Whispered and voiced speech: Database development, speech production and perception analysis

Luis M. T. Jesus [1,2]

[1] School of Health Sciences of Aveiro (ESSUA), University of Aveiro, Portugal; [2]Institute of Electronics and Informatics Engineering of Aveiro (IEETA), University of Aveiro, Portugal

lmtj@ua.pt

Cofinanciado por:

Whisper in normal voices

"Whispering is wide-spread across human cultures and - like laughing, crying and screaming - regarded as an universal form of vocalization" (Cirillo and Todt 2005: 114).

"Whispering, however is not a regular component of human communication, and is applied rather seldomly" (Cirillo and Todt 2005: 114).

Whispered speech can be used for quiet and private communication over mobile devices (Ito, Takeda and Itakura 2005: 139), and to mediate tenderness and support social bonding (Cirillo and Todt 2005: 115).

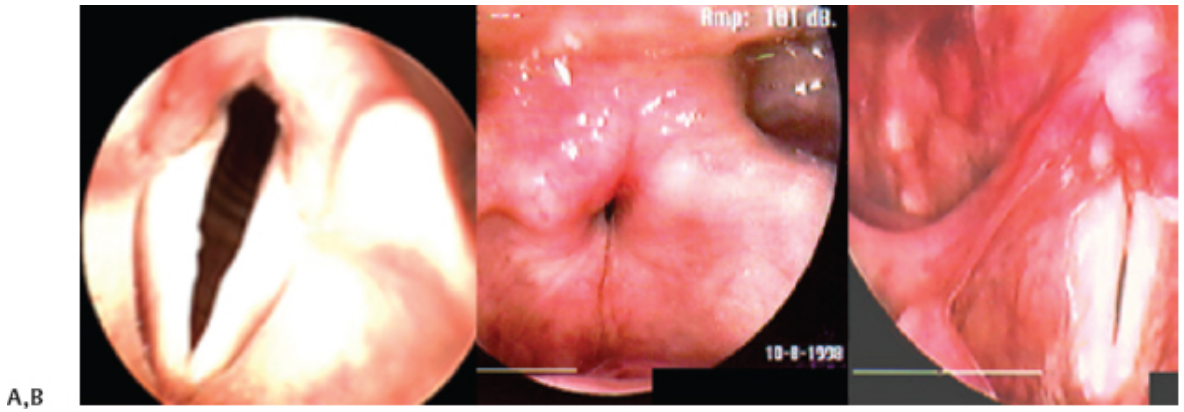
Whisper in disordered voices

Vocal fold paralysis

Weak and breathy whisper (Macdonell and Holmes 2007: 159)

Functional voice disorders: Psychogenic voice disorders

Aphonia with normal or tight (sharp and strained) whisper (Baker 2016: 393; Macdonell and Holmes 2007: 159)



(A) The true folds are abducted (B); The false folds are hyperadducted, obscuring the underlying true folds; (C) The true vocal folds are adducted but not vibrating. Adapted from Leonard (2016).

Whispered and voiced speech database

30 normal participants (15 male 15 female; 17 participants aged 18 to 45 years and 13 participants aged 46 to 65 years) were asked to produce

- 4 sustained sibilants × 3 (repetitions)
- 9 sustained European Portuguese oral vowels × 3 (repetitions)
- 28 dissyllabic words × 3 (repetitions)
- 6 Portuguese CAPE-V phrases × 3 (repetitions)
- 1 phonetically balanced text (“North Wind and the Sun”)

92 files per participant (48 voiced speech and 48 whispered speech)

Whispered and voiced speech database

4 sustained sibilants

/s, z, ʃ, ʒ/

9 sustained European Portuguese oral vowels

/i, e, ɛ, a, ɐ, ɔ, o, u, ɨ/

28 dissyllabic words

With fricative [s]: <sala>, <assa> and <face>

With fricative [z]: <zaro>, <asa> and <vaze>

With fricative [ʃ] <chama>, <acha> and <ache>

With fricative [ʒ] <jarra>, <haja> and <laje>

…

Whispered and voiced speech database

Portuguese CAPE-V phrases (Jesus, Belo, Machado and Hall 2017)

<A Marta e o avô vivem naquele casarão rosa velho> [ɐ ˈmaɾtɐ i u ɐˈvo ˈvivɐ̃j nɐˈkeliɬ kɐzɐˈɾɐ̃w ˈʁɔzɐ ˈvɛʌu] - **Production of every Portuguese oral vowel**.

<Sofia saiu cedo da sala> [suˈfiɐ sɐˈiw ˈsedu dɐ ˈsalɐ] - **Easy onset with /s/**.

<A asa do avião andava avariada> [ɐ ˈazɐ du ɐviˈɐ̃w ɐ̃ˈdavɐ ɐvɐriˈadɐ] - **All voiced**.

<Agora é hora de acabar> [ɐˈgɔɾɐ ɛ ˈɔɾɐ dɨ ɐkɐˈbaɾ] - **Elicits hard glottal attack**.

<Minha mãe mandou-me embora> [ˈmiɲɐ mɐ̃j mɐ̃ˈdomɨ ɐ̃ˈbɔɾɐ] - **Nasal sounds**.

<O Tiago comeu quatro peras> [u tiˈagu kuˈmew kuˈatru ˈpeɾɐʃ] - **Weighted with voiceless stops**.

Whispered and voiced speech database

Phonetically balanced text (Jesus, Valente and Hall 2015)

<O vento norte e o sol discutiam qual dos dois era o mais forte, quando sucedeu passar um viajante envolto numa capa. Ao vê-lo, põem-se de acordo em como aquele que primeiro conseguisse obrigar o viajante a tirar a capa seria considerado o mais forte. O vento norte começou a soprar com muita fúria, mas quanto mais soprava, mais o viajante se aconchegava à sua capa, até que o vento norte desistiu. O sol brilhou então com todo o esplendor, e imediatamente o viajante tirou a capa. O vento norte teve assim de reconhecer a superioridade do sol> - **98 words and 196 syllables**.

Whispered and voiced speech database

What was annotated?

All the phones

- 13 sustained fricatives and oral vowels
- 28 dissyllabic words

Phones [s, z, ʃ, ʒ, i, a, ɔ, u]

- Phrases
  - All the phones of some phrases
- Text

Speech production

Acoustic analysis of speech

- Source-filter models of vowels
- Parametric models fricative noise source spectra

Feature characterisation

The purpose of our parameter selection was to analyse speech production of voiced and whispered speech in terms of the source and filter characteristics.

Since the ultimate objective of the evidence gathered through this analysis process is to synthesise speech, we have collected data that will allow us to

control **models of vowels** based on the Source-Filter Theory of Speech Production (Fant, 1970)

control **models of fricative noise source** stylised dipole spectra (Narayanan & Alwan, 2000).

Feature characterisation

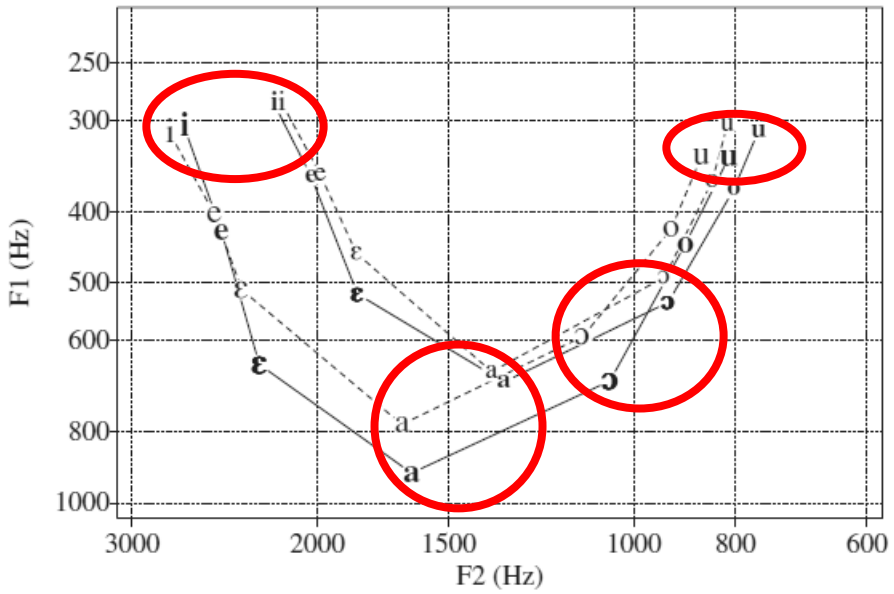
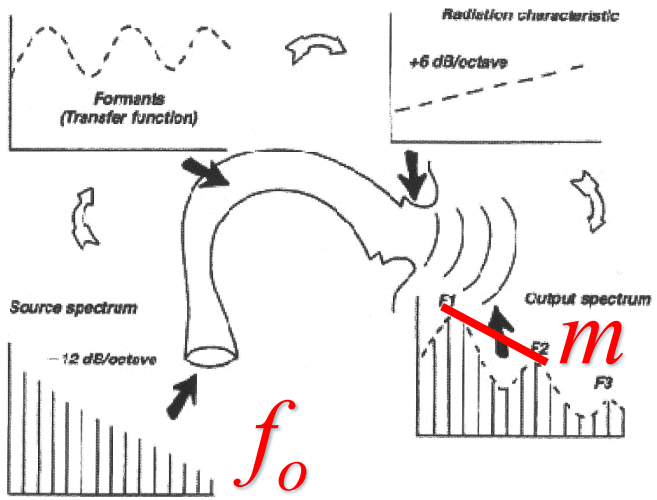


FIG. 6. The vowel spaces of the four groups. Solid lines and bold symbols=BP; dashed lines=EP. Large font: women; small font: men.

