Journal of Phonetics 97 (2023) 101223

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Research Article

Discriminative segmental cues to vowel height and consonantal place and voicing in whispered speech

Luis M.T. Jesus [a,*], Sara Castilho [b], Aníbal Ferreira [c], Maria Conceição Costa [d]

[a] School of Health Sciences (ESSUA), Institute of Electronics and Informatics Engineering of Aveiro (IEETA), Intelligent Systems Associate Laboratory (LASI), University of Aveiro, Aveiro, Portugal
[b] Hospital Arcebispo João Crisóstomo, Cantanhede, Portugal
[c] Department of Electrical and Computer Engineering, University of Porto, Portugal
[d] Department of Mathematics (DMat) and Centre of Research and Development in Mathematics and Applications (CIDMA), University of Aveiro, Portugal

ARTICLE INFO

Article history:
Received 30 July 2021
Received in revised form 27 December 2022
Accepted 22 January 2023
Available online 28 February 2023

Keywords:
Speech production
Acoustic phonetics
Whispered speech
Vowels
Fricatives

ABSTRACT

Purpose: The acoustic signal attributes of whispered speech potentially carry sufficiently distinct information to define vowel spaces and to disambiguate consonant place and voicing, but what these attributes are and the underlying production mechanisms are not fully known. The purpose of this study was to define segmental cues to place and voicing of vowels and sibilant fricatives and to develop an articulatory interpretation of acoustic data.
Method: Seventeen speakers produced sustained sibilants and oral vowels, disyllabic words, sentences and read a phonetically balanced text. All the tasks were repeated in voiced and whispered speech, and the sound source and filter analysed using the following parameters: Fundamental frequency, spectral peak frequencies and levels, spectral slopes, sound pressure level and durations. Logistic linear mixed-effects models were developed to understand what acoustic signal attributes carry sufficiently distinct information to disambiguate /i, a/ and /s, ʃ/.
Results: Vowels were produced with significantly different spectral slope, sound pressure level, first and second formant frequencies in voiced and whispered speech. The low frequencies spectral slope of voiced sibilants was significantly different between whispered and voiced speech. The odds of choosing /a/ instead of /i/ were estimated to be lower for whispered speech when compared to voiced speech. Fricatives' broad peak frequency was statistically significant when discriminating between /s/ and /ʃ/.
Conclusions: First formant frequency and relative duration of vowels are consistently used as height cues, and spectral slope and broad peak frequency are attributes associated with consonantal place of articulation. The relative duration of same-place voiceless fricatives was higher than voiced fricatives both in voiced and whispered speech. The evidence presented in this paper can be used to restore voiced speech signals, and to inform rehabilitation strategies that can safely explore the production mechanisms of whispering.

1. Introduction

Whispered speech is used by speakers with normal laryngeal status to communicate across various languages and cultures, and can be used for quiet and private communication, to mediate tenderness and support social bonding (Cirillo & Todt, 2005). However, whispered speech is not used on a regular basis because it does not "carry the communication effectiveness that normal voice allows" (Boone et al., 2020, p. 5), so there are no large corpora of production data from a wide range of whispered speech tasks.

Some individuals with voice impairments use whispered speech to communicate and can still express complex linguistic information, but they usually do so with reduced intelligibility / understandability of speech, as well as loss of individual voice traits and sonority properties (Boone et al., 2020). Voice disorders such as vocal fold paralysis result in a voice quality that can be described as weak and breathy whispered speech (MacDonell & Holmes, 2007), and functional voice disorders, more specifically, a psychogenic voice disorder such as aphonia results in a normal or tight (sharp and strained) whispered speech (Baker, 2016; Boone et al., 2020; MacDonell & Holmes, 2007). Individuals with voice impairments feel limited

[*] Corresponding author at: Universidade de Aveiro, Campus Universitário de Santiago, 3810-193 Aveiro, Portugal.
E-mail address: lmt@ua.pt (L.M.T. Jesus).

https://doi.org/10.1016/j.wocn.2023.101223
0095-4470/© 2023 The Author(s). Published by Elsevier Ltd.

in different voice activities such as speaking and having a conversation, with consequences for their social integration and socialisation (Ma et al., 2007; Mertl et al., 2018). Whispering research is pertinent to create alternatives to restore the speech signal of people with vocal fold paralysis or aphonia, improving their social participation and their quality of life. This has motivated the current study and the development of an assistive technology (Silva et al., 2021) whose objective is to reconstruct natural speech sounds from whispered speech in real time (Oliveira, 2022), so to facilitate effective and comfortable communication by patients while using their speech production system seamlessly.

Also, some of the production mechanisms involved in whispering are not fully understood. Some aphonic patients who produce intelligible whispered speech (Boone et al., 2020, p. 81) use laryngeal hyperfunction. Although these patients use whispered speech, there is not sufficient experimental evidence "to state unequivocally that whispering is not harmful" (Colton et al., 2011, p. 357). In contrast, there is evidence that whispered speech can be helpful in improving voice quality and be safely used as part of voice rehabilitation (Scherer et al., 2016).

1.1. Production mechanisms of whispered speech

The physiology and physics of whispered speech have been the focus of various studies (Benninger et al., 1988; Fleischer et al., 2007; Hufnagle & Hufnagle, 1983; Monoson & Zemlin, 1984; Rubin et al., 2006; Solomon et al., 1989; Stathopoulos et al., 1991; Sundberg et al., 2010; Swerdlin et al., 2010; Tsunoda et al., 1994) mostly discussing, over the last 40 years, the role of laryngeal mechanisms and configurations in vocal fold tissue or supraglottic structures compression (Rubin et al., 2006). See Appendix A for a list of publications and details about the studies. Tables A.1, A.2, A.3 and A.4 are not exhaustive regarding studies prior to 1980 but these can be reviewed in Monoson and Zemlin (1984).

Whispered speech is produced with a partially adducted glottis, the vocal folds do not vibrate, and there is possibly some narrowing of the vocal tract around the ventricular folds and raising of the larynx (Matsuda & Kasuya, 1999). Sundberg et al. (2010) described four types of whispering: Three distinct modes, according to laryngeal adduction levels (hypofunctional, hyperfunctional and neutral) and one directly after a brief phonation. In this study we will be analysing neutral whispering, as perceptually assessed by a voice specialist.

Evidence from whispered speech has shown that the posterior cricoarytenoid (PCA) muscle has a central role in spreading the glottis for voiceless speech segments, that during whispered speech the PCA still contributes to the distinction between voiceless and voiced consonants, and that the thyropharyngeus (TP) muscle activity contributes to the supraglottic constriction adjustment (impacting the first and second formant frequencies). There was also a positive correlation (similar activity) between glottal (PCA activity) and supralaryngeal (TP activity) adjustments, suggesting that turbulence is generated between the glottis and a constriction above (Tsunoda et al., 1994).

Whispered speech is acoustically and aerodynamically different from voiced speech (Scherer et al., 2016). In whispered speech there are various spectral differences compared to voiced speech, such as wider bandwidth / less peaky spectral structure, loss of energy at low frequencies, a flattening of high frequencies, lowering of speech rate and intensity, and lengthening of syllables or other segments (Jovičić & Šarić, 2008; Meynadier & Gaydina, 2013; Zhang & Hansen, 2007). The sound source in whispered speech is a broad-band noise source generated by the exhaled air passing through a constriction, causing turbulent aperiodic airflow (Sharifzadeh et al., 2012; Sundberg et al., 2010). This noise source excites the vocal tract resonators, so the spectra of whispered vowels show amplitude peaks that are wider but comparable in magnitude to voiced vowels, and it is expected that formant frequency differences across vowels will be maintained (Maurer, 2016; Sundberg et al., 2010). Zhang and Hansen (2007) have previously used a regression line to compute the slope of the inverse filtered (Hansen, 1989) glottal source spectrum from 1000 Hz to 4000 Hz (Hansen & Varadarajan, 2009, p. 370) showing that the spectral slope of whispered speech is lower (less steep) than that of voiced speech.

1.2. Previous whispered speech production studies

Studies have focused on whispered speech production by speakers of 10 different languages (Arabic, Dutch, English, French, German, Japanese, Polish, Serbian, Spanish and Swedish), reporting a wide range of data for various phonemes. See Appendix A for a full list of publications and details about the studies. Most studies have only analysed audio signals but some have also focused on airflow and/or air pressure (Konnai et al., 2017; Meynadier, 2015; Meynadier & Gaydina, 2013; Murry & Brown, 1976; Schwartz, 1972; Stathopoulos et al., 1991; Weismer & Longstreth, 1980) during whispered speech production. However, a larger number of studies have only analysed sustained phones, isolated phones and syllables or nonsense words, and most speech production studies have recruited a limited number (1 to 12) of heterogenous speakers. The only exceptions were Kallail and Emanuel's (1984) study of 20 speakers that produced sustained /i, æ, ɑ, ʌ, u/, Schwarz's (1972) analysis of air pressure in syllables with /i, æ, ɑ, u, p/, and a previous study (Zygis et al., 2017) of 16 Polish speakers focusing on segmental cues to intonation.

Acoustic studies have shown that there are differences in formant patterns of voiced and whispered vowels: For example, the first and second formant frequencies are higher in whispered speech than in voiced speech (Maurer, 2016; Swerdlin et al., 2010). Matsuda and Kasuya (1999) have shown that electrical circuit speech production models that do not introduce a significant zero in the transfer function, incorporating weak acoustic coupling effects between the subglottal system and a constriction between the false vocal folds, can simulate the raising of the frequency of lower formants observed in whispered speech. Furthermore, Sharifzadeh et al. (2012) found that whispered /ə/ and /ʌ/ formant frequency shifts from voiced reference values were more pronounced than for other vowels. In whispered vowels there was also more convergence of adjacent vowels, for example, /i/ and /ɪ/ first formant (F_1) and second formant (F_2) frequency values were more similar in whispered speech than in voiced speech (Sharifzadeh et al., 2012).

Duration and fundamental frequency (f_o) are also used as complementary (to formant frequencies) features to discriminate vowels (Heeren, 2015). Intrinsic f_o has been shown (Jacewicz & Fox, 2015) to be positively correlated to vowel height, a phenomenon that plays out across more than 30 languages (Whalen & Levitt, 1995). Open vowels have been shown to be longer than close vowels, and height-related vowel duration differences are used in different languages as a secondary feature to enhance contrast (Cho, 2015). Vowels' intrinsic duration has also been shown to be conditioned by physiological factors (Holt et al., 2015): Vowels that are produced with a low jaw are longer than those produced with high jaw position. Durational differences have also been observed in consonants of different languages: Voiced obstruents are typically shorter than voiceless obstruent with the same place of articulation (Pape & Jesus, 2015; Smith, 1997; Stevens et al., 1992).

Jovičić and Šarić (2008) studied the acoustics of whispered consonants, measuring duration and intensity. Results revealed that whispered phonologically voiced consonants were longer and had lower intensity than their voiced cognates (reduced in intensity as much as 25 dB), but phonologically voiceless consonants were produced with almost unchanged intensity.

Heeren (2015) studied whispered /f/ and /s/ with different pitch targets for intonational (as opposed to tonal) purposes and compared them to the acoustic characteristics of fricatives produced in voiced speech, searching for compensatory cues to pitch. She found that there was no difference between normal and whispered speech in the relative duration of the fricatives, and that the fricatives' intensity was lower in whispered speech. She concluded that acoustic correlates of pitch targets were of a secondary nature. Heeren (2015) also estimated the centre of gravity (CoG) of time-averaged spectra over the whole duration of fricatives, concluding that the CoG was systematically lower for whispered speech than voiced speech.

Zygis et al. (2017) have shown that spectral features of fricatives are used as segmental cues to intonation both in voiced and whispered speech. They used the CoG, its standard deviation, skewness and kurtosis, two spectral regression slopes (m1 – the slope of the spectral regression line for the frequency range between 0.5 and 3 kHz; m2 – the slope of the spectral regression line for the range between 3 and 11 kHz); and the frequency of the highest spectral peak of the frication noise in the range of 2–4 kHz.

According to Jesus and Shadle (2002, pp. 447–449), the broad peak for fricatives with a localised source (which is the case of the sibilant fricatives that we have analysed in this paper) will interact with m1; the predicted effects are that an increase in m1 will be correlated with either a more posterior place or greater source strength, because a more posterior place (or rounding) lowers the frequency of the broad peak (F_{BP}) and a greater source strength increases the broad peak level (L_{BP}). The predicted effects of a higher source strength on m2 are that it will became less negative (high frequencies spectral slope will increase).

There have been some studies focusing on the laryngeal articulation used to contrast phonologically voiced with phonologically voiceless obstruents in whispered speech. To the best of our knowledge, Slis and Cohen (1969a, 1969b) were the first to explore the acoustical differences between voiced and whispered productions of /p, b/, /t, d/, /f, v/ and /s, z/ pairs, and to propose an articulatory model (Slis & Cohen, 1969b) supporting the view that the action of the pharyngeal constrictors differs in voiced vs voiceless pairs in both speech modes. Their "elaborate" model of articulation (Slis & Cohen, 1969b) was mostly based on what the literature, at the time, allowed them to identify distinct pharyngeal constrictor muscle, pharyngeal wall and laryngeal attributes in voiced/voiceless pairs. Slis and Cohen (1969b) concluded that the voiced-voiceless distinction works similarly in voiced and whispered speech.

The voiced/voiceless contrast in whispered obstruents has also been studied in various aerodynamic studies, such as Murry and Brown's (1976) analysis of intraoral pressure, finding that the peak values for /p, b/ and /t, d/, were not significantly different in voiced and whispered speech. Weismer and Longstreth (1980) also analysed /pa/ and /ba/ syllables produced by American English (AE) speakers in a carrier sentence, with results showing the intraoral peak pressure was not significantly different in the voiced and whispered speech but the oral peak airflow at stop release was significantly lower in whispered than in voiced speech.

More recently, Meynadier and Gaydina (2013) observed that in French /p, b, f, v, s, z, ʃ, ʒ/, the [±voice] feature is associated with distinct glottal constrictions, even when the vocal folds do not vibrate (Meynadier, 2015). Voicing assimilation in /sb/ Spanish consonant clusters has also been shown to be maintained in whispered speech (Kohlberger & Strycharczuk, 2015).

1.3. This study's aims

The current study takes a fresh look at how speakers cope with the partial information available in whispered speech. We aim to compare the characteristics of voiced and whispered speech in different speech tasks, produced by a large number of speakers (when compared to previous studies), from the same dialectal region (in Portugal) and age group (young adults to control for voice changes with age). Acoustic differences between sustained vowels /i, a, ɔ, u/ and fricatives /s, z, ʃ, ʒ/, and the same segments in isolated words, sentences and a phonetically balanced text, were analysed. Various tasks have been used in previous studies (see Appendix A for a list of speech tasks), so in order to be able to compare these with our results, a range of speech tasks has been explored. The vowels were chosen to cover the European Portuguese (EP) vowel space previously described by Escudero et al. (2009). The sibilants /s, z, ʃ, ʒ/ were studied in search of changes/adaptations in place of articulation and voicing cues in different tasks.

We envision that the evidence discussed in this paper will impact the way whispered speech is used by voice specialists in clinical practice (Colton et al., 2011), and improve reconstructed voiced speech from whispered speech samples (Konno et al., 2016; Morris & Clements, 2002; Sharifzadeh et al., 2010) and automatic whispered speech recognition accuracy (Ito et al., 2005; Marković et al., 2013; Zhang & Hansen, 2007; Zhou et al., 2019). We have therefore collected data that will help define whispered speech production models based on the Source-Filter Theory of Speech Production

(Fant, 1970) and control acoustic models of noise sources (Narayanan & Alwan, 2000).

2. Research questions

This study's overarching research question stems from Lindblom's (1996) view that speech production is a highly adaptable process, and that plasticity and economy are key to speech production and perception, especially when the acoustic speech signal is deprived of some of its cues (as in whispered speech). Therefore, our question could be set to Lindblom's (1996) analysis of multidimensional clusters of acoustic cues:

> Which acoustic signal attributes carry sufficiently distinct information to differentiate vowel height and to disambiguate consonant place and voicing in whispered speech?

