Estudios de Fonética Experimental
Journal of Experimental Phonetics
ISSN: 1575-5533 · ISSN-e: 2385-3573

Voice disguise and foreign accent: Prosodic aspects of English produced by Brazilian Portuguese speakers

Leônidas Silva Jr. [1] 0000-0002-3728-9851
Plínio A. Barbosa [2] 0000-0001-6317-3548

[1] State University of Paraíba / Universidade Estadual da Paraíba (Brazil)
[2] University of Campinas / Universidade de Campinas (Brazil)

DOI: 10.1344/efe-2023-32-195-226

Corresponding address: leonidas.silvajr@gmail.com

Received: 08/02/2023 Accepted: 24/10/2023 Published: 28/11/2023

Silva Jr., L., & Barbosa, P. A. (2023). Voice disguise and foreign accent: prosodic aspects of English produced by Brazilian Portuguese speakers. *Estudios de Fonética Experimental*, *32*, 195–226. https://doi.org/10.1344/efe-2023-32-195-226

Abstract

This study aims to explore how much a speaker of English as a foreign language can mask one's accent by performing voice disguise toward a native-like accent, and what implications it might bring forward. For the Methods, we conducted three experiments, (*a*) both the native and the foreign groups produced authentic voices, (*b*) the native group performed authentic voice and the foreign group performed disguised voice, and (*c*) a group of native listeners of English rated the degree of foreign accent from experiments *a* and *b*. The results showed evidence of correlation between the ratings and authentic voices. Nevertheless, they remained somewhat inconclusive for the disguised voice. It may be concluded that the more the Brazilian speaker approaches the L1-English target accent when voice disguising, the harder it is for the listener to recognize his/her L2-English accent.

Keywords

voice disguise; foreign accent; prosodic-acoustic parameters

Disfressa de veu i accent estranger: Aspectes prosòdics de l'anglès produït per parlants de portuguès brasiler

RESUM

Aquest estudi té l'objectiu d'explorar en quina mesura un parlant d'anglès com a llengua estrangera pot emmascarar el seu accent perquè soni com un parlant nadiu i quines implicacions podria tenir. Pel que fa a la metodologia, es van realitzar tres experiments: (*a*) tant el grup nadiu com l'estranger van produir veus autèntiques, (*b*) el grup de parlants nadius va produir una veu autèntica i el grup estranger va produir una veu disfressada, i (*c*) un grup d'oients nadius d'anglès va qualificar el grau d'accent estranger dels experiments *a* i *b*. Els resultats van mostrar una correlació entre les qualificacions i les veus autèntiques. No obstant això, els resultats obtinguts per a la veu disfressada no són concloents. Es pot concloure que com més s'apropa a l'accent de l'anglès-L1 el parlant brasiler quan disfressa la veu, més difícil li resulta a l'oient reconèixer el seu accent d'anglès-L2.

MOTS CLAU

veu disfressada; accent estranger; paràmetres prosòdics-acústics

1. Introduction

This study aims to explore how much a speaker of English as a foreign language (L2) can mask one's accent toward the target L2. The target accent will be analyzed via prosodic-acoustic parameters/features, and gauged as which of these features are candidates to influence the lay native listener's perception. The present study made use of two different speaking styles: *authentic voice* (recorded from a story-reading task and extracted from an interview), and *disguised voice* (recorded from a story-reading imitation as well as, an interview impersonation toward the target accent). Hereon, the following research questions are proposed:

I. As a (Brazilian) foreign speaker of English converges toward the prosodic structure of the L2 when voice disguising, will it be more difficult to the lay listener to recognize this L2-English accent?
II. Which prosodic parameters, from fundamental frequency, intensity, duration and voice quality, best correlate with listeners' judgments of L2 speakers' degree of foreign accent?

The use of prosodic parameters based on fundamental frequency (F0), duration, intensity and voice quality have shown promising results (on foreign accent recognition) in studies such as, Munro and Derwing (2001), Keating and Esposito (2007), Engelbert (2014), Keating et al. (2015), Niebuhr et al. (2018) among others. For instance, Munro and Derwing (2001) found a U-curve pattern for speech rate, where both speeches — too fast and too slow — caused problems for listeners at identifying accent degree. In the studies of Gut (2012), Urbani (2012), Gonzales et al. (2013), and Silva Jr. and Barbosa (2019, 2023), speech rate played an important role on the identification of German, Italian, Japanese and Brazilian speakers of English, respectively.