From Escudero, Boersma, Rauber and Bion (2009, p. 1385)

Absolute and relative durations



SPL

F_1

F_2

F_3

m

f_o

From Ludlow, Kent, & Gray (2019, p. 163)

Feature characterisation

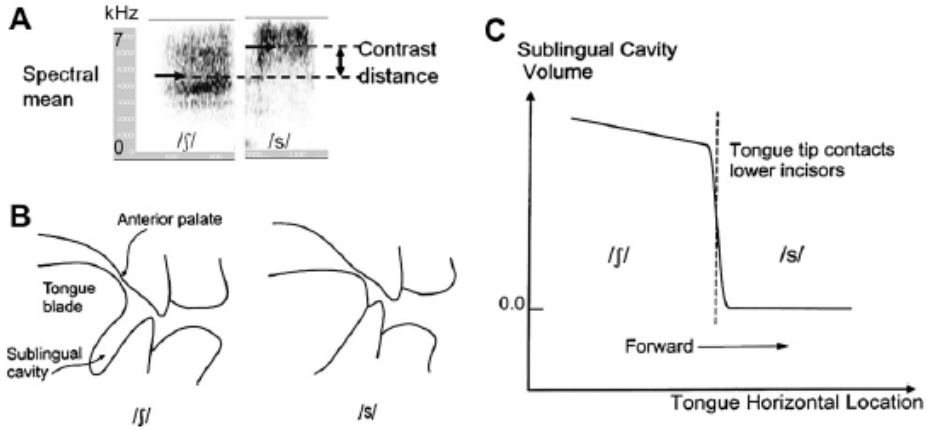


Fig. 2. A: Spectrograms of productions of the sounds /ʃ/ and /s/, illustrating the approximate frequencies of the spectral mean and the use of the spectral mean as a measure of the contrast distance between the two sounds. B: Schematic midsagittal drawings of articulations of /ʃ/ and /s/. C: Schematic illustration of a biomechanical saturation effect.

From Perkell (2012, p. 387)

SPL

Absolute and relative durations

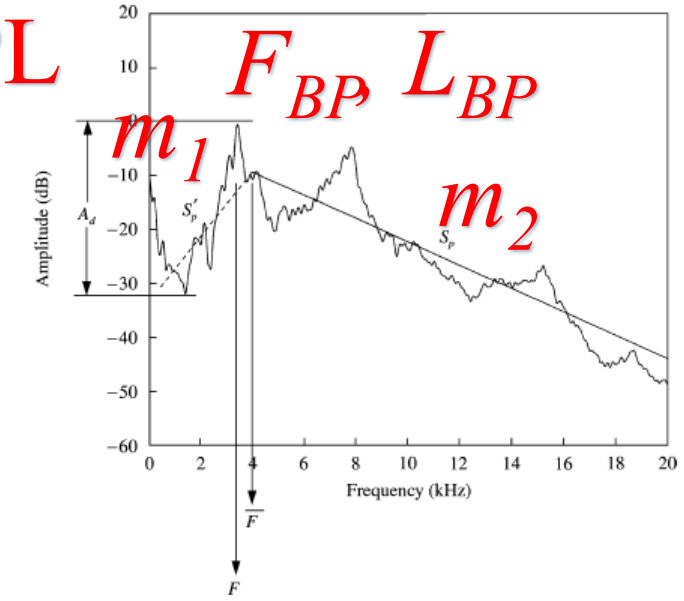
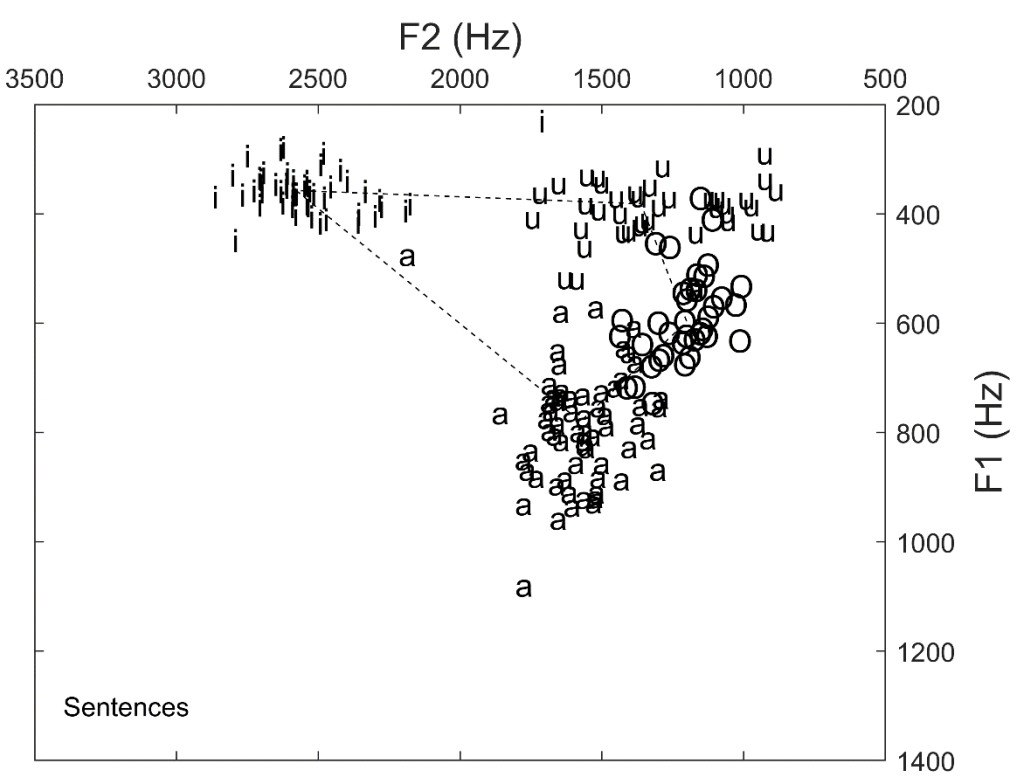
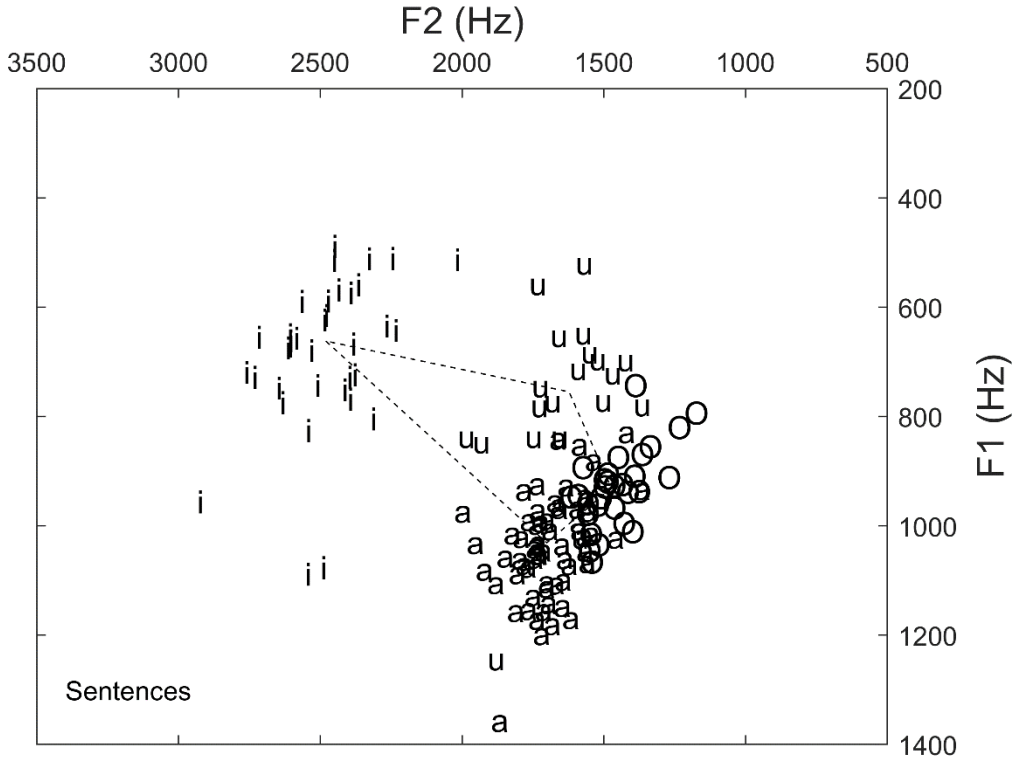


Figure 3. Dynamic amplitude A_d, and regression lines used to calculate low-frequency (500 Hz to $\bar{F}$ kHz) slope S'_p (dashed line) and high-frequency ($\bar{F}$ kHz to 20 kHz) slope S_p (solid line). Sustained fricative /ʃ/ (Corpus 1a) produced by Speaker ISSS.