In this study, speech segments that allow the exploration of such a question included phonemes that define the extremes of the vowel space; in EP they are /i, a, ɔ, u/. No other vowels were analysed since it has been shown that the remaining vowels falling within the vowel space area can be perceived using the available signal-plus-knowledge-mechanism[1] (Lindblom, 1996). The sibilants /s, z, ʃ, ʒ/ were also included in search of changes/ adaptations in place of articulation during whispered speech (Zygis et al., 2017) and to explore cues to voicing when the vocal folds do not vibrate.

The first three formant frequencies of close-front unrounded /i/ and open-front unrounded /a/ vowels can be reliably affiliated with specific front or back cavity resonances (Kent & Rountrey, 2020), and the broad spectral peak of alveolar /s/ and postalveolar /ʃ/ fricatives results from a front cavity resonance (Perkell, 2012). Therefore, the frequency of these spectral peaks will be used to propose an articulatory interpretation of the acoustic data, based on a mixed-effects logistic regression (MELR) model.

Specific research questions were:

1. Are vowels produced with significantly different formant frequencies, spectral slopes and sound pressure levels in voiced and whispered speech?

   Hypothesis H1.1: The back cavity used to produce whispered vowels is shorter than in voiced speech.

   Hypothesis H1.2: Whispered vowels are produced with a flatter spectrum and a lower sound pressure level than voiced vowels.

   Significantly different F_1 and F_2 formant frequencies (more specifically higher formant frequencies), spectral slope and sound pressure level have been previously observed in whispered speech, for example, in English (Sharifzadeh et al., 2012) and Polish (Zygis et al., 2017). The back cavity is likely to be shorter in whispered speech since, as in a previous study (Matsuda & Kasuya, 1999), raising of the larynx has been observed during whispered speech production.

2. Are the vowels' F_1 frequency shifts in whispered speech (relative to same-sex reference voiced F_1 frequency values) correlated to the voiced speech f_o values?

   Hypothesis H2: There is a positive correlation between f_o values and F_1 frequency shifts in whispered speech.

   A perceptual compensation process for the missing fundamental has been shown (Higashikawa & Minifie, 1999) to be at play in whispered speech, i.e., formant frequency shifts may be used to replace source information in whispered speech.

3. Are the vowel space areas of voiced and whispered speech significantly different?

   Hypothesis H3: There is a compression of the vowel space in whispered speech, when compared to voiced speech.

   It is not clear from previous research if whispered vowel space areas are smaller than in voiced speech, but shifts in the position of the vowel spaces have been observed, due to changes in both F_1 and F_2 frequencies, that point to that being a hypothesis (Sharifzadeh et al., 2012).

4. Are close /i, u/ vowel durations significantly shorter than close/open-mid vowels /a, ɔ/ both in voiced and whispered speech?

   Hypothesis H4: Intrinsic vowel durations (Abramson & Ren, 1990; Holt et al., 2015) are observed both in voiced and whispered speech.

5. Are the fricatives produced with significantly different spectral slopes, sound pressure level, broad peak (frequency and level) and durations in voiced and whispered speech?

   Hypothesis H5.1: Fricatives are produced in whispered speech with a lower noise source strength to that used for voiced speech. It has been previously shown (Jesus & Shadle, 2002) that greater noise source strength is related to higher spectral slopes, sound pressure level and broad spectral peak level.

   Hypothesis H5.2: The relative duration of same-place and speech mode phonologically voiceless fricatives is significantly higher than phonologically voiced fricatives (Pape & Jesus, 2015; Smith, 1997; Stevens et al., 1992) both in voiced and whispered speech.

6. What acoustic signal attributes (in words, sentences and text tasks) carry sufficiently distinct information to disambiguate /i/ from /a/ (vowel height) and /s/ from /ʃ/ (fricative place of articulation), both in voiced and whispered speech? F_1, F_2 and relative durations of vowels, will be considered as predictors in a MELR model for the vowels, and the low frequencies spectral slope (m1), broad peak frequency (F_{BP}) and relative durations of fricatives will be considered as predictors in a MELR model for the fricatives.

   Hypothesis H6.1: F_1, F_2 and relative durations carry sufficiently distinct information to disambiguate /i/ from /a/.

   Hypothesis H6.2: m1, F_{BP} and the relative durations carry sufficiently distinct information to disambiguate /s/ from /ʃ/.

The explanatory power (Eisenhauer, 2009) of each predictor or independent variable will also be estimated for each of the MELR models developed. We are not considering vowels /u, ɔ/ because they involve rounding and protrusion of the lips. These are likely to be confounding effects, or even the main effects on formant frequencies (Titze, 2000).

The reason why we will be focusing on /i, a, s, ʃ/ as the speech materials to explore question 6 has to do with one of the aims of this paper: To produce new evidence regarding whispered speech, based on materials that voice specialists use routinely in clinical assessment (Jesus, Belo et al., 2017). These include sustained vowels and fricatives, and the same phonemes in sentences and a phonetically balanced text. Since most reference values used in acoustic voice evaluation are based on sustained productions (Jesus, Belo et al., 2017), determining which speech patterns hold as we use increasingly realistic tasks (sustained → words → sentences → text) could help define what phonetic correlates constitute the multidimensional "cloud" (Lindblom, 1996) of sufficiently distinct informa-

[1] "The reason why speech perception succeeds in coping with partial information is that percepts are never completely ''raw'' records of the physical signal" (Lindblom, 1996, p. 1684). They are the product of the *information* available in the acoustic signal, *plus* the *knowledge* each "listener invokes to modulate the stimulus" (Lindblom, 1996, p. 1686).

tion to disambiguate between /i/ and /a/, and between /s/ and /ʃ/, both in voiced and whispered speech.

3. Method

Ethical permission (*Parecer* N° P523-10/2018, dated 21/11/2018) was obtained from an independent ethics committee (*Comissão de Ética da Unidade Investigação em Ciências da Saúde – Enfermagem da Escola Superior de Enfermagem de Coimbra*, Coimbra, Portugal), and informed consent was collected from all participants prior to data collection.

3.1. Participants and the recording conditions

Seventeen (17) participants (9 male speakers and 8 female speakers; 22 to 33 years of age; Mean = 26 years; Standard Deviation of the Mean (Std) = 3 years) were recruited using convenience sampling in the districts of Aveiro and Coimbra in Portugal. They were all from the same dialectal region in Portugal (*Dialetos Setentrionais* / North-western Dialects) and had not spent extended periods of time in other regions (Pape & Jesus, 2015).

The following inclusion criteria were used: No history of voice disorders; no vocal pathology at the time of the recordings as assessed by a Voice Specialist using a standardised case history form (Ferreira et al., 2014); no upper respiratory tract infection on recording day; Portuguese as first language and from the centre of Portugal, where the *Dialetos Setentrionais Setentrionais* (North-western Dialects) are spoken according to Segura (2013). Exclusion criteria included: Impairments in oro-motor structure and function; use of orthodontic (correction) devices; respiratory pathology; laryngopharyngeal reflux; fluency disorders; having been submitted to vocal laryngeal surgery; not being able to produce all the vocal tasks (particularly whispering).

Participants were seated in a quiet room, with a background noise level of 15.1 dB (A-weighted time-averaged / equivalent sound pressure level – LAeq), and recorded using a head-mounted Sennheiser Ear Set 1 condenser microphone. Acoustic data was sampled at 48000 Hz with 16 bits per sample.

A similar screening and training procedure to that previously used (Konnai et al., 2017) to ensure participants can discriminate and produce voiced and whispered speech was adopted in this study. Since no images of the glottal configurations were available at the time of data acquisition a Voice Specialist perceptually monitored and identified deviations from the targeted neutral whispering, also described by Konnai et al. (2017) as normal adduction and medium loudness of whispered speech.

3.2. Corpus

Materials included four sustained sibilants, four sustained oral vowels, 12 disyllabic words, six sentences used by clinicians to evaluate voice quality (Jesus, Tavares et al., 2017, p. 223) and a phonetically balanced text (Jesus et al., 2015).

For the purpose of the present study, we only analysed the four sibilant fricatives /s, z, ʃ, ʒ/ and the four oral vowels /i, a, ɔ, u/ that define the corners of the EP vowel space (Escudero et al., 2009). The twelve (12) CVCV disyllabic real words, shown in Table 1, had the fricatives in initial, mid and final word positions. In EP, the syllable type CV is the most frequent (Vigário et al., 2006), so the four sibilants were combined with /a/ and /ɐ/ to maintain a stable vowel height environment (open-mid to open) across the syllables. The same vowels and fricatives (/i, a, ɔ, u, s, z, ʃ, ʒ/) were also analysed in six sentences and a phonetically balanced text that are part of the speech materials used regularly in Portugal to evaluate voice quality (Jesus et al., 2022).

All the tasks were repeated three times in voiced speech and three times in whispered speech (Zygis et al., 2017), except the text that was read once in each speech mode. For some speakers, whispering can be more traumatic to the larynx than voiced speech (Rubin et al., 2006), so the tasks were carefully selected to provide information about voicing versus whispering mechanisms, without causing vocal fatigue and all recordings were made in the presence of a Voice Specialist. Changing speech mode frequently and in a short period of time could be difficult and confusing (Zygis et al., 2017) so all the tasks, one at a time, were recorded first in voiced speech and then in whispered speech.

More specifically, the full database has 54 files per participant (27 in voiced speech and 27 in whispered speech) containing:

- Four (4) sustained sibilants – /s, z, ʃ, ʒ/
- Four (4) sustained EP oral vowels – /i, a, ɔ, u/
- Twelve (12) disyllabic real words with sibilant fricatives in initial, medial and final word positions (shown in Table 1)
- Six (6) CAPE-V phrases (Jesus, Tavares et al., 2017, p. 223)
  - <A Marta e o avô vivem naquele casarão rosa velho> [ɐ ˈmaɾtɐ i u ɐˈvo ˈvivɐ̃j nɐˈkelɨ kɐzɐˈɾɐ̃w ˈʁɔzɐ ˈvɛʎu] – Production of every Portuguese oral vowel
  - <Sofia saiu cedo da sala> [suˈfiɐ sɐˈiw ˈsedu dɐ ˈsalɐ] – Easy onset with /s/ (words with /s/ at syllable onset)
  - <A asa do avião andava avariada> [ɐ ˈazɐ du ɐviˈɐ̃w ɐ̃ˈdavɐ ɐvɐɾiˈadɐ] – All voiced
  - <Agora é hora de acabar> [ɐˈgɔɾɐ ɛ ˈɔɾɐ dɨ ɐkɐˈbaɾ] – Elicits hard glottal attack
  - <Minha mãe mandou-me embora> [ˈmiɲɐ mɐ̃j mɐ̃ˈdomɨ ẽˈbɔɾɐ] – Nasal sounds
  - <O Tiago comeu quatro peras> [u tiˈagu kuˈmew kuˈatɾu ˈpeɾɐʃ] – Weighted with voiceless stops
- European Portuguese phonetically balanced text (“The North Wind and the Sun”) with 98 words and 196 syllables (Jesus et al., 2015)

3.3. Data segmentation and annotation

The start and end of all the phones from sustained and word materials (8 sustained fricatives and oral vowels; all phones in 12 disyllabic words) were manually annotated using previously described criteria (Lousada et al., 2010; Pape & Jesus, 2015) based on perceptual and acoustic analysis. All of the productions of /i, a, ɔ, u/ and /s, z, ʃ, ʒ/ in sentences and the phonetically balanced text were also annotated.

Table 1
Portuguese disyllabic words with fricatives.

Fricative	Word initial	Word medial	Word final
[s]	<sala> [ˈsalɐ]	<assa> [ˈasɐ]	<face> [ˈfas]
[z]	<zaro> [ˈzaɾu]	<asa> [ˈazɐ]	<vaze> [ˈvaz]
[ʃ]	<chama> [ˈʃɐmɐ]	<acha> [ˈaʃɐ]	<ache> [ˈaʃ]
[ʒ]	<jarra> [ˈʒaʁɐ]	<haja> [ˈaʒɐ]	<laje> [ˈlaʒ]

Evidence used to annotate voiced vowel boundaries included: Both waveform and spectrogram in Praat's 6.0.47 `SoundEditor` (wideband spectrogram, with default settings, e.g., view range 0 to 5000 Hz) were used to analyse the periodicity of the acoustic signal, F_2 amplitude and the f_o track, together with constant auditory monitoring (over headphones) of all recordings. Fricatives produced in voiced speech mode were annotated using spectrograms with a wider view range (0 to 16000 Hz).

The segmentation process in whispered speech is distinct from voiced speech (Jovičić & Šarić, 2008), demanding manual / laborious processes (Meynadier & Gaydina, 2013; Sharifzadeh et al., 2010). Segmentation involves the visual inspection of waveforms and formant structure in spectrograms (F_2 and F_3 onset and offset) and changes in intensity (Heeren, 2015). Praat's default spectrogram settings were only changed regarding the view range, that was set to 0 to 16000 Hz, both for vowels and fricatives. The main acoustic anchors used to annotate whispered speech were the waveforms and spectrograms of frication noise. Phones produced with a hard or abrupt glottal attack were not annotated.

The second author carried out all phonetic annotations and transcriptions. In addition, the productions of two randomly selected participants were annotated and transcribed, by a trained phonetician not involved in the study and blind to its aims. Point-to-point reliability was 92.34 %, a value that was considered adequate for the objective of this study. Two participants represent 12 % of speech samples, and this percentage is equivalent to what is reported when checking reliability in other whispered speech studies (Heeren & Heuven, 2014; Jovičić & Šarić, 2008).

3.4. Acoustic analysis

The purpose of our acoustic analysis was to characterise the laryngeal and supralaryngeal sound sources, and filter during voiced and whispered speech production. The first author designed, programmed and carried out all the acoustic analysis. All of the measured parameters are defined in detail below.

The parameters used to analyse the source of vowels were: Fundamental frequency (f_o) – Hz [only for voiced vowels to control for reported formant frequency biased estimation especially when f_o is high and/or the F_1 frequency is low (Shadle et al., 2016, p. 713) and to analyse possible correlations between f_o in voiced speech and first/second formant frequency values in whispered speech]; spectral slope (m) – dB/kHz^2; sound pressure level (SPL) – dB.

The parameters used to analyse the filter characteristics of vowels were: First formant frequency (F_1) – Hz; second formant frequency (F_2) – Hz; third formant frequency (F_3) – Hz. For a close-front unrounded vowel such as /i/, F_1 is the visible outcome of a Helmholtz resonance, F_2 can be affiliated with the first back-cavity resonance and F_3 is a consequence of the first front-cavity resonance. For an open-front unrounded vowel such as /a/, F_1 is affiliated with the first back cavity resonance, F_2 with the first front cavity resonance and F_3 with the second back cavity resonance (Titze, 2000; Whalen et al., 2022).

The parameters used to analyse the source of fricatives were (Jesus & Shadle, 2002; Zygis et al., 2017): Low frequencies spectral slope (m1) – dB/kHz^2; high frequencies spectral slope (m2) – dB/kHz^2; sound pressure level (SPL) – dB.

The broad peak frequency (F_{BP}) – kHz and the broad peak level (L_{BP}) – dB/kHz were also analysed since they are expected to correspond to the first front cavity resonance (Jesus & Shadle, 2002, p. 447), i.e., fricative filter characteristics.

We also extracted absolute durations of /i, a, ɔ, u/ and /s, z, ʃ, ʒ/ as in previous studies (Escudero et al., 2009; Jesus & Shadle, 2003), and calculated the following relative durations to control for possible speech-rate effects: Phone to word-length ratio of the word task; phone to sentence-length ratio in the sentence reading task; phone to text-length ratio (including pauses) in the phonetically-balanced text reading task.

Summarising, the following parameters were extracted: Vowel /i, a, ɔ, u/ fundamental frequency (f_o), first formant frequency (F_1), second formant frequency (F_2), third formant frequency (F_3), spectral slope (m), sound pressure level (SPL) and absolute and relative durations; fricative /s, z, ʃ, ʒ/ low frequencies spectral slope (m1), high frequencies spectral slope (m2), sound pressure level (SPL), broad peak frequency (F_{BP}), broad peak level (L_{BP}), absolute and relative durations.

