intelligibility.

We have introduced the interest of focusing on the behaviour of ***segmental spectral measures in healthy speech*** before using these objective intelligibility measures in specific speech-disordered populations. Therefore, the objective of this study is to systematically review papers investigating the link between spectral acoustic measures and perceived speech intelligibility in 'natural' (that is, not consciously altered) speech in healthy talkers, as rated by healthy listeners without hearing loss or cognitive impairment and considering a 'normal' sound wave transfer (Fontan, 2012).

Methods

Protocol and registration

This systematic review has been carried out according to the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement and checklist. These recommendations help the researcher to carry out a rigorous and transparent review of the scientific literature (Liberati et al., 2009; Moher, Liberati, Tetzlaff, & Altman, 2009), by providing procedures on how to search for, how to select and how to analyse the retrieved papers from scientific databases.

This study was registered on *PROSPERO* under the registration number CRD42019129597.

Eligibility criteria

In order to be included in this review, articles had to:

- address both notions of intelligibility[1] and speech-related spectral acoustics (excluding papers addressing environmental acoustics);
- investigate natural speech of healthy adult speakers over 18 years of age (thus also excluding papers studying modified or vocoded speech when no data about the unprocessed speech was also provided);
- use segmental acoustics (not only global acoustic measures, such as the long-term average spectrum over a whole sentence);
- be written in English;
- be original articles (oral presentations, case studies, author letters, conference proceedings, and reviews were excluded);
- include at least six healthy speakers.

[1] Studies using perceptual assessment methods that fitted the umbrella-term 'intelligibility' rather than the more specific definition focusing on low-level segmental units were not excluded *a priori* but differentiated in the Discussion.

Other exclusion criteria were:

- the exclusive investigation of voice/phonation (dysphonia, voice quality measures), and not speech per se;
- addressing tonal languages, for which intelligibility analyses additionally rely on lexical tone and prosody (Ding, McLoughlin, & Tan, 2003; Yiu, van Hasselt, Williams, & Woo, 1994);
- the exclusive use of durational measures (such as vowel length or speaking rate);
- the study of the perception of speech by hearing impaired listeners;
- the application of automatic speech processing techniques, such as deep neural networks.

All eligibility criteria had to be met in order for the papers to be included in this review.

Information sources and search strategy

The literature search was carried out on the fourth of December 2018 in two biomedical databases: Embase and PubMed. No date-related exclusion criterion was used, as some relevant sources known to the authors date back to the mid-1950s. All references of the included papers were checked for additional relevant articles. The search terms and syntax are listed in Table 1.

Table 1. Search strategy for the two databases

Database	Search Terms (subject headings and free text words)	Number of Records
Embase:	((speech intelligibility/) OR (Intelligibil*.ab. OR Intelligibil*.ti. OR comprehensibil*.ab. OR comprehensibil*.ti. OR understandabil*.ab. OR understandabil*.ti.)) AND (acoustics/ OR speech analysis/ OR acoustic analysis/ OR sound analysis/ OR phonetics/ OR signal processing/ OR fourier analysis/ OR sound detection/ OR sound/ OR frequency/ OR frequency analysis/ OR pitch/ OR noise/ OR signal noise ratio/)	3326

PubMed:	(("Speech Intelligibility"[Mesh]) OR (intelligibil*[Title/Abstract] OR comprehensibil*[Title/Abstract] OR understandabil*[Title/Abstract])) AND ("Acoustics"[Mesh] OR "Speech Acoustics"[Mesh] OR "Speech Production Measurement"[Mesh] OR "Phonetics"[Mesh] OR "Signal Processing, Computer-Assisted"[Mesh] OR "Fourier Analysis"[Mesh] OR "Sound Spectrography"[Mesh] OR "Sound"[Mesh] OR "Signal-To-Noise Ratio"[Mesh] OR "Noise"[Mesh])	3393
		Total: 6719
		Total after exclusion of duplicates: 4818

The titles and abstracts were retrieved via EndNote X9 and screened by two independent raters (TP and MB), applying the aforementioned selection criteria. In view of the large number of abstracts, the whole set was divided into two. Each rater thus reviewed half of the whole set, plus a randomly selected set of 20% abstracts, taken from the other half. Hence, 40% of the abstracts were read by both raters, allowing for a weighted Kappa to be measured to assess the inter-rater agreement. Agreement interpretation guidelines (Landis & Koch, 1977) are: <.00: poor; .00-.20: slight; .21-.40: fair; .41-.60: moderate; .61-.80: substantial; .81-1.00: almost perfect. Differences in the eligibility ratings were resolved by reaching a consensus. The full-text articles of the selected papers were then retrieved and reviewed by each rater. A flowchart illustrating the article selection process according to the PRISMA guidelines (Liberati et al., 2009) is shown in Figure 1 in the Results section.

Critical appraisal of methodological quality and level of evidence

The methodological quality of the selected papers was rated using the QualSyst tool (Kmet, Lee, & Cook, 2004). This tool was developed as a scoring system in order to methodologically assess the quality of quantitative as well as of qualitative research papers, by analysing, among others, the study design, the research question, the study group selection and description, and the control of confounding factors. As interpretation guidelines, a score

>80% was considered as strong methodological quality, 60-79% as good, 50-59% as appropriate and <50% as poor quality. The latter was considered as an exclusion criterion.

The National Health and Medical Research Council Hierarchy (NHMRC, 1999) was used to assess the level of evidence. Six levels are described: Level I Highest level, systematic reviews of randomized controlled trials, Level II Randomized controlled trials, Level III-1 Pseudo-randomized controlled trials, III-2 Comparative studies with concurrent controls and allocation not randomized (cohort studies, case control studies, or interrupted time series with a control group), Level III-3 Comparative study without concurrent controls, with historical controls, two or more single-arm studies, or interrupted time series without a parallel control group, and Level IV Lowest level, case series. Research reports of level IV and expert opinions were not further analysed, as well as systematic reviews.

Data items

After selection based on the eligibility criteria and the methodological quality assessment, the following information was extracted for each article: the study population (number, age, gender, language), the speech sample used for the acoustic measure(s) (targeted phonemes), the acoustic parameter(s), the perceptual intelligibility measure(s), the main conclusion regarding the link between acoustics and intelligibility and the descriptive data if available.

No contact was sought with authors to inquire about unreported data.

Results

Study selection

A total of 4818 titles and abstracts were retrieved from the databases (after automatic removal of most of the duplicates). Each of the two independent raters screened half of these

records (2405), as well as 20% (964) of the other half. The raters agreed on the eligibility criteria for 1792/1928 (93%) abstracts, with a weighted Kappa of .89 – corresponding to an 'almost perfect' agreement according to the guidelines of Landis and Koch (1977).

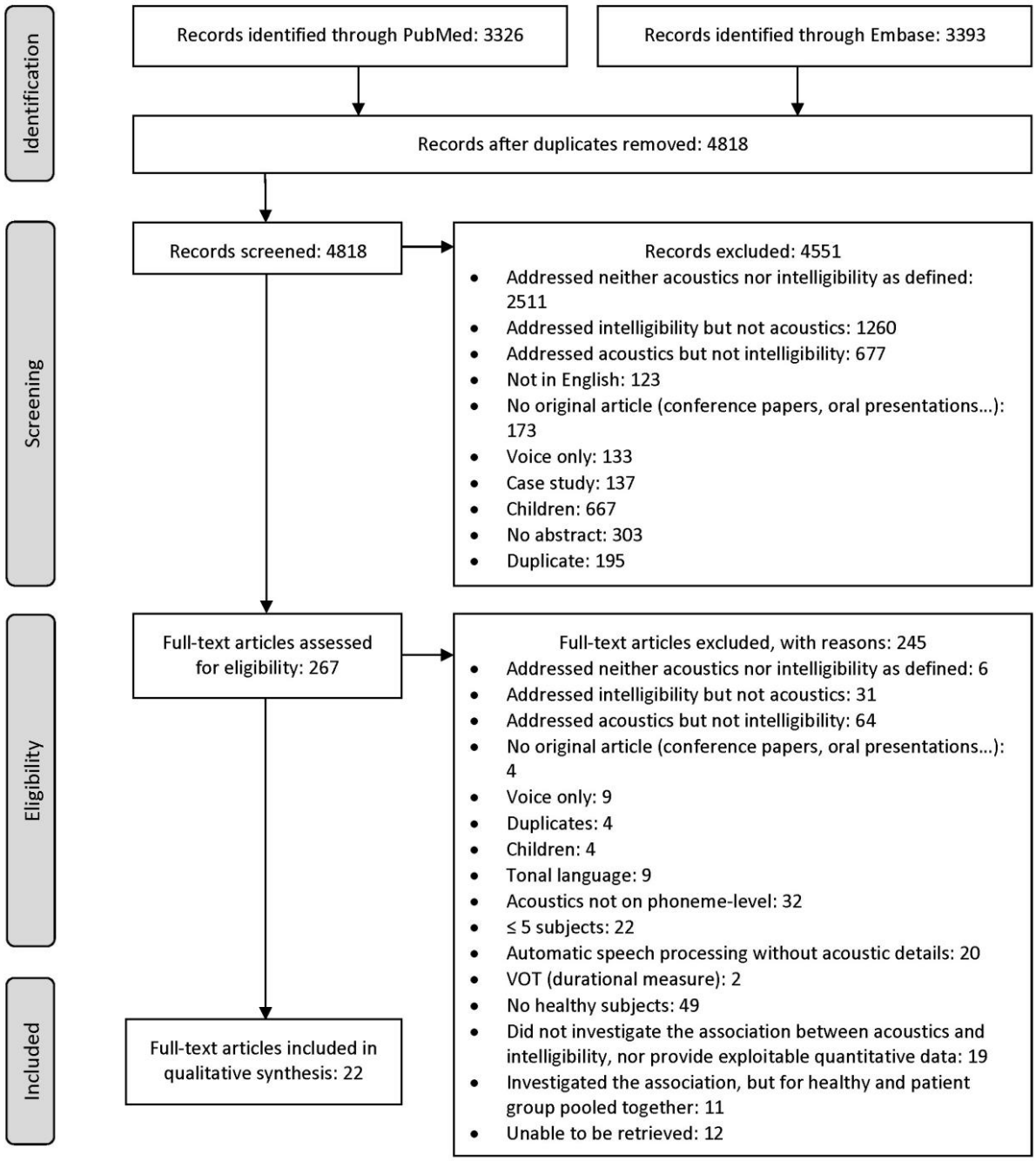

Figure 1 - Flow diagram illustrating the selection process according to the PRISMA guidelines. Adapted from Moher et al. (2009).

Two hundred and sixty-seven full-text articles were reviewed, of which 22 were retained. Nine of these studies addressed the association between spectral acoustic and perceptual measures (A01-A09). The remaining 13 papers, albeit not assessing the link per se, were retained because they provided quantitative data for both acoustic measures and perceptual ratings in healthy speakers, which provides useful information.

The study selection process is illustrated in Figure 1. A detailed synthesis of the 22 included studies is available in Appendix A. For readability purposes, an identification code has been assigned to each of the 22 papers (see Table 2), which will be used for the in-text citations throughout this article.

Quality assessment

The QualSyst scores of the 22 papers ranged from 71% (good methodological quality) to 100% (strong quality). Only one article's methodological quality was graded as 'good', the other 21 were rated as 'strong'.

According to the NHMRC hierarchy for the level of evidence assessment, 14 papers were categorized as level III-2 evidence (comparative study with concurrent controls), the other eight papers were classified as level III-3 evidence ('comparative study without concurrent controls'). The rating for each individual paper can be found in Table 2.

Table 2. Methodological quality ratings for the 22 included articles using the Qualsyst critical appraisal tool by Kmet et al. and level of evidence according to the National Health and Medical Research Council (NHMRC) hierarchy

Reference	Qualsyst score[1] (%)	Methodology quality	NHMRC Level of Evidence[2]
A01. McRae et al., 2002	20/24 (83)	Strong	III-2
A02. Hazan et al., 2004	21/24 (88)	Strong	III-3
A03. Neel, 2008	18/22 (82)	Strong	III-3
A04. Ferguson et al., 2014	20/22 (91)	Strong	III-3
A05. Whitfield et al., 2017	21/24 (88)	Strong	III-3
A06. Katz et al., 1991	20/24 (83)	Strong	III-3
A07. Flege et al., 1992	21/24 (88)	Strong	III-3
A08. Bunton et al., 2001	21/24 (88)	Strong	III-2
A09. Ferguson et al., 2007	20/24 (83)	Strong	III-3
A10. Weismer et al., 1992	17/24 (71)	Good	III-2

A11. Hohoff et al., 2003	20/24 (83)	Strong	III-3
A12. Yunusova et al., 2005	20/24 (83)	Strong	III-2
A13. de Bruijn et al., 2009	21/24 (88)	Strong	III-2
A14. Van Lierde et al., 2012	20/24 (83)	Strong	III-2
A15. Skodda et al., 2013	21/24 (88)	Strong	III-2
A16. Whitfield et al., 2014	23/24 (96)	Strong	III-2
A17. Neel et al., 2015	22/22 (100)	Strong	III-2
A18. Dwivedi et al., 2016	24/24 (100)	Strong	III-2
A19. Connaghan et al., 2017	21/24 (88)	Strong	III-2
A20. Fletcher et al., 2017	22/24 (92)	Strong	III-2
A21. Kim et al., 2017	22/24 (92)	Strong	III-2
A22. Martel-Sauvageau et al., 2017	23/24 (96)	Strong	III-2

[1] Methodological quality: strong > 80%; good 60–79%; adequate 50–59%; poor < 50%.
[2] NHMRC hierarchy: Level I Systematic reviews; Level II Randomized control trials; Level III–1 Pseudo-randomized control trials; Level III–2 Comparative studies with concurrent controls and allocation not randomized (cohort studies), case control studies, or interrupted time series with a control group; Level III–3 Comparative studies with historical control, two or more single-arm studies, or interrupted time series without a control group; Level IV Case series.
Note: The studies were ordered according to 1) the type of outcome: A01-A05 = direct correlation between acoustics and perceptual ratings; A06-A09 = indirect investigation of the link between acoustics and perceptual ratings; A10-A22: quantitative data for both acoustics and perceptual ratings, without investigation of the link; 2) the chronological order

Study characteristics

Study populations

Out of the 22 studies, 14 originally included both a subject group and a healthy control group, of which only the healthy control group was kept for the present analysis. The remaining eight studies only included healthy speakers as a study group. Keeping in mind that only studies including more than five subjects were retained, the median size of the study sample was 15 (min.: 8, max.: 93), with an interquartile range of 18.5. Regarding the gender distribution in the samples, most of the studies (20/22, 91%) included both men and women. In 13 of these studies (65%), the men/women ratio was 1:1 (i.e. perfect gender balance). Four studies showed a small gender imbalance (i.e. less than 20% difference between both gender groups), while three showed a preponderance of men (>20% difference). Of the two remaining studies, one included only men (A10), and the other did not mention the subjects' gender(s) (A14). With regards to the age factor, half of the studies were carried out on groups aged more than 50 years, 10 on subjects aged less than 50 years, and one did not report the study population's age (A03). Regarding the investigated languages, seventeen out of the 22

studies (73%) were carried out in English. Eleven of these used American English (of which three specified an Upper Midwest dialect), one used British English, one used New Zealand English, and the remaining four did not specify the English variant. Two studies were carried out in Dutch, two in French (of which one in Quebec French), one in German, and one both in Korean and in English.