At this point, it is important clarify which concepts were adopted for the definition of "authentic" and "disguised" voices applied in this research. "Authentic" (or modal) voice is the most commonly occurring form of voicing in the world's languages. It is also language-dependent. For Ladefoged and Maddieson, (1996/2008, p. 48–50) authentic voice is "the regular vibration of the vocal folds at any frequency within the speaker's normal range […] in which arytenoid cartilages are in neutral position for speech." Temporal, pitch and spectral parameters of the authentic voice are usually preserved. Authentic voice refers to the resonant mode of vocal folds; i.e., the optimal combination of airflow and glottal tension that yields maximum vibration in vowels, in sonorants consonants, and in syllables during speech (Ladefoged & Maddieson, 1996/2008; Rojczyk, 2010).

In authentic voice, pitch range usually presents an extension window in between 60 to 500 Hz – 60 to 220 Hz for men's range, while women and children are in the range of 150 to 500 Hz. These measures are considered very robust during speaker recognition task (He, 2017). Authentic voice also presents a spectral slope of –12 dB per octave as an amplitude difference between the F0 and the 2nd harmonic (Rose, 2002).

On the other hand, "disguised" voice refers to an intentional act of a speaker who wants to hide one's identity by several means (for imitating a target dialectical/foreign accent, for entertainment performing, for criminal purposes inter alia). In the disguised voice, temporal, pitch, amplitude and voice quality features are the most affected ones and the preferred of the imitators (Masthoff, 1996; Rose, 2002; Neuhauser & Simpson, 2007; Eriksson, 2010; Rojczyk, 2015; Kaur & Kaur, 2021 among others). Zheng et al. (2020) also claim for the use of those features in artificially-based voice disguising for testing automatic speaker recognition (ASR) systems.

From what was put forward in this section, this study will refer to "authentic voice" as the modal or neutral way that a speaker sets one's larynx for native language (L1) or L2 speech, and will refer to "disguised voice" as the shift in any one of the temporal, pitch, amplitude or voice quality parameters by means of converging to the L2 target accent.

2. Literature review

In the domain of foreign language, Perrot et al. (2007), Clark and Foulkes, (2007), Eriksson (2010), Neuhauser (2011) states that voice disguise by accent imitation may be used for at least two reasons: to conceal one's own voice, or to pretend that the language used is the speaker's native language. The author also addresses that phonation-level disguises (whisper or raised/lowered pitch) are commonly chosen. Farrús et al. (2008) reported results from durational parameters (voiced/unvoiced portions of word-internal elements) as being particularly difficult to disguise due to the intrinsic duration of a given segment or syllable.

According to Rojczyk (2015), accent imitation is not new in speech research and has been used for various purposes accounting for a positive effect on the perceived social attractiveness of the speaker. For his study, accent imitation of non-native sounds would be shaped both in perception and production by already established native sound categories (sound transfer) such as the ones modeled by the Perceptual Assimilation Model (Best, 1995; Best & Tyler, 2007), as well as a number of other perceptually-based statistical learning models during L2 acquisition.

Andrews (2019) advocates that, L2 acquisition is a lifelong, dynamic process with periods of more intensive acquisition and loss, where production and perception are not realized consecutively, but simultaneously. This sets the stage for a re-evaluation of how one studies L2 acquisition, processing, and proficiency achievements. Automatization is the end result of such achievements that happens through a process of repeated sessions of rehearsal and evaluation, which rely heavily on conscious and supervision. For Andrews (2019), segmental and prosodic aspects in multilingual speakers, if treated equally, will be acquired simultaneously.

Concerning the L2 prosody acquisition, phonetic literature has laid down on the use of acoustic features for promoting some nativelikeness degree in L2 speech production and perception, and this includes duration, F0 and intensive features as mentioned by Fletcher (2010), and Jackson and O'Brien (2011). Yet, Nooteboom (1997), Wrembel (2007), Levis (2018) and Das et al. (2020) spotlight the relevance of voice quality and its interaction with rhythmic and intonational aspects in the recognition of a foreign accent.

In L2 speech perception, Neuhauser and Simpson (2007) conducted a study which the main question was not whether listeners were able to identify the presence/absence of a foreign accent, but whether they could judge if the accent they were hearing was authentic (a native accent), or imitated (a foreign accent). Rojczyk (2015) addresses that some speech features may be less or highly exposed depending on the task specificity during foreign accent imitation. According to Rojczyk's (2015) study, when asked to imitate foreign accent in their L1, the learners will be pressed to reveal the features of a foreign-accented pronunciation that they have already acquired. Such imitation might be a cognitively demanding task.