3.4.1. Analysis of vowels

All annotated segments of vowels /i, a, ɔ, u/ for each participant were analysed automatically with a Praat script written specifically for this purpose. Formant data (Burris et al., 2014; Derdemezis et al., 2016; Kent & Rountrey, 2020; Kent & Vorperian, 2018; Shadle et al., 2016; Whalen et al., 2022) was extracted using the following Praat function: `To Formant (burg)...` 0.01 5 5500 0.025 50 [Time step(s); Maximum number of formants; Maximum formant (Hz); Window length (s); Pre-emphasis from (Hz)]. The parameterisation of the Praat scripting language function used to extract f_o data was: `To Pitch (cc)...` 0 75 15 no 0.03 0.45 0.15 0.35 0.14 500 [Time step (s); Pitch floor (Hz); Max. number of candidates; Very accurate; Silence threshold; Voicing threshold; Octave cost; Octave-jump cost; Voiced / unvoiced cost; Pitch ceiling (Hz)]. The pitch-tracking parametrisation was the same as Praat's `SoundEditor` window defaults except for the Octave cost that has been increased from 0.01 to 0.15 so that the current f_o results would be comparable to those available from a large open access resource for voice clinicians and speech research (Jesus, Belo et al., 2017) that includes sustained / a, i, u/, and the same sentences and phonetically balanced text as those used to extract the current data. The same cross-correlation method used by Escudero et al. (2009) to estimate EP vowel formant frequencies was used here to allow comparisons between studies and because it is adequate for measuring short vowels (Escudero et al., 2009, p. 1381) such as those produced in words, sentences and the phonetically balanced text by the speakers recruited for the current study.

All vowels with a duration of less than 50 ms were excluded from the analysis process (Lee et al., 1999, p. 1457), corresponding to 17 % of the total number of the vowels that fell below this criterion. Various reduction processes (Bisol & Veloso, 2016) were observed and it was not possible to obtain reliable estimates of f_o, F_1, F_2 and F_3.

For all vowels, a section of Praat's `SoundEditor` window corresponding to the segment that had just been analysed was visually monitored manually. Automatic measures resulting from Praat's formant tracker were verified against greyscale wide and narrow band spectrograms (Sharifzadeh et al., 2012). Those samples that did not allow the correct identification of the periods, yield accurate formants estimates in Praat or show clear formant structure in the spectrogram, were dropped out of the database (Lee et al., 1999, p. 1457). No manual estimations of formant frequency values were used.

Median f_o and formant frequency values were calculated over a 46 ms window centred in the middle of the vowel, a process that has previously been used to produce representative f_o and formant frequencies (Lee et al., 1999, p. 1457). Using the median instead of the mean has been claimed to reduce f_o estimation errors (Escudero et al., 2009, p. 1381) in a reference study of the acoustics of Portuguese vowels and the global medians of the f_o, F_1, F_2 and F_3 tracks have been used to analyse the largest American English acoustic developmental database (Lee et al., 1999).

The SPL was calculated with Matlab 9.5.0.944444 (R2018b) using: SPL = 10 × log10((Pa/Pa$_{ref}$)^2), where Pa$_{ref}$ = 20 × 10^{-6} Pa is the reference pressure for SPL. Since this equation provides instantaneous pressure values, in order to produce a time-smoothed SPL track (along with the f_o, F_1, F_2 and F_3 provided by a Praat script), the envelope of the pressure signal was extracted by rectifying and low-pass filtering the signal with a first order lowpass Butterworth filter: `butter`(1,8/(48000/2),'low'). Again, median SPL values over the same 46 ms window centred in the middle of the vowel were obtained in order to allow further statistical analysis over the same time-frame as f_o, F_1, F_2 and F_3.

The spectral slope (m) was also calculated from Welch's power spectral density (PSD) estimates over 46 ms (2048 samples) Hamming windows at the centre of every vowel using Matlab's Signal Processing Toolbox Version 8.1 (R2018b) `pwelch` function (segments with1024 samples, with 128 overlapped samples and a 1024 samples Discrete Fourier Transform were used). A regression line was then fit to the PSD from 300 Hz (above the highest f_o registered in the database) to 3000 Hz (just above the highest F_2 value for all speakers) using Matlab's `polyfit` function. The slope of this regression line (m) was extracted in order to capture previously reported (Sharifzadeh et al., 2012, pp. e49–e50) equivalent levels of the first and second formant peaks in whispered speech, hypothesized to contrast with the usual -12 dB/octave spectral slope of voiced speech (Titze, 2000, p. 131).

Since one of the aims of this study was to compare Portuguese whispered formant data with that previously reported for English (Kallail & Emanuel, 1984), Japanese (Matsuda & Kasuya, 1999), Polish (Zygis et al., 2017) and Swedish (I. Eklund & Traunmuller, 1996), the raw F_1, F_2 and F_3 data in Hz was used to render all graphical representations and statistical analysis. Normalisation procedures were not employed because previous studies that compared vowel formant frequencies in voiced and whispered speech have not used normalisation across multiple tokens. Formant frequencies have been typically reported in Hz (Kallail & Emanuel, 1984; Matsuda & Kasuya, 1999; Zygis et al., 2017), and the two papers (Higashikawa et al., 1996; Sharifzadeh et al., 2012) that also discuss vowel spaces have done so in a linear Hz scale.

Vowel space plotting and area calculations were performed with phonR 1.0.7 (McCloy, 2016) for R version 4.0.0 running in RStudio Version 1.3.959.

3.4.2. Analysis of fricatives

Multitaper spectrum estimates (Shadle, 2012; Thomson & Haley, 2014) were produced using Matlab Signal Processing Toolbox, with a 12 ms (512 samples) Hamming window centred in the middle of the fricative. More specifically, PSD estimates were calculated via the Thomson multitaper method (linear combination with unity weights of individual spectral estimates and the default FFT length) using the `pmtm` function.

A first regression line (low frequencies) was fit to the PSD from 500 Hz to 6000 Hz for /s, z/ and from 500 Hz to 4000 Hz for /ʃ, ʒ/ (Jesus & Shadle, 2002, p. 447) using Matlab's `polyfit` function. The slope of this regression line (m1) was then calculated.

A second regression line (high frequencies) was fitted to the PSD from 6000 Hz to 20000 Hz for /s, z/ and from 4000 Hz to 20000 Hz for /ʃ, ʒ/ (Jesus & Shadle, 2002, p. 447) using Matlab's `polyfit` function. The slope of this regression line (m2) was then calculated.

The broad peak frequency (F_{BP}) and broad peak level (L_{BP}) of /s, z, ʃ, ʒ/ were extracted manually from a visual inspection of the PSD, using lower and higher bound frequencies (5000 Hz to 9000 Hz for /s, z/ and 2000 Hz to 5000 Hz for /ʃ, ʒ/) previously reported for EP by Jesus (2001), overlaid on the PSD.

The fricative's median SPL over a 46 ms window was calculated with Matlab, using the same time-smoothing technique previously described for vowels.

3.5. Data visualisation, statistical analysis and modelling

The boxplots presented in this paper were generated using R version 4.0.0 and include a central mark that indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the first quartile (Q1) − 1.5 × Interquartile Range (IQR) and the third quartile (Q3) + 1.5 × IQR (Winter, 2020). The boxplots were combined with univariate scatterplots (strip plots) using the beeswarm 0.2.3 package. The datapoints are shifted horizontally to avoid superposition and the overlap between adjacent boxplots is circumvented with the "random corral" method (A. Eklund, 2021). Transparent datapoints have been used so the reader could differentiate datapoints that had to be superimposed and so that the boxplot could still be visible. We therefore combined two visualisations which show the full distribution of data (Politzer-Ahles & Piccinini, 2018). No voiced or whispered /ʒ/ were produced in sentences, so only /s, z, ʃ/ are visualised for this dataset. The ellipses used in vowel space representations were calculated using the phonR 1.0.7 (McCloy, 2016) package based on the covariance of the tokens and a 68 % confidence level corresponding to ± 1 standard deviation of the normal density contour estimated from the data. Probability density distributions of vowel and fricative durations were generated using the `density` function contained within the base R version 4.0.0.

IBM SPSS Statistics 25 was used to estimate Tukey and chi-squared distances, analyse data through Kruskal-Wallis tests, Dunn's nonparametric comparisons for post hoc testing after Kruskal-Wallis tests, Mann-Whitney U tests and Wilcoxon-signed ranks tests. IBM SPSS Statistics 25 was also used to calculate the Pearson's and Spearman's correlation coefficients, as well as to run Student's t tests and Analysis of Variance (ANOVA) with Bonferroni corrections. A detailed description of all the statistical analysis tasks, dependent and independent variables, and normality tests that were used is provided in Appendix B.

Binary logistic regressions within the generalised linear mixed model context were developed using IBM SPSS Statistics 25, both for fricatives and vowels. Models were developed in order to try to disambiguate /i, a/ for each of the speech tasks separately for the vowels, and /s, ʃ/ for each of the speech tasks separately for the fricatives. Accordingly, data had to be previously re-coded as being a success (being a voiced or whispered /a/, for the vowels or being a voiced or whispered /ʃ/ for the fricatives) or unsuccess (being a voiced or whispered /i/, for the vowels or being a voiced or whispered /s/ for the fricatives). The three MELR models for the vowels were computed considering the outcome variable as the recoded dichotomous variable that distinguishes between /i/ and /a/, thus allowing us to estimate the conditional probability that the outcome variable equals one, at a particular value of the predictor variables. The three MELR models for the fricatives were computed with the outcome variable being the recoded dichotomous variable that distinguishes between /s/ and /ʃ/. Since correlated errors are expected to occur given the fact that there are multiple productions from the same speaker and there were 17 different speakers, these analyses were done under the mixed effects model context, taking the speaker id as the random effect (intercept only). Speech mode, sex and relative durations were taken as fixed effects both for the vowels and the fricatives, and additionally for the vowels F_1 and F_2, and for the fricatives m1 and F_{BP}, were also considered.

The choice of predictors is related to the research questions and the reduced number of predictors (two frequency domain predictors and one time-domain predictor) that allowed us to understand what the model was learning (interpretability). Also, the first part of the statistical analysis provided an insight into the effects of all variables in the study, allowing us to include in the mixed-effects logistic regression models only those variables that seemed reasonable to have explanatory power on the discriminations considered.

4. Results

We first present the results of acoustic signal attributes of vowels (subsections 4.1. to 4.5), then those acoustic parameters used to characterise the source and filter of fricatives (subsections 4.6. to 4.9), and finally, in subsection 4.10, six MELR models are presented. A model was developed for each speech material separately: Disyllabic words, sentences and phonetically balanced text.

4.1. Vowel formant frequencies (F_1, F_2 and F_3)

The boxplots of F_1 frequencies (presented in Fig. 1 and Fig. 2) revealed a very similar behaviour in female and male speakers: Whispered vowels were produced with a significantly higher F_1 frequency than voiced vowels both for female (see Table B.1.1 in Appendix B) and male (see Table B.1.2) speakers in all speech tasks. Female F_1 frequencies increased by 192-961 Hz and male F_1 frequencies shifted by 198-281 Hz on average, varying by vowel quality (/i, a, ɔ, u/) and speech tasks (sustained, words, sentences and text).

Separate correlation analysis for men and women with all the data revealed a significant positive correlation between voiced and whispered F_1 frequencies of female (Spearman's correlation coefficient = 0.924, p = 0.000; two-tailed p-value) and male (Spearman's correlation coefficient = 0.947, p = 0.000; two-tailed p-value) speakers. The same positive correlation was found to be significant between voiced and whispered F_2 frequencies of female (Spearman's correlation coefficient = 0.994, p = 0.000; two-tailed p-value) and male (Spearman's correlation coefficient = 0.979, p = 0.000; two-tailed p-value) speakers. A significant positive correlation was also found between voiced and whispered F_3 frequencies, both for female (Spearman's correlation coefficient = 0.921, p = 0.000; two-tailed p-value) and male (Spearman's correlation coefficient = 0.691, p = 0.003; two-tailed p-value) speakers.

Female's F_2 frequency values were significantly different in voiced and whispered speech (see Table B.2.1) for all vowels, except for sustained /i, ɔ/. Male speakers' F_2 frequency values were also significantly different in voiced and whispered speech (see Table B.2.2), except for all sustained, words and sentences tasks of vowel /i/. Female /a, ɔ, u/ F_2 frequencies increased by 142-314 Hz and male /a, ɔ, u/ F_2 frequencies increased by 99-338 Hz on average.

Female's F_3 frequency values of whispered vowels were higher than voiced vowels, except for sustained /i, ɔ/, /i, a, ɔ/ in words and sentences, and /i, u/ in text (see Table B.3.1). Male's F_3 frequency values of whispered vowels were also higher than voiced vowels, except for sustained /i/ and /i/ in words, /i, u/ in sentences, and /u/ in text (see Table B.3.2). F_3 mean frequency shifts varied widely from negative (ranging from -11 Hz to -228 Hz) to positive (mean 4-157 Hz increase) ranges of values.

4.2. Vowels' spectral slope (m) and sound pressure level (SPL)

Female (see Table B.4.1) and male (see Table B.4.2) speakers' spectral slope values of all vowels increased significantly for whispered speech (relative to voiced speech), and slope findings were consistent across tasks. The SPL of all of female's and male's (see Tables B.5.1 and B.5.2) whispered vowels was significantly lower than in voiced exemplars, with a mean downward shift between 19 and 25 dB, that was very stable across speech tasks (Female SPL $_{\text{sustained}}$ = -25 dB; Female SPL $_{\text{words}}$ = -24 dB; Female SPL $_{\text{sentences}}$ = -23 dB; Female SPL $_{\text{text}}$ = -22 dB; Male SPL $_{\text{sustained}}$ = -19 dB; Male

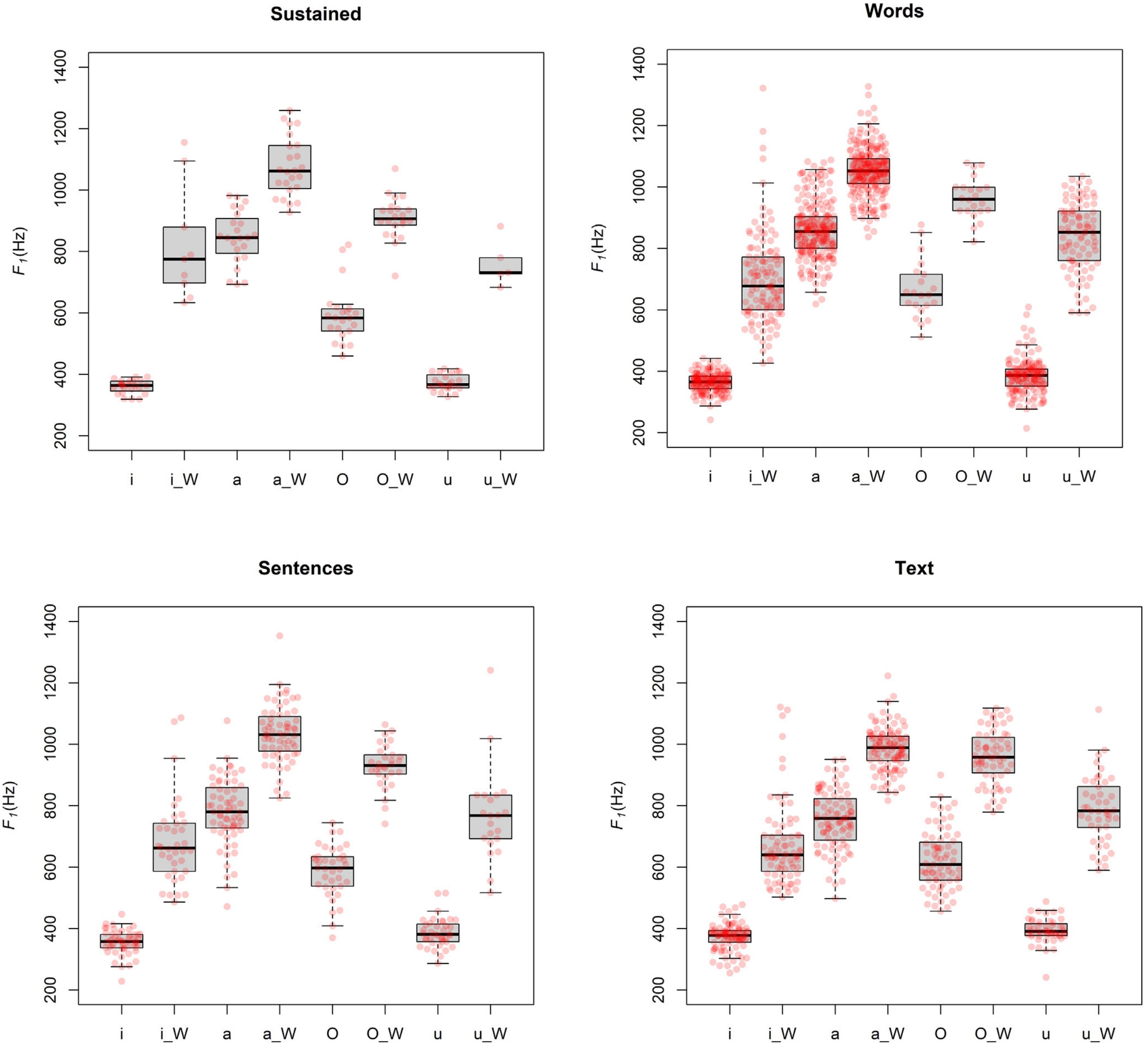

Fig. 1. Female speakers' F_1 frequencies. The Speech Assessment Methods Phonetic Alphabet – SAMPA (Wells, 1997) is used to represent voiced vowels /i, a, O, u/ and their whispered counterparts /i_W, a_W, O_W, u_W/.