Speech samples and spectral measures

The different phonemes analysed in the studies were extracted from isolated words or from words in sentences. Two studies analysed isolated phonemes (sustained vowel [i] in A14, and [s]-sound in A18).

Vowels. Eighteen out of the 22 papers (82%) studied vowel acoustics. The corner vowels [i, u, a, æ] are the most investigated (8/18 studies). One paper (A20) studied the New Zealand English corner vowels [aː, iː, oː]. Only three studies analysed an extensive panel of vowels [i, ɪ, e, ɛ, æ, ɑ, ʌ, o, ʊ, u]. Three studies did not explicitly mention the vowels used for the analyses. None of the studies investigated nasal vowels.

Regarding the spectral analysis of vowels, seventeen out of the 18 studies (94%) used steady-state formant measures, four studies examined dynamic formant measures.
For a list with definitions and formulas of the acoustic measures used in the retained studies and reported in the outcome table (Appendix A), please refer to Appendix B.

Glides. One article (A22) studied the two glides [w, j] in addition to vowels, using the F2 slopes as a measure of the rate of phonatory tract modification.

Consonants. Eight articles (36%) investigated consonants. The most investigated consonants are the fricatives [s, ʃ] (6 studies). The other two papers studied the plosives [t, d] (voiced-voiceless contrast, A07) and the velar [x] (in Dutch, A13), respectively. None of the studies investigated nasal consonants nor liquids.

Among these eight papers, five used spectral moment analyses. Four of them used the first moment, while the fifth used the second moment. The remaining acoustic measures were studied in single studies and are reported in the outcome table (Appendix A).

Perceptual measures

Percent correct identification. Ten studies used the percentage of correctly identified stimuli. One paper did not describe the identification task (A02). The remaining nine all used a multiple-choice task, six in which the listener had to choose the target in a list of words, two in which the listener had to choose between two targets (A06 and A07), and one (A19) in which the listener had to choose the target vowel among 12 vowels (monophthongs or diphthongs). None of the studies used a transcription task.

Ordinal scales. Seven studies used Likert-type equal appearing interval scales, out of which five asked the listeners to rate the 'overall intelligibility', three asked them to rate the 'articulation', one the 'speech clarity', one the 'speech precision' and one the 'speech severity'. Two studies used rating scales where a high score indicated a good speech rating ('positive scales'); four studies used 'negative scales' (a high score meaning a negative rating). One study used both types of scales (A13).

Visual analogue scales (VAS). Five papers used visual analogue scales, out of which two asked the raters to judge the 'speech clarity' (A05, A16) and the others respectively the

'overall intelligibility' (A17), the 'speech precision' (A17), the 'articulatory precision' (A20), the 'ease of understanding' (A20) and 'how much [the listener] understood of what the person said' (A22).

Three of the studies used positive VAS scales (a high score meaning a good overall intelligibility), the other two used negative VAS scales (a high score indicating a low overall intelligibility).

Direct magnitude estimation (DME). Two studies used direct magnitude estimation with a modulus of 100. In one study, listeners were asked to rate 'overall severity' (negative scale) (A01), in the other they were asked to rate 'overall intelligibility' (instruction: 'ease to understand') on a positive scale (A12).

Outcome measure

Nine of the 22 retained articles analysed the link between spectral acoustic and perceptual measures. Two different methodologies can be identified. Five articles (A01-A05) directly addressed the correlation between acoustic and perceptual measures (VAS, DME and Likert scales or percent-identification scores). Four other articles (A06-A09) indirectly investigated the link between acoustics and perceptual ratings, by investigating acoustic differences between groups that had been created based on their intelligibility (A09), or by analysing acoustic differences between two correctly perceived phonemes/syllables: [sɪ] *vs.* [sʊ] (A06); [t] *vs.* [d] (A07); [ɛ] *vs.* [æ] and [ɪ] *vs.* [ɛ] (A08). The remaining 13 articles (A10-A22) analysed spectral measures as well as perceptual measures but did not directly address the association between both.

Summary of findings

The conclusions of the different studies are reported in the outcome table (Appendix A), sorted into three categories: the studies directly addressing the link between spectral and perceptual measures; the studies indirectly investigating this link; and the studies only providing descriptive data for acoustics and perceptual ratings without analysing the link. Regarding the first category, the significant and non-significant correlations are shown in Table 3. Significant correlations between spectral measures and perceptual ratings have been measured in vowels only, for steady-state F1 and F2 measures (A04), the F1 range in men (A03), the [i] vs [u] F2 difference (A02), the vowel space area (A03), the relative change in the acoustic-articulatory vowel space area (A05), the mean amount of formant movement in women (A03) and the dynamic vector length measure (A04).

Table 3. Significant and non-significant correlations between acoustic measures and perceptual ratings of speech

	Vowels														Consonants
	F1	F2	F1 range	F2 range	Euclidean distance F1-F2	F1≠ [i-æ]	F2≠ [i-u]	Vowel distance	VSA	AAVS	Formant movement	Dynamic ratio	Vector Length	Trajectory length	1st moment
DME									∅ (A01)						∅ (A01)
Likert	∅ (A02)	∅ (A02)			∅ (A02)	∅ (A02)	✓ (A02)								
VAS										✓ (A05)					
%corr	∅ (A02,A03) ✓ (A04)	∅ (A02, A03) ✓ (A04)	✓ (M, A03) ∅ (F, A03)	∅ (A03)	∅ (A02)	∅ (A02)	✓ (A02)	∅ (A03)	✓ (A03)		✓ (F, A03) ∅ (M, A03)	∅ (A03)	✓ (A04)	∅ (A04)	

Note : ✓: significant correlation; ∅: non-significant correlation
Abbreviations: F1/F2 = first and second formant; F1≠ [x,y] = F1 difference between vowels x and y; AAVS = articulatory-acoustic vowel space; DME = direct magnitude estimation; Likert: Likert-type equal-appearing interval scale; VAS = visual analog scale; %corr = percent correct identification score

Among the studies that indirectly addressed the link between spectral acoustics and perceptual estimates (i.e. without correlations), A06 and A07 targeted consonant measures, whereas A08 and A09 focused on vowels. In A06, the fricative centroid energy and the fricative spectral peak in the [s]-sound in [si] and [su] were found to be acoustic underliers of the coarticulation effect, the values being significantly higher for the [s] in the syllables identified as [si]. A07 found significantly higher steady-state F1 offset frequencies in vowels

preceding [t] than for [d], in native English speakers. The authors concluded that this acoustic measure is a good indicator of the correct perception of the voiced/voiceless contrast in apico-alveolar stop consonants. Regarding the measures targeting vowels, significant F0-F1 and F1-F2 differences were found in A08, for the correctly identified vowels in the pairs [ɛ-æ] and [ɪ-ɛ]. Hence, the authors concluded that these measures are related to the speech intelligibility, as they seem to be linked to the perception of the tongue-height contrast. The F1-F2 difference was considered to be the primary cue, whereas the F0-F1 difference was interpreted as a secondary cue, linked to the F2-F1 difference. In A09, the 'spectral change' measure was found to be significantly larger for speakers with a high clear speech word identification benefit.

Discussion

The data from this review confirms the highly variable nature of speech in healthy adult speakers. In light of the differing rating tasks and instructions (e.g. rating on visual analog scales of intelligibility vs articulatory precision) and targeted speech units (e.g. percent correct identification of phonemes vs words), no aggregated variability measure could be computed across the studies in this review. Among the studies using percent correct identification, for example, while four found values higher than 90% (on words, isolated vowels and vowels in CVC syllables), four others found mean scores between 60.6% and 71% (on phonemes in CVC syllables and on syllables). The speech variability in healthy speakers is also found between subjects in the different studies. For example, while three of the studies using percent correct identification scores report a relatively low standard deviation (ranging from 1.12% to 4%), the studies using ordinal scales show a higher variability: if all the results are normalized to percentages, the standard deviations range from 6.25% to 12%. These results illustrate that even in healthy talkers, the physiological limits do

not always allow the speech production system to meet the many demands of spontaneous speech. The resulting 'imprecisions' are mainly found at the phoneme level (Rossi & Peter-Defare, 1998; Schiller, 2006), leading to a certain overlap of speech sound categories, i.e. vowel and consonant reductions, as well as phoneme omissions (Benzeguiba et al., 2007; Guenther, 1995; Meunier, 2007; Van Son & Pols, 1996, 1999).

The aim of this review was to investigate further how the variations in healthy speech can be measured in order to be taken into account when analysing speech in patient populations. Indeed, the publication dates of the retained papers – of which only three date back to the 1990s – illustrate that the rise of technology has led to an increasing interest in the acoustic investigation of speech. This is mainly due to the fact that acoustic measures do not have to be carried out manually anymore and are thus faster to obtain as well as more reliable. In the next section, we will thus focus on the spectral acoustic underpinnings of intelligibility.

Spectral measures of speech intelligibility in healthy speakers

In our review, most of the studies using spectral measures focused on vowels. Vowel reduction in informal speech is a well-described, universal phenomenon (Van Son & Pols, 1996). Two types of reduction are found (Maurová Paillereau, 2016): vowel centralization and contextual assimilation. Vowel centralization is observed when formant frequencies tend to those of a neutral vowel, whereas contextual assimilation occurs when a vowel's formant frequencies change toward the acoustic loci of neighbouring consonants. The data in this review shows that steady-state formant measures (F1, F2, F1 range, F2 difference between /i/ and /u/, F1-F2 difference in [ɛ-æ] and [ɪ-ɛ], vowel space area [VSA]) are linked with vowel identification scores (A02, A03, A04, A08). The VSA is commonly used to account for vowel centralization, often in pathological speech (Liu, Tsao, & Kuhl, 2005; Sapir, Połczyńska, & Tobin, 2009; Weismer, Jeng, Laures, Kent, & Kent, 2001), but has also been shown to be

sensitive to intelligibility differences in healthy speech (Bond & Moore, 1994) and to articulatory changes in clear speech (Lam, Tjaden, & Wilding, 2012; Smiljanić & Bradlow, 2009). The VSA is to some extent related to the size and shape of the resonance cavities created by the jaw and tongue positions (Sandoval, Berisha, Utianski, Liss, & Spanias, 2013), and thereby provides a global overview of the articulatory working space. However, it has shown inconsistent results (Lansford & Liss, 2014; Sapir et al., 2009) and might not be sensitive enough to subtle vowel articulation changes, both in healthy speech (Ferguson & Kewley-Port, 2007) and in motor speech disorders (Whitfield & Goberman, 2014). Sapir et al. (2009) explained that all Euclidean distances of the vowel space do not equivalently contribute to the differentiation between healthy and pathological speakers. In light of this asymmetry of the vowel formant sensitivity to articulatory changes, they suggested the use of the Euclidean distance between /i/ and /u/ instead, which was found to be the most sensitive marker. The F2 difference between /i/ and /u/ was also shown to be related to vowel intelligibility in A02. Furthermore, Lam et al. (2012) found that in clear speech, high tense and lax vowels (/i, ɪ, u, ʊ/) contributed most to the vowel space expansion. These observations indicate that the formant measures in these vowels should be prioritized for diagnostic purposes. Several alternatives to the VSA have been suggested, such as the vowel articulation index (VAI) (A15) and its inverse, the formant centralization ratio (FCR) (A20), designed to minimize inter-speaker variability and maximize the sensitivity to vowel reduction (Sapir, Ramig, Spielman, & Fox, 2010, 2011). However, all of the above measures only use the midpoint of three to four corner vowels of the vowel space. Whitfield et al. (2014) therefore suggested another alternative measure, the acoustic-articulatory vowel space (AAVS), which interestingly uses formant measures across the voiced portions of a whole utterance in continuous speech and thus provides a more global, also supposedly more sensitive measure (Whitfield & Goberman, 2014, 2017). Furthermore, the AAVS has been

shown to be significantly larger in clear speech (A05)(Whitfield & Goberman, 2017). It would therefore be interesting to further investigate how the AAVS correlates with segmental perceptual intelligibility estimates and accounts for variations in healthy speech.

Regarding dynamic formant measures, the 'formant movement' (A03), 'vector length' (A04) and 'spectral change' (A09) measures show that vowels with larger changes in the F1xF2 space are significantly better identified. Lam et al. (2012) showed that dynamic vowel formant measures also showed increased values in clear speech. These measures are related to intra-vowel antero-posterior tongue movements and changes in tongue height. They could thus also be useful in the investigation of imprecisions due to motor constraints in informal speech and subsequently in pathological speech.

Studies targeting the spectral features of consonants are rarer in our review, although consonants are reduced as much as vowels in informal speech and this articulatory reduction affects their intelligibility (Van Son & Pols, 1999). In this review, the fricative centroid energy and the fricative spectral peak in the [s]-sound in [s i] and [s u] were found to be acoustic underliers of the coarticulation effect (A06). The fricative centroid energy (or 'centre of gravity' [CoG]) is the first of the four spectral moment measures (Jongman, Wayland, & Wong, 2000) and corresponds to the 'frequency that divides the spectrum into two halves' (Yoon, 2015). It has been shown to be decreased in non-plosives in spontaneous speech of healthy talkers (Van Son & Pols, 1996, 1999), making it a relevant acoustic measure of consonant reduction. Spectral moment measures consider and describe the whole spectrum as a statistical distribution. Evers et al. argued that it is wiser to consider the global aspect of sibilant spectra rather than specific frequency regions (Evers, Reetz, & Lahiri, 1998). Indeed, sibilants are characterized by two sound sources, one at the tongue constriction and one at the teeth (Fant, 1960), which makes spectral peak locations difficult to predict. Also, the spectral shape of consonants is less defined than the clear vowel formant structure. Therefore, the

description of the overall spectral shape of consonants should be preferred to the use of specific frequency regions ('formant patterns') (Fant, 1960; Stevens & Blumstein, 1978). Another argument in favour of using spectral moments is that they are said to be correlated with the length and shape of the cavity in front of the articulatory constriction (Behrens & Blumstein, 1988; Kay, 2012; Stevens, 1998; Yoon, 2015). Hence, they can lead to an articulatory interpretation. However, study A06 demonstrates that spectral moments are likely to vary according to the vowel context/to coarticulation.