From Jesus & Shadle (2002, p. 447)

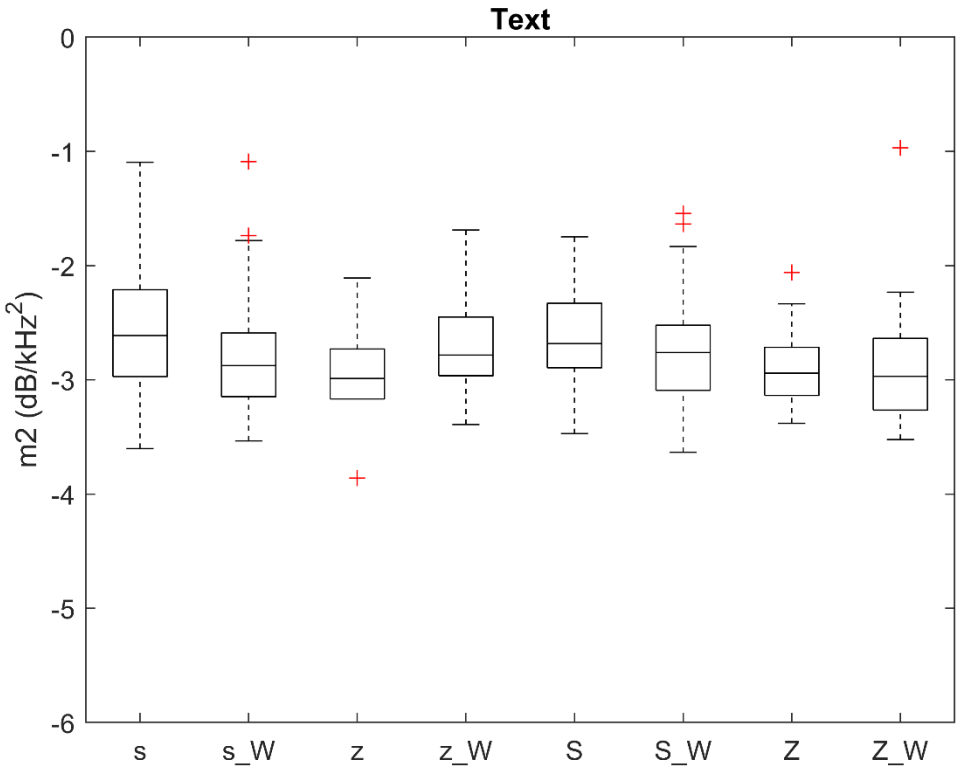
Feature characterisation

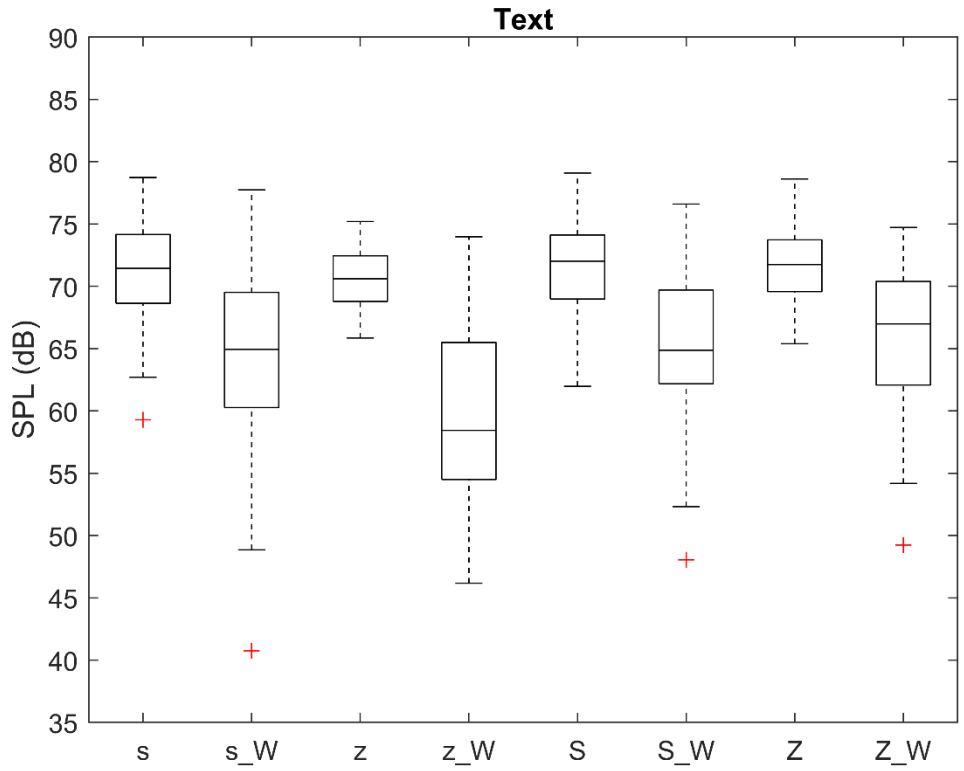




Young adulthood (18-45 years of age) # 8 Female speakers: SPF01; SPF02; SPF03; SPF04; SPF05; SPF06; SPF07; SPF10