SPL $_{\text{words}}$ = -23 dB; Male SPL $_{\text{sentences}}$ = -23 dB; Male SPL $_{\text{text}}$ = -23 dB).

4.3. Vowels' fundamental frequency and first formant frequency shifts

The fundamental frequency (f_o) used by the 8 female speakers (shown in Fig. 3) and the 9 male speakers, to produce vowels /i, a, ɔ, u/ in four speech tasks (sustained, words, sentences and text) was analysed, and the $F_{1(whispered)}$ - $F_{1(voiced)}$ frequency differences were calculated to answer the second research question: Are the vowels' first formant frequency shifts in whispered speech (relative to same-sex reference voiced first formant frequency values) correlated to the voiced speech fundamental frequency?

A significant positive correlation was found for f_o and $F_{1(whispered)}$ - $F_{1(voiced)}$ of female speakers (Pearson's correlation coefficient = 0.660, p = 0.005; two-tailed p-value). The correlation between f_o and $F_{1(whispered)}$ - $F_{1(voiced)}$ of male speakers was not significant (Pearson's correlation coefficient = 0.071, p = 0.795; two-tailed p-value). We also ran separate correlations analysis for each task (sustained, words, sentences and text), but only one significant, negative correlation was found for male speakers' sustained vowels (Pearson's correlation coefficient = -0.985, p = 0.015; two-tailed p-value).

4.4. Vowel spaces

The vowel space area (VSA) calculated using a polygon with vertices at the mean value for each vowel (McCloy, 2016), shown in Fig. 4 and Fig. 5, revealed a compression in whispered speech, when compared to an equivalent voiced speech task, both for female and male speakers. Wilcoxon Signed Ranks Test indicated that voiced ranks were not statis-

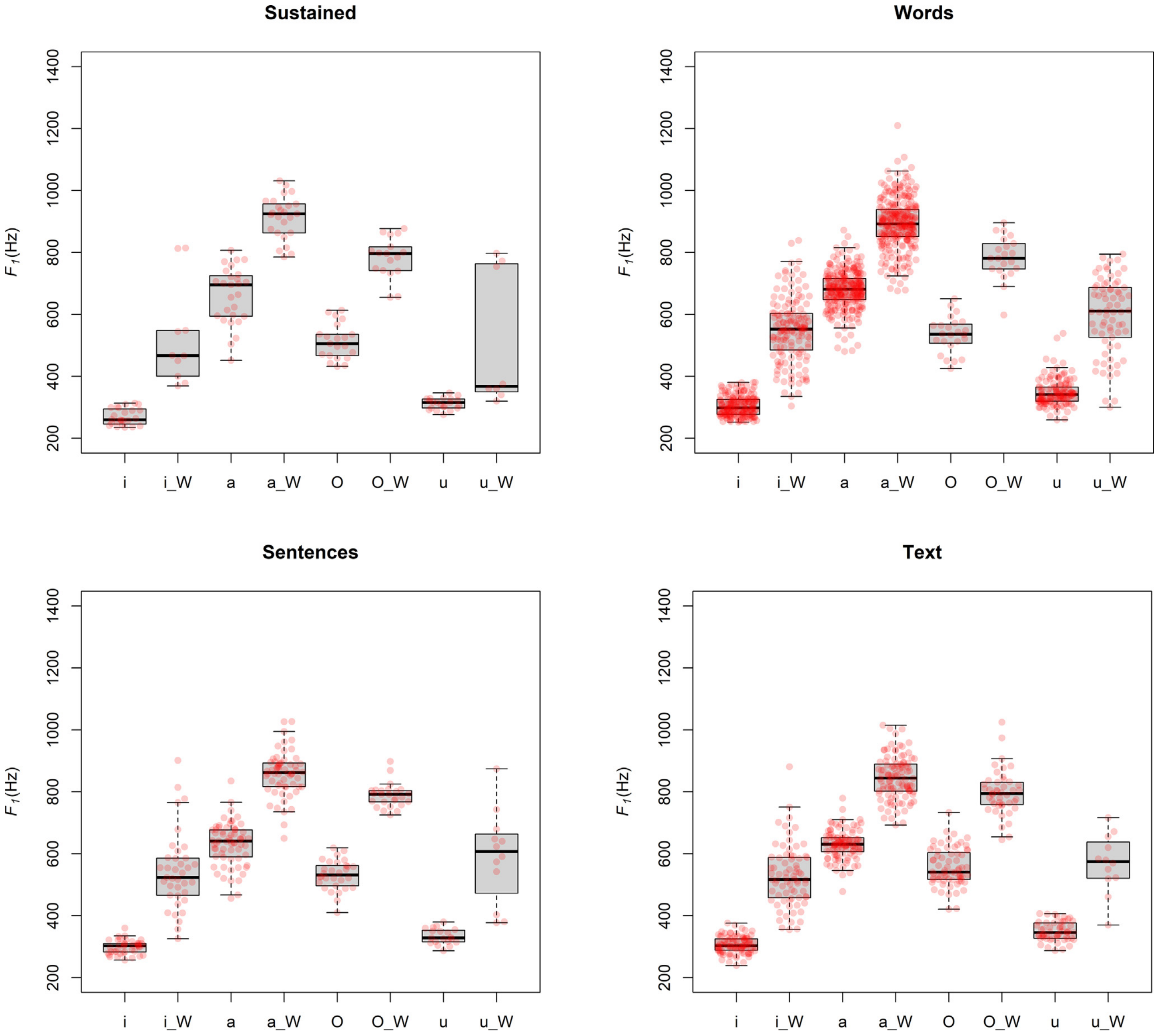

Fig. 2. Male speakers' F_1 frequencies. The Speech Assessment Methods Phonetic Alphabet – SAMPA (Wells, 1997) is used to represent voiced vowels /i, a, O, u/ and their whispered counterparts /i_W, a_W, O_W, u_W/.

tically significantly different from whispered ranks (Z = -1.826, p = 0.063; for both male and female speakers). When a convex hull enclosing all vowels was used to calculate the VSA (McCloy, 2016), voiced and whispered results were also not significantly different (Z = 0.000, p = 0.563 for female speakers; Z = -1.826, p = 0.063 for male speakers). Nevertheless, a clear downward shift (relative to voiced speech) of vowel spaces in whispered speech could be observed for all speech tasks.

4.5. Vowels' absolute and relative durations

Absolute durations of female and male, voiced and whispered speech were used to differentiate close /i, u/ from open/open-mid /a, ɔ/ vowels, the only exception being the values of male /i/ when compared to /ɔ/. A Kruskal-Wallis test provided evidence of a difference (p = 0.000) between the mean ranks of at least one pair of groups of all the different possible multi-comparisons. Dunn's pairwise tests of female and male, voiced and whispered speech were carried out for the four pairs (/i/-/a/; /i/-/ɔ/; /u/-/ɔ/; /u/-/a/), showing significantly different durations between close and open/open-mid vowels, except for male voiced /i/-/ɔ/ (p = 0.152; two-tailed p-value) and /u/-/ɔ/ (p = 0.195; two-tailed p-value) pairs.

However, when probability density distributions of durations were analysed, the overlap between close and open-mid vowels was extensive, both in voiced and whispered speech, as shown in Fig. 6 for /i/ and /a/ produced by female speakers. The average duration of close vowels was lower than open-mid vowels, the height categories (close versus open-mid) were bimodal but overlapping.

The relative durations (shown in Fig. 7 for female speakers) correspond to the phone to word-length ratio of the word task, the phone to sentence-length ratio of the sentence task and

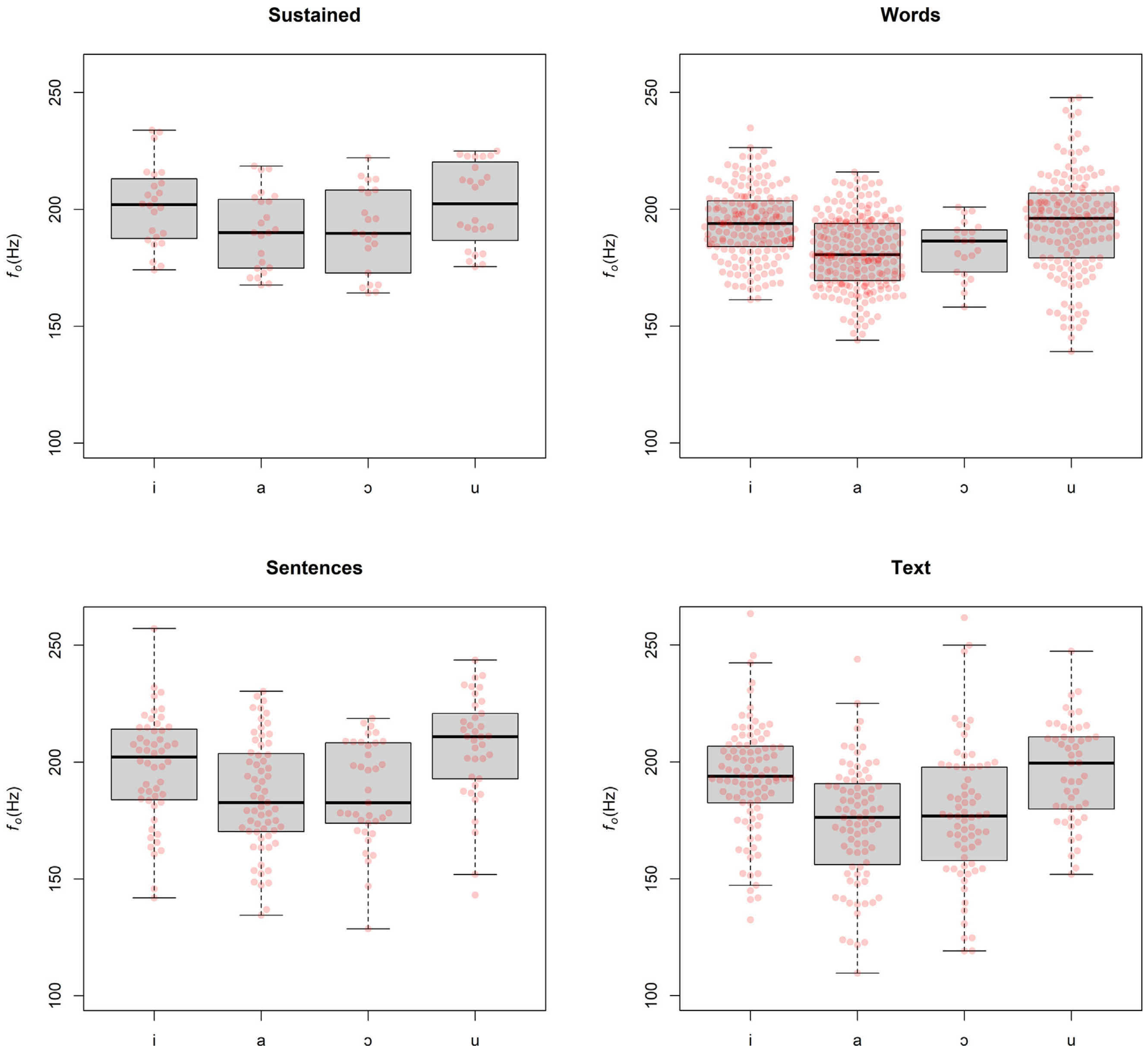

Fig. 3. Comparing f_o used by female speakers to produce vowels in four speech tasks.

the phone to text-length ratio in the text reading task. Both female and male relative, voiced and whispered speech, durations unveiled a new pattern that had only just surfaced when looking at the absolute values: Close /i, u/ vowels were significantly shorter than open-mid vowels /a, ɔ/. A Kruskal-Wallis test provided evidence of a significant difference (p = 0.000; two-tailed p-value) between the mean ranks of at least one pair of groups. Dunn's pairwise tests were carried out for the four pairs (/i/-/a/; /i/-/ɔ/; /u/-/ɔ/; /u/-/a/). There was evidence (p = 0.000, adjusted using the Bonferroni correction; two-tailed p-value), that intrinsic vowel durations were at play here, even when the speakers whispered the vowels.

4.6. Fricatives' low (m1) and high (m2) frequencies spectral slope

Both female and male m1 results in the two speech modes revealed no significant differences between voiceless fricatives (the exceptions being male's sustained /s/ and /s, ʃ/ in words), and significantly higher m1 values in whispered than voiced speech modes for phonologically voiced fricatives (see Table B.6.1 and B.6.2), except for the alveolar fricative /z/ in female's text and male's sentences. Place of articulation had a significant effect on m1 values (the more posterior place of articulation had a steeper slope, i.e., higher m1 values), both in voiced and whispered speech (see Table B.6.3.1, B.6.3.2, B.6.4.1 and B.6.4.2).

Results for m2 (shown in Fig. 8 for female speakers) were not significantly different between the two speech modes, the only exceptions being (see Table B.6.5 and B.6.6): Females'

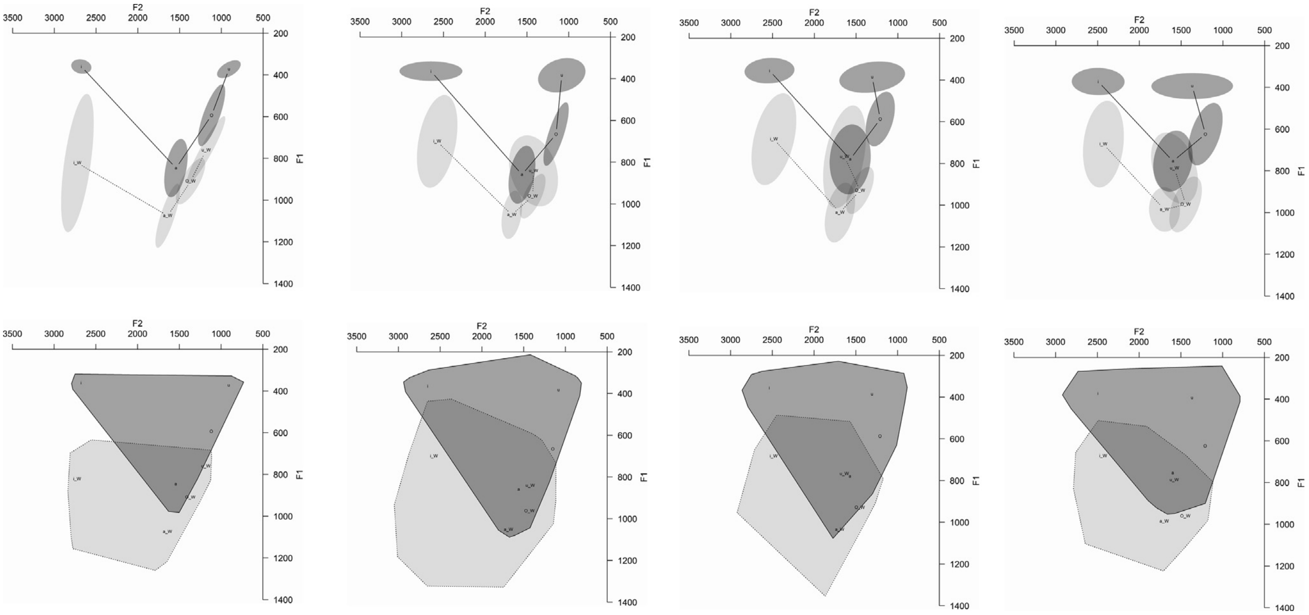

Fig. 4. Vowel spaces of female speakers in two speech modes: Sustained (first column), words (second column), sentences (third column) and text (forth column) speech tasks. The vowel spaces are represented using a polygon (top row) and a convex hull (bottom row). The Speech Assessment Methods Phonetic Alphabet – SAMPA (Wells, 1997) is used to represent voiced vowels /i, a, O, u/ and their whispered counterparts /i_W, a_W, O_W, u_W/ at F_1 and F_2 mean values in Hz.