Just as in vowels, another type of measure that has been used in the retained papers are the dynamic formant transitions, among which the F2 slope. The F2 slope measure, used in glides in A22, is 'a dynamic measure that reflects the rate at which speech movements can be performed' (R. D. Kent et al., 1989) and is thus related to speaking rate. Van Son & Pols (1999), investigating acoustic correlates of consonant reduction in healthy speech, found that the F2 slope difference (i.e. difference between the F2 slope in the VC- and CV-boundaries in VCV syllables) is lower in spontaneous than in read speech. This reduced F2 slope difference indicated a lower consonant-induced coarticulation in the VCV syllable, thus a reduced consonant articulation. The use of formant transition measures is all the more noteworthy since it has been shown that in healthy ageing a decrease in intelligibility can be partly attributed to slower tongue movements (Kuruvilla-Dugdale et al., 2020).

To summarize this discussion, we highlighted the importance of investigating variations at the phoneme level in healthy speech, using acoustic measures to analyse both vowel and consonant reductions. Various spectral acoustic measures, mainly on vowels, proved to be related to perceived speech intelligibility in healthy speakers. However, the results show that none of these measures account for a large percentage of the variance in the perceptual intelligibility scores. While acoustic measures allow for a more objective investigation of speech, they do not comprehensively represent the speech signal, but rather

target specific cues that are believed to be theoretically relevant. One should also keep in mind that the accurate perception of phonemes relies on several phonemic features (Jakobson, Fant, & Halle, 1951) and it is not one sole feature, but the whole set of speech units that makes up the notion of intelligibility (Flanagan, 1972, p. 311). Hence, a combination of acoustic measures, taking into account various phonemic traits and spectral aspects, could be a first way to a more comprehensive assessment of speech intelligibility (e.g. Bradlow et al., 1996; Ray D. Kent et al., 1989; J. Lee, Hustad, & Weismer, 2014; Lindblom, 1990; Weismer, 2008). Furthermore, there is a complex entanglement of segmental acoustic features with factors at other levels of granularity such as intonation, stress (e.g. acoustic differences between stressed and unstressed vowels in A19), voice quality and speech rate. This has been demonstrated in connected speech (Metz et al., 1990) as well as in clear speech (Kuruvilla-Dugdale et al., 2020; Smiljanić & Bradlow, 2009; Whitfield & Goberman, 2017). Eventually, before using segmental acoustic measures on specific patient populations, extensive research is still needed to get a better understanding of their behaviour in the healthy speakers, to identify relevant acoustic combinations that could account for perceived speech variations and to provide normative data from a large set of healthy speakers.

Further perspectives and future directions of research

From the analyses made throughout this review, a few leads for further studies can be considered. First, the diversity of the methodologies used in the retained papers demonstrates that speech can be investigated in many different ways at a perceptual as well as at an acoustic level. Of the 22 retained papers in our review, only five addressed the definition of the targeted speech-related concept(s), of which four (A08, A12, A20, A22) provided a definition of intelligibility. In light of the various terms used to refer to speech production – each of which refers to a specific concept – unambiguous definitions should be provided in

research papers. Also, the rating tasks and the acoustic measures should be extensively described, so as to allow the reader to interpret the results accordingly, as well as for the methods to be replicable. It can be observed that even if several studies use the same measure, the study population, the phonemic sample, the computing method and the reporting of the results are very different and sometimes not reported (according to the aim of each study), which makes it difficult to relate the resulting values. To illustrate this point, an attempt to compare the results of similar acoustic measures used in the different studies is shown in Appendix C.

In this review, we have observed a majority of studies focusing on vowels when it comes to spectral cues. Vowels play an important role in speech intelligibility (Chen, Wong, & Wong, 2013; Cole, Yan, Mak, & Fanty, 1996; Kewley-Port, Burkle, & Lee, 2007) and are also more convenient to analyse spectrally, as they are by definition voiced and composed of periodic waveforms and can be sustained (in contrast to plosive consonants). However, consonants also significantly contribute to speech intelligibility. Lindblom (1990) already postulated that despite the coarticulation effects, a combination of spectral features could allow for a good distinction between stop consonants. Furthermore, while vowels were found to have a more important effect on talker identity discrimination, consonants are essential for word identification (Bonatti, Peña, Nespor, & Mehler, 2005; Owren & Cardillo, 2006). The consonant intelligibility, their variability and reductions in healthy speech, as well as related spectral cues (in addition to the more investigated time-domain cues), should therefore be further explored.

Some considerations can also be highlighted with regard to the study populations. The majority of the studies included both men and women in a balanced ratio. However, very few of them actually differentiated the results by gender, especially in the control groups, for which the results are very often pooled. It is well known that vowel formant values, for

example, vary between men and women (Bradlow et al., 1996; Coleman, 1971; Yang, 1996). Generally speaking, greater account still needs to be taken of this factor, and the study group's gender information should systematically be specified. One possible way to address the issue of across-sex value comparisons is to use Bark scales (Fletcher, McAuliffe, Lansford, & Liss, 2017), as could be observed in some of the studies in this review. Also, while half of the studies were carried out on study groups aged more than 50 years, none of the studies investigated the impact of age in adults on the spectral measures or on perceived speech intelligibility. It would be noteworthy to take the age factor into account in order to analyse the evolution of speech-related acoustics and perceived intelligibility in normal ageing (Kuruvilla-Dugdale et al., 2020). Indeed, speech has been shown to vary across the lifespan due to physiological and neuromuscular modifications (Benjamin, 1997; Bilodeau-Mercure & Tremblay, 2016; Hazan, 2017; Hazan et al., 2018; Hooper & Cralidis, 2009; Tremblay et al., 2017). The study of speech modifications in 'normal' ageing as compared to pathological ageing might help further understand speech production strategies in healthy speech.

Limitations

The studies discussed in this systematic review have been retrieved from two databases (PubMed and Embase) that were thought to include papers from the targeted topic. We are, however, aware that there might be studies from other sources that address the subject but that are not referenced in these two databases.

Regarding the acoustic measures considered in this review, we would like to underline that time-domain measures were not taken into account in order to limit the noise in the initial database search (e.g. studies about the speaking rate, prosody and pauses in fluency disorders…). As explained in the introduction, only frequency-domain measures were

included. In a future study, it would, however, also be interesting to investigate the link between time-domain measures (such as the voice onset time) and perceived intelligibility, as time- and frequency-domain measures provide complementary data (Floegel, Fuchs, & Kell, 2020; Li et al., 2008). The resulting higher number of studies focusing on vowels might also stem from this methodological decision. Further studies on time-domain measures could clarify if this is a general trend among phoneme-level measures, or if it is limited to spectral measures.

Last but not least, while this review focused on studies written in English, it would also be informative to review studies written in – and thus focusing on - other languages. The most contrastive example to illustrate the interest of investigating other languages are tonal languages. In the latter, the acoustic and perceptual underliers of speech intelligibility might be very different from those in Western languages. Suprasegmental measures (eg. F0 contour) might for example contribute to a higher degree to intelligibility, as compared to phoneme-level measures (Chen & Loizou, 2011).

Conclusions

Our results highlight that speech is highly variable within and across healthy adult speakers, which stresses the need for further studies regarding the acoustic underpinnings of speech intelligibility in healthy speech. Healthy speech shows inherent imprecisions and is thus not, as often presumed, 100% accurate. A better understanding of the imprecisions in healthy spontaneous speech will provide a more realistic baseline for the investigation of disordered speech.

The direct investigation of the correlation between spectral cues and speech intelligibility estimates remains scarce, especially in consonants. In this review, for vowels, the following measures were shown to be linked to sub-lexical perceived speech intelligibility

ratings: steady-state F1 and F2 measures, the F1 range, the [i]-[u] F2 difference, F0-F1 and F1-F2 differences in [ɛ-æ] and [ɪ-ɛ], the vowel space area, the mean amount of formant movement, the vector length and the spectral change measure. For consonants, only the fricative centroid energy and the fricative spectral peak in the [s]-sound, as well as the steady-state F1 offset frequencies in vowels preceding [t] and [d] have shown a significant link with phoneme identification scores.

An important question is raised by this review: Can perceived intelligibility be quantified by single acoustic measures? It indeed appears that, to date, no acoustic measure is able to predict speech intelligibility to a large extent. There is still extensive research to be carried out to identify relevant acoustic combinations that could account for perceived speech variations (e.g. vowel and consonant reductions) in healthy speech. Subsequently, normative data will have to be gathered from a large number of healthy speakers in order to then investigate these measures in specific patient populations. To that end, speech-related terms (e.g. intelligibility, comprehensibility, severity) need to be clearly defined and methodologies described in sufficient details to allow for replication, cross-comparisons/meta-analyses and pooling of data.

Acknowledgments

This project was supported by the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie Grant 766287.
The authors declare that there is no conflict of interest.

References

Aichert, I., & Ziegler, W. (2004). Syllable frequency and syllable structure in apraxia of speech. *Brain and Language*, *88*(1), 148–159. https://doi.org/10.1016/S0093-934X(03)00296-7

Behrens, S. J., & Blumstein, S. E. (1988). Acoustic characteristics of English voiceless fricatives: A descriptive analysis. *Journal of Phonetics*, *16*(3), 295–298.

https://doi.org/10.1016/s0095-4470(19)30504-2

Benjamin, B. (1997). Speech production of normally ageing adults. *Seminars in Speech and Language, 18*(2), 135–141. https://doi.org/10.1055/s-2008-1064068

Benzeguiba, M., De Mori, R., Deroo, O., Dupon, S., Erbes, T., Jouvet, D., … Wellekens, C. (2007). Automatic speech recognition and speech variability: A review. *Speech Communication, 49*(10–11), 763–786. https://doi.org/10.1016/j.specom.2007.02.006

Bilodeau-Mercure, M., & Tremblay, P. (2016). Age differences in sequential speech production: Articulatory and physiological factors. *Journal of the American Geriatrics Society, 64*(11), e177–e182. https://doi.org/10.1111/jgs.14491

Bohland, J. W., & Guenther, F. H. (2006). An fMRI investigation of syllable sequence production. *NeuroImage, 32*(2), 821–841. https://doi.org/10.1016/j.neuroimage.2006.04.173

Bonatti, L. L., Peña, M., Nespor, M., & Mehler, J. (2005). Linguistic constraints on statistical computations: The role of consonants and vowels in continuous speech processing. *Psychological Science, 16*(6), 451–459. https://doi.org/10.1111/j.0956-7976.2005.01556.x

Bond, Z. S., & Moore, T. J. (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech Communication, 14*(4), 325–337. https://doi.org/10.1016/0167-6393(94)90026-4

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication, 20*(3–4), 255–272. https://doi.org/10.1016/S0167-6393(96)00063-5

Bunton, K., & Weismer, G. (2001). The relationship between perception and acoustics for a high-low vowel contrast produced by speakers with dysarthria. *Journal of Speech, Language, and Hearing Research, 44*(6), 1215–1228. https://doi.org/10.1044/1092-4388(2001/095)

Carmichael, J., & Green, P. (2004). Revisiting dysarthria assessment intelligibility metrics. *8th International Conference on Spoken Language Processing (ICSLP)*, 742–745.

Chen, F., & Loizou, P. C. (2011). Predicting the intelligibility of vocoded and wideband Mandarin Chinese. *The Journal of the Acoustical Society of America*, *129*(5), 3281–3290. https://doi.org/10.1121/1.3570957

Chen, F., Wong, L. L. N., & Wong, E. Y. W. (2013). Assessing the perceptual contributions of vowels and consonants to Mandarin sentence intelligibility. *The Journal of the Acoustical Society of America, 134*(2), EL178–EL184. https://doi.org/10.1121/1.4812820

Cole, R., Yan, Y., Mak, B., & Fanty, M. (1996). The contribution of consonants versus vowels to word recognition in fluent speech. *IEE Proceedings - Vision, Image, and Signal Processing, 2*, 853–856. https://doi.org/10.1121/1.417028

Coleman, R. O. (1971). Male and female voice quality and its relationship to vowel formant frequencies. *Journal of Speech and Hearing Research, 14*(3), 565–577. https://doi.org/10.1044/jshr.1403.565

Connaghan, K. P., & Patel, R. (2017). The impact of contrastive stress on vowel acoustics and intelligibility in dysarthria. *Journal of Speech, Language, and Hearing Research, 60*(1), 38–50. https://doi.org/10.1044/2016_JSLHR-S-15-0291

Cox, R. M., Alexander, G. C., & Gilmore, C. (1987). Intelligibility of average talkers in typical listening environments. *Journal of the Acoustical Society of America*, *81*(5), 1598–1608. https://doi.org/10.1121/1.394512

De Bruijn, M. J., Ten Bosch, L., Kuik, D. J., Quené, H., Langendijk, J. A., Leemans, C. R., & Verdonck-De Leeuw, I. M. (2009). Objective acoustic-phonetic speech analysis in patients treated for oral or oropharyngeal cancer. *Folia Phoniatrica et Logopaedica, 61*(3), 180–187. https://doi.org/10.1159/000219953

Derdemezis, E., Vorperian, H. K., Kent, R. D., Fourakis, M., Reinicke, E. L., & Bolt, D. M. (2016). Optimizing vowel formant measurements in four acoustic analysis systems for diverse speaker groups. *American Journal of Speech-Language Pathology, 25*(3), 335–354. https://doi.org/10.1044/2015_AJSLP-15-0020

Ding, Z. Q., McLoughlin, I. V., & Tan, E. C. (2003). Extension of proposal of standards for intelligibility tests of Chinese speech: CDRT-tone. *IEE Proceedings - Vision, Image, and Signal Processing, 150*(1), 1. https://doi.org/10.1049/ip-vis:20030161

Dubno, J. R., Dirks, D. D., & Schaefer, A. B. (1989). Stop-consonant recognition for normal-hearing listeners and listeners with high-frequency hearing loss. II: Articulation index predictions. *The Journal of the Acoustical Society of America*, *85*(1), 355–364. https://doi.org/10.1121/1.397687

Dwivedi, R. C., St.Rose, S., Chisholm, E. J., Clarke, P. M., Kerawala, C. J., Nutting, C. M., … Harrington, K. J. (2016). Acoustic parameters of speech: Lack of correlation with perceptual and questionnaire-based speech evaluation in patients with oral and oropharyngeal cancer treated with primary surgery. *Head & Neck, 38*(5), 670–676. https://doi.org/10.1002/hed.23956

Enderby, P., & Palmer, R. (2008). *FDA-2: Frenchay Dysarthria Assessment (2nd ed.)*. Pro-Ed.