According to Costa (2017), short and, especially, long-term-parameters of vocal effort reflect the speaker's cognitive and vocal load during the task of foreign accent imitation. For the author, the bilingual brain differentiates authentic (modal) from disguised voices in the proficient L2 speaker's accent.

In terms of voice quality (long-term spectral features), authentic and disguised voices from L2 speech usually present higher slopes than the native speech. It suggests that the cognitive load for performing L2 vocal activities seems to be intense. In addition, Costa's (2017) study yet describes that when speaking a L2, the cerebral and cognitive bases suffer a deficit in memory, attention, and emotion reflecting in the speaker's phonetic performance. These aspects foster a decline of the prosodic features, which generates more vocal load during the L2 production.

Moreover, Hernandez (2012) pinpoints that voice imitation tasks in a foreign language demand a great deal of vocal and high-level cognitive effort from specific cortical areas of the brain (very small and concentrated in language-designated areas). Both Hernandez (2012) and Costa (2017) studies suggest that individual variability is likely to represent how well a person uses the brain zones of foreign language, i.e., the smaller and highly concentrated the brain area, the better the performance of one's L2 speech, it is to say, that less ability for performing pronunciation in a foreign language is a result from a more diffuse brain representation (see Appendix E a for functional magnetic resonance image – fMRI, for different brain cortex areas in (non)skilled L2 speakers).

Thus, Hernandez (2012) concludes that in multilingual speakers, the poorer language is disrupted over a larger area than the better one. Hernandez (2012) yet highlights that, the foreign accent is much less likely to improve past a certain point even with continued increases in vocabulary and other higher-level forms of language in L2 speakers.

For Costa (2017), the cognitive processes related to emotion, attention, among others, work independently from one another, and are of difficult convergence from L1 to L2. The author points out that their interactions work in very complex way making high-level speaking tasks in L2 (voice disguising, imitations, impersonations) harder to be accessed.

3. Methods

This research consists of three different experiments. The first and second ones are speech production experiments, and the third one is a perception experiment. In sections 3.1 and 3.2, we detail the experiments 1 and 2 respectively, as well as participants, data collection (corpus, task and recording procedures), acoustic and statistical analyses and the (discussed) results for each experiment. In section 3.3, we detail the experiment 3 (the perception one) with its participants, listening task, data collection (corpus) and perceptual analysis, statistical analysis and the (discussed) results.

3.1. Experiment 1

For Experiment 1 both groups produced speech in authentic voice.

3.1.1. Participants

For this experiment, the participants were a group of L1-English speakers (four Americans who lived in Brazil for about two years when the experiment was run), and a group of L2-English speakers (twenty Brazilians).

The L1-English group consisted of participants who were 50% female/male with ages between 26 and 50 years (M = 38.1, SD = 13.2). All of the American participants were submitted to the Certificate of Proficiency in Portuguese Language for Foreigners (CELPE-Bras; *Certificado de Proficiência em Língua Portuguesa para Estrangeiros*).[1]

The group qualified as "high-intermediate level" (B1–B2), "low-advanced level" (B2–C1), and "advanced level" (C1) according to the Brazilian National Institute of Educational Studies and Research (*INEP; Instituto Nacional de Estudos e Pesquisas Educacionais*), based on the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001).

The L2-English group consisted of 50% female/male participants with ages between 22 and 44 years (M = 27.6, SD = 7.6). The L2-English group was composed by fifteen undergraduate students from the State of Paraíba, Brazil, and five graduate professionals (two from Paraíba and three from the State of Pernambuco, Brazil). All of the Brazilian participants were submitted to the Oxford Online Placement Test (Purpura, 2009) for proficiency level assessment. The group qualified as "low-advanced level" (B2–C1), "advanced level" (C1), and

[1] https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/celpe-bras/provas.

"master's advanced level" (C2) based on the CEFR (Council of Europe, 2001).

3.1.2. Data collection

Corpus. Participants from the L1- and L2-English groups read a phonetically-balanced version of the Aesop's fable, "The Lion and the Mouse" (see Appendix A).