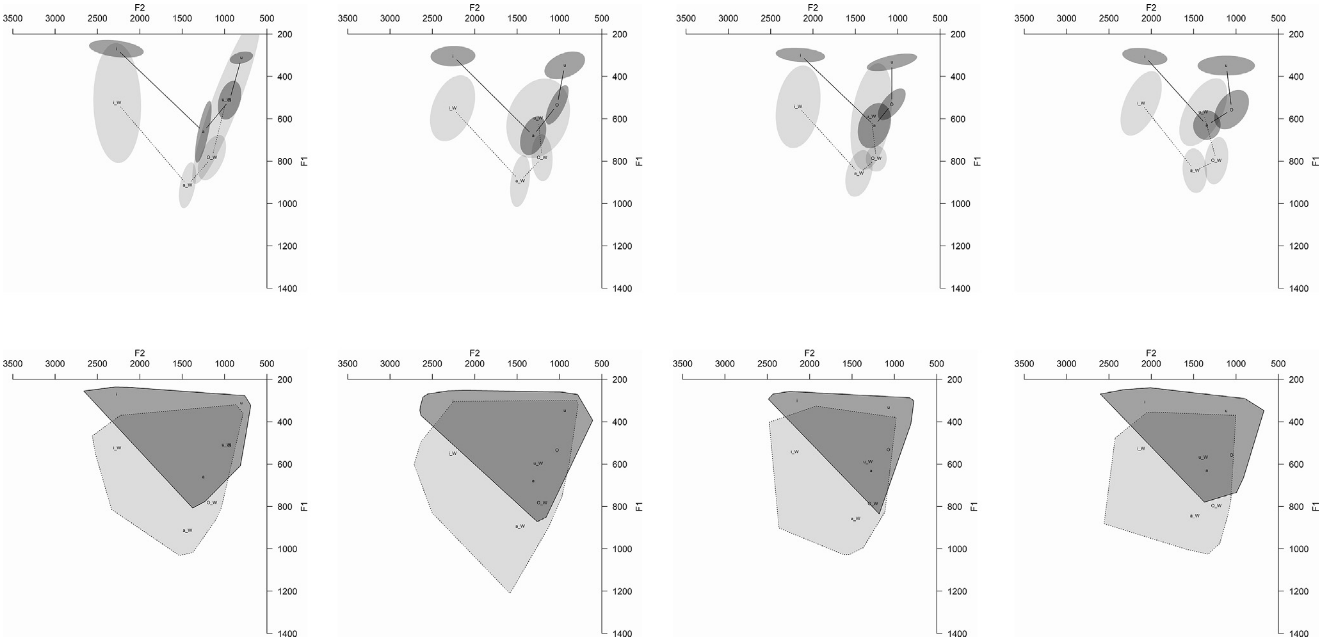

Fig. 5. Vowel spaces of male speakers in two speech modes: Sustained (first column), words (second column), sentences (third column) and text (forth column) speech tasks. The vowel spaces are represented using a polygon (top row) and a convex hull (bottom row). The Speech Assessment Methods Phonetic Alphabet – SAMPA (Wells, 1997) is used to represent voiced vowels /i, a, O, u/ and their whispered counterparts /i_W, a_W, O_W, u_W/ at F_1 and F_2 mean values in Hz.

and males sustained /s/, and /s/ in words and text; female's sustained /z/; male's /ʃ/ in sentences; when /ʒ/ was produced in words by both female and male speakers. It should be noted in Table B.6.5 and B.6.6 that 6/8 results for /s/ showed a significant change in the same direction (the m2 parameter values were lower in whispered speech than in voiced speech).

4.7. Fricative spectral broad peak frequency (F_{BP}) and level (L_{BP})

The values of F_{BP} for both female's and male's whispered and voiced speech modes alveolar fricatives /s, z/ were significantly higher (p = 0.000; ANOVA with Bonferroni correction and Dunn's nonparametric comparison for post hoc testing

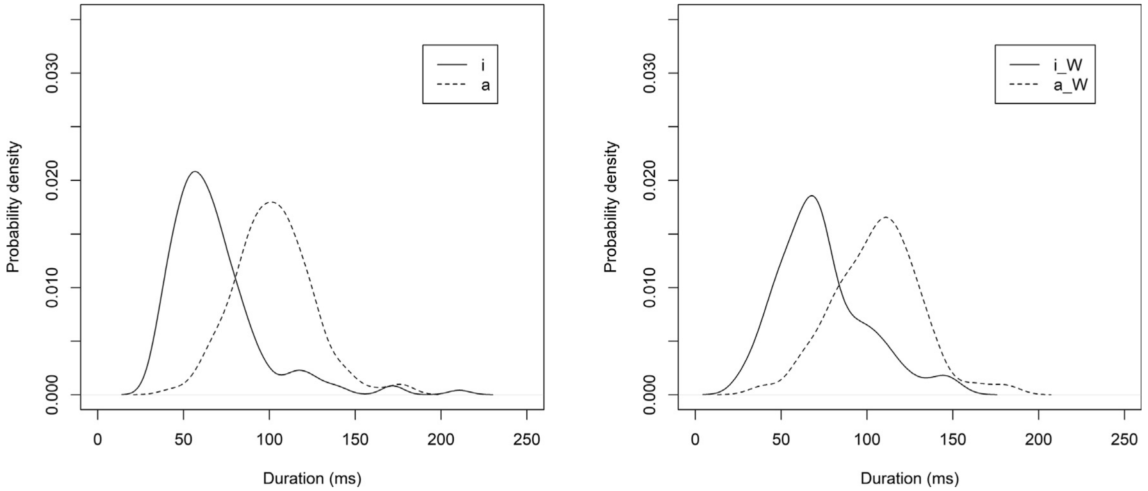

Fig. 6. Probability density distributions of durations of voiced (left) and whispered (right) vowels /i, a/ produced by female speakers in a phonetically balanced text.

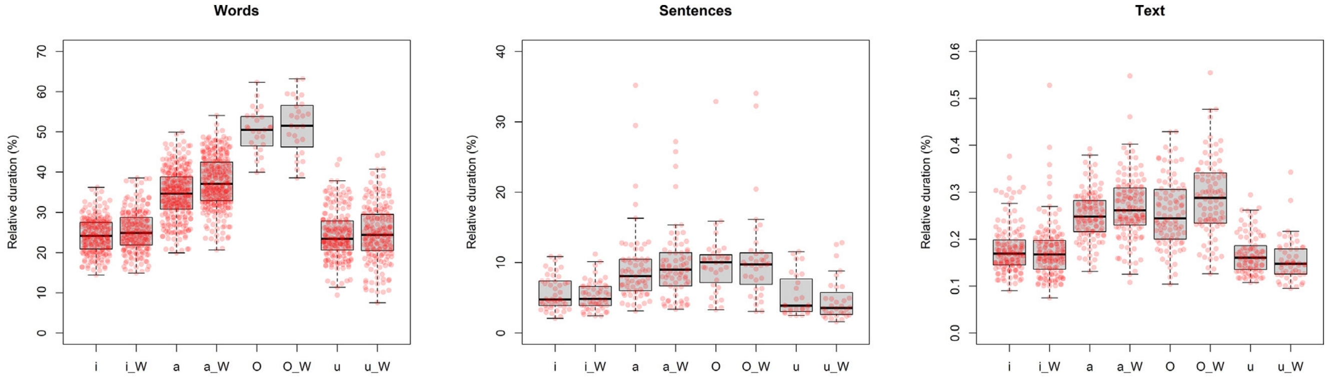

Fig. 7. Female vowel relative durations. The Speech Assessment Methods Phonetic Alphabet – SAMPA (Wells, 1997) is used to represent voiced vowels /i, a, O, u/ and their whispered counterparts /i_W, a_W, O_W, u_W/.

after a Kruskal-Wallis tests; one-tailed p-value) than for postalveolar fricatives /ʃ, ʒ/, in all four speech tasks. There was not a significant difference between same-place whispered and voiced fricatives produced in sentences, but some voiceless fricatives were produced with significantly lower F_{BP} values in whispered speech than in voiced speech (see Tables B.7.1 and B.7.2): Both female's and male's sustained /s/; female's /s/ in text; female's /ʃ/ in words.

Fricatives' spectral broad peak level (L_{BP}) results (shown in Fig. 9 for female speakers) did not reveal consistent differences between places of articulation. Nevertheless, more posterior fricatives in text tended to be produced with significantly higher L_{BP} values both for female ($p_{/s/-/ʃ/\ \mathrm{voiced}} = 0.000$; $p_{/s/-/ʃ/\ \mathrm{whispered}} = 0.000$; $p_{/z/-/ʒ/\ \mathrm{whispered}} = 0.001$; ANOVA with Bonferroni correction; one-tailed p-values) and male ($p_{/s/-/ʃ/\ \mathrm{voiced}} = 0.003$; $p_{/z/-/ʒ/\ \mathrm{voiced}} = 0.002$; $p_{/s/-/ʃ/\ \mathrm{whispered}} = 0.000$; $p_{/z/-/ʒ/\ \mathrm{whispered}} = 0.004$; ANOVA with Bonferroni correction; one-tailed p-values) speakers. More posterior fricatives in words were produced with significantly higher L_{BP} values only by female speakers in whispered speech ($p_{/s/-/ʃ/} = 0.000$; $p_{/z/-/ʒ/} = 0.001$; ANOVA with Bonferroni correction; one-tailed p-values).

The fricative spectral broad peak level is maximised for a higher source strength (Jesus & Shadle, 2002, p. 448) so we could predict that it would be higher in voiced speech mode than in whispered speech. Voiceless fricatives /s, ʃ/ were indeed produced with a significantly higher L_{BP} value in voiced than in whispered speech (see Tables B.7.3 and B.7.4), with the exception of /ʃ/ produced by male speakers in sentences. Voiced fricative's L_{BP} results were not significantly different in the two speech modes, the only exceptions were /z/ produced in words by male speakers (see Table B.7.4) and /z, ʒ/ produced in words by female speakers (see Table B.7.3).

4.8. Fricative sound pressure level (SPL)

Results for female and male speakers, and for the four speech tasks (see Table B.8.1 and B.8.2) revealed that whispered speech SPL was significantly lower than voiced speech,

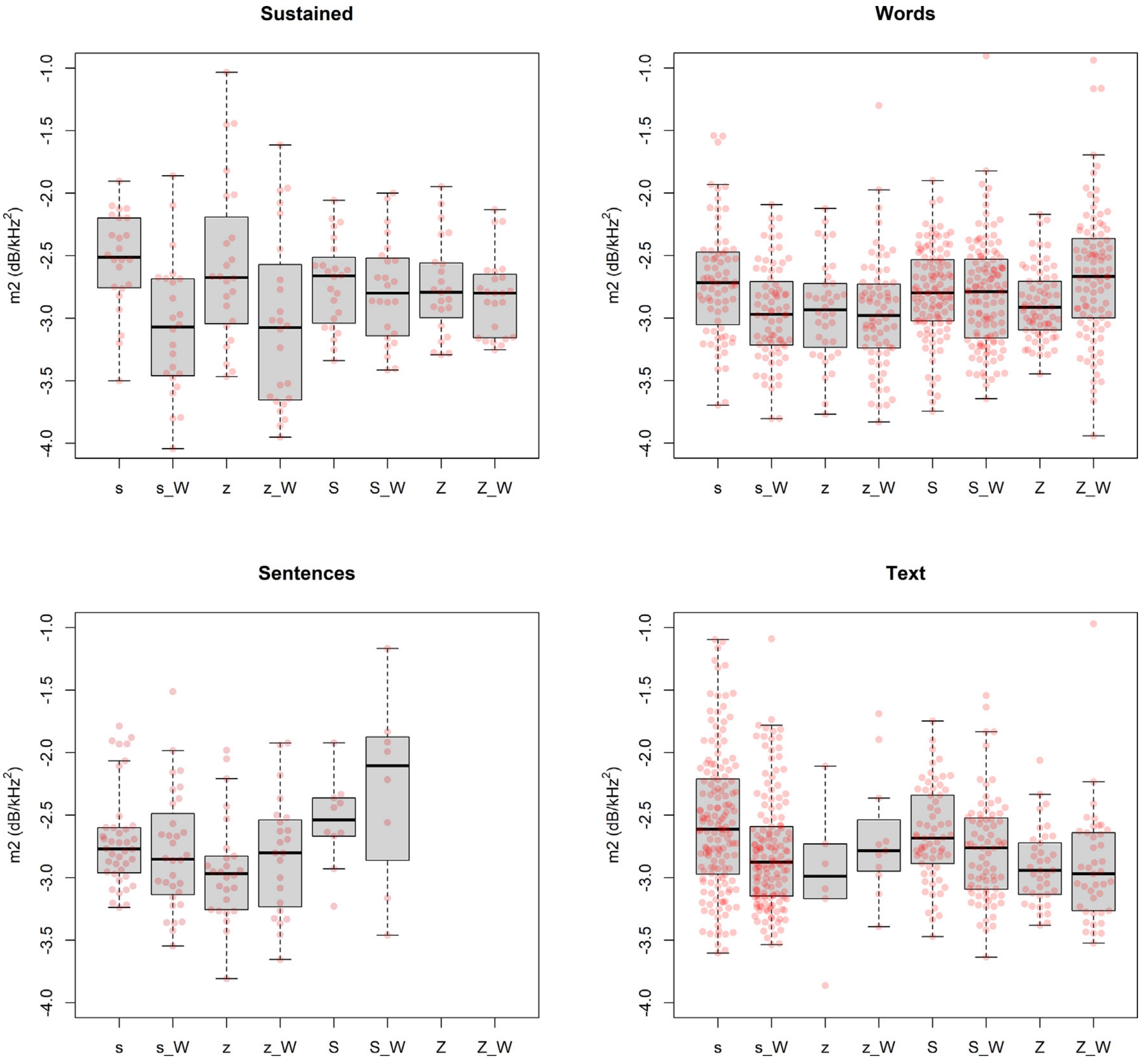

Fig. 8. High frequencies spectral slope (m2) of fricatives produced by female speakers. The Speech Assessment Methods Phonetic Alphabet – SAMPA (Wells, 1997) is used to represent voiced fricatives /s, z, S, Z/ and their whispered counterparts /s_W, z_W, S_W, Z_W/.

when the same fricative was compared in the two speech modes, except for /ʃ/ produced by male speakers in sentences.

The SPL was not significantly different between voiceless and voiced fricatives with the same place of articulation (between /s/ and /z/, and between /ʃ/ and /ʒ/, both for whispered and voiced speech modes), with the exception of male ($p_{/s/-/z/}$ = 0.007; $p_{/ʃ/-/ʒ/}$ = 0.000; Dunn's nonparametric comparison for post hoc testing after a Kruskal-Wallis test; one-tailed p-values) and female ($p_{/s/-/z/}$ = 0.029; $p_{/ʃ/-/ʒ/}$ = 0.000; ANOVA with Bonferroni correction; one-tailed p-values) sustained fricatives in voiced speech mode, female sustained fricatives in whispered speech ($p_{/s/-/z/}$ = 0.014; $p_{/ʃ/-/ʒ/}$ = 0.001; ANOVA with Bonferroni correction; one-tailed p-values), and whispered /s/-/z/ produced in sentences by male speakers (p = 0.009; ANOVA with Bonferroni correction; one-tailed p-value).