Eringis, D., & Tamulevičius, G. (2014). Improving speech recognition rate through analysis parameters. *Electrical, Control and Communication Engineering, 5*(1), 61–66. https://doi.org/10.2478/ecce-2014-0009

Evers, V., Reetz, H., & Lahiri, A. (1998). Crosslinguistic acoustic categorization of sibilants independent of phonological status. *Journal of Phonetics, 26*(4), 345–370. https://doi.org/10.1006/jpho.1998.0079

Fant, G. (1960). *Acoustic theory of speech production*. The Hague: Mouton.

Ferguson, S. H., & Kewley-Port, D. (2007). Talker differences in clear and conversational speech: acoustic characteristics of vowels. *Journal of Speech, Language, and Hearing Research, 50*(5), 1241–1255. https://doi.org/10.1044/1092-4388(2007/087)

Ferguson, S. H., & Quené, H. (2014). Acoustic correlates of vowel intelligibility in clear and conversational speech for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America, 135*(6), 3570–3584. https://doi.org/10.1121/1.4874596

Fitch, W. T. (2000). The evolution of speech: a comparative review. *Trends in Cognitive Science, 4*(7), 258–267. https://doi.org/https://doi.org/10.1016/S1364-6613(00)01494-7

Flanagan, J. L. (1972). *Speech analysis, synthesis and perception (2nd ed.)*. Springer.

Flege, J. E., Munro, M. J., & Skelton, L. (1992). Production of the word-final English /t/–/d/ contrast by native speakers of English, Mandarin, and Spanish. *The Journal of the Acoustical Society of America, 92*(1), 128–143. https://doi.org/10.1121/1.404278

Fletcher, A. R., McAuliffe, M. J., Lansford, K. L., & Liss, J. M. (2017). Assessing vowel centralization in dysarthria: A comparison of methods. *Journal of Speech, Language, and Hearing Research, 60*(2), 341–354. https://doi.org/10.1044/2016_JSLHR-S-15-0355

Floegel, M., Fuchs, S., & Kell, C. A. (2020). Differential contributions of the two cerebral

hemispheres to temporal and spectral speech feedback control. *Nature Communications*, *11*(1), 1–12. https://doi.org/10.1038/s41467-020-16743-2

Fontan, L. (2012). *De la mesure de l'intelligibilité à l'évaluation de la compréhension de la parole pathologique en situation de communication* [Doctoral dissertation, Université Toulouse 2 Le Mirail]. https://tel.archives-ouvertes.fr/tel-00797883

Fontan, L., Magnen, C., Tardieu, J., Ferrané, I., Pinquier, J., Farinas, J., … Aumont, X. (2014). Comparaison de mesures perceptives et automatiques de l'intelligibilité: Application à de la parole simulant la presbyacousie. *Revue Traitement Automatique Des Langues, 55*(2), 151–174.

Ghio, A., Lalain, M., Giusti, L., Pouchoulin, G., Robert, D., Rebourg, M., … Woisard, V. (2018). Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique. *XXXIIe Journées d'Études Sur La Parole*, 285–293. https://doi.org/10.21437/jep.2018-33

Guenther, F. H. (1995). Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production. *Psychological Review*, *102*(3), 594–621. https://doi.org/10.1037/0033-295X.102.3.594

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, *96*(3), 280–301. https://doi.org/10.1016/j.bandl.2005.06.001

Hazan, V. (2017). Speech communication across the lifespan. *Acoustics Today, 13*(1), 36–43.

Hazan, V., & Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America, 116*(5), 3108–3118. https://doi.org/10.1121/1.1806826

Hazan, V., Tuomainen, O., Tu, L., Kim, J., Davis, C., Brungart, D., & Sheffield, B. (2018). How do ageing and age-related hearing loss affect the ability to communicate effectively in challenging communicative conditions? *Hearing Research, 369*, 33–41. https://doi.org/10.1016/j.heares.2018.06.009

Hohoff, A., Seifert, E., Fillion, D., Stamm, T., Heinecke, A., & Ehmer, U. (2003). Speech performance in lingual orthodontic patients measured by sonagraphy and auditive analysis. *American Journal of Orthodontics and Dentofacial Orthopedics, 123*(2), 146–152. https://doi.org/10.1067/mod.2003.12

Honda, K. (2008). Physiological processes of speech production. In J. Benesty, M. M. Sondhi, & Y. A. Huang (Eds.), *Springer Handbook of Speech Processing* (pp. 7–26). Springer. https://doi.org/https://doi.org/10.1007/978-3-540-49127-9_2

Hood, J. D., & Poole, J. P. (1980). Influence of the speaker and other factors affecting speech intelligibility. *Audiology, 19*, 434–455.

Hooper, C. R., & Cralidis, A. (2009). Normal changes in the speech of older adults: You've still got what it takes; it just takes a little longer! *Perspectives on Gerontology, 14*(2), 47. https://doi.org/10.1044/gero14.2.47

Hustad, K. C. (2008). The relationship between listener comprehension and intelligibility scores for speakers with dysarthria. *Journal of Speech, Language, and Hearing Research, 51*(3), 562–573. https://doi.org/10.1044/1092-4388(2008/040)

Jakobson, R., Fant, C. G. M., & Halle, M. (1951). *Preliminaries to speech analysis: The distinctive features and their correlates*. The MIT Press.

Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. *The Journal of the Acoustical Society of America, 108*(3), 1252. https://doi.org/10.1121/1.1288413

Katz, W. F., Kripke, C., & Tallal, P. (1991). Anticipatory coarticulation in the speech of adults and young children: Acoustic, perceptual, and video data. *Journal of Speech, Language, and Hearing Research, 34*(6), 1222–1232. https://doi.org/10.1044/jshr.3406.1222

Kay, T. S. (2012). *Spectral analysis of stop consonants in individuals with dysarthria secondary to stroke* [Master's thesis, Louisiana State University]. LSU Digital Commons. https://digitalcommons.lsu.edu/gradschool_theses/2632/

Kent, R. D., Kent, J. F., Weismer, G., Martin, R. E., Sufit, R. L., Brooks, B. R., & Rosenbek, J. C. (1989). Relationships between speech intelligibility and the slope of second-formant transitions in dysarthric subjects. *Clinical Linguistics and Phonetics*, *3*(4), 347–358. https://doi.org/10.3109/02699208908985295

Kent, R. D., & Kim, Y. J. (2003). Toward an acoustic typology of motor speech disorders. *Clinical Linguistics and Phonetics*, *17*(6), 427–445. https://doi.org/10.1080/0269920031000086248

Kent, R. D., Weismer, G., Kent, J. F., & Rosenbek, J. C. (1989). Toward phonetic intelligibility testing in dysarthria. *Journal of Speech and Hearing Disorders*, *54*(4), 482–499.

https://doi.org/10.1044/jshd.5404.482

Kent, R. D. (1992). *Intelligibility in speech disorders: Theory, measurement and management*. John Benjamins Publishing Company. https://doi.org/10.1075/sspcl.1

Kewley-Port, D., Burkle, T. Z., & Lee, J. H. (2007). Contribution of consonant versus vowel information to sentence intelligibility for young normal-hearing and elderly hearing-impaired listeners. *The Journal of the Acoustical Society of America, 122*(4), 2365–2375. https://doi.org/10.1121/1.2773986

Kim, Y., & Choi, Y. (2017). A cross-language study of acoustic predictors of speech intelligibility in individuals with Parkinson's Disease. *Journal of Speech, Language, and Hearing Research, 60*(9), 2506–2518. https://doi.org/10.1044/2017_jslhr-s-16-0121

Kmet, L. M., Lee, R. C., & Cook, L. S. (2004). Standard quality assessment criteria for evaluating primary research papers from a variety of fields. *HTA Initiative, 13,* 1–22.

Kuruvilla-Dugdale, M., Dietrich, M., McKinley, J. D., & Deroche, C. (2020). An exploratory model of speech intelligibility for healthy aging based on phonatory and articulatory measures. *Journal of Communication Disorders, 87,* 105995. https://doi.org/10.1016/j.jcomdis.2020.105995

Lalain, M., Ghio, A., Giusti, L., Robert, D., Fredouille, C., & Woisard, V. (2020). Design and development of a speech intelligibility test based on pseudowords in French: Why and how? *Journal of Speech, Language, and Hearing Research, 63*(7), 2070–2083. https://doi.org/10.1044/2020_JSLHR-19-00088

Lam, J., Tjaden, K., & Wilding, G. (2012). Acoustics of clear speech: Effect of instruction. *Journal of Speech, Language, and Hearing Research, 55*(6), 1807–1821. https://doi.org/10.1044/1092-4388(2012/11-0154)

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174.

Lansford, K. L., & Liss, J. M. (2014). Vowel acoustics in dysarthria: Mapping to perception. *Journal of Speech, Language, and Hearing Research, 57*(1), 68–80. https://doi.org/10.1044/1092-4388(2013/12-0263)

Lavoie, L. M. (2002). Subphonemic and suballophonic consonant variation: The role of the phoneme inventory. *ZAS Papers in Linguistics, 28,* 39–54.

Lee, J., Hustad, K. C., & Weismer, G. (2014). Predicting speech intelligibility with a multiple speech subsystems approach in children with cerebral palsy. *Journal of Speech, Language, and Hearing Research*, *57*(5), 1666–1678. https://doi.org/10.1044/2014_JSLHR-S-13-0292

Lee, T., Liu, Y., Huang, P. W., Chien, J. T., Lam, W. K., Yeung, Y. T., … Law, S. P. (2016). Automatic speech recognition for acoustical analysis and assessment of Cantonese pathological voice and speech. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 6475–6479. https://doi.org/10.1109/ICASSP.2016.7472924

Levelt, W. J. M. (1995). The ability to speak: From intentions to spoken words. *European Review*, *3*(1), 13–23. https://doi.org/10.1017/S1062798700001290

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–75. https://doi.org/10.1017/S0140525X99001776

Li, X., Ning, Z., Brashears, R., & Rife, K. (2008). Relative contributions of spectral and temporal cues for speech recognition in patients with sensorineural hearing loss. *Journal of Otology*, *3*(2), 84–91. https://doi.org/10.1016/S1672-2930(08)50019-5

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., … Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *Journal of Clinical Epidemiology*, *62*(10), e1-34. https://doi.org/10.1016/j.jclinepi.2009.06.006

Lindblom, B. (1990). Explaining phonetic variation: A sketch of the H&H theory. In Hardcastle W.J., Marchal A. (Eds.), *Speech production and speech modelling* (pp. 403–439). Springer. https://doi.org/10.1007/978-94-009-2037-8_16

Liu, H.-M., Tsao, F.-M., & Kuhl, P. K. (2005). The effect of reduced vowel working space on speech intelligibility in Mandarin-speaking young adults with cerebral palsy. *The Journal of the Acoustical Society of America*, *117*(6), 3879–3889. https://doi.org/10.1121/1.1898623

Maniwa, K., Jongman, A., & Wade, T. (2009). Acoustic characteristics of clearly spoken English fricatives. *The Journal of the Acoustical Society of America*, *125*(6), 3962–3973.

https://doi.org/10.1121/1.2990715

Martel-Sauvageau, V., & Tjaden, K. (2017). Vocalic transitions as markers of speech acoustic changes with STN-DBS in Parkinson's Disease. *Journal of Communication Disorders, 70*, 1–11. https://doi.org/10.1016/j.jcomdis.2017.10.001

Maurová Paillereau, N. (2016). Do isolated vowels represent vowel targets in French? An acoustic study on coarticulation. *SHS Web of Conferences*, *27*, 09003. https://doi.org/10.1051/shsconf/20162709003

McRae, P. A., Tjaden, K., & Schoonings, B. (2002). Acoustic and perceptual consequences of articulatory rate change in parkinson disease. *Journal of Speech, Language, and Hearing Research, 45*(1), 35–50. https://doi.org/10.1044/1092-4388(2002/003)

Metz, D. E., Schiavetti, N., Samar, V. J., & Sitler, R. W. (1990). Acoustic dimensions of hearing-impaired speakers' intelligibility: Segmental and suprasegmental characteristics. *Journal of Speech and Hearing Research*, *33*(3), 476–487. https://doi.org/10.1044/jshr.3303.476

Meunier, C. (2007). Phonétique acoustique. In P. Auzou, V. Rolland-Monnoury, S. Pinto, & C. Ozsancak (Eds.), *Les dysarthries* (pp. 164–173). Solal.

Middag, C., Martens, J. P., Van Nuffelen, G., & De Bodt, M. (2009). Automated intelligibility assessment of pathological speech using phonological features. *Eurasip Journal on Advances in Signal Processing*, 2009. https://doi.org/10.1155/2009/629030

Miller, N. (2013). Measuring up to speech intelligibility. *International Journal of Language and Communication Disorders, 48*(6), 601–612. https://doi.org/10.1111/1460-6984.12061

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine, 6*(7), e1000097. https://doi.org/10.1371/journal.pmed.1000097

Molis, M. R., & Leek, M. R. (2011). Vowel identification by listeners with hearing impairment in response to variation in formant frequencies. *Journal of Speech, Language, and Hearing Research, 54*(4), 1211–1223. https://doi.org/10.1044/1092-4388(2010/09-0218)

National Health and Medical Research Council. (1999). *A guide to the development, implementation and evaluation of clinical practice guidelines*. NHMRC.

Neel, A. T. (2008). Vowel space characteristics and vowel identification accuracy. *Journal of*

Speech, Language, and Hearing Research, 51(3), 574–585. https://doi.org/10.1044/1092-4388(2008/041)

Neel, A. T., Palmer, P. M., Sprouls, G., & Morrison, L. (2015). Muscle weakness and speech in oculopharyngeal muscular dystrophy. *Journal of Speech, Language, and Hearing Research, 58*(1), 1–12. https://doi.org/10.1044/2014_JSLHR-S-13-0172

Owren, M. J., & Cardillo, G. C. (2006). The relative roles of vowels and consonants in discriminating talker identity versus word meaning. *The Journal of the Acoustical Society of America, 119*(3), 1727–1739. https://doi.org/10.1121/1.2161431

Peeters, G. (2004). A large set of audio features for sound description (similarity and classification) in the CUIDADO project. *CUIDADO IST Project Report, 54*, 1–25.

Preminger, J., & Wiley, T. L. (1985). Frequency selectivity and consonant intelligibility in sensorineural hearing loss. *Journal of Speech, Language, and Hearing Research, 28*(2), 197–206. https://doi.org/10.1044/jshr.2802.197

Rossi, M., & Peter-Defare, É. (1998). *Les lapsus ou Comment notre fourche a langué*. Presses Universitaires de France.