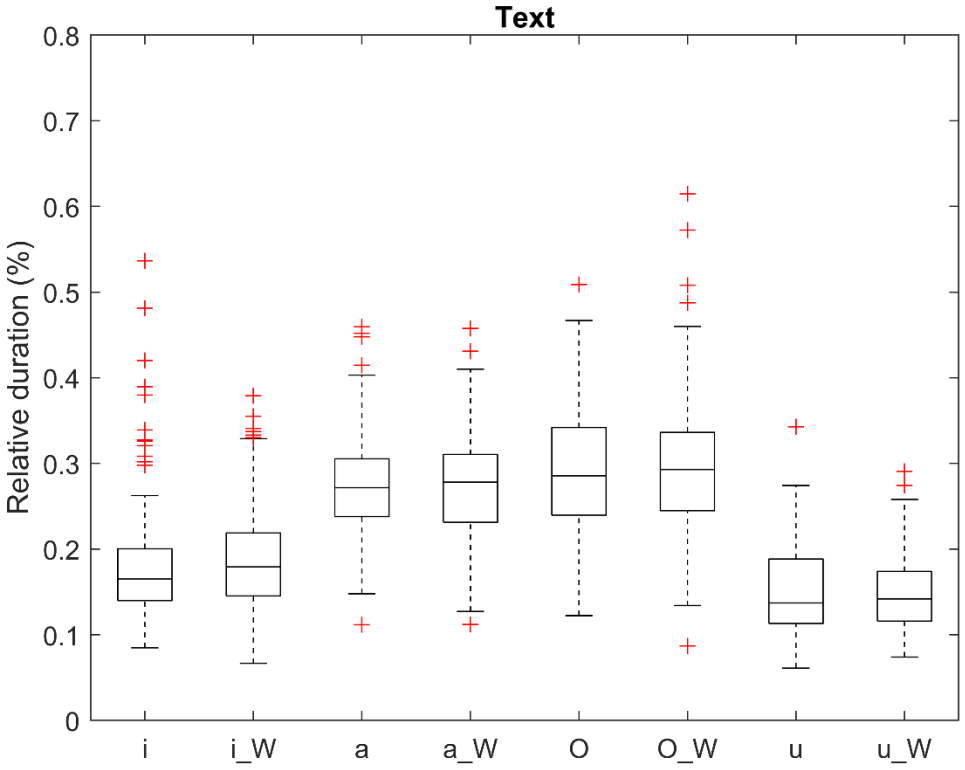
Feature characterisation

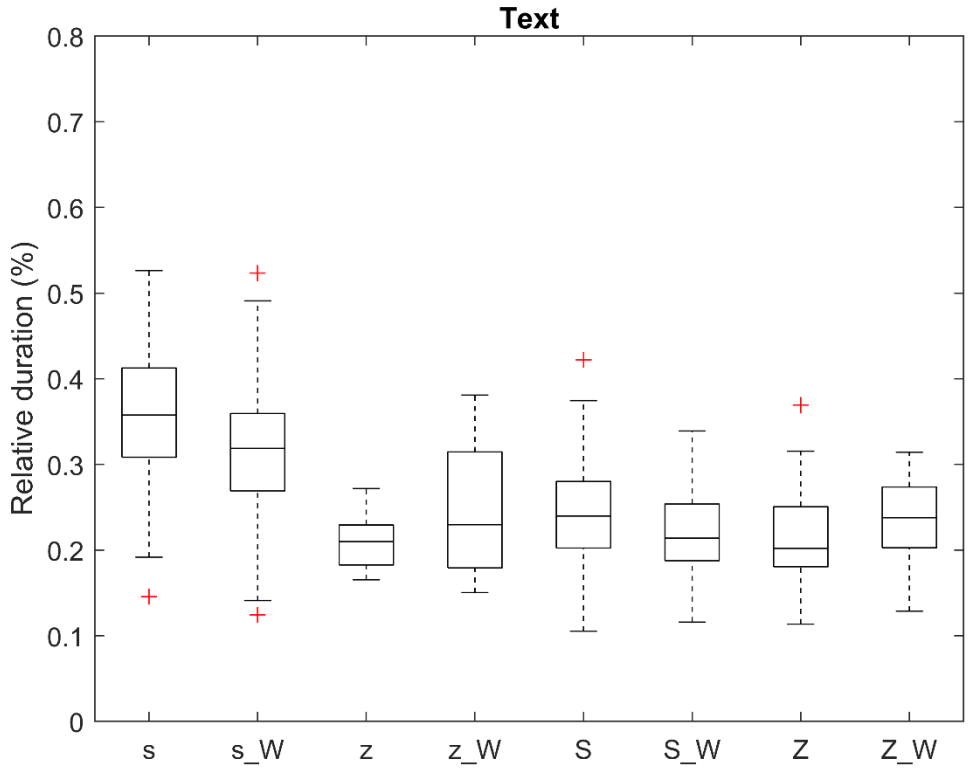




Young adulthood (18-45 years of age) # 8 Female speakers: SPF01; SPF02; SPF03; SPF04; SPF05; SPF06; SPF07; SPF10

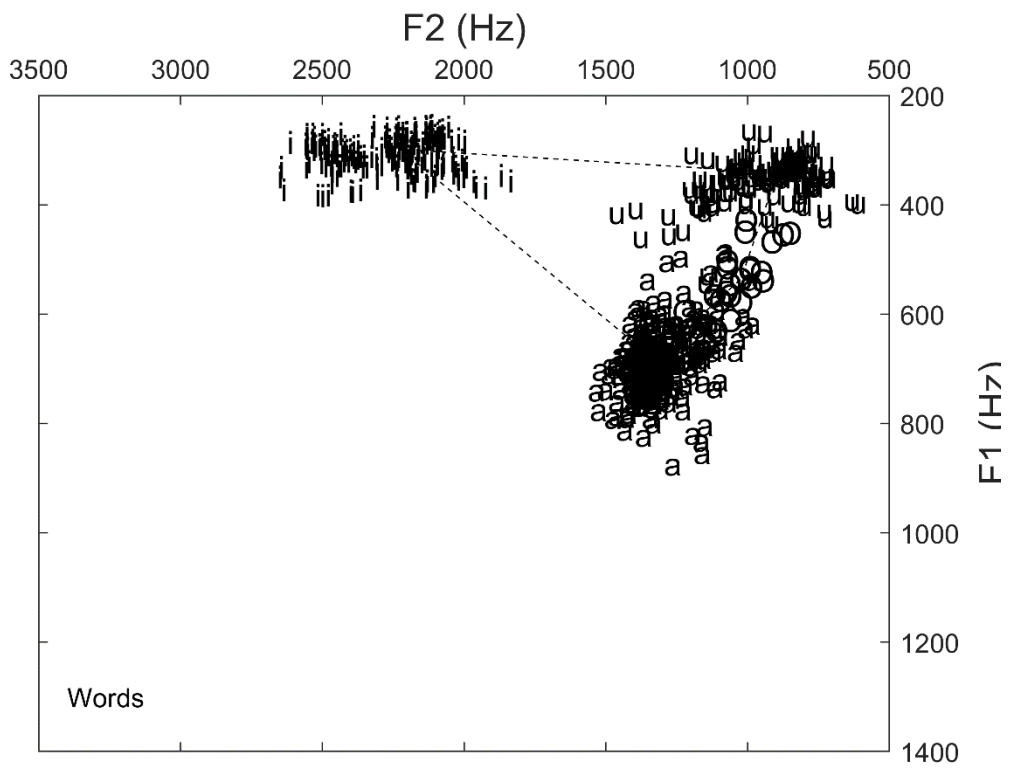
Feature characterisation

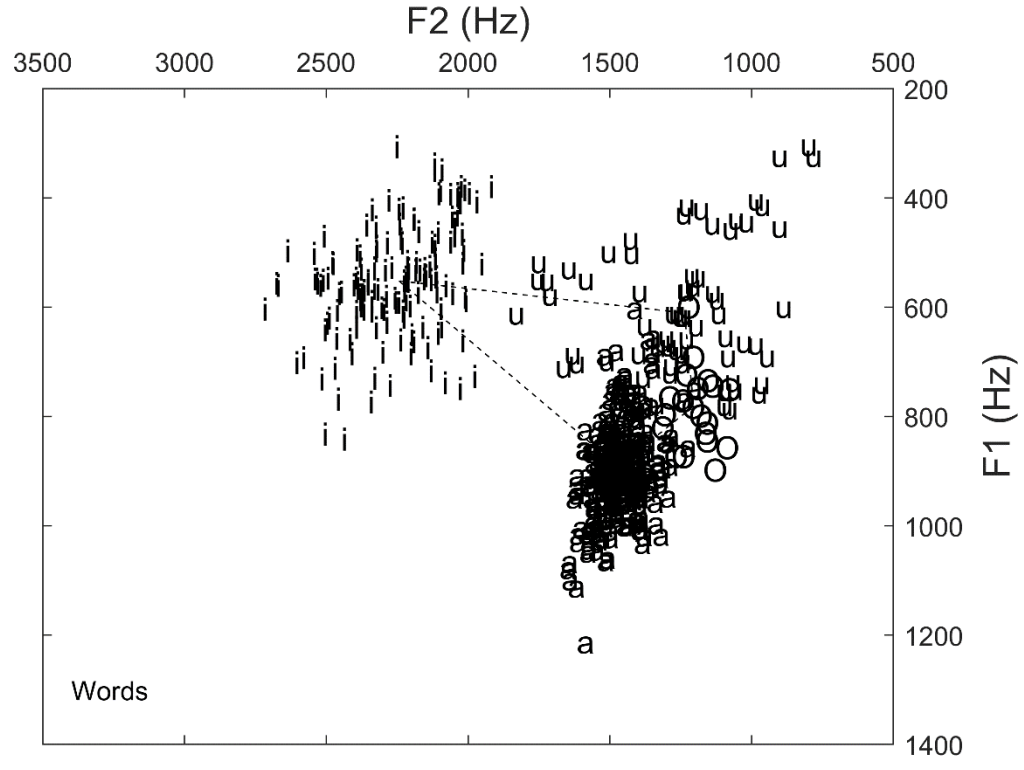




Young adulthood (18-45 years of age) # 8 Female speakers: SPF01; SPF02; SPF03; SPF04; SPF05; SPF06; SPF07; SPF10