4.9. Fricatives' absolute and relative durations

Female and male absolute durations of same-place voiceless fricatives were only significantly different (p = 0.000; Dunn's nonparametric comparison for post hoc testing after a Kruskal-Wallis test; two-tailed p-values) from voiced fricatives (/s/ versus /z/ and /ʃ/ versus /ʒ/) for the voiced speech mode. Females produced, in voiced speech mode, voiceless fricatives /s, ʃ/ that were significantly longer (see Table B.9.1) than the same fricatives produced in whispered speech, for all speech tasks. Male /s, ʃ/ in words and /s/ in text were also significantly longer (see Table B.9.2) when produced in voiced than in whispered speech. Both female and male speakers produced significantly shorter /z, ʒ/ (see Tables B.9.1 and B.9.2) in voiced speech mode than in whisper, except for female's and male's /z/ in sentences and text, and male's /ʒ/ in words.

When probability density distributions of durations were also analysed, there was an overlap between voiceless and voiced same-place fricatives for all tasks, both in voiced and whispered speech, as shown in Fig. 10 for /s, z/ produced by female speakers in sentences.

The relative duration of same-place and speech mode voiceless fricatives was significantly higher (see Tables B.9.3 and B.9.4) than voiced fricatives, except for female /ʃ/-/ʒ/ produced in text (both speech modes) and /s/-/z/ produced in whispered words, and male voiced /ʃ/-/ʒ/ in text.

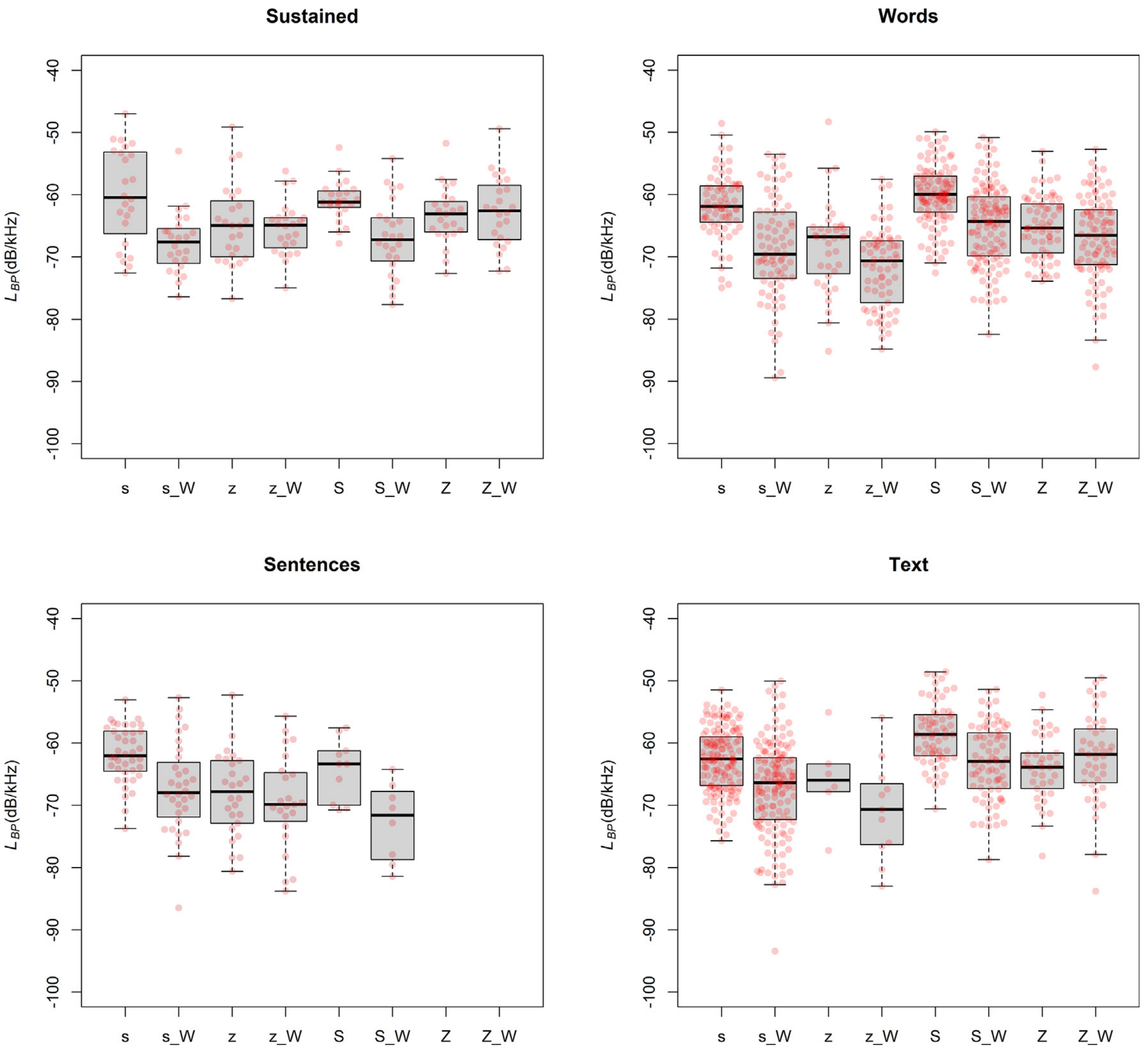

Fig. 9. Female speakers' fricative L_{BP}. The Speech Assessment Methods Phonetic Alphabet – SAMPA (Wells, 1997) is used to represent voiced fricatives /s, z, S, Z/ and their whispered counterparts /s_W, z_W, S_W, Z_W/.

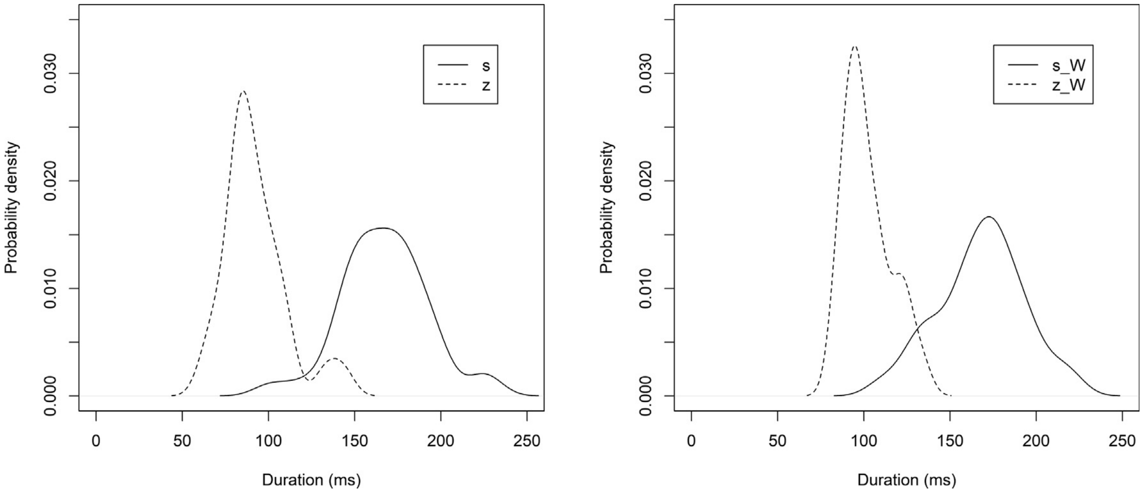

Fig. 10. Probability density distributions of voiced (left) and whispered (right) fricatives /s, z/ produced by female speakers in sentences.

4.10. Mixed-effects logistic regression (MELR) models

Considering the MELR model for the vowels in words, the overall percentage of correct classification was 99.8 %. From the outputs of the tests of fixed effects one could conclude that predictors F_1 (F = 33.860, p = 0.000), F_2 (F = 90.232, p = 0.000) and relative duration (F = 5.454, p = 0.020) had a significant effect on the disambiguation between /i/ and /a/. Estimates (and other statistics) for the fixed coefficients of the MELR model are summarised in Table C.1.1 (Appendix C). Speech mode (a factor with two levels) had a statistically significant effect (F = 8.468, p = 0.004) and the odds of choosing /a/ instead of /i/ were estimated to be 68 % lower for whispered speech, all other things being equal.

As to the MELR model for the vowels in sentences, the overall percentage of correct classification was 99.5 %. From the outputs of the tests of fixed effects it could be concluded that predictors F_1 (F = 13.454, p = 0.000) and F_2 (F = 34.503, p = 0.000) had a significant effect on the disambiguation between /i/ and /a/. Estimates for the fixed coefficients, presented in Table C.1.2, showed that the odds of choosing /a/ instead of /i/ were 72 % lower for whispered speech when compared to voiced speech.

In relation to the MELR model for vowels in text, the overall percentage of correct classification was 99.6 %. From the outputs of the tests of fixed effects it could be concluded that predictors F_1 (F = 29.082, p = 0.000) and F_2 (F = 51.346, p = 0.000) had a significant effect on the disambiguation between /i, a/. Speech mode also had a significantly different effect (F = 7.243, p = 0.007). Estimates for the fixed coefficients, presented in Table C.1.3, showed that the odds of choosing /a/ instead of /i/ were 78 % lower for whispered speech than for voiced speech.

Regarding the MELR model for fricatives in words, the overall percentage of correct classification was 99.5 %. From the outputs of the tests of fixed effects one could conclude that the predictors F_{BP} (F = 94.373, p = 0.000) and m1 (F = 4.238, p = 0.040) had a significant effect on the disambiguation between /s/ and /ʃ/. Different levels of factor sex had significantly different effects (F = 31.743, p = 0.000) and estimates for the fixed coefficients, presented in Table C.2.1, showed that the odds of choosing /ʃ/ instead of /s/ were 20 times higher for female than male speakers.

The overall percentage of correct classification of the MELR model for fricatives in sentences was 98.3 %. From the outputs of the tests of fixed effects one could conclude that predictors F_{BP} (F = 20.131, p = 0.000) and relative duration (F = 7.155, p = 0.008) had a significant effect on the disambiguation /s, ʃ/. Estimates for the fixed coefficients, presented in Table C.2.2, revealed that the odds of choosing /ʃ/ instead of /s/ were 2.79 times higher for female than male speakers.

For the MELR model for fricatives in text, the overall percentage of correct classification was 99.0 %. From the outputs of the tests of fixed effects one could conclude that predictors F_{BP} (F = 75.555, p = 0.000), m1 (F = 28.465, p = 0.000) and relative duration (F = 22.842, p = 0.000) had a significant effect on the disambiguation /s, ʃ/. Speech mode also had a significantly different effect (F = 4.923, p = 0.027). Estimates for the fixed coefficients, presented in Table C.2.3, show that the odds of choosing /ʃ/ instead of /s/ were 71 % lower for whispered speech when compared to voiced speech.

5. Discussion

This section provides a discussion of the evidence produced to explore each one of the research questions, presented in the same order (research questions 1 to 6) as they were introduced in section 2. It will also focus on synthesising the evidence that supports (or rebuts) the nine hypothesis (H1.1, H1.2, H2, H3, H4, H5.1, H5.2, H6.1 and H6.2) that were defined in the second section of this paper, and includes some implications for clinical practice and a brief account of the limitations of the study.

Table 2
Direction of change for the significant differences, with the symbol ▲ meaning that the parameter values are higher in whispered than in voiced speech, and the symbol ▼ meaning that the parameter values are lower in whispered than in voiced speech, for vowels produced by female (♀) and male (♂) speakers.

		F_1		F_2		F_3		m		SPL	
		♀	♂	♀	♂	♀	♂	♀	♂	♀	♂
Task	Vowel										
Sustained	/i/	▲	▲					▲	▲	▼	▼
	/a/	▲	▲	▲	▲	▲	▲	▲	▲	▼	▼
	/ɔ/	▲	▲		▲		▲	▲	▲	▼	▼
	/u/	▲	▲	▲	▲	▲	▲	▲	▲	▼	▼
Words	/i/	▲	▲	▼				▲	▲	▼	▼
	/a/	▲	▲	▲	▲		▲	▲	▲	▼	▼
	/ɔ/	▲	▲	▲	▲		▲	▲	▲	▼	▼
	/u/	▲	▲	▲	▲	▲	▲	▲	▲	▼	▼
Sentences	/i/	▲	▲	▼				▲	▲	▼	▼
	/a/	▲	▲	▲	▲		▲	▲	▲	▼	▼
	/ɔ/	▲	▲	▲	▲		▲	▲	▲	▼	▼
	/u/	▲	▲	▲	▲	▲		▲	▲	▼	▼
Text	/i/	▲	▲	▼	▲		▲	▲	▲	▼	▼
	/a/	▲	▲	▲	▲	▲	▲	▲	▲	▼	▼
	/ɔ/	▲	▲	▲	▲	▲	▲	▲	▲	▼	▼
	/u/	▲	▲	▲	▲			▲	▲	▼	▼

Table 3
Synthesis of formant frequency results related to front and back cavity sizes, with the symbol ▲ meaning that the parameter values are higher in whispered than in voiced speech, and the symbol ▼ meaning that the parameter values are lower in whispered than in voiced speech.

/i/	F_1	F_2	F_3
Cavity affiliation	Helmholtz	Back cavity	Front cavity
Formant results	all F_1 ▲	F_2 ▼ not significantly different in 7/8 cases	Not significantly different in in 7/8 cases
Suggests	Shorter back cavity	Not clear	Stable front cavity
/a/	F_1	F_2	F_3
Cavity affiliation	Back	Front	Back
Formant results	all F_1 ▲	F_2 ▲ in 7/8 cases	F_3 ▲ in 6/8 cases
Suggests	Shorter back cavity	Shorter front cavity	Shorter back cavity

6. Research questions and Hypothesis

Regarding the first research question, clear evidence, summarised in Table 2, has been presented supporting that vowels are produced with significantly different F_1 and F_2 formant frequencies, spectral slope and sound pressure level in Portuguese voiced and whispered speech, as previously observed for English (Sharifzadeh et al., 2012), Japanese (Higashikawa et al., 1996), Polish (Zygis et al., 2017), Serbian (Jovičić & Šarić, 2008), and Swedish (I. Eklund & Traunmuller, 1996). Results indicate that in whispered speech there is a much flatter spectrum (H1.2) due to a lowering of F_1 amplitude into the range of F_2 and an increase in high frequency content (mean slope differences between 2 and 8 dB/kHz^2).

Hypothesis H1.1 seems plausible: The back cavity is likely to be shorter in whispered speech because the close-front unrounded vowels' Helmholtz resonance and the open-front unrounded back cavity resonance frequency are both significantly higher in whispered speech than in voiced speech mode. This may result from raising of the larynx and narrowing of the vocal tract around the ventricular folds (Matsuda & Kasuya, 1999) for whispered speech production. The close-front unrounded vowels' F_3 frequency values were only significantly different ($p = 0.029$) in voiced and whispered speech for men's /i/ in text, so the front cavity is likely to have the same cross-sectional area and length in the two speech modes (voiced and whispered). This is further supported by the fact that the close-front unrounded vowels' F_2 (affiliated with a back-cavity resonance) frequency was significantly different in the two speech modes for female's and male's productions in text, and female's /i/ in words and sentences. The data, synthesised in Table 3, do not provide a firm support for H1.1.

Vowels' F_1 frequency shifts in whispered speech and their correlation to the same speaker's voiced speech f_o were at the core of our second research question, so f_o was estimated for all voiced vowels, and a positive correlation between f_o values and F_1 shifts, relative to same-sex reference voiced F_1, in whispered speech was only found when analysing all of the female tasks together. Hypothesis H2 regarding perceptual compensation for the missing fundamental could not be confirmed in general across sexes. We cannot extrapolate from our data conclusions that are related to intrinsic f_o (Heeren, 2015; Jacewicz & Fox, 2015). The speakers imposed an intonational contour on all the tasks that we did not control for rigorously, so it is not possible to know what was the actual effect of intrinsic f_o, usually a small effect when compared to the magnitude of intonational f_o changes.