Sandoval, S., Berisha, V., Utianski, R. L., Liss, J. M., & Spanias, A. (2013). Automatic assessment of vowel space area. *The Journal of the Acoustical Society of America*, *134*(5), EL477–EL483. https://doi.org/10.1121/1.4826150

Sapir, S., Połczyńska, M., & Tobin, Y. (2009). Why does the vowel space area as an acoustic metric fail to differentiate dysarthric from normal vowel articulation and what can be done about it? *Poznan Studies in Contemporary Linguistics, 45*(2), 301–311. https://doi.org/10.2478/v10010-009-0018-2

Sapir, S., Ramig, L. O., Spielman, J. L., & Fox, C. (2010). Formant centralization ratio: A proposal for a new acoustic measure of dysarthric speech. *Journal of Speech, Language, and Hearing Research, 53*(1), 114–125. https://doi.org/10.1044/1092-4388(2009/08-0184)

Sapir, S., Ramig, L. O., Spielman, J. L., & Fox, C. (2011). Acoustic metrics of vowel articulation in Parkinson's disease: Vowel space area (VSA) vs. Vowel articulation index (VAI). *Models and Analysis of Vocal Emissions for Biomedical Applications - 7th International Workshop, MAVEBA 2011*, *9*, 173–175.

Schiller, N. O. (2006). Phonological encoding in speech production. *Proceedings of ISCA*

Tutorial and Research Workshop on Experimental Linguistics. https://doi.org/10.13140/2.1.3238.6249

Skodda, S., Grönheit, W., Mancinelli, N., & Schlegel, U. (2013). Progression of voice and speech impairment in the course of Parkinson's Disease: A longitudinal study. *Parkinson's Disease, 2013*, 1–8. https://doi.org/10.1155/2013/389195

Smiljanić, R., & Bradlow, A. R. (2009). Speaking and hearing clearly: Talker and listener factors in speaking style changes. *Language and Linguistics Compass*, *3*(1), 236–264. https://doi.org/10.1111/j.1749-818X.2008.00112.x

Souza, P. E., Wright, R. A., Blackburn, M. C., Tatman, R., & Gallun, F. J. (2015). Individual sensitivity to spectral and temporal cues in listeners with hearing impairment. *Journal of Speech, Language, and Hearing Research*, *58*(2), 520–534. https://doi.org/10.1044/2015_JSLHR-H-14-0138

Stevens, K. N. (1998). *Acoustic Phonetics*. Cambridge: MIT Press.

Stevens, K. N., & Blumstein, S. E. (1978). Invariant cues for place of articulation in stop consonants. *The Journal of the Acoustical Society of America*, *64*(5), 1358–1368. https://doi.org/10.1121/1.382102

Stipancic, K. L., Tjaden, K., & Wilding, G. (2016). Comparison of intelligibility measures for adults with Parkinson's Disease, adults with Multiple Sclerosis, and healthy controls. *Journal of Speech, Language, and Hearing Research*, *59*(2), 230–238. https://doi.org/10.1044/2015_JSLHR-S-15-0271

Sussman, J. E., & Tjaden, K. (2012). Perceptual measures of speech from individuals with Parkinson's Disease and Multiple Sclerosis: Intelligibility and beyond. *Journal of Speech, Language, and Hearing Research*, *55*(4), 1208–1219. https://doi.org/10.1044/1092-4388(2011/11-0048)

Tremblay, P., Sato, M., & Deschamps, I. (2017). Age differences in the motor control of speech: An fMRI study of healthy aging. *Human Brain Mapping*, *38*(5), 2751–2771. https://doi.org/10.1002/hbm.23558

Van Lierde, K., Browaeys, H., Corthals, P., Mussche, P., Van Kerkhoven, E., & De Bruyn, H. (2012). Comparison of speech intelligibility, articulation and oromyofunctional behaviour in subjects with single-tooth implants, fixed implant prosthetics or conventional removable prostheses. *Journal of Oral Rehabilitation, 39*(4), 285–293.

https://doi.org/10.1111/j.1365-2842.2011.02282.x

Van Nuffelen, G., Middag, C., De Bodt, M., & Martens, J. (2009). Speech technology-based assessment of phoneme intelligibility in dysarthria. *International Journal of Language and Communication Disorders, 44*(5), 716–730. https://doi.org/10.1080/13682820802342062

Van Son, R. J. J. H., & Pols, L. C. W. (1996). Acoustic profile of consonant reduction. *International Conference on Spoken Language Processing, ICSLP, Proceedings*, *3*, 1529–1532. https://doi.org/10.1109/icslp.1996.607908

Van Son, R. J. J. H., & Pols, L. C. W. (1999). Acoustic description of consonant reduction. *Speech Communication*, *28*(2), 125–140. https://doi.org/10.1016/S0167-6393(99)00009-6

Voiers, W. D. (1983). Evaluating processed speech using the Diagnostic Rhyme Test. *Speech Technology*, 30–39.

Weismer, G. (2008). Speech Intelligibility. In M. J. Ball, M. R. Perkins, N. Müller, & S. Howard (Eds.), *The handbook of clinical linguistics* (pp. 568–582). Blackwell Publishing.

Weismer, G., Jeng, J.-Y., Laures, J. S., Kent, R. D., & Kent, J. F. (2001). Acoustic and intelligibility characteristics of sentence production in neurogenic speech disorders. *Folia Phoniatrica et Logopaedica*, *53*(1), 1–18. https://doi.org/10.1159/000052649

Weismer, G., Martin, R., Kent, R. D., & Kent, J. F. (1992). Formant trajectory characteristics of males with amyotrophic lateral sclerosis. *The Journal of the Acoustical Society of America, 91*(2), 1085–1098. https://doi.org/10.1121/1.402635

Whitfield, J., & Goberman, A. (2014). Articulatory-acoustic vowel space: Application to clear speech in individuals with Parkinson's Disease. *Journal of Communication Disorders, 51*, 19–28. https://doi.org/10.1016/j.jcomdis.2014.06.005

Whitfield, J., & Goberman, A. (2017). Articulatory-acoustic vowel space: Associations between acoustic and perceptual measures of clear speech. *International Journal of Speech-Language Pathology, 19*(2), 184–194. https://doi.org/10.1080/17549507.2016.1193897

Yang, B. (1996). A comparative study of American English and Korean vowels produced by male and female speakers. *Journal of Phonetics, 24*(2), 245–261. https://doi.org/10.1006/jpho.1996.0013

Yiu, E. M., van Hasselt, C. A., Williams, S. R., & Woo, J. K. S. (1994). Speech intelligibility in tone language (Chinese) laryngectomy speakers. *International Journal of Language & Communication Disorders, 29*(4), 339–347. https://doi.org/10.3109/13682829409031287

Yoon, T.-J. (2015). A corpus-based study on the effects of gender on voiceless fricatives in American English. *Phonetics and Speech Sciences*, *7*(1), 117–124. https://doi.org/10.13064/ksss.2015.7.1.117

Yorkston, K. M., Strand, E. A., & Kennedy, M. R. T. (1996). Comprehensibility of dysarthric speech. *American Journal of Speech-Language Pathology, 5*(1), 55–66. https://doi.org/10.1044/1058-0360.0501.55

Yunusova, Y., Weismer, G., Kent, R. D., & Rusche, N. M. (2005). Breath-group intelligibility in dysarthria. *Journal of Speech, Language, and Hearing Research, 48*(6), 1294–1310. https://doi.org/10.1044/1092-4388(2005/090)

APPENDIX A – Outcome table: Description of the 22 included studies

A. Studies describing associations between acoustic variables in healthy speakers and auditory perception

Reference	Study design[1]	QualSyst (by Kmet et al.)[2]	Healthy population [N, Gender, Age (years), Language]	Speech sample for acoustics (target phoneme)	Acoustic parameters (Definitions)	Perceptual measure(s)	Conclusions
A.1. Outcome measure: Correlation between acoustic measure(s) and perceptual rating scale(s)							
A01. McRae et al., 2002	III-2	20/24 83% strong	N=13 (9 M, 4 F) Age: µ=67 (range 59-80) Language: American English	- Vowels [i, ɑ, u, æ] bounded by obstruent consonants - Fricatives [s, ʃ] in word-initial and -final positions	- Vowels: Vowel space area (VSA; quadrilateral, using F1 and F2 frequencies at temporal midpoint) - Consonants: 1st moment coefficient difference [ʃ]-[s]; lower 1st moment coefficient suggests more posterior constriction, looser constriction, or increased lip rounding in [ʃ]	Overall speech severity: direct magnitude estimation (DME) using a modulus with the value of 100 (= moderately severe)	*Association:* - Regression between vowel space area and overall speech severity: not significant - Regression between first moment difference and overall speech severity: not significant *Descriptive data:* Mean (range) - Overall speech severity (DME): 28 (2-51) - VSA: N.R. (vowel quadrilateral graphics) - First moment difference: N.R. (graphics)
A02. Hazan et al., 2004	III-3	21/24 88% strong	N=33 (15 M, 18 F) Age F µ=33.11 (SD=10.9); M µ=30.7 (SD=10.5) Language: British English	Vowels [i, u, æ] in CVC monosyllabic words	- Vowel formant measures: • F1 and F2 at the steady-state vowel region • Euclidean distance between F1 and F2 for each vowel - Vowel space measures: • Difference between F1 frequencies for [i] and [æ] • Difference between F2 frequencies for [i] and [u]	- Percent-correct identification (task N.R.) - Subjective ratings (7-point scale, 1-7: 7= highest score on the positive attribute of the pair): mumbly–precise, unpleasant–pleasant, muffled–clear, husky–not husky, creaky–not creaky, nasal–not nasal, high for a (fe)male-low for a (fe)male, thin–rich, weak–powerful, and harsh–smooth	*Association:* - Vowel formant measures: • F1, F2 and Euclidean distance F1-F2: no significant correlation with percent-correct identification, nor with subjective rating scales - Vowel space measures: • [i]-[æ] F1 difference: not correlated with percent-correct identification, nor with any subjective rating • [i]-[u] F2 difference: Significant correlation with percent-correct identification (r=0.401, p=0.006) and with two rating scales (rs= N.R., p<0.01): mumbly–precise, unclear–clear scales *Descriptive data:* • Percent-correct word identification: N.R. (graphics) • Subjective rating scales: N.R. (graphics) • Vowel formant and space measures: N.R.

Reference	Study design[1]	QualSyst (by Kmet et al.)[2]	Healthy population [N, Gender, Age (years), Language]	Speech sample for acoustics (target phoneme)	Acoustic parameters (Definitions)	Perceptual measure(s)	Conclusions
A03. Neel, 2008	III-3	18/22 82% strong	N=93 (45 M, 48 F) Age: N.R. Language: American English (Michigan/Upper Midwest dialect)	Vowels [i, ɪ, e, ɛ, æ, ɑ, ʌ, o, ʊ, u] in [hVd] context	- Global measures: • Mean F1 and mean F2 across the 10 vowels • Mean amount of formant movement, averaged across the 10 vowels: Sum of the Euclidean distance in the F1xF2 space from the vowel onset to the steady state, and the Euclidean distance from the vowel steady state to the offset - Fine-grained measures: • Vowel space area (VSA, quadrilateral; two triangles: [i, æ, u] and [æ, u, ɑ]) • Mean distance among vowels (vowel dispersion): average Euclidean distance between each vowel pair • F1 and F2 ranges: subtraction of the lowest F1/F2 value from the highest • Dynamic ratio (distinctiveness among vowels with dynamic and static trajectories): average Euclidean distance (from vowel onsets to steady states to offsets in the F1 x F2 space) covered by the 3 most dynamic vowels ([æ, ʌ, ʊ]) divided by the distance covered by the 3 most static ones ([i, ɛ, u])	Percent-correct identification scores across the 10 vowels for each talker	*Association:* - None of the acoustic measures individually or combined accounted for more than 18% of variance in vowel identification - Significant predictors of vowel identification: • Men: VSA (r^2=0.12, p<0.02) and F1 range (r^2=0.14, p<0.01) • Women: VSA (r^2=0.09, p< 0.02) and mean amount of formant movement (r^2 =0.11, p<0.02) - Non-significant predictors: F1, F2, formant movement in men, distance among vowels, F1 range in women, F2 range, dynamic ratio *Descriptive data:* Mean (SD): • Vowel identification scores: M=95.6% (4.0%); F=96.8% (2.6) • F1 (Bark): M=5.04 (0.20); F=5.88 (0.30) • F2 (Bark): M=13.05 (0.37); F=14.70 (0.53) • Formant movement: M=1.43 (0.29); F=1.99 (0.47) • VSA: M=18.57 (4.13); F=25.07 (6.55) • Distance among vowels: M=4.54 (0.50); F=5.46(0.66) • F1 range (Bark): M=3.83 (0.59); F=4.32 (0.80) • F2 range (Bark): M=9.37 (1.04); F=11.15 (1.06) • Dynamic ratio: M=1.97 (0.53); F=2.24 (0.59)

Reference	Study design[1]	QualSyst (by Kmet et al.)[2]	Healthy population [N, Gender, Age (years), Language]	Speech sample for acoustics (target phoneme)	Acoustic parameters (Definitions)	Perceptual measure(s)	Conclusions
A04. Ferguson et al., 2014	III-3	20/22 91% strong	N=41 (20 M, 21 F) Age: range 18-45 Language: English	Vowels [i, ɪ, e, ɛ, æ, ɑ, ʌ, o, ʊ, u] in [bVd] context, in conversational (CON) and in clear (CL) speech	- Steady-state F1 and F2 values - Dynamic values: • Vector length (VL) = Euclidean distance of the vector in the F1xF2 space connecting the formant values at 20% and 80% of the vowels • Trajectory length (TL) = Sum of the lengths of four temporally equidistant vowel sections: 20%-35%, 35%-50%, 50%-65%, 65%-80%	Percent-correct vowel identification: selection out of ten sets of three keywords [e.g. '(1) feet, thief, bead, (2) sit, rib, bid' etc.]	*Association:* - Steady-state F1: Significant and strong correlation with vowel identification (Z=4.5, p<0.0001) - Steady-state F2: Significant and strong correlation with vowel identification (Z=4.6, p<0.0001) - VL: Significantly positive regression slopes with vowel identification (Z=5.1, p<0.001): vowels with larger change in F1xF2 space are better identified - TL: Non-significant effect on the accuracy of vowel identification: a more curved trajectory in the vowel space does not affect the vowel identification *Descriptive data:* - Vowel identification: N.R. - F1: N.R. (graphics) - F2: N.R. (graphics) - Mean VL (Barks): CL=1.27, CON=1.1, ratio=0.09 - Mean TL (Barks): CL=0.50, CON=0.49, ratio=1
A05. Whitfield et al., 2017	III-3	21/24 88% strong	N=10 5 M (Age: μ=24.40, range 20-36) 5 F (Age: μ=24.30, range 18-29) Language: Standard American English	Corner vowels [i, ɑ, u, æ] in sentences, in conversational (CON) and in clear (CL) speech.	- Articulatory-acoustic vowel space (AAVS): square root of the generalized variance of all sampled vowel formants in the F1xF2 coordinate plot This measure was carried out on one sentence containing all the corner vowels, and on two sentences from the Rainbow Passage. - Traditional vowel space area (VSA): formant values measured during the steady state of the vocalic nuclei of the words 'stack', 'key', 'blue' and 'box' (in the sentence containing all the corner vowels)	Rating of speech clarity on a 100mm visual analogue scale (0-100: unclear - very clear)	*Association:* Significant correlation between perceptual difference scores (conversational vs. clear) and relative change in AAVS (r=0.67, r^2=0.45, p<0.01) *Descriptive data:* Mean (SD): A. Sentence with corner vowels - *Speech clarity rating* (mm): • CON: M=66.33 (2.14); F=71.97 (8.82) • CL: M=80.10 (4.85); F=85.80 (3.72) - *AAVS* (kHz): • CON: M=31.59 (3.77); F=77.17 (16.43) • CL: M= 43.43 (6.67); F=107.17 (19.87) - *VSA* (kHz): M=200.81(23.65); F= 577.74(94.11) B. Sentence 1 from the Rainbow Passage - *Speech clarity rating* (mm): • CON: M=61.80 (7.55); F=70.73 (4.93) • CL: M= 74.20 (3.72); F=80.57 (8.77) - *AAVS* (kHz):