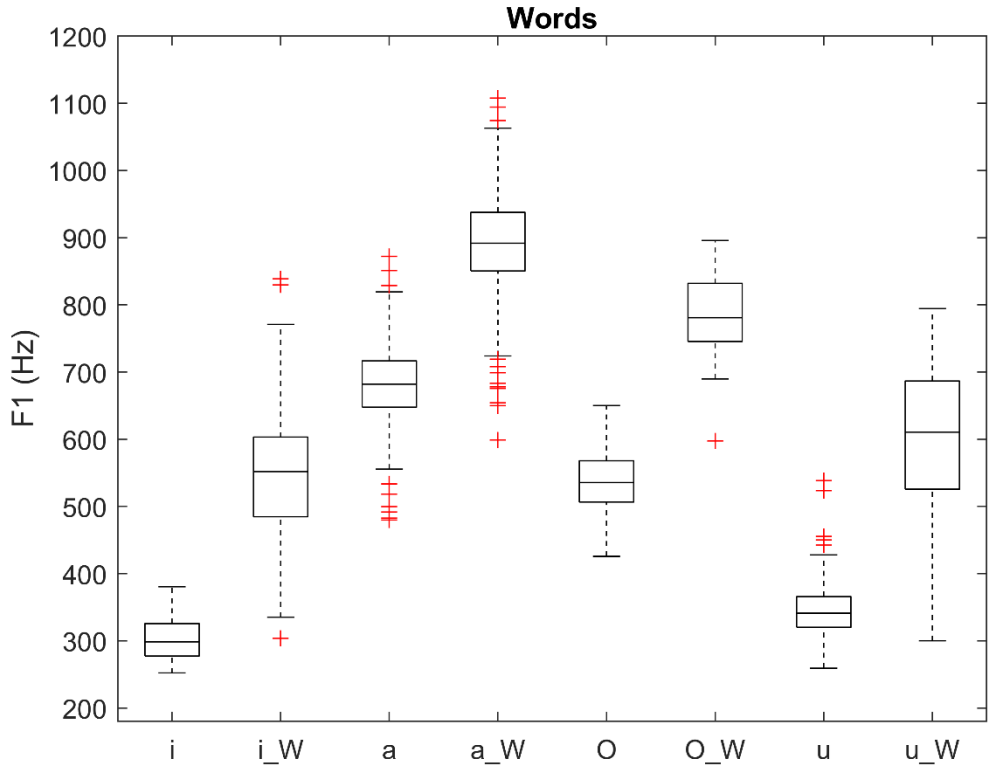
Feature characterisation ♂

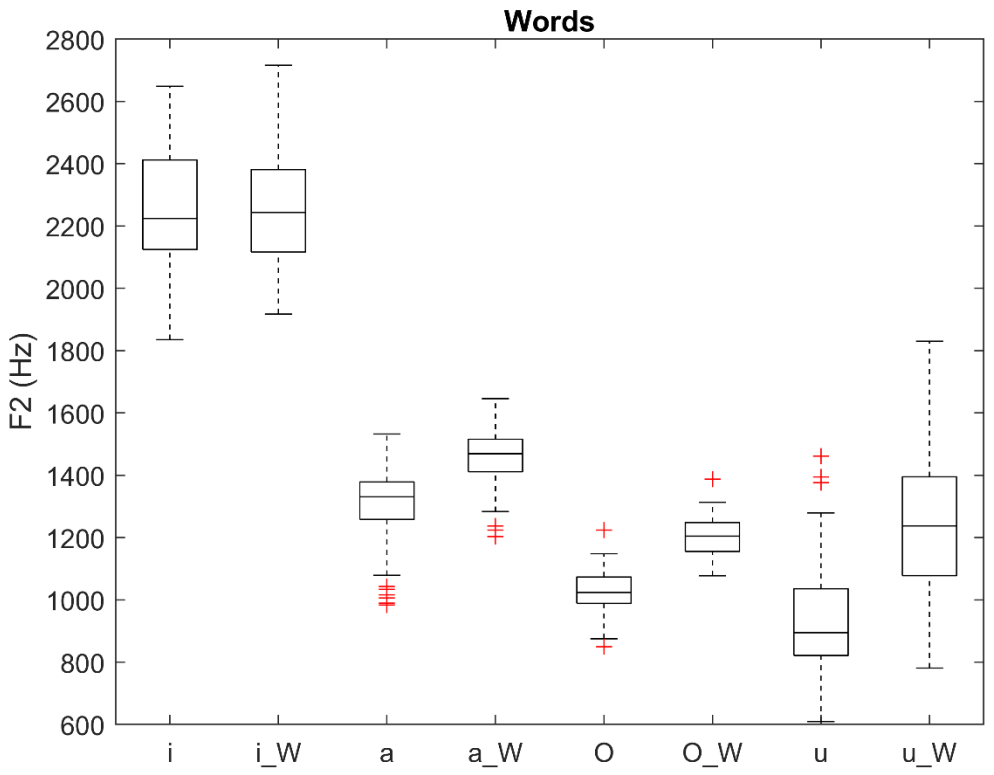




Young adulthood (18-45 years of age) ♂ # 9 Male speakers: SPM01; SPM02; SPF04; SPF05; SPF06; SPF07; SPF08; SPF09; SPF18

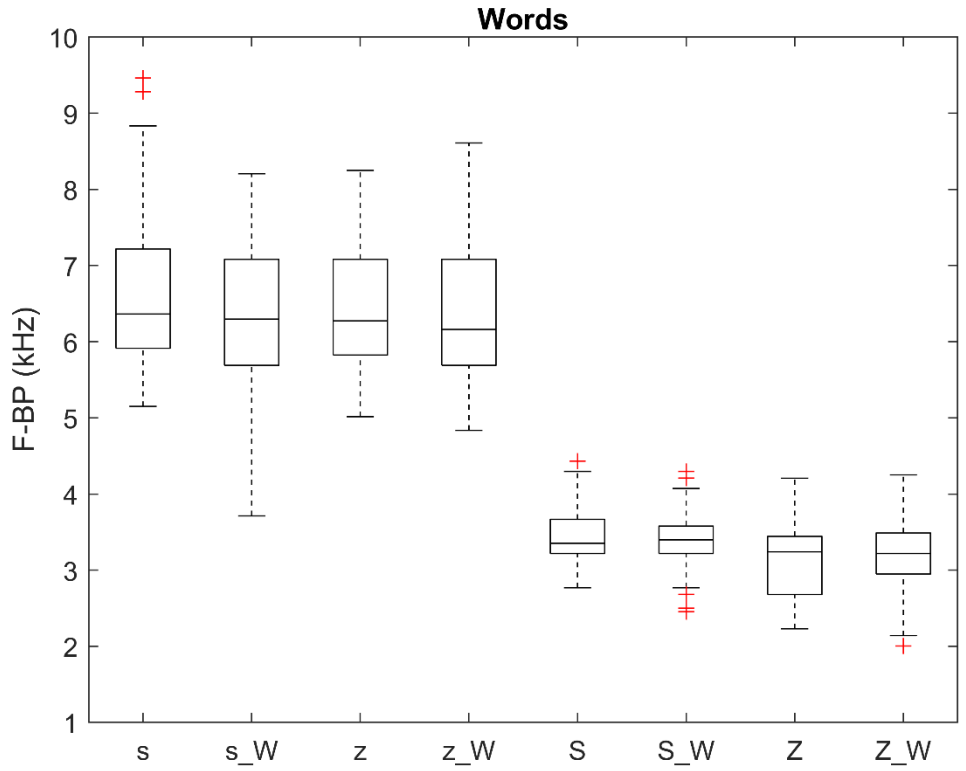
Feature characterisation ♂

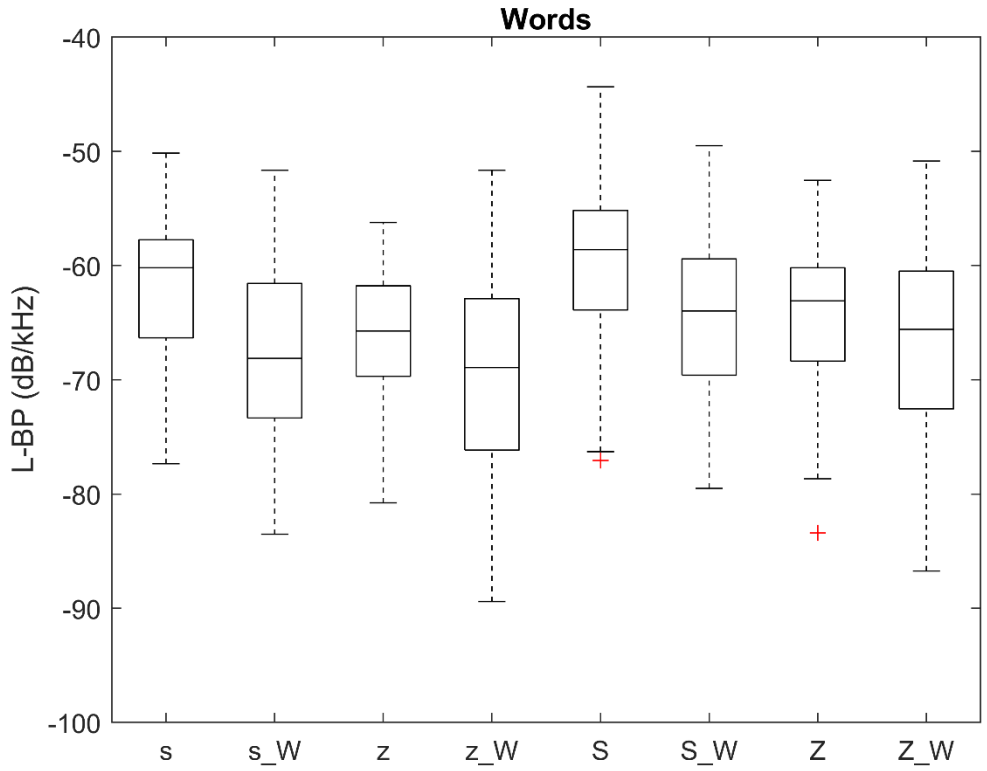




Young adulthood (18-45 years of age) ♂ # 9 Male speakers: SPM01; SPM02; SPF04; SPF05; SPF06; SPF07; SPF08; SPF09; SPF18