The third research question, that pertains to the VSA, was tackled using an analysis for each speech task and sex (given well known anatomical differences that shift formant frequency

Table 4
Direction of change for the significant differences, with the symbol ▲ meaning that the parameter values are higher in whispered than in voiced speech, and the symbol ▼ meaning that the parameter values are lower in whispered than in voiced speech, for fricatives produced by female (♀) and male (♂) speakers.

		m1		m2		F_{BP}		L_{BP}		SPL	
		♀	♂	♀	♂	♀	♂	♀	♂	♀	♂
Task	Fricative										
Sustained	/s/		▼	▼	▼	▼	▼	▼	▼	▼	▼
	/z/	▲	▲	▼						▼	▼
	/ʃ/							▼	▼	▼	▼
	/ʒ/	▲	▲							▼	▼
Words	/s/		▼	▼	▼			▼	▼	▼	▼
	/z/	▲	▲					▼	▼	▼	▼
	/ʃ/		▼			▼		▼	▼	▼	▼
	/ʒ/	▲	▲	▲	▲			▼		▼	▼
Sentences	/s/							▼	▼	▼	▼
	/z/	▲								▼	▼
	/ʃ/				▲			▼		▼	
Text	/s/			▼	▼	▼		▼	▼	▼	▼
	/z/		▲							▼	▼
	/ʃ/							▼	▼	▼	▼
	/ʒ/	▲	▲							▼	▼

values). Our results show a more compressed VSA in whispered speech than in equivalent voiced speech tasks, but the areas were not significantly different, both for female and male speakers (H3). This pattern (compression of the VSA) had not previously observed in studies of English (Sharifzadeh et al., 2012) and Japanese (Higashikawa et al., 1996) due to limitations in the size of the corpora and breadth of speech tasks.

Evidence gathered from this study's data regarding the fourth research question showed that close /i, u/ vowels durations were significantly shorter than close/open-mid vowels /a, ɔ/ both in voiced and whispered speech. Absolute and relative durations of vowels calculated for voiced and whispered speech revealed that vowel intrinsic durations (Abramson & Ren, 1990) were used as secondary (to F_1 and F_2 discussed above) acoustic cues to differentiate close /i, u/ from open/open-mid /a, ɔ/ vowels. We could therefore corroborate H4: Intrinsic vowel durations are observed both in voiced and whispered speech. Furthermore, two acoustic cues (intrinsic vowel durations and formant frequencies) have a clear role in the identification of vowels, i.e., a correlation between vowel duration and quality has been observed (Abramson & Ren, 1990).

The fifth research question tapped into acoustic parameters (shown in Table 4) estimated from frication signals that reveal acoustic characteristics in voiced and whispered speech. Results showed that the low frequencies slope (m1) of voiced sibilants was significantly different in the two speech modes (H5.1), the m2 values were lower in whispered /s/ than in voiced /s/, and that the whispered speech SPL was significantly lower than voiced speech mode, as summarised in Table 4. A greater noise source strength is related to higher m1 values (Jesus & Shadle, 2002), as seen in voiced speech mode. Voiceless fricatives' spectral broad peak frequency was significantly lower for female' /s/ in sustained and text speech tasks and /ʃ/ in words, and male's sustained /s/. The spectral broad peak level of both voiceless and voiced fricatives was lower in whispered than in voiced speech mode, except for females' and males' /z/ in sustained, sentences and text tasks, females' and males' /ʒ/ in sustained and text tasks, and males' /ʒ/ in words. Voiced fricatives were significantly shorter in voiced speech mode than in whispered speech (Jovičić & Šarić, 2008), and voiceless fricatives' durations shifted in the opposite direction: /s, ʃ/ were longer when produced in voiced speech mode than the same fricatives produced in whispered speech, except for women's /ʃ/ in words.

Summarising, results for m1 revealed that the source strength is not generally significantly different between voiceless fricatives produced in the two speech modes and is significantly different for voiced fricatives. Place of articulation had a significant effect on m1 values (Jesus & Shadle, 2002), both in voiced and whispered speech (see Tables B.6.4.1 and B.6.4.2 in Appendix B). Most high frequencies spectral slopes (m2) of whispered fricatives were not significantly different from the slopes observed in voiced speech mode, not supporting H5.1, i.e., that fricatives are produced in whispered speech with a lower source strength than in voiced speech mode.

The sibilant fricatives' spectral broad peak results from a front cavity resonance that shifts in frequency with the place of articulation (Jesus & Shadle, 2002, p. 447). This was observed both for whispered and voiced speech modes, when comparing the F_{BP} values of alveolar fricatives /s, z/ with those of postalveolar fricatives /ʃ, ʒ/. There was not a significant difference between same-place whispered and voiced fricatives produced in sentences, which is likely due to a very stable place of articulation in both speech modes. However, some voiceless fricatives, in specific tasks, were produced with a significantly larger front cavity (lower F_{BP} values) in whispered speech than in voiced speech.

The relative duration of same-place and speech mode voiceless fricatives was longer than voiced fricatives both in voiced and whispered speech, thus supporting H5.2 When comparing the probability density functions of /s/ durations with those of /z/, and those of /ʃ/ with /ʒ/, voicing categories are far from bimodal as previously pointed out by Crystal and House (1988). The use of durational cues "must be probabilistic, since the distributions for individual classes are" extensively overlapped (Crystal & House, 1988, p. 1935). Nevertheless, one possible cue for voicing in whispered speech is frication duration as previously suggested by Tartter (1989).

The sixth research question, about the acoustic signal attributes that may be able to disambiguate vowel height (H6.1) and fricative place of articulation (H6.2), was addressed using the MELR models (Cho, 2015; Holt et al., 2015; Jesus & Shadle, 2002). The odds of choosing /a/ instead of /i/ were estimated to be lower for whispered speech when compared to voiced speech, also decreasing with an increase of F_2 frequency values. The probability of choosing /a/ instead of /i/, on the other hand, was estimated to increase for higher F_1 frequency values, and to also increase when the vowels were longer. Also, even though the effects were not statistically significant with this data, the odds of choosing /a/ instead of /i/, seem to be much higher for female speakers when compared to male speakers.

The results for fricatives (H6.2) were not as consistent as for vowels, when comparing different speech tasks (Jesus & Shadle, 2002; Zygis, et al., 2017). Nevertheless, a few remarks can be made and the first one is that F_{BP} is always statistically significant when discriminating between /s/ and /ʃ/, for all speech tasks, and it is estimated that the odds of choosing /ʃ/ instead of /s/ decreases with an increase in F_{BP}. Considering disyllabic words and sentences, the probability of choosing /ʃ/ instead of /s/ was generally estimated to be higher for female speakers when compared to male speakers.

6.1. Implications for clinical practice

People with vocal fold paralysis or psychogenic dysphonia, resulting in a whispered voice, may benefit from alternative communication resources based on voice reconstruction. Since decreased loudness is often observed in their speech, those people may experience lack of normal conversational interactions and limited prosodic variation, leading to an atypical voice with professional and social effects (Boone et al., 2020), so electronic devices to project or reconstruct the voice could help with that difficulty (Stewart & Allen, 2006). Furthermore, the mood of a person, with changes in prosodic and rhythm patterns of vocalisation, can be heard in a healthy voice (Boone et al., 2020).

The results of this study were used to define the heuristics of a voice reconstruction system (Silva et al., 2021), based on whispered speech samples, that is capable of producing a more natural voice and reflecting the person's identity (e.g., segmental durations in both voiced and whispered speech realisations).

When reconstructing Portuguese voiced sounds from whispered speech, in real time, in addition to a careful phoneme-oriented segmentation, the algorithm operation uses segmental cues to vowel height and consonantal place and voicing in whispered speech to define strategies for the conversion process.

The technology that we are developing takes as input whispered speech, identifies those regions in the signal that would be voiced in natural speech, and implants there synthetic voicing. This reconstructed signal is carefully shaped in frequency and time, considering acoustic differences between the two speech modes (voiced and whispered) as reported in this paper, enhancing the linguistic content of the resulting synthetic speech, to improve voice projection.

6.2. Limitations of the study

The limited number of speakers and data loss conditioned the power of statistical analysis and generalisation of predictions. The reduced naturalness of speech samples and lack of real-world communication scenarios has not yet allowed us to test whispered speech in settings where it spontaneously used. The links between whispered speech production and perception were not explored, but hold great potential as a source of new evidence.

7. Conclusions

The segmental cues to place and voicing in two very distinct speech modes (voice and whisper) and four increasingly realistic speech tasks (sustained → words → sentences → text), were at the core of this study. In this section, conclusions regarding speech patterns that hold across these conditions and could help define what phonetic correlates constitute sufficiently distinct information to disambiguate vowel height (/i/ and /a/), fricative place (/s/ and /ʃ/; /z/ and /ʒ/) and voicing (/s/ and /z/; /ʃ/ and /ʒ/), will be synthesised. Conclusions regarding source and filter adjustments during whispered speech are also presented, providing new interpretations for production mechanisms that can be helpful in improving voice rehabilitation processes.

The parameters used to analyse the source of vowels (f_o, m and SPL) revealed that whispered vowels are produced with a weaker (around 20 dB) sound source than voiced vowels. The parameters used to analyse the filter characteristics of vowels (F_1, F_2 and F_3) revealed some evidence of a stable front cavity in both speech modes and a shorter back cavity used to produce whispered speech. Intrinsic vowel durations along with formant frequencies had a role in differentiating vowel height in both speech modes, and there was some evidence for a raised larynx and narrowing of the posterior vocal tract in whispered speech F_1 and F_2 frequency values. A significant positive correlation was found between voiced and whispered F_1, F_2 and F_3 frequencies of female and male speakers.

The fricatives' source strength was not always significantly different between fricatives produced in the two speech modes, but L_{BP}, and SPL results represent strong evidence that the source strength of whispered speech is lower than voiced speech; place of articulation had a significant effect on source strength, both in voiced and whispered speech. The parameters (F_{BP} and L_{BP}) expected to correspond to the first front cavity resonance (fricative filter characteristics) revealed the same shifts in frequency (F_{BP}) with the place of articulation in whispered and voiced speech modes; voiceless fricatives /s, ʃ/ were produced with a significantly higher spectral broad peak level L_{BP} value in voiced than in whispered speech, but voiced fricative's L_{BP} results were not significantly different (the only exception being females' and males' /z/ in words, and females' /ʒ/ in words) in the two speech modes. Since L_{BP} is maximised for a higher source strength this constitutes further evidence that voiceless fricatives are produced with a weaker source in whispered speech. The relative duration of same-place and speech mode voiceless fricatives was higher than voiced fricatives both in voiced and whispered speech, constituting the only viable cue for voicing in whisper.

The MELR models revealed that the acoustic signal attributes F_1 and F_2 frequencies, and intrinsic durations carry sufficiently distinct information to disambiguate /i/ from /a/ in all speech tasks. Fricatives' F_{BP} was statistically significant when discriminating /s/ from /ʃ/, for all speech tasks.

This study, with data collected from different speech tasks, shows that changes during whispered speech production can be observed both in the laryngeal (source) and vocal tract (filter) configurations. Therefore, clinicians who use the whispered speech technique for voice rehabilitation, usually centred on the absence of vocal fold vibration, should also consider changes in the vocal tract configuration. Given the fact that the most significant supraglottic configuration adjustments were observed, in this study, for vowel production, clinicians could base their whispered speech techniques on voice samples that combine these with fricatives, shown in this paper to be produced with a very stable vocal tract configuration. This has the potential of smoothing out the transition into voice therapy that gradually engages the larynx in the rehabilitation process.

We are currently reconstructing voiced speech from whispered signals, and will be integrating the evidence presented in this paper in speech synthesis processes based on the Source-Filter Theory of Speech Production.

CRediT authorship contribution statement

Luis M.T. Jesus: Conceptualization, Methodology, Software, Investigation, Writing – review & editing. **Sara Castilho:** Investigation, Data curation, Writing – original draft. **Aníbal Ferreira:** Funding acquisition, Project administration. **Maria Conceição Costa:** Formal analysis.

Acknowledgements

This work was financially supported by Project PTDC/EMD-EMD/29308/2017 - POCI-01-0145-FEDER-029308 - funded by FEDER funds through COMPETE2020 - Programa Operacional Competitividade e Internacionalização (POCI) and by national funds (PIDDAC) through FCT/MCTES. Support was also received from National Funds through the FCT - Foundation for Science and Technology, in the context of the projects UIDB/00127/2020 and UIDB/04106/2020.

Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.wocn.2023.101223.

References

Abramson, A. S., & Ren, N. (1990). Distinctive vowel length: Duration vs. spectrum in Thai. *Journal of Phonetics, 18*(2), 79–92. https://doi.org/10.1016/S0095-4470(19)30395-X.

Baker, J. (2016). Functional voice disorders. In M. Hallett, J. Stone, & A. Carson (Eds.), *Handbook of Clinical Neurology* (pp. 389–405). Elsevier. https://doi.org/10.1016/B978-0-12-801772-2.00034-5.

Benninger, M. S., Finnegan, E. M., Kraus, D. H., Sterman, B. M., Miller, R., Carwell, M. A., & Levine, H. L. (1988). The whisper and the whistle: The role in vocal trauma. *Medical Problems of Performing Arts, 3*(4), 151–154.

Bisol, L., & Veloso, J. (2016). Phonological processes affecting vowels. In W. L. M. Wetzels, J. Costa, & S. Menuzzi (Eds.), *The Handbook of Portuguese Linguistics* (pp. 69–85). Wiley. https://doi.org/10.1002/9781118791844.ch5.

Boone, D., McFarlane, S., Berg, S., & Zraick, R. (2020). *The Voice and Voice Therapy* (10th ed.). Pearson.

Burris, C., Vorperian, H. K., Fourakis, M., Kent, R. D., & Bolt, D. M. (2014). Quantitative and descriptive comparison of four acoustic analysis systems: Vowel measurements. *Journal of Speech, Language, and Hearing Research, 57*(1), 26–45. https://doi.org/10.1044/1092-4388(2013/12-0103).

Cho, T. (2015). Language effects on timing at the segmental and suprasegmental levels. In M. A. Redford (Ed.), *The Handbook of Speech Production* (pp. 505–529). Wiley. https://doi.org/10.1002/9781118584156.ch22.

Cirillo, J., & Todt, D. (2005). Perception and judgement of whispered vocalisations. *Behaviour, 142*(1), 113–129. https://doi.org/10.1163/1568539053627758.

Colton, R., Casper, J., & Leonard, R. J. (2011). *Understanding Voice Problems: A Physiological Perspective for Diagnosis and Treatment* (4th Ed.). Lippincott Williams & Wilkins.

Crystal, T. H., & House, A. S. (1988). A note on the durations of fricatives in American English. *The Journal of the Acoustical Society of America, 84*(5), 1932–1935.

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.

Eisenhauer, J. G. (2009). Explanatory power and statistical significance. *Teaching Statistics, 31*(2), 42–46.

Eklund, A. (2021). *The Bee Swarm Plot: An Alternative to Stripchart*. Available from https://github.com/aroneklund/beeswarm.

Eklund, I., & Traunmuller, H. (1996). A comparative study of male and female whispered and phonated versions of the long vowels of Swedish. *TMH-QPSR, 2*, 131–134.

Escudero, P., Boersma, P., Rauber, A. S., & Bion, R. A. H. (2009). A cross-dialect acoustic description of vowels: Brazilian and European Portuguese. *The Journal of the Acoustical Society of America, 126*(3), 1379–1393.

Fant, G. (1970). *Acoustic Theory of Speech Production* (2nd Ed.). Mouton.

Ferreira, M., Jesus, L. M. T., Couto, P., & Vilarinho, H. (2014). University of Aveiro's standardised voice case history form. *Revista de Saúde Pública, 48*, 297.

Fleischer, S., Kothe, C., & Hess, M. (2007). Die Kehlkopfkonfiguration beim Flüstern [glottal and supraglottal configuration during whispering]. *Laryngo-Rhino-Otologie, 86*(4), 271–275. https://doi.org/10.1055/s-2006-945000.

Hansen, J. (1989). Evaluation of acoustic correlates of speech under stress for robust speech recognition. *Proceedings of the Fifteenth Annual Northeast Bioengineering Conference*, 31–32. https://doi.org/10.1109/NEBC.1989.36683.

Hansen, J., & Varadarajan, V. (2009). Analysis and compensation of Lombard speech across noise type and levels with application to in-set/out-of-set speaker recognition. *IEEE Transactions on Audio, Speech, and Language Processing, 17*(2), 366–378. https://doi.org/10.1109/TASL.2008.2009019.

Heeren, W. F. L. (2015). Coding pitch differences in voiceless fricatives: Whispered relative to normal speech. *The Journal of the Acoustical Society of America, 138*(6), 3427–3438. https://doi.org/10.1121/1.4936859.

Heeren, W. F. L., & Heuven, V. J. (2014). The interaction of lexical and phrasal prosody in whispered speech. *The Journal of the Acoustical Society of America, 136*(6), 3272–3289. https://doi.org/10.1121/1.4901705.