Task. Participants were previously shown the text, and instructed to read it in normal pace using authentic (modal) voice. Before the recording process began, participants were allocated in a room and presented the text. They could read the text silently or aloud as long as they needed. This task is referred to as "story reading". The whole task took a total of 2:30 h (1 h for familiarization to the text + 1:30 h for the reading-recording process).

Recording procedures. Participants were recorded from a studio room with appropriate acoustic conditions. For the recordings, we used a Tascam DR 100 MKII digital recorder, and a unidirectional electromagnetic-isolated cardioid Shure SM7B dynamic microphone, at a sampling rate of 48 kHz and 16-bit quantization rate to ensure high quality and noise-interference reduction in order to guarantee the preservation of the intensity and voice quality features used for later acoustic analysis.

3.1.3. Acoustic analysis

Data segmentation was performed in *Praat* software (Boersma & Weenink, 1992–2021) as presented in *a* and *b*:

a) Segmentation was performed into the following units: vowels (V), consonants (C), phonetic syllables (i.e., vowel onset to the next vowel onset, V-to-V), silent and filled pauses (#), and higher units (i.e., chunks, CH) of speech (see Appendix D for an arrangement of the segmentation procedure);

b) Automatic extraction of the prosodic-acoustic parameters was performed through a script for *Praat* ('ProsodyDescriptorExtractor'; Barbosa, 2020) over the segmented units (see Appendix C for a detailed table of the acoustic parameters, units of measure and description of the acoustic feature).

For the segmentation and annotation protocols, this research is supported by Barbosa's (2006) study for the phonetic syllable (V-to-V) and pause (#) units, Ramus et al. (1999) for phonemic-sized (V/C) and pause (#) units, as well as Caroll (1994), Silva Jr. and Barbosa (2019, 2023), and Ortega-Llebaria et al. (2023) for chunk (CH) units.

The recordings were segmented into four chunks per participant (Appendix A). A total of 96 chunks were computed after data segmentation (4 chunks from the text × 24 participants = 96 chunks).

3.1.4. Statistical analysis

For the statistical analysis, a 1-way analysis of variance (ANOVA) test was performed in order to assess the effect of the factor 'Accent' (native or foreign) on the prosodic parameters. The effect size of the factor was determined by the coefficient of determination (the adjusted R^2). The coefficient of determination represents a measure of the proportion of variance for the factor 'Accent'.

With regards to the coefficient of determination, the statistic literature is controversial when determining a landmark for the R^2 power of effect values. In order to normalize these values in the analyses of the present research, the R^2 was described based on the Mean and the SD values from Cohen's (1992), Chin's (1998), and Henseler's et al. (2009) studies, in addition to the concepts supported by Cohen (1992) and Moksony (1999). The normalized R^2 values adopted in the present work are referred to as: strong effect size ($R^2 \geq .50$), weak-to-satisfactory effect size: ($.21 \leq R^2 \leq .49$), and weak effect size: ($R^2 \leq .20$).

		Native accent		Foreign accent				
Acoustic correlate	**Prosodic parameters**	*M*	*SD*	*M*	*SD*	*F*(1,94)	*p*	R^2
F0	Minimum	78.86	4.42	81.78	4.81	3.92		.31
	Semi-amplitude between quartiles	26.38	11.20	18.81	5.45	14.44	***	.12
	SD	5.12	1.16	3.56	0.79	35.44	***	.37
	Negative slope	–6.29	2.02	–5.15	1.52	5.42		.46
	Positive slope SD	5.78	1.97	4.80	1.45	4.26		.34
	Negative slope SD	6.09	1.41	4.97	1.47	6.07	***	.52
	Total slope SD	8.59	1.72	7.19	1.82	6.22		.54
	Peaks SD	49.00	11.30	37.00	10.60	13.11	***	.12
Intensity	Variation coefficient (Varco-I)	20.79	4.97	17.73	1.62	17.90	***	.62
Voice quality	HNR	10.61	4.32	13.22	2.16	11.64	***	.58
	LTAS (0–1:4–8 kHz)	–24.94	3.15	–27.76	3.29	8.53		.17
	Jitter	2.34	0.66	1.83	0.42	12.98	***	.21
	Shimmer	9.68	2.87	7.35	1.33	22.19	***	.57
Duration	Speech rate	3.70	0.39	3.16	0.40	79.02	***	.77
	Articulation rate	4.48	0.43	4.04	0.40	36.23	***	.51

Table 1. Means, SDs, and One-way ANOVA measures: *F* values (degrees of freedom), and R^2 values for the effect size for the prosodic-acoustic parameters of F0, intensity, voice quality and duration (*** = $p < .001$).