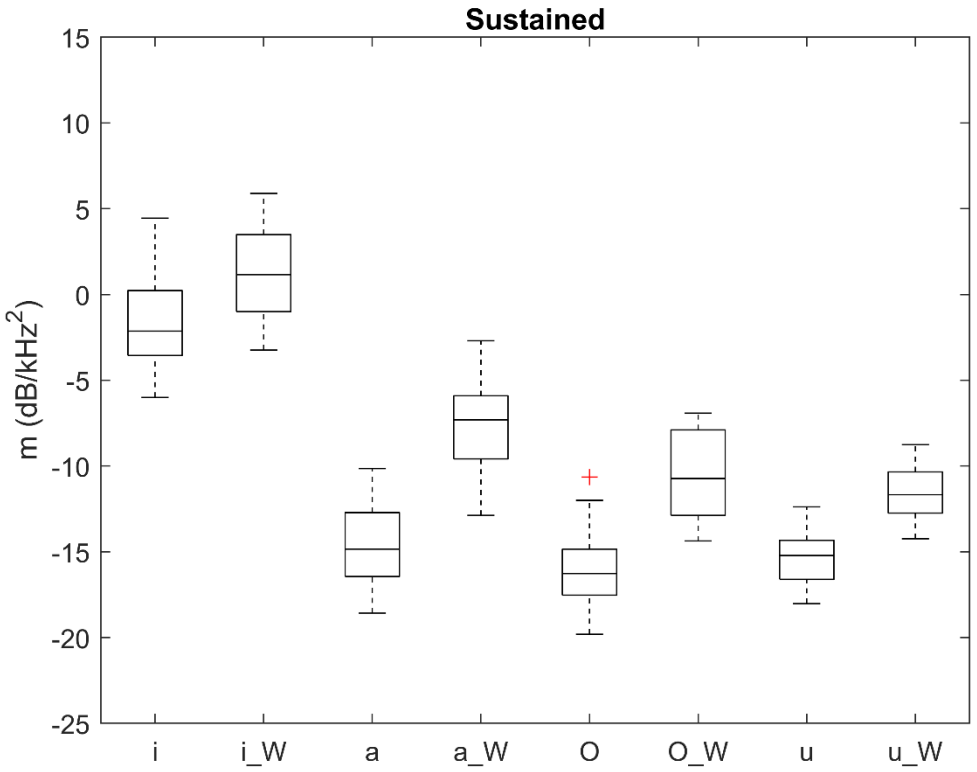
Feature characterisation ♂

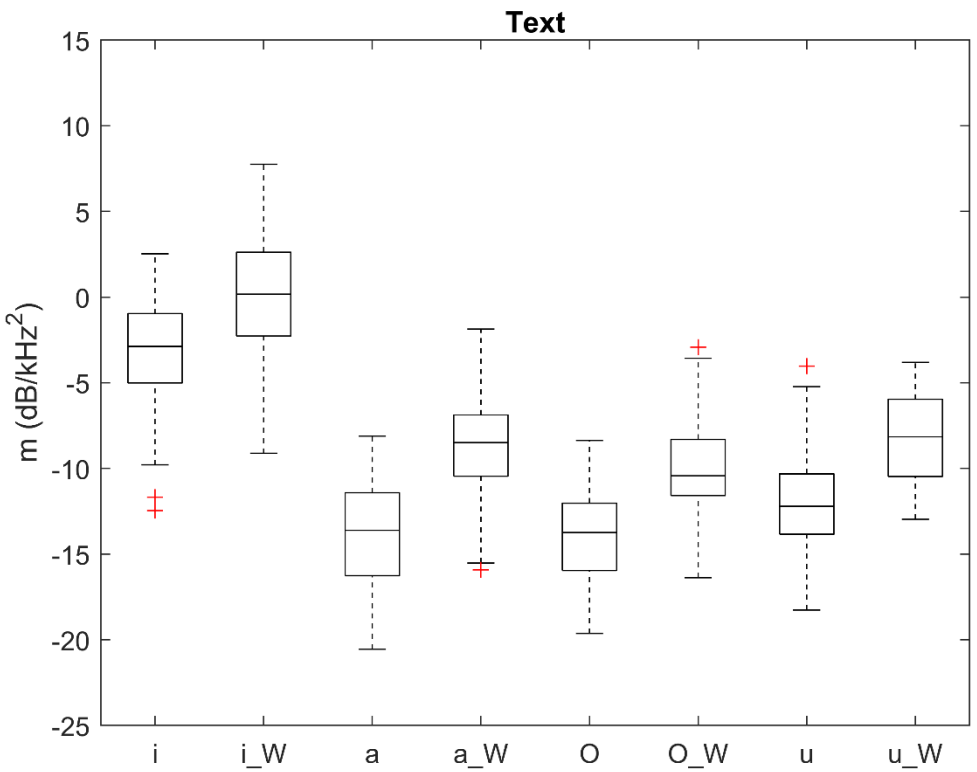




Young adulthood (18-45 years of age) ♂ # 9 Male speakers: SPM01; SPM02; SPF04; SPF05; SPF06; SPF07; SPF08; SPF09; SPF18

Feature characterisation ♂





Young adulthood (18-45 years of age) ♂ # 9 Male speakers: SPM01; SPM02; SPF04; SPF05; SPF06; SPF07; SPF08; SPF09; SPF18

Multiple comparisons of acoustic features

We aim to compare the characteristics of whispered speech in sustained segments and read text.

Acoustic differences between whispered sustained segments **/s, z, ʃ, ʒ, i, a, ɔ, u/**, the same segments in isolated words, sentences and a phonetically balanced text, were analysed.

The vowels **/i, a, ɔ, u/** were chosen to cover the European Portuguese vowel space previously described by Escudero, Boersma, Rauber, & Bion (2009).

Multiple comparisons of acoustic features

The sibilants **/s, z, ʃ, ʒ, i, a, ɔ, u/** were chosen in search of changes/ adaptations in place and manner of articulation in different tasks.

To understand if those differences are secondary or compensatory cues, the same tasks were analysed in normal speech.

There are few studies about whispered fricatives (Heeren, 2015; Meynadier & Gaydina, 2013), so the acoustic analysis of whispered sibilants is a central aim of this study.

The production and perception of voicing

The theoretical framework of this study is grounded on views of the **laryngeal feature of contrast** for **obstruents** that have been considerably enriched over the last decade by new acoustic, aerodynamics and articulatory evidence which strengthened arguments that in some languages, obstruent voicing is **phonologically active** and in others, it is **passive** (Beckman, Jessen, & Ringen, 2013).

Our long term plan is to contribute complementary evidence to a recent paper (Jesus e Costa, In Press) that supports the view that the feature of contrast in Portuguese is privative **[spread glottis]** and suggests that the laryngeal feature of contrast in EP is not **[voice]**, as recently shown for German and English (Beckman et al., 2013).

Jesus, L., and M. Costa (In Press). The Aerodynamics of Voiced Stop Closures. *EURASIP Journal on Audio Speech and Music Processing*. doi: 10.1186/s13636-019-0162-z

The [spread glottis] feature of contrast: Evidence from whispered speech

Tsunoda, Niimi and Hirose (1994) have shown that the posterior cricoarytenoid (PCA) muscle has a central role in **spreading the glottis** for voiceless speech segments.

- During whispered speech the PCA still contributes to the distinction between voiceless and voiced consonants.

Thyropharyngeus (TP) muscle activity contributed to the supraglottic constriction adjustment.

- There was also a relationship between glottal (PCA activity) and supra-laryngeal (TP activity) adjustments, suggesting that turbulence was generated between the glottis and a constriction above.

Speech perception

Perception of whispered and voiced speech

- Speech perception as a phonetic mode of listening (Johnson 2012: 101): Sounds of speech rather than words.
- Identity perception

Natural versus synthesised stimuli

"… the behavior of the organism in an artificial environment and in response to artificial stimuli may not tell us how the particular mechanism being investigated evolved to behave in response to natural stimuli occurring in the real world".

"It is important, then, that there be a thorough understanding of the perception of natural speech stimuli" (Dannenbring 1980: 983).

Perception of whispered and voiced speech

Stimuli [asɐ], [azɐ], [aʃɐ] and [aʒɐ]:

[ˈasɐ] from file 30 [s] <assa> SPM01_30_02.wav

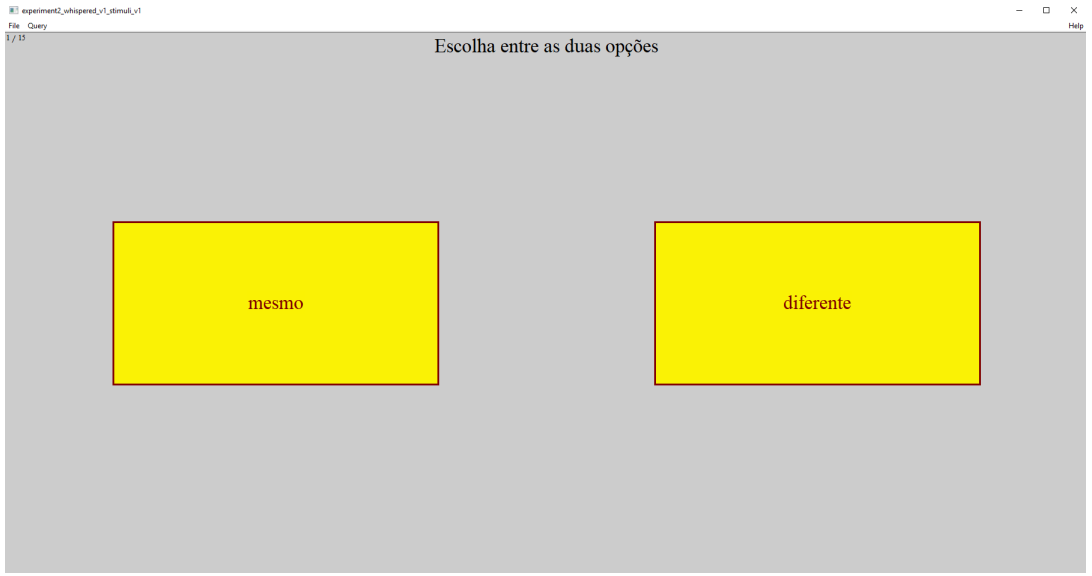
[ˈazɐ] from file 34 [z] <asa> SPM01_34_02.wav