Higashikawa, M., & Minifie, F. D. (1999). Acoustical-perceptual correlates of "whisper pitch" in synthetically generated vowels. *Journal of Speech, Language, and Hearing Research, 42*(3), 583–591. https://doi.org/10.1044/jslhr.4203.583.

Higashikawa, M., Nakai, K., Sakakura, A., & Takahashi, H. (1996). Perceived pitch of whispered vowels relationship with formant frequencies: A preliminary study. *Journal of Voice, 10*(2), 155–158.

Holt, Y. F., Jacewicz, E., & Fox, R. A. (2015). Variation in vowel duration among southern African American english speakers. *American Journal of Speech-Language Pathology, 24*(3), 460–469. https://doi.org/10.1044/2015_AJSLP-14-0186.

Hufnagle, J., & Hufnagle, K. (1983). Is quiet whisper harmful to the vocal mechanism? A research note. *Perceptual and Motor Skills, 57*(3), 735–737. https://doi.org/10.2466/pms.1983.57.3.735.

Ito, T., Takeda, K., & Itakura, F. (2005). Analysis and recognition of whispered speech. *Speech Communication, 45*(2), 139–152. https://doi.org/10.1016/j.specom.2003.10.005.

Jacewicz, E., & Fox, R. A. (2015). Intrinsic fundamental frequency of vowels is moderated by regional dialect. *The Journal of the Acoustical Society of America*, *138*(4), EL405–EL410. https://doi.org/10.1121/1.4934178.

Jesus, L. M. T. (2001). *Acoustic Phonetics of European Portuguese Fricative Consonants*. University of Southampton, UK. Ph.D. Thesis,.

Jesus, L. M. T., Belo, I., Machado, J., & Hall, A. (2017). The Advanced Voice Function Assessment Databases (AVFAD): Tools for voice clinicians and speech engineering research. In F. Fernandes (Ed.), *Advances in Speech-Language Pathology* (pp. 237–255). InTech. https://doi.org/10.5772/intechopen.69643.

Jesus, L. M. T., Castilho, S., Alves, M., & Hall, A. (2022). An Open Access Standardised Voice Evaluation Protocol. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2021.09.010.

Jesus, L. M. T., & Shadle, C. H. (2002). A parametric study of the spectral characteristics of European Portuguese fricatives. *Journal of Phonetics, 30*(3), 437–464. https://doi.org/10.1006/jpho.2002.0169.

Jesus, L. M. T., & Shadle, C. H. (2003). Temporal and devoicing analysis of European Portuguese fricatives. In *15th International Congress of Phonetic Sciences (ICPhS 2003)* (Vol. 1, pp. 779–782).

Jesus, L. M. T., Tavares, A. I., & Hall, A. (2017). Cross-cultural adaption of the GRBAS and CAPE-V scales for Portugal and a new training programme for perceptual voice evaluation. In F. Fernandes (Ed.), *Advances in Speech-language Pathology* (pp. 221–236). InTech. https://doi.org/10.5772/intechopen.69644.

Jesus, L. M. T., Valente, A. R. S., & Hall, A. (2015). Is the Portuguese version of the passage "The North Wind and the Sun" phonetically balanced? *Journal of the International Phonetic Association, 45*(1), 1–11. https://doi.org/10.1017/S0025100314000255.

Jovičić, S. T., & Šarić, Z. (2008). Acoustic analysis of consonants in whispered speech. *Journal of Voice, 22*(3), 263–274. https://doi.org/10.1016/j.jvoice.2006.08.012.

Kallail, K. J., & Emanuel, F. W. (1984). Formant-frequency differences between isolated whispered and phonated vowel samples produced by adult female subjects. *Journal of Speech, Language, and Hearing Research, 27*(2), 245–251. https://doi.org/10.1044/jshr.2702.251.

Kent, R. D., & Rountrey, C. (2020). What acoustic studies tell us about vowels in developing and disordered speech. *American Journal of Speech-Language Pathology, 29*(3), 1749–1778.

Kent, R. D., & Vorperian, H. K. (2018). Static measurements of vowel formant frequencies and bandwidths: A review. *Journal of Communication Disorders, 74*, 74–97. https://doi.org/10.1016/j.jcomdis.2018.05.004.

Kohlberger, M., & Strycharczuk, P. (2015). Voicing assimilation in whispered speech. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*.

Konnai, R., Scherer, R. C., Peplinski, A., & Ryan, K. (2017). Whisper and phonation: Aerodynamic comparisons across adduction and loudness. *Journal of Voice, 31*(6), 773.e11–773.e20. https://doi.org/10.1016/j.jvoice.2017.02.016.

Konno, H., Kudo, M., Imai, H., & Sugimoto, M. (2016). Whisper to normal speech conversion using pitch estimated from spectrum. *Speech Communication, 83*, 10–20. https://doi.org/10.1016/j.specom.2016.07.001.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America, 105*(3), 1455–1468. https://doi.org/10.1121/1.426686.

Lindblom, B. (1996). Role of articulation in speech perception: Clues from production. *The Journal of the Acoustical Society of America, 99*(3), 1683–1692. https://doi.org/10.1121/1.414691.

Lousada, M., Jesus, L. M. T., & Hall, A. (2010). Temporal Acoustic Correlates of the Voicing Contrast in European Portuguese Stops. *Journal of the International Phonetic Association, 40*(3), 261–275. https://doi.org/10.1017/S0025100310000186.

Ma, E., Yiu, E., & Verdolini, K. (2007). Application of the ICF in voice disorders. *Seminars in Speech and Language, 28*(4), 343–350. https://doi.org/10.1055/s-2007-986531.

MacDonell, R., & Holmes, R. (2007). Motor speech and swallowing disorders. In A. H. V. Schapira, E. Byrne, R. S. J. Frackowiak, R. T. Johnson, Y. Mizuno, M. A. Samuels, S. D. Silberstein, ... & Z. K. Wszolek (Eds.), *Neurology and Clinical Neuroscience* (pp. 155–170). Elsevier: Mosby.

Marković, B., & Jovic˘ić, S. T., Galić, J., & Grozdić, ã. (2013). Whispered speech database: Design, processing and application. In I. Habernal & V. Matousek (Eds.), *TSD 2013, LNAI 8082* (pp. 591–598). Springer-Verlag. https://doi.org/10.1007/978-3-642-40585-3_74.

Matsuda, M., & Kasuya, H. (1999). Acoustic nature of the whisper. *Proceedings of Eurospeech, 99*, 133–136.

Maurer, D. (2016). Acoustics of the Vowel: Preliminaries. *Peter Lang*. https://doi.org/10.1519/JSC.0000000000000372.

McCloy, D. R. (2016). *Normalizing and Plotting Vowels with phonR 1.0.7*. USA: University of Washington.

Mertl, J., Žáčková, E., & Řepová, B. (2018). Quality of life of patients after total laryngectomy: The struggle against stigmatization and social exclusion using speech synthesis. *Disability and Rehabilitation: Assistive Technology, 13*(4), 342–352. https://doi.org/10.1080/17483107.2017.1319428.

Meynadier, Y. (2015). Aerodynamic tool for phonology of voicing. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*.

Meynadier, Y., & Gaydina, Y. (2013). Aerodynamic and durational cues of phonological voicing in whisper. *Proceedings of Interspeech, 2013*, 335–339. https://doi.org/10.1038/nchem.120.

Monoson, P., & Zemlin, W. R. (1984). Quantitative study of whisper. *Folia Phoniatrica et Logopaedica, 36*(2), 53–65. https://doi.org/10.1159/000265721.

Morris, R. W., & Clements, M. A. (2002). Reconstruction of speech from whispers. *Medical Engineering & Physics, 24*(7–8), 515–520. https://doi.org/10.1016/S1350-4533(02)00060-7.

Murry, T., & Brown, W. S. (1976). Peak intraoral air pressures in whispered stop consonants. *Journal of Phonetics, 4*(3), 183–187. https://doi.org/10.1016/S0095-4470(19)31242-2.

Narayanan, S. S., & Alwan, A. A. H. (2000). Noise source models for fricative consonants. *IEEE Transactions on Speech and Audio Processing, 8*(2), 328–344.

Oliveira, M. A. (2022). Machine Learning Approaches for Whisper to Normal Speech Conversion. *U.Porto. Journal of Engineering, 8*(2), 202–212. https://doi.org/10.24840/2183-6493_008.002_0016.

Pape, D., & Jesus, L. M. T. (2015). Stop and fricative devoicing in European Portuguese. *Italian and German. Language and Speech, 58*(2), 224–246. https://doi.org/10.1177/0023830914530604.

Perkell, J. S. (2012). Movement goals and feedback and feedforward control mechanisms in speech production. *Journal of Neurolinguistics, 25*(5), 382–407. https://doi.org/10.1016/j.jneuroling.2010.02.011.

Politzer-Ahles, S., & Piccinini, P. (2018). On visualizing phonetic data from repeated measures experiments with multiple random effects. *Journal of Phonetics, 70*, 56–69. https://doi.org/10.1016/j.wocn.2018.05.002.

Rubin, A. D., Praneetvatakul, V., Gherson, S., Moyer, C. A., & Sataloff, R. T. (2006). Laryngeal hyperfunction during whispering: Reality or myth? *Journal of Voice, 20*(1), 121–127. https://doi.org/10.1016/j.jvoice.2004.10.007.

Scherer, R. C., Sundberg, J., & Konnai, R. (2016). Whisper. In R. T. Sataloff & M. S. Benninger (Eds.). *Sataloff's Comprehensive Textbook of Otolaryngology: Head and Neck Surgery: Laryngology* (Vol. 4, pp. 81–87). Jaypee Brothers Medical Publishers.

Schwartz, M. F. (1972). Bilabial closure durations for /p/, /b/, and /m/ in voiced and whispered vowel environments. *The Journal of the Acoustical Society of America, 51*(6B), 2025–2029. https://doi.org/10.1121/1.1913063.

Segura, L. (2013). Variedades dialetais do Português Europeu. In E. B. P. Raposo, M. F. B. Nascimento, M. A. C. Mota, L. Segura, A. Mendes, G. Vicente, & R. Veloso (Eds.). *Gramática do Português* (Vol. 1, pp. 85–142). Fundação Calouste Gulbenkian.

Shadle, C. H. (2012). The acoustics and aerodynamics of fricatives. In A. Cohn, C. Fougeron, & M. Huffman (Eds.), *The Oxford Handbook of Laboratory Phonology* (pp. 511–526). Oxford: Oxford University Press.

Shadle, C. H., Nam, H., & Whalen, D. H. (2016). Comparing measurement errors for formants in synthetic and natural vowels. *The Journal of the Acoustical Society of America, 139*(2), 713–727. https://doi.org/10.1121/1.4940665.

Sharifzadeh, H. R., McLoughlin, I., & Ahmadi, F. (2010). Reconstruction of normal sounding speech for laryngectomy patients through a modified CELP codec. *IEEE Transactions on Biomedical Engineering, 57*(10), 2448–2458. https://doi.org/10.1109/TBME.2010.2053369.

Sharifzadeh, H. R., McLoughlin, I., & Russell, M. J. (2012). A comprehensive vowel space for whispered speech. *Journal of Voice, 26*(2), e49–e56. https://doi.org/10.1016/j.jvoice.2010.12.002.

Silva, J. P., Cardoso, C. F., Oliveira, M. A., Jesus, L. M. T., & Ferreira, A. J. S. (2021). A comparative study of European Portuguese stop consonants and fricatives in whispered speech and normal speech for real-time operation of voice conversion. *Proceedings of the 12th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA 2021)*, 53–56.

Slis, I. H., & Cohen, A. (1969a). On the complex regulating the voiced-voiceless distinction I. *Language and Speech, 12*(2), 80–102.

Slis, I. H., & Cohen, A. (1969b). On the complex regulating the voiced-voiceless distinction II. *Language and Speech, 12*(3), 137–155.

Smith, C. L. (1997). The devoicing of /z/ in American English: Effects of local and prosodic context. *Journal of Phonetics, 25*, 471–500.

Solomon, N. P., McCall, G. N., Trosset, M. W., & Gray, W. C. (1989). Laryngeal configuration and constriction during two types of whispering. *Journal of Speech, Language, and Hearing Research, 32*(1), 161–174. https://doi.org/10.1044/jshr.3201.161.

Stathopoulos, E. T., Hoit, J. D., Hixon, T. J., Watson, P. J., & Solomon, N. P. (1991). Respiratory and laryngeal function during whispering. *Journal of Speech, Language, and Hearing Research, 34*(4), 761–767. https://doi.org/10.1044/jshr.3404.761.

Stevens, K., Blumstein, S., Glicksman, L., Burton, M., & Kurowski, K. (1992). Acoustic and perceptual characteristics of voicing in fricatives and fricative clusters. *Journal of the Acoustical Society of America, 91*(5), 2979–3000.

Stewart, C., & Allen, E. (2006). Voice therapy for unilateral vocal fold paralysis. In L. Sulica & A. Blitzer (Eds.), *Vocal Fold Paralysis* (pp. 87–96). Springer-Verlag.

Sundberg, J., Scherer, R., Hess, M., & Müller, F. (2010). Whispering - A single-subject study of glottal configuration and aerodynamics. *Journal of Voice, 24*(5), 574–584. https://doi.org/10.1016/j.jvoice.2009.01.001.

Swerdlin, Y., Smith, J., & Wolfe, J. (2010). The effect of whisper and creak vocal mechanisms on vocal tract resonances. *The Journal of the Acoustical Society of America, 127*(4), 2590–2598. https://doi.org/10.1121/1.3316288.

Tartter, V. C. (1989). What's in a whisper? *The Journal of the Acoustical Society of America, 86*(5), 1678–1683. https://doi.org/10.1121/1.398598.

Thomson, D. J., & Haley, C. L. (2014). Spacing and shape of random peaks in non-parametric spectrum estimates. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 470*(2167). https://doi.org/10.1098/rspa.2014.0101.

Titze, I. R. (2000). *Principles of Voice Production* (2nd Print). *National Center for Voice and Speech*.

Tsunoda, K., Niimi, S., & Hirose, H. (1994). The roles of the posterior cricoarytenoid and thyropharyngeus muscles in whispered speech. *Folia Phoniatrica et Logopaedica, 46*(3), 139–151. https://doi.org/10.1159/000266306.

Vigário, M., Freitas, M., & Frota, S. (2006). Grammar and frequency effects in the acquisition of prosodic words in European Portuguese. *Language and Speech, 49*(2), 175–203.

Weismer, G., & Longstreth, D. (1980). Segmental gestures at the laryngeal level in whispered speech. *Journal of Speech, Language, and Hearing Research, 23*(2), 383–392. https://doi.org/10.1044/jshr.2302.383.

Wells, J. (1997). SAMPA computer readable phonetic alphabet. In D. Gibbon, R. Moore, & R. Winski (Eds.), *Handbook of Standards and Resources for Spoken Language Systems* (pp. 684–732). Mouton de Gruyter.

Whalen, D. H., Chen, W.-R., Shadle, C. H., & Fulop, S. A. (2022). Formants are easy to measure; resonances, not so much: Lessons from Klatt (1986). *The Journal of the Acoustical Society of America, 152*(2), 933–941. https://doi.org/10.1121/10.0013410.

Whalen, D. H., & Levitt, A. G. (1995). The universality of intrinsic F0 of vowels. *Journal of Phonetics, 23*(3), 349–366. https://doi.org/10.1016/S0095-4470(95)80165-0.

Winter, B. (2020). *Statistics for Linguists: An Introduction Using R*. Routledge.

Zhang, C., & Hansen, J. (2007). Analysis and classification of speech mode: Whispered through shouted. *Proceedings of Interspeech, 2007*, 2289–2292. https://doi.org/10.21437/Interspeech.2007-621.

Zhou, J., Hu, Y., Lian, H., Pang, C., Wang, H., & Tao, L. (2019). An audio-visual whisper database in Chinese. *Proceedings of ICSP, 2019*. https://doi.org/10.1088/1742-6596/1237/2/022106.

Zygis, M., Pape, D., Koenig, L. L., Jaskula, M., & Jesus, L. M. T. (2017). Segmental cues to intonation of statements and polar questions in whispered, semi-whispered and normal speech modes. *Journal of Phonetics, 63*, 53–74. https://doi.org/10.1016/j.wocn.2017.04.001.