3.1.5. Results and discussion

This section presents the (significant) statistical results from the analyzed prosodic-acoustic parameters. Table 1 presents the Mean and SD values for each group in addition to the *F* values and R^2 values. Figure 1 presents the violin plots (along with the boxplots and mean values) showing the performance of each group.

Table 1 presents the parameters for experiment 1. The choice for durational parameters (Speech and Articulation rate), and the F0 parameters of centrality and variability (Minimum, Semi-amplitude interquartile, SD) are aligned with San Segundo et al. (2019) in the constitution of a (prosodic and segmental) multiparametric system for forensic speaker comparison, as well as the protocols determined by the European Network of Forensic Sciences Institute (ENFSI, 2021).

Features of F0 modulation and their variability (the parameters related to the dynamics of the F0 trajectory such as, Negative slope, SD of the positive/negative/total slope and SD of the peaks, are suggested by Eriksson and Wretling (1997), Harrington (2010), Mennen et al. (2011), Tremblay et al. (2016), and Silva and Arantes (2021) for speaker comparison and recognition tasks (in forensics).

As for the choice of variation coefficient of intensity, the present research is aligned with He (2017), that compared speaker recognition strengths based on intensity. Pellegrino et al. (2021) used variation coefficient of intensity for age-related rhythmic variation comparison between younger and older speakers of Zurich German.

As for voice quality, Alcaraz (2023) used the long-term average spectrum (LTAS) in the forensic domain for comparison and discrimination of different speakers.

For other voice quality parameters such as, HNR, jitter and shimmer, and duration parameter such as, Articulation rate, San Segundo et al.'s (2019)

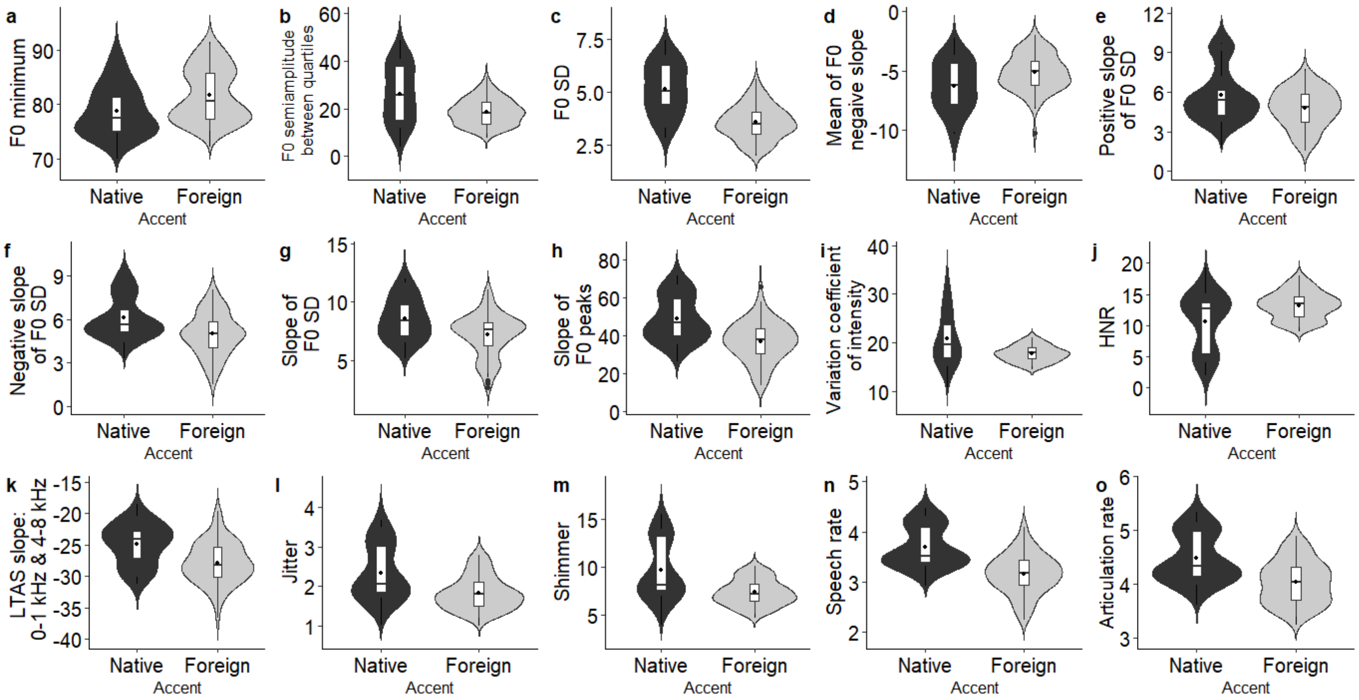


Figure 1. Violin plots (accompanied with boxplots and mean values) for the parameters that best discriminated between groups from Experiment 1 (authentic voices produced by both groups). Melodic (1a to 1h), intensive (1i to 1m), and durational parameters (1n and 1o).