							• CON: M=24.45 (6.54); F=61.83 (6.44) • CL: M=31.10 (7.05); F=83.23 (18.60) C. Sentence 3 from the Rainbow Passage - *Speech clarity rating* (mm): • CON: M=64.60(11.16); F=66.57(11.88) • CL: M=75.63 (11.12); F=87.03 (9.05) - *AAVS* (kHz): • CON: M=27.98 (5.06); F=68.83 (6.86) • CL: M=35.37 (8.13); F=93.81 (20.21)

A.2. Outcome measure: Indirect association between perceptual ratings and acoustic measure(s)

Reference	Study design[1]	QualSyst (by Kmet et al.)[2]	Healthy population [N, Gender, Age (years), Language]	Speech sample for acoustics (target phoneme)	Acoustic parameters (Definitions)	Perceptual measure(s)	Conclusions
A06. Katz et al., 1991	III-3	20/24 83% strong	N=10 (5 M, 5 F) Age: μ=32 (SD=6.7; range 26-45) Language: English	[s] in [s u] and [s i]	- Fricative centroid energy at 30 ms and 100 ms prior to fricative offset: thought to indicate front cavity resonances (indication of the degree of anticipatory labial movement) - Fricative spectral peaks at 30 ms prior to fricative offset, anticipating F2 of the vowel	Hearing only the [s] sound, identification of the original syllable	- Syllable identification scores: N.R. - Mean centroid energy (Hz) 30ms prior to offset: • [s i]: 5524 • [s u]: 5134 • [s i]/[s u] ratio: 1.08 - Mean centroid energy (Hz) 100ms prior to offset: • [s i]: 6806 • [s u]: 6182 • [s i]/[s u] ratio: 1.10 Values for [i] significantly higher than for [u], (reflecting labial coarticulation), F=30.9, p<.001; Values for 100-ms window significantly higher than for 30-ms window (F=107.2, p<.001) - Mean fricative spectral peaks at fricative offset: • [s i]: 1999 • [s u]: 1866 • [s i]/[s u] ratio: 1.07 Strong vowel context effects, values for [s i] greater than for [s u], (reflecting lingual and labial coarticulation effects), F=101.2, p<.001

Reference	Study design[1]	QualSyst (by Kmet et al.)[2]	Healthy population [N, Gender, Age (years), Language]	Speech sample for acoustics (target phoneme)	Acoustic parameters (Definitions)	Perceptual measure(s)	Conclusions
A07. Flege et al., 1992	III-3	21/24 88% strong	N=30 (15 M, 15 F) 10 native English (Age: μ=27.9, SD=6.5) 10 native Spanish 'experienced' (Age: μ=28.2, SD=5.7) 10 native Mandarin 'experienced' (Age: μ=28.1, SD=3.0) Language: American English	Word-final [t]-[d] contrast in minimal pairs of CVCs ([bVt]-[bVd] and [sVt]-[sVd]) containing either [i, ɪ, ɛ, æ].	F1 offset frequency in the final 45ms of the vowels	Percent-correct identification scores for the word-final stops ([t]/[d]); computation of A' scores based on number of correct identifications of [t] and false alarms (incorrect identification of [d]'s as voiceless), on [bVC] words	- Native English speakers: • In general, [t]-final words significantly higher F1 offset frequency than [d]-final words • For [bVt]-[bVd] pairs (F values from 11.9 to 20.8, p < 0.01), F1 offset on average 68 Hz higher for [t] • For [sVt]-[sVd] pairs (F= 76.7, p <0.01), F1 offset on average 73 Hz higher for [t] - Nonnative experienced English speakers: • Acoustic differences not significant for [bVt]-[bVd] • For [sVt]-[sVd], significant difference (543 vs 518 Hz) for Mandarin subjects (F= 11.6, p <0.01), not significant for Spanish subjects In nonnative experienced English speakers, F1 offset frequency accounted for 1.5% of the variance in the [t]-[d] identification (F=4.67, p=0.032) *Descriptive data:* - Percent-correct identification: overall rate from 68% to 71%. Rates higher for [t] than [d] (82% vs 65%). Native English speakers' stops significantly higher correct scores (A'=0.953) than experienced and inexperienced Spanish and Mandarin subjects (A'=0.751, 0.720 and 0.668, 0.679, respectively). - F1 offset frequency: N.R. (graphics)
A08. Bunton et al., 2001	III-2	21/24 88% strong	N=10 (5 M, 5 F) Age: range 68-77 Language: American English (Upper Midwest dialect)	High vs low vowels (tongue-height): vowel pairs [ɛ-æ] and [ɪ-ɛ] in words	Differences between F0-F1 and between F1-F2 (measured at 50% of the total vowel duration)	Percent-correct word identification, multiple-choice format (minimal or near minimal pairs)	In the correctly perceived tokens - Statistically significant F0-F1 differences ([ɛ]-[æ]: U = 222.0, p<0.001; [ɪ]-[ɛ]: U = 194.0, p<0.001) - Statistically significant F1-F2 differences ([ɛ]-[æ]: U = 179.5, p<0.001; [ɪ]-[ɛ]: U = 116.0, p<0.001) These measures are thus considered to be linked to perceived tongue-height contrast. *Descriptive data:* - Percent-correct identification: 96.44% (94.38%-98.38%) - F0-F1, F1-F2: N.R. (graphics)

Reference	Study design[1]	QualSyst (by Kmet et al.)[2]	Healthy population [N, Gender, Age (years), Language]	Speech sample for acoustics (target phoneme)	Acoustic parameters (Definitions)	Perceptual measure(s)	Conclusions
A09. Ferguson et al., 2007	III-3	20/24 83% strong	N=12 (6 M, 6 F) Age: μ=32.75 (range 20-45) Language: American English	Vowels [i, ɪ, e, ɛ, æ, ɑ, ʌ, o, ʊ, u] in [bVd] context, in conversational and in clear speech	- Steady-state formant values: • Perimeter: overall dimensions of the vowel space; sum of 4 Euclidean distances between adjacent vowels ([i]-[æ], [æ]-[ɑ], [ɑ]-[u], and [u]-[i]) • F1 range: difference between the average F1 for low vowels [æ, ɑ], and F1 for high vowels [i, u] • F2 front: average F2 value for [i, ɪ, e, ɛ, æ] • F2 back: average F2 value for [ɑ, ʌ, o, ʊ, u] - Dynamic formant movement: • Spectral change (λ): Sum of the absolute formant frequency shift (from 20% to 80%) for F1 and F2 • Spectral angle (Ω): Sum, in radians, of the absolute values of F1 and F2 angles, calculated as the arctangents of the difference between the formant frequencies at the 20% and 80% points, divided by the duration between these two points	Percent-correct vowel in clear (CL) and conversational speech (CON): Creation of 'No Benefit' (NB) and 'Big Benefit' (BB) groups according to vowel intelligibility gain (BB = large clear speech effect for listeners, relative to conversational speech)	*Association:* Clear speech benefit differences (differences between NB and BB talker groups): - Steady-state formant values: • Perimeters: No sign. difference, F = 0.66, p = 0.44 • F1 range: No significant difference, F = 0.497, p = 0.5 • F2 front & back range: No sign. difference, F = 0.29, p = 0.59; F = 2.1, p = 0.15 - Dynamic formant movement: • Spectral change: Sign. larger for NB, F =10.86, p<0.01 • Spectral angle: No sign. difference, F = 3.06, p = 0.08 *Descriptive data:* - Percent correct vowel identification: ▪ BB: CL=79.2, CON=60.6, difference = 18.6 ▪ NB: CL= 67.0, CON=68.1, difference = -1.1 - Steady-state formant values (all measures in Barks): • Perimeters: ▪ BB: CL=13.77, CON=12.65, difference=1.12 ▪ NB: CL=14.26, CON=13.85, difference=0.41 • F1 range: ▪ BB: CL=3.07, CON=2.79, difference=0.28 ▪ NB: CL=3.12, CON=3.16, difference=-0.04 • F2 front range: ▪ BB: CL=13.13, CON=12.82, difference=0.31 ▪ NB: CL=13.2, CON=13.06, difference=0.14 • F2 back range: ▪ BB: CL=10.13, CON=10.26, difference=-0.13 ▪ NB: CL=9.75, CON=9.86, difference=-0.11 - Dynamic formant movement: • Spectral change (Barks): ▪ BB: CL=1.88, CON=1.56, difference=0.32 ▪ NB: CL=2.27, CON=2.12, difference=0.15 • Spectral angle (radians):

- BB: CL=0.89, CON=1.04, difference=-0.15
- NB: CL=1.03, CON=1.12, difference=-0.09

B. Studies presenting descriptive data on acoustic variables in healthy speakers

Reference	Design	Qualsyst (by Kmet et al.)	Healthy population (Gender, Age, Language)	Speech sample for acoustics (target phoneme)	Acoustic parameters (Definitions)	Perceptual measure(s)	Descriptive data in healthy speakers
A10. Weismer et al., 1992	III-2	17/24 71% good	N=15 (M) Age: μ=72 (range 68-80) Language: American English	Vowels (N.R.) in 12 monosyllabic words (CV, CVC, CCVC, CVCC, VC): wax, sigh, sip, ship, sew, coat, row, cash, hail, ate, shoot, and blend.	F1-F2 formant trajectories: • Transition extent (TE): amount of frequency change along the transitional segment of a trajectory • Averaged transition rate or slope: TE/TD (TD: duration of the transitional segment) • Starting frequency (SF): onset frequency of the transitional segment	Percent-correct word identification: selection among four possible words	*Perceptual* - Percent-correct identification: N.R. *Acoustic* - Descriptive data available for all acoustic measures: for each of the 12 words, for F1 (Table 2, page 1094) and for F2 (Table 3, page 1095)
A11. Hohoff et al., 2003	III-3	20/24 83% strong	N=23 (6 M, 17 F) Age: μ=35.1 (SD=10.3; range 19.6-57.1) Language: Standard French	[s] in the word 'soleil'	Upper boundary frequency (UBF) of the fricative sound: Maximum frequency of the bandwidth, maximum greyness range in the wide-band spectrogram	5-point Likert scale (1-5: non-pathological - highly pathological [s] sound production)	Mean (SD; range): *Perceptual* - Likert scale for [s] sound: 1.43 (0.46; 1.00-3.20) *Acoustic* - UBF [s] sound: 12961.48 Hz (585.77; 11454 - 13898)
A12. Yunusova et al., 2005	III-2	20/24 83% strong	N=10 (7 M, 3 F) Age M: μ=56.9, F: μ=57.3 Language: American English (Upper Midwest dialect)	- All vocalic segments of an oral reading - Fricatives [s, ʃ] in word-initial position	- Vocalic segments: F2 interquartile ranges (IQR) for each breath group - Fricatives: First moment differences between [s]-[ʃ]	- Sentence-level: direct magnitude estimation, modulus of 100 (ease to understand) - Word-level: Percent-correct word identification	Mean (SD): *Perceptual* - Sentence intelligibility (DME): 222.38 (23.81; 175-244) - Word intelligibility (%): 98.66 (1.12; 96.6-99.9) *Acoustic* - F2 IQR (Hz): 500 (73) - 1st moment difference [s]-[ʃ] (Hz): 1.3 (0.7)
A13. de Bruijn et al., 2009	III-2	21/24 88% strong	N=18 ['gender and age matched' to study group; study group M=55%, F=45%, Age μ=53.8 (SD=8.7)] Language: Dutch	Corner vowels [a,i,u] and velar consonant [x] in words	- F1 and F2, and vowel space (VS) - Spectral slope for [x]	Ratings on intelligibility (10-point scale, 1-10: poor - good intelligibility), articulation and nasal resonance (4-point scale, 1-4: normal - deviant speech quality)	Mean (SD): *Perceptual* - Intelligibility ratings: N.R. *Acoustic* - F1 [i]: 296 Hz (49) - F2 [i]: 2325 Hz (248) - Vowel triangle size: 0.213 Hz^2 (0.11)