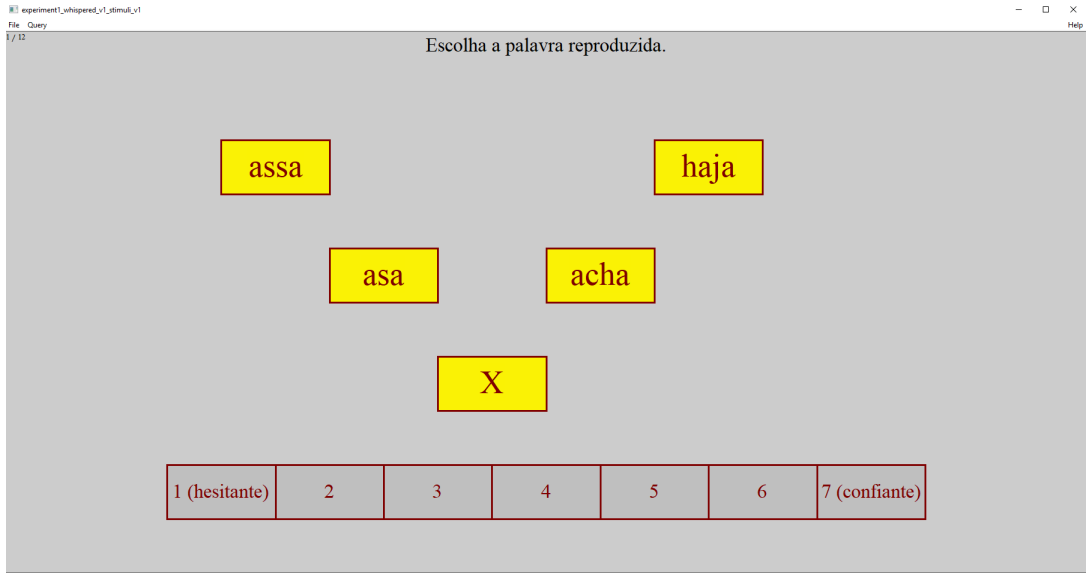
[ˈaʃɐ] from file 28 [S] <acha> SPM01_28_02.wav

[ˈaʒɐ] from file 36 [Z] <haja> SPM01_36_02.wav

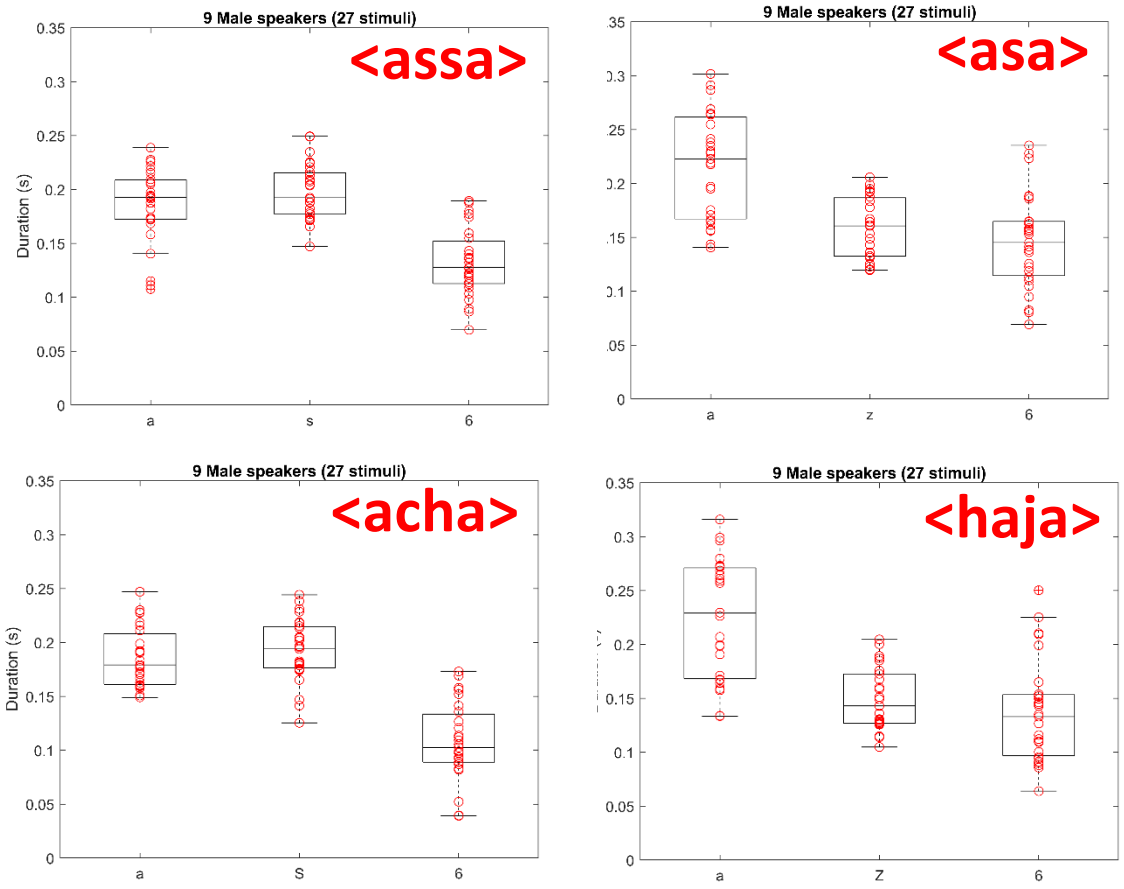
4 different VCV sequences ([asɐ]; [azɐ]; [aʃɐ]; [aʒɐ]) × 17 young adults × 3 repetitions = **204 stimuli** in each speech mode (voiced and whispered)

Discrimination and Identification





Correlate responses of listeners with acoustic properties of whispered stimuli?



Discrimination of whispered speech (Dannenbring 1980)

Dannenbring (1980) studied the perceptual discrimination of naturally produced voiced/voiceless consonant pairs (including the [s]/[z] pair) in whispered speech.

Listeners were asked which consonant pair they heard and indicated their confidence on a 7-point scale.

Responses were combined on a 14 point scale that represented responses from voiceless (1) to voiced (14).

Discrimination of whispered speech (Allen and Haggard 1977; Dannenbring 1980)

These were then converted to a D score from -1 to 1 that indicates from systematically incorrect judgments (-1), through random judgments (0) to perfect discrimination (1)

The [s]/[z] presented the lowest D scores (Dannenbring 1980) .

Compute Allen and Haggard's (1977) **discriminability** ($\log_e \alpha$ / d') and **bias** ($\log_e v$ / $\log\beta$) estimates for **voicing** and **place of articulation**.

The A' score used by Pape and Jesus (2014: 1339) should also be considered in the analysis of the results.

Also a 100 ms interval between the [s]/[z] and [ʃ]/[ʒ] pairs of stimuli to guarantee low memory load (Pape and Jesus 2014: 1337) .

Identification of phonemes in whispered and voiced speech (Tartter 1989, 1991)

Experimental design used by Tartter (1989, 1991)

Perception of voicing and place of articulation

Tartter (1989) used American English CV syllables where V was [a] and C one of 18 American English vowels

We currently evaluating the possibility of using Portuguese VCV words or VC syllables spliced from them.

Identifiability of vowels

Tartter (1991) used American English [hVd] syllables where V was one of 10 American vowels

Could we use one of CAPE-V's sentences, that includes all European Portuguese vowels?

<a Marta e o avô vivem naquele casarão rosa velho>

[ɐ ˈmaɾtɐ i u ɐˈvo ˈvivɐ̃j nɐˈkeɫɨ kɐzɐˈɾɐ̃w ˈʁɔzɐ ˈvɛʌu]

Identification of phonemes in whispered and voiced speech (Tartter 1991)

- Interesting features of Tartter's (1991) experimental design
  - Experiment 1 by Tartter (1991)
    - "Half of the subjects from each experiment … received the whispered vowel test, followed by the normally phonated vowel test. For the other half, test order was reversed" (Tartter 1991: 367)

Identification of phonemes in whispered and voiced speech (Tartter 1989, 1991)

Produce confusion matrices and pool information out of them that will allow us to interpret the perceptual processes that caused the confusions (Johnson 2012: 115-123) based on **maps of distances** using multidimensional scaling (MDS).

	/s/	/z/	/ʃ/	/ʒ/
[ˈasɐ]				
[ˈazɐ]				
[ˈaʃɐ]				
[ˈaʒɐ]				

Compute **similarity** and derive **distance** from it.