							- F1 and F2 [a, u]: N.R. - Spectral slope for [x]: N.R.
Reference	**Design**	**Qualsyst (by Kmet et al.)**	**Healthy population (Gender, Age, Language)**	**Speech sample for acoustics (target phoneme)**	**Acoustic parameters (Definitions)**	**Perceptual measure(s)**	**Descriptive data in healthy speakers**
A14. Van Lierde et al., 2012	III-2	20/24 83% strong	N=9 (M/F ratio N.R.) µ=47.6 (22-61) Language: Dutch	Sound [s]	2nd spectral moment: dispersion (standard deviation) of the frequencies around the centre of gravity	- Phonetic transcription - Assessment of overall speech intelligibility on a 4-point ordinal scale (0-3: normal - severely impaired)	*Perceptual* - Phonetic transcription: 'All subjects were capable of producing all Dutch vowels and consonants… the phonetic characteristics were normal' - Speech intelligibility: 'The control group had a normal speech intelligibility' *Acoustic* - 2nd spectral moment in [s]: 2081 Hz
A15. Skodda et al., 2013	III-2	21/24 88% strong	N=60 (30 M, 30 F) Age: µ=66.87 (median=67.5; SD=7.1; range 55-80) Language: German	Corner vowels [a, i, u] in words	Vowel Articulation Index: $\frac{F2i + F1a}{F1i + F1u + F2u + F2a}$	Rating on 4-point scale: intelligibility (0-3: good-poor intelligibility) and 5-point scale: articulation (0-4: normal articulation - markedly reduced intelligibility)	Mean (SD): *Perceptual* - Intelligibility rating: 0.08 (0.28) - Articulation rating: 0.07 (0.25) *Acoustic* - VAI: M=0.767 (0.058); F=0.874 (0.062)
A16. Whitfield et al., 2014	III-2	23/24 96% strong	N=10 5 M (Age: µ=65.8, range 57-73), 5 F (Age: µ=71.8, range 58-81) Language: Standard American English	All voiced segments from the first paragraph of the Rainbow Passage	Articulatory-acoustic vowel space (AAVS): square root of the generalized variance of the F1-F2 data, resulting in an elliptical space representing the average bivariate variability in F1xF2 space	Rating of the speech clarity on a 100mm visual analogue scale (0-100: unclear-very clear)	Mean (SD): *Perceptual* - Rating of speech clarity (mm): M=63mm (9.92); F=64.60mm (17.96) *Acoustic* - AAVS (kHz^2): M=38.45 (5.20); F= 64.59 (9.77)
A17. Neel et al., 2015	III-2	22/22 100% strong	N=12 (4 M, 8 F) Age: range 52-69 Language: English	- Initial [s] in the words 'sip', 'seep,' and 'see' - vowels [a,i,u] in the words 'heed', 'hod', 'who'd'	- Consonant: first spectral moment at the centre of the fricative [s] = weighted average of the spectral peak frequencies (measure of tongue placement accuracy) - Vowels: • F1 range (lowest F1 value of the 3 vowels subtracted from the highest value) • F2 range • Vowel space area (VSA)	Using a 100mm visual analogue scale (0-100: no impairment - severe impairment): overall intelligibility, articulatory precision	Mean (SD; range): *Perceptual* - Intelligibility rating: 4.4 - Articulatory precision rating: N.R. (graphics) *Acoustic* - Consonant 1st spectral moment [s] (Hz): 6962.6 (1282.6; 4700–8756) - Vowels: • F1 range (Hz): 448.9 (83.9; 286–532) • F2 range (Hz): 1552.8 (197.8; 1309–1899) • VSA (Hz^2): 334262 (98,557; 192980–526903)

Reference	Design	Qualsyst (by Kmet et al.)	Healthy population (Gender, Age, Language)	Speech sample for acoustics (target phoneme)	Acoustic parameters (Definitions)	Perceptual measure(s)	Descriptive data in healthy speakers
A18. Dwivedi et al., 2016	III-2	24/24 100% strong	N=51 (32 M, 19 F) Age: μ=54.4 (SD=9.3) Language: English	The sustained vowel [i] (mid-stable portion)	F1 and F2	London Speech Evaluation 4-point scale (0-3: normal-severe impairment): intelligibility ('auditory-perceptual impression of understandability'), articulation, overall grade	Mean (SD): *Perceptual* - London Speech Evaluation: N.R. *Acoustic* - F1: M=315.9 Hz (170.7); F=353.8 Hz (78.3) - F2: M=1782.6 (846.2); F=2111.5 (986.7)
A19. Connaghan et al., 2017	III-2	21/24 88% strong	N=15 (9 M, 6 F) Age: μ=36 (range 22-59) Language: American English	High vowels [i], [ɪ] and low vowel [æ] in stressed and in unstressed words.	- F1 and F2 - Euclidean distance (ED) between the F1xF2 vowel centroids of each vowel pair ([æ-i], [æ-ɪ], [i-ɪ]), as a measure of vowel dispersion - Probability density function (PDF): relative probability that each vowel token came from the target vowel F1 × F2 area	Percentage of correct identification for each vowel	Mean (SD): *Perceptual* - Percent-correct vowel identification: N.R. *Acoustic (average over all vowels)* *Stressed words*: - F1(Hz): 545.0 (208.4); F2(Hz): 2178.8 (336.3) - ED(Hz): 536.1 (129.6) - PDF: 0.99 (0.06) *Unstressed words*: - F1: 507.2 (169.9); F2: 2068.0 (320.1) - ED: 476.9 (113.6) - PDF: 0.87 (0.30) Data for each vowel is also available (Table 3, p.44).
A20. Fletcher et al., 2017	III-2	22/24 92% strong	N=17 (11 M, 6 F) Age: μ=66 (SD=N.R.) Language: New Zealand English	Corner vowels [aː], [iː], [oː] in words	- Vowel space area (VSA): - Formant centralization ratio (FCR): $\frac{F2[o:] + F2[a:] + F1[i:] + F1[a:]}{F2[i:] + F1[o:]}$ Both measures were made in Barks and in Hertz, at the vowel midpoint and at the point where there was the least movement in the formant tracks (= flexible point)	Ratings using visual analogue scales: Listener group 1: Intelligibility = Ease to understand the speaker (0-100: easy-difficult) Listener group 2: Speech precision (0-100: precise-imprecise)	Mean (SD): *Perceptual* - Intelligibility rating: 0.877 (0.110) - Speech precision rating: 0.987 (0.198) *Acoustic* - VSA($Bark^2$) flexible: M=10.91(2.64); F=13.87(2.64) - VSA($Bark^2$) midpoint: M=7.76(2.16); F=10.79(1.68) - VSA(Hz^2) flexible: M=243.21(69.88); F=385.82(103.11) - VSA(Hz^2) midpoint: M=174.73(55.40); F=295.13(61.61) - FCR(Bark) flexible: M=1.19(0.07); F=1.19(0.03) - FCR(Bark) midpoint: M=1.28(0.07); F=1.25(0.03) - FCR(Hz) flexible: M=0.97(0.07); F=0.92(0.03) - FCR(Hz) midpoint: M=1.06(0.08); F=0.98(0.03)

Reference	Design	Qualsyst (by Kmet et al.)	Healthy population (Gender, Age, Language)	Speech sample for acoustics (target phoneme)	Acoustic parameters (Definitions)	Perceptual measure(s)	Descriptive data in healthy speakers
A21. Kim et al., 2017	III-2	22/24 92% strong	N=24 (14 M, 10 F) 12 American English Age: median=59 (range 49-85) 12 Korean Age: median=N.R. (range 52-72)	Vowels [a,i,u] in words	Acoustic vowel space (AVS) derived from F1 and F2 frequencies at the temporal midpoint of the vowel: $AVS = \left\lvert \frac{\frac{F1i\ x\ (F2a-F2u)+}{F1a\ x\ (F2u-F2i)+}}{2} \right\rvert$	Intelligibility rating on a 10-point Equal Appearing Interval scale (1-10: totally unintelligible-completely intelligible)	*Perceptual* - Intelligibility rating: N.R. *Acoustic* - AVS: Mean in log (SD): • English talkers: 5.21 (0.09) • Korean talkers: 5.48 (0.19)
A22. Martel-Sauvageau et al., 2017	III-2	23/24 96% strong	N=8 (3 M, 5 F) Age: 'Age matched +- 2 years' (Ages N.R.) Language: Quebec French	- Glide contexts: [wa, ja, ɛj] in words - CVCV tokens in a carrier phrase, with the target vowels [i,u,a] and the consonants [b,d,g]	- Glides: F2 slopes = overall frequency change divided by the transition duration - CVCV tokens: • Locus equations (LE): linear regression function using F2 vowel onset and F2 midpoint; for [b, d, g] $F2_{onset} = k \times F2_{mid} + c$ (k and c: constants) • LE distinctiveness: Distinctiveness between LE of the three places of articulation, measured using the constant parameters (k, c) of the equations as dimensions of a triangular locus space. The area of this space is then calculated using Euclidean distances between [b]-[d]-[g] coordinates	Overall speech intelligibility rated on a 229mm visual analogue scale (0-229: understood none - understood all')	Mean (SD): *Perceptual* - Overall speech intelligibility: N.R. (graphics) *Acoustic* - F2 slopes (Hz/ms): • [wa]: 12.95(1.97) • [ja]: -5.78(1.53) • [ɛj]: 5.03(1.06) - Locus equations: • [b]: slope=0.67, intercept=506.6 • [d]: slope=0.29, intercept=1331.2 • [g]: slope=1.04, intercept=134.2 - LE distinctiveness: 0.05

[1]The study designs are reported according to the NHMRC hierarchy: Level 1 Systematic reviews; Level II Randomized control trials; Level III–1 Pseudo-randomized control trials; Level III–2 Comparative studies with concurrent controls and allocation not randomized (cohort studies), case control studies, or interrupted time series with a control group; Level III–3 Comparative studies with historical control, two or more single-arm studies, or interrupted time series without a control group; Level IV Case series.
[2]The QualSyst methodological quality score interpretation guidelines are: strong > 80%; good 60–79%; adequate 50–59%; poor < 50%.

APPENDIX B – Definitions and formulas (if applicable) of the acoustic measures used in the studies of this review

Vowel measures

Steady-state formant measures

(1) Vowel Space Area (Fletcher et al., 2017): the first and second formant values of the corner vowels of the investigated language are used as coordinates in an F1/F2 space to construct a vowel triangle or quadrilateral. The area of the resulting triangle or quadrilateral are then computed using classic formulas such as:

$$Hz^2 = 0.5\ x\ |F1[v_1]x(F2[v_2] - F2[v_3]) + F1[v_3]x(F2[v_1] - F2[v_2]) + F1[v_2]x(F2[v_3] - F2[v_1])|$$

(where v1, v2 and v3 are the corner vowels of the vowel triangle)

(2) Articulatory–Acoustic Vowel Space (A05, A16): 'This space is calculated as the square root of the generalized variance of all sampled vowel formants in the F1–F2 co-ordinate plot. The generalized variance for the AAVS is calculated as the product of the variance of the F1 data, the variance of the F2 data, and the portion of the unshared variance between them. The square root of the generalized variance provides a measure of formant variability that is the equivalent to a bivariate standard deviation in F1–F2 space. Therefore, an increase in the range or spread of F1 or F2 values in an utterance would yield a larger AAVS.' (Whitfield & Goberman, 2017)

(3) Steady-state F1 and F2 measures: the first and second formants are extracted, usually at temporal midpoint. They can then be compared for example between vowels, or between speaker groups.

(4) F1 and F2 ranges: subtraction of the lowest F1/F2 value from the highest

(5) F0-F1 difference (A08): Euclidean distance between the fundamental frequency and the first formant

(6) F1-F2 difference (A02, A08): Euclidean distance between the first and second formants

(7) Euclidean distance between vowel pairs in the F1xF2 space (A03, A19): Euclidean distance between the F1xF2 vowel centroids of each vowel pair, as a measure of vowel dispersion

(8) Vowel Articulation Index (A15): a 'surrogate parameter of the first and second formant frequencies ($F1$ and $F2$) of the three corner vowels /α/, /i/, and /u/':

$$VAI = \frac{F2[i] + F1[a]}{F1[i] + F1[u] + F2[u] + F2[a]}$$

(9) Formant Centralization Ratio (A20): the VAI's reciprocal value (Skodda, Grönheit, & Schlegel, 2012), a measure that 'weighs formants that are likely to increase as a result of vowel centralization against formants that are expected to lower':

$$FCR = \frac{\mathrm{F2[o:]} + \mathrm{F2[ɐ:]} + \mathrm{F1[i:]} + \mathrm{F1[ɐ:]}}{\mathrm{F2[i:]} + \mathrm{F1[o:]}}$$

(using New Zealand English corner vowels)

(10) Probability Density Function (A19): relative probability that a vowel token came from the target vowel F1xF2 area

(11) Onset Frequency (A10): the starting frequency (in Hertz) of the transitional segment (see 18.)

Dynamic formant measures

(12) Spectral Change (A09): 'the sum, in Barks, of the absolute formant frequency shift for F1 and F2. Thus, λ is calculated as

$$\lambda\ (Barks) = |F1_{80} - F1_{20}| + |F2_{80} - F2_{20}|$$

where $F1_{20}$, $F1_{80}$, $F2_{20}$, and $F2_{80}$ are the F1 and F2 values in Barks at 20% and 80% of the vowel duration.'

(13) Spectral Angle (A09): The spectral angle (or tilt) is computed for each vowel by comparing both F1 and F2 at 80% of the vowel duration to the frequency measured at 20% of the vowel duration. The angle θ in radians for each formant *n* is first computed as the arctangent of the difference between the frequency of the formant at 80% and 20% of the vowel duration, divided by the duration separating these two points scaled to deciseconds. The spectral angle is the sum in radians of the absolute values of the two formant angles:

$$\theta\ (\text{radians}) = \left|\arctan\left(\frac{F1_{80} - F1_{20}}{\frac{time_{80} - time_{20}}{100}}\right)\right| + \left|\arctan\left(\frac{F2_{80} - F2_{20}}{\frac{time_{80} - time_{20}}{100}}\right)\right|$$

where $F1_{20}$, $F1_{80}$, $F2_{20}$, and $F2_{80}$ are the F1 and F2 values in Barks at 20% and 80% of the vowel duration

(14) Mean Formant Movement across vowels (A03): For each vowel, the sum of 'the Euclidean distance in the F1×F2 bark space from the vowel onset (20% of vowel duration) to the steady state […] and the Euclidean distance from the vowel steady state to the offset (80% of vowel duration)' is computed. These distances are then averaged across the different vowels for each speaker.

$$Mean\ formant\ movement(Barks) =$$

$$\sqrt{(F1_{50} - F1_{20})^2 + (F2_{50} - F2_{20})^2} + \sqrt{(F1_{80} - F1_{50})^2 + (F2_{80} - F2_{50})^2}$$

where $F1_{20}$, $F1_{50}$, $F1_{80}$, $F2_{20}$, $F2_{50}$ and $F2_{80}$ are the F1 and F2 values in Barks at 20%, 50% and 80% of the vowel duration.'

(15) Dynamic Ratio (A03): a composite indicator based on dynamic measures; distinctiveness in Barks among vowels with dynamic and static trajectories; average Euclidean distance (from vowel onsets to steady states to offsets in the F1 x F2 bark space, see 13.) covered by the three most dynamic vowels ([æ, ʌ, ʊ]) divided by the distance covered by the three most static ones ([i, ɛ, u])

(16) Vector Length (A04): the Euclidean distance in the F1×F2 space from the vowel onset (20% of vowel duration) to the offset (80% of vowel duration):

$$VL\ (Barks) = \sqrt{(F1_{80} - F1_{20})^2 + (F2_{80} - F2_{20})^2}$$

where $F1_{20}$, $F1_{80}$, $F2_{20}$ and $F2_{80}$ are the F1 and F2 values in Barks at 20% and 80% of the vowel duration

(17) Trajectory Length (A04): the sum in Barks of the four Euclidean distances between the vowel sections 20%-35%, 35%-50%, 50%-65% and 65%-80%:

$$vowel\ section\ length\ VSL_n\ (Barks) = \sqrt{(F1_n - F1_{n+1})^2 + (F2_n - F2_{n+1})^2}$$

$$trajectory\ length\ TL\ (Barks) = \sum_{n=1}^{4} VSL_n$$

(18) Transition Extent (A10): the amount of frequency change (in Hertz) along the transitional segment of a trajectory. The onset and offset of this segment are 'the first and last time-frequency pairs, respectively, for which the following 20-ms increment [is] associated with at least a 20-Hz change'.