Identity perception from speech signals (Lavan et al. 2019; Tartter 1991)

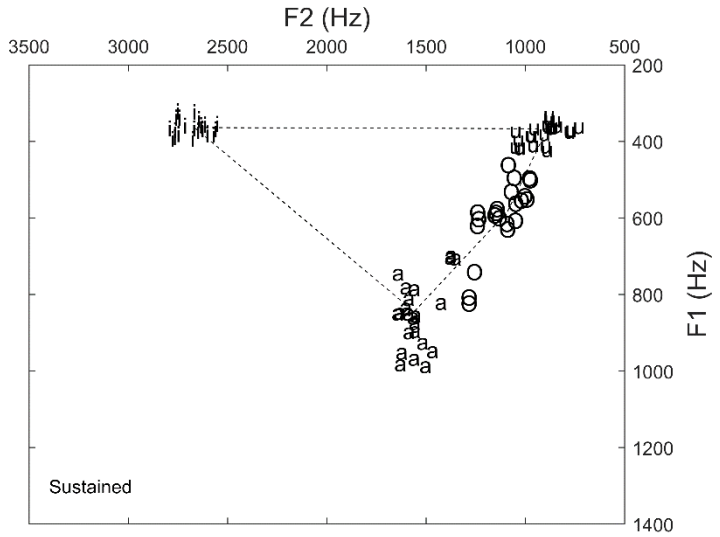
Discrimination task (Tartter 1991)

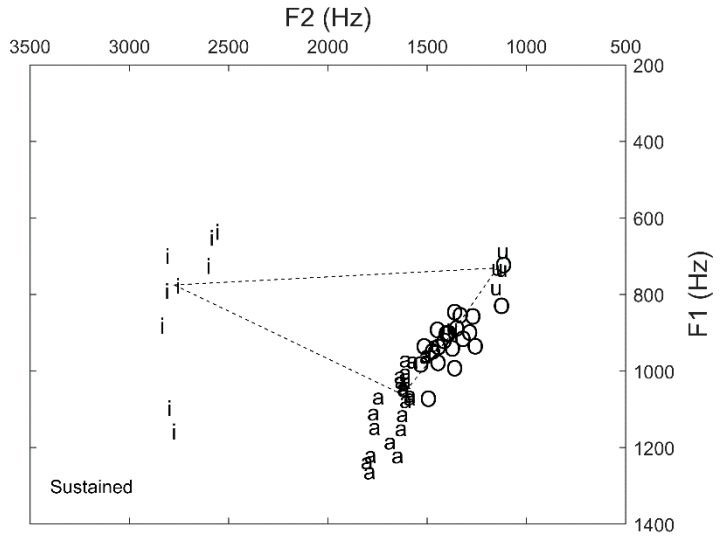
We will be trying to address a source of **within-person variability** (whispering) and attempting to determine if "there is sufficient information in whispered speech that allows processing a speaker's identity" (Lavan, Burton, Scott and McGettigan 2019: 96)

Identical experimental design to Tartter (1991) using the <a Marta e o avô vivem naquele casarão rosa velho> CAPE-V sentence

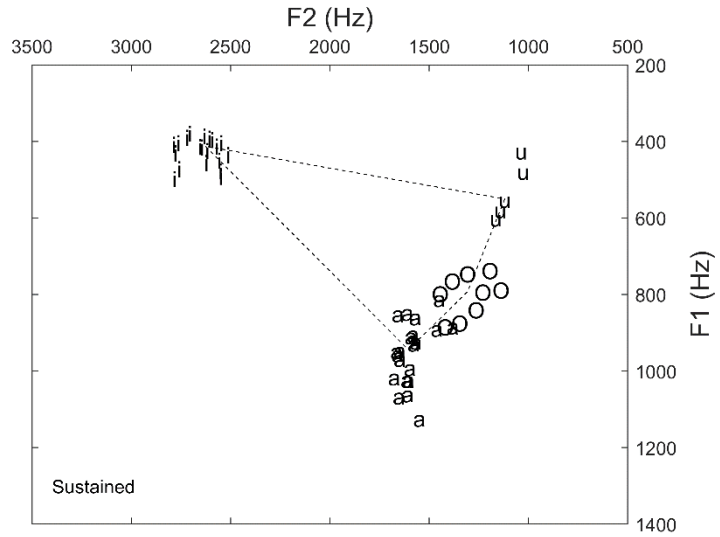
- Experiment 2 by Tartter (1991)
  - "a normally phonated syllable was presented and, after a 1-sec pause, the same syllable whispered by 3 speakers of the same sex as the speaker of the normally phonated one were presented in random order" (Tartter 1991: 369)

Voiced, Whispered and Residuals



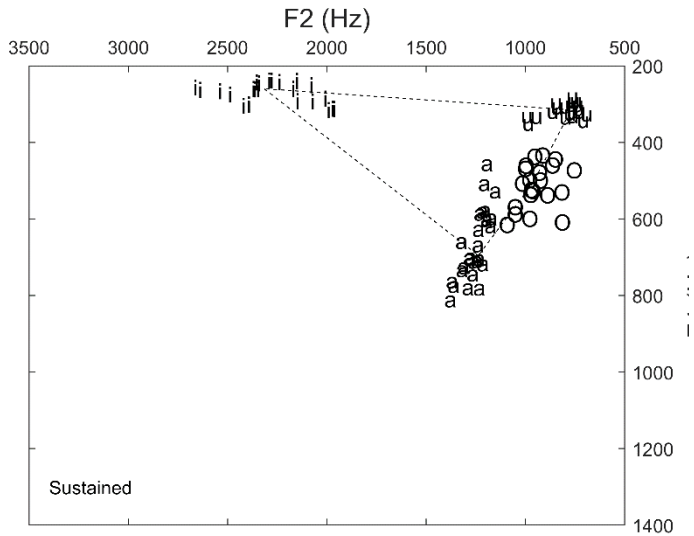


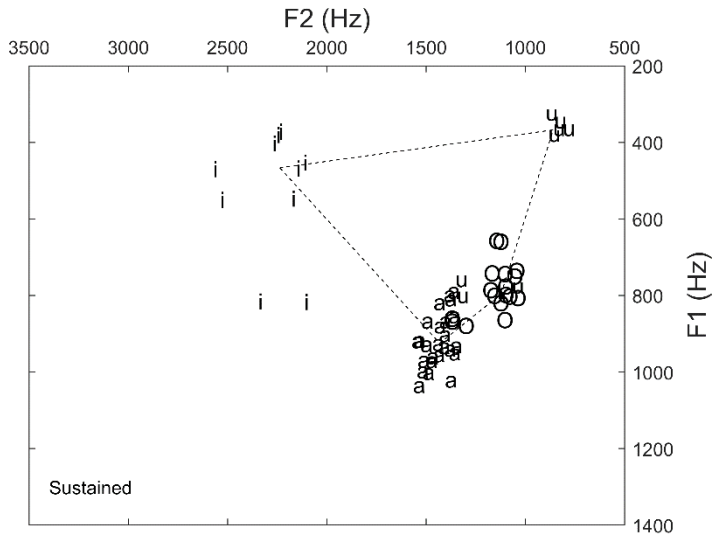
Young adulthood (18-45 years of age)
8 Female speakers: SPF01; SPF02; SPF03; SPF04; SPF05; SPF06; SPF07 ; SPF10



Young adulthood (18-45 years of age)
7 Female speakers: SPF01; SPF02; SPF03; SPF04; SPF05; SPF06; SPF07

Voiced, Whispered and Residuals ♂

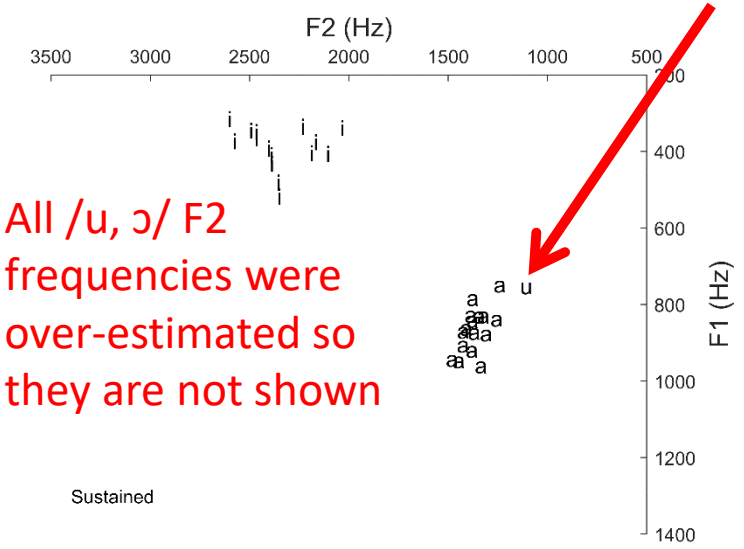




Young adulthood (18-45 years of age) ♂
9 Male speakers: SPM01; SPM02 ;
SPF04; SPF05; SPF06; SPF07; SPF08;
SPF09; SPF18

This is the only exception

All /u, ɔ/ F2 frequencies were over-estimated so they are not shown



Young adulthood (18-45 years of age) ♂
8 Male speakers: SPM01; SPM02; SPF05;
SPF06; SPF07; SPF08; SPF09; SPF18

SLH LAB
SPEECH, LANGUAGE AND HEARING LABORATORY

Thank you!

Follow us on:

https:// www.facebook.com/slhlab/

https://twitter.com/SLH_Lab