(19) Transition Rate or slope (A10): the division of the transition extent (in Hertz) by the duration (in ms) of the transitional segment

Glide measure

(1) F2 slope (A20): the overall frequency shift in Hertz (transition extent) in a glide, divided by the transition duration (in ms), as a measure of the rate of phonatory tract modification

Consonant measures

(1) Spectral moments: when using the spectral moment analysis, the consonant spectrum (in fricatives or plosives) is considered as a statistical distribution, which can be

described by four measures: the first moment (centre of gravity) is the 'frequency that divides the spectrum into two halves such that the amount of energy in the higher frequency regions is equal to that in the lower frequency region' (Yoon, 2015). The second moment (standard deviation) measures the dispersion of the spectral energy around this centre of gravity. The third spectral moment (SKEW) refers to the asymmetry of the energy distribution with respect to the mean, e.g. a skewness of 0 is measured in symmetrical distributions, while positive values indicate that the distribution is skewed to the right, i.e., the right tail extends further than the left one (Jongman et al., 2000). The fourth moment (KURT) is a measure of peakedness.

(2) Fricative spectral peak of the [s]-sound (A06): determined from short-term LPC spectra at 30 ms prior to the fricative offset, anticipating F2 of the vowels [i] and [u]

(3) Upper Boundary Frequency (A11): the highest frequency of the friction noise, the greyest range in the wide-band spectrogram

(4) Spectral slope for the fricative [x] (B04): a measure of the decline of the spectral energy from the low to the high frequencies in the spectrum, computed by linear regression (Peeters, 2004)

(5) F1 offset frequency (A07): first formant frequency measured in the final 45 ms of the vowels [i, ɪ, ɛ, æ] before the plosives [t] and [d]

(6) Locus Equations (A22): An 'alternative acoustic metric for characterizing segmental transition characteristics', as the linear regression function using F2 at vowel onset and F2 at midpoint:

$$F2_{onset} = k \times F2_{mid} + c$$

(7) Locus Equation distinctiveness (A22): the distinctiveness between locus equations corresponding to the three places of articulation [b, d, g], measured using the

constant parameters *k* and *c* of the equations as dimensions of a triangular locus space. The area of this space is then calculated using Euclidean distances between [b]-[d]-[g] coordinates.

APPENDIX C – Attempted cross-comparison of acoustic results

The measures that can be compared across studies are mainly vowel acoustics, except for the centroid frequency (first spectral moment) on the fricative [s] used in studies A06 and A17. In the other papers investigating consonants, the incompletely reported data for the control groups, the various measures and the different methodologies do not allow for a comparative analysis.

An attempt to compare the results of similar acoustic measures used in the different studies is shown in Table C.1. It can be observed that even if several studies use the same measure, the study population, the phonemic sample, the computing method and the reporting of the results are very different and sometimes not reported (according to the aim of each study), which makes it difficult to relate the resulting values. If we look at steady-state first and second formant measures, for example, study A03 uses the Bark scale, whereas studies A13, A18 and A19 use Hertz. The formant extraction method is only reported in study A18. However, formant values may differ depending on the extraction method (e.g. linear predictive coding, Fast Fourier Transform, cepstral analysis), and on extraction/analysis parameters (such as window type, frame size, time step and parameters specific to each method) (Derdemezis et al., 2016; Eringis & Tamulevičius, 2014). Hence, this lack of information does not allow for the replication of the study's methodology, nor for comparative analyses. Moreover, study A13 was carried out on Dutch samples, while studies A18 and A19 used English samples, which can have an impact on the vowel pronunciation. Also, the study population in study A19 is almost 20 years younger than the ones of studies A13 and A18.

Furthermore, for the AVS measures, incoherence is noticed in the units used by the different authors: four studies use squared Hertz units, but the values are very dissimilar. Study A13 yields values lower than 1Hz2, whereas study A17 shows values above 30000Hz2,

and studies A20 and A21 report values between 150 and 300Hz2. For the AAVS values, despite the fact that the main author is the same in both papers: study A05 uses kHz, while study A16 used kHz2, nonetheless both report values between 25 and 65. This underlines the necessity to be precise when describing and reporting acoustic measurements.

Table C.1 Acoustic measures that have been used in different studies and their results for comparison purposes

Measure	Study	Result	Unit	Sample	Extraction	Language	Age	N
F1 & F2	A03	F1: M=5.04(0.20); W=5.88(0.30) F2: M=13.05(0.37); W=14.70(0.53)	Bark	[i, ɪ, e, ɛ, æ, ɑ, ʌ, o, ʊ, u] (pooled)	?	American English (Michigan/Upper Midwest dialect)	NR	N=93 (45 M, 48 W)
	A13	F1: 296(49) F2: 2325(248)	Hz	[i]	?	Dutch	Matched to study group: μ=53.8 (SD=8.7)	N=18 (Matched to study group: M=55%, W=45%)
	A18	F1: M=315.9(170.7); W=353.8(78.3) F2: M=1782.6(846.2); W=2111.5(986.7)	Hz	[i]	LPC	English	μ=54.4 (SD=9.3)	N=51 (32 M, 19 W)
	A19	F1[i]: stressed= 345.7(47.2) unstressed= 350.8(50.3) F2[i]: stressed= 2508.4(243.0) unstressed= 2377.3(239.5)	Hz	[i], [ɪ], [æ]	? (30 ms window)	American English	Age: μ=36 (range 22-59)	N=15 (9 M, 6 W)
F1 & F2 range	A03	F1 range: M=3.83(0.59); W=4.32(0.80) F2 range: M=9.37(1.04); W=11.15(1.06)	Bark	[i, ɪ, e, ɛ, æ, ɑ, ʌ, o, ʊ, u] (pooled)	?	American English (Michigan/Upper Midwest dialect)	NR	N=93 (45 M, 48 W)
	A17	F1 range: 448.9 (83.9; 286–532) F2 range:	Hz	[i, ɑ, u]	?	English	range 52-69	N=12 (4 M, 8 W)

		1552.8 (197.8; 1309–1899)						
VSA	A03	M=18.57 (4.13); W=25.07 (6.55)	?	quadrilateral [i, ɑ, u, æ]	?	American English (Michigan/Upper Midwest dialect)	NR	N=93 (45 M, 48 W)
	A05	M=200.81(23.65); W=577.74(94.11)	kHz	quadrilateral [i, ɑ, u, æ]	LPC (Burg method, window length= 50 ms; time step= 1 ms)	Standard American English	M: μ=24.40, range 20-36 F: μ=24.30, range 18-29	N=10 (5 M, 5 W)
	A13	0.213(0.11)	Hz^2	triangle [i, ɑ, u]	?	Dutch	Matched to study group: μ=53.8 (SD=8.7)	N=18 (Matched to study group: M=55%, W=45%)
	A17	334262 (98,557; 1929&80–526903)	Hz^2	triangle [i, ɑ, u]	?	English	range 52-69	N=12 (4 M, 8 W)
	A20	flexible point: M=10.91(2.64); W=13.87(2.64) temporal midpoint: M=7.76(2.16); W=10.79(1.68)	$Bark^2$	triangle [aː, iː, oː]	LPC (Burg method, window length= 25 ms, time step= 6.25 ms)	New Zealand English	μ=66	N=17 (11 M, 6 W)
		flexible point: M=243.21(69.88); W=385.82(103.11) temporal midpoint: M=174.73(55.40); W=295.13(61.61)	Hz^2					
	A21	English talkers: 5.21 (0.09) Korean talkers: 5.48 (0.19)	Log (Hz^2)	triangle [i, ɑ, u]	?	American English Korean	English: median=59 (range 49-85) Korean: median=N.R. (range 52-72)	N=24 (14 M, 10 W) 12 English 12 Korean
AAVS	A05	Conversational: M=27.98(5.06); W=68.83(6.86) Clear: M=35.37(8.13); W=93.81(20.21)	kHz	quadrilateral [i, ɑ, u, æ]	LPC (Burg method, window length= 50 ms; time step= 1 ms)	Standard American English	M: μ=24.40, range 20-36 F: μ=24.30, range 18-29	N=10 (5 M, 5 W)
	A16	M=38.45 (5.20); W= 64.59 (9.77)	kHz^2	All voiced segments from the 1st paragraph of the Rainbow Passage	LPC (Burg method, window length= 50 ms; time step= 1 ms)	American English	M: μ=65.8, range 57-73 F: μ=71.8, range 58-81	N=10 (5 M, 5 W)
Centroid frequency	A06	At 30 ms prior to fricative offset: [si]=5524; [su]=5134	Hz	[s] in [s u] and [s i]	DFT (Window length= 20 ms; 30 ms and 100 ms prior to fricative offset)	English	μ=32 (SD=6.7; range 26-45)	N=10 (5 M, 5 W)

	At 100 ms prior to fricative offset: [si]=6806; [su]=6182						
A17	6962.6 (1282.6; 4700–8756)	Hz	Initial [s] in the words ‘sip’, ‘seep,’ and ‘see’	? (Window length= 20 ms; at the centre of the fricative)	English	range 52-69	N=12 (4 M, 8 W)

Abbreviations: M = men; W = women; F1/F2 = first and second formant; VSA = vowel space area; AAVS = articulatory-acoustic vowel space; LPC = linear predictive coding; DFT = discrete Fourier transform; NR = not reported; μ = mean

Table 1. Search strategy for the two databases

Database	Search Terms (subject headings and free text words)	Number of Records
Embase:	((speech intelligibility/) OR (Intelligibil*.ab. OR Intelligibil*.ti. OR comprehensibil*.ab. OR comprehensibil*.ti. OR understandabil*.ab. OR understandabil*.ti.)) AND (acoustics/ OR speech analysis/ OR acoustic analysis/ OR sound analysis/ OR phonetics/ OR signal processing/ OR fourier analysis/ OR sound detection/ OR sound/ OR frequency/ OR frequency analysis/ OR pitch/ OR noise/ OR signal noise ratio/)	3326
PubMed:	(("Speech Intelligibility"[Mesh]) OR (intelligibil*[Title/Abstract] OR comprehensibil*[Title/Abstract] OR understandabil*[Title/Abstract])) AND ("Acoustics"[Mesh] OR "Speech Acoustics"[Mesh] OR "Speech Production Measurement"[Mesh] OR "Phonetics"[Mesh] OR "Signal Processing, Computer-Assisted"[Mesh] OR "Fourier Analysis"[Mesh] OR "Sound Spectrography"[Mesh] OR "Sound"[Mesh] OR "Signal-To-Noise Ratio"[Mesh] OR "Noise"[Mesh])	3393
		Total: 6719
		Total after exclusion of duplicates: 4818

Table 2. Methodological quality ratings for the 22 included articles using the Qualsyst critical appraisal tool by Kmet et al. and level of evidence according to the National Health and Medical Research Council (NHMRC) hierarchy

Reference	**Qualsyst score[1] (%)**	**Methodology quality**	**NHMRC Level of Evidence[2]**
A01. McRae et al., 2002	20/24 (83)	Strong	III-2
A02. Hazan et al., 2004	21/24 (88)	Strong	III-3
A03. Neel, 2008	18/22 (82)	Strong	III-3
A04. Ferguson et al., 2014	20/22 (91)	Strong	III-3
A05. Whitfield et al., 2017	21/24 (88)	Strong	III-3
A06. Katz et al., 1991	20/24 (83)	Strong	III-3
A07. Flege et al., 1992	21/24 (88)	Strong	III-3
A08. Bunton et al., 2001	21/24 (88)	Strong	III-2
A09. Ferguson et al., 2007	20/24 (83)	Strong	III-3
A10. Weismer et al., 1992	17/24 (71)	Good	III-2
A11. Hohoff et al., 2003	20/24 (83)	Strong	III-3
A12. Yunusova et al., 2005	20/24 (83)	Strong	III-2
A13. de Bruijn et al., 2009	21/24 (88)	Strong	III-2
A14. Van Lierde et al., 2012	20/24 (83)	Strong	III-2
A15. Skodda et al., 2013	21/24 (88)	Strong	III-2
A16. Whitfield et al., 2014	23/24 (96)	Strong	III-2
A17. Neel et al., 2015	22/22 (100)	Strong	III-2
A18. Dwivedi et al., 2016	24/24 (100)	Strong	III-2
A19. Connaghan et al., 2017	21/24 (88)	Strong	III-2
A20. Fletcher et al., 2017	22/24 (92)	Strong	III-2
A21. Kim et al., 2017	22/24 (92)	Strong	III-2
A22. Martel-Sauvageau et al., 2017	23/24 (96)	Strong	III-2

[1] Methodological quality: strong > 80%; good 60–79%; appropriate 50–59%; poor < 50%.

[2] NHMRC hierarchy: Level I Systematic reviews; Level II Randomized control trials; Level III–1 Pseudo-randomized control trials; Level III–2 Comparative studies with concurrent controls and allocation not randomized (cohort studies), case control studies, or interrupted time series with a control group; Level III–3 Comparative studies with historical control, two or more single-arm studies, or interrupted time series without a control group; Level IV Case series.

Note: The studies were ordered according to 1) the type of outcome: A01-A05 = direct correlation between acoustics and perceptual ratings; A06-A09 = indirect investigation of the link between acoustics and perceptual ratings; A10-A22: quantitative data for both acoustics and perceptual ratings, without investigation of the link; 2) the chronological order

Table 3. Significant and non-significant correlations between acoustic measures and perceptual ratings of speech

	Vowels														Consonants
	F1	F2	F1 range	F2 range	Euclidean distance F1-F2	F1≠ [i-æ]	F2≠ [i-u]	Vowel distance	VSA	AAVS	Formant movement	Dynamic ratio	Vector Length	Trajectory length	1st moment
DME									∅ (A01)						∅ (A01)
Likert	∅ (A02)	∅ (A02)			∅ (A02)	∅ (A02)	✓ (A02)								
VAS										✓ (A05)					
%corr	∅ (A02,A03) ✓ (A04)	∅ (A02, A03) ✓ (A04)	✓ (M, A03) ∅ (W, A03)	∅ (A03)	∅ (A02)	∅ (A02)	✓ (A02)	∅ (A03)	✓ (A03)		✓ (W, A03) ∅ (M, A03)	∅ (A03)	✓ (A04)	∅ (A04)	

Note: ✓: significant correlation; ∅: non-significant correlation; M: men; W: women

Abbreviations: F1/F2 = first and second formant; F1≠ [x,y] = F1 difference between vowels x and y; AAVS = articulatory-acoustic vowel space; DME = direct magnitude estimation; Likert: Likert-type equal-appearing interval scale; VAS = visual analogue scale; %corr = percent correct identification score

Figure 1. Flow diagram illustrating the selection process according to the PRISMA guidelines. Adapted from Moher et al. (2009)