study used such features in voice disguise for forensic speaker recognition along with short-term cepstral features (MFCCs) on 12 monozygotic twin pairs. The authors' findings show promising trends for the mentioned features applied in instances they refer to as "forensically-realistic characteristics", that is, background noise, reverberation, within-speaker variability, as well as signal compression.

Figure 1 presents the violin plots, the boxplots and mean values of each group for the parameters in Table 1.

As for the effect size shown in Table 1 and the groups' performance for the acoustic features presented in Figure 1, it may be concluded that:

a) up to 47% of the prosodic-acoustic parameters presented a strong effect size ($R^2 \geq .50$, for F0 (negative slope, SD of the slope; Figures 1f, 1g), for intensity (variation coefficient; Figure 1i), for voice quality (HNR, shimmer; Figures 1j, 1m) and for duration (speech and articulation rates; Figures 1n, 1o);

b) up to 33% of the prosodic-acoustic parameters presented a weak-to-satisfactory effect size ($.21 \leq R^2 \leq .49$, for F0 (minimum, SD, mean of the negative slope, SD of the positive slope); Figures 1a, 1c, 1d, 1e) and for voice quality (jitter; Figure 1l), and

c) up to 20% of the prosodic-acoustic parameters presented a weak effect size ($R^2 \leq .20$, for F0 (semi-amplitude interquartile, SD of the peaks); Figures 1b, 1h) and for voice quality (LTAS slope in the frequencies 0–1:4–8 kHz; Figure 1k).

Results presented in Table 1 will be detailed in this section '3.1.5.1' to '3.1.5.4'.

3.1.5.1. F0 results

F0 minimum (Figure 1a) values showed to be lower for the native group in comparison to the foreign group, $F(1, 94) = 3.92$, $p = .050$, $R^2 = .31$. Similar results can be found in Mennen et al. (2011), Urbani (2012), Järvinen and Laukkanen (2015), Tremblay et al. (2016), Idemaru et al. (2018), Silva Jr. and Barbosa (2019), Silva and Arantes (2021) studies where vocal load is attested to cause higher values of F0 in foreign speech during story reading due to cognitive effort.

When dispersion parameters come to scene, the native group presented higher variability than the foreign group as one may observe in the semi-amplitude interquartile (Figure 1b), $F(1, 94) = 14.44$, $p < .001$, $R^2 = .12$, and the standard deviation (Figure 1c), $F(1, 94) = 35.44$, $p < .001$, $R^2 = .37$.

The F0 modulation parameters also presented higher variability for the native group. This can be noticed by the mean of the negative slope (Figure 1d), $F(1, 94) = 5.42$, $p = .022$, $R^2 = .46$, the standard deviation of the positive, the negative, and the total slopes (Figure 1e), $F(1, 94) = 4.26$, $p = .041$, $R^2 = .34$, (Figure 1f), $F(1, 94) = 6.07$, $p < .001$, $R^2 = .52$, and (Figure 1g), $F(1, 94) = 1.80$, $p = .014$, $R^2 = .54$, respectively, as well as the standard deviation of F0 peaks (Figure 1h), $F(1, 94) = 13.11$, $p < .001$, $R^2 = .12$.

Magen's (1998) study on the perception of F0 by English native speakers reveals that, along with the F0 minimum, the variability parameters were considered the most relevant acoustic features on foreign accent discrimination in speaker's comparison. As aforementioned for the F0 minimum, our findings for F0 variability corroborate Urbani (2012), in which Italian speakers of L2-English speak with a narrower F0 range and less variation. This might have happened because of the difficulties that foreign speaker presents to process the L2 speech, and as a consequence, it might increase the cognitive load.

3.1.5.2. Intensity results

Regarding the intensity parameters, the presented results corroborate classic L2 prosody studies that analyzed the relation between F0, intensity and duration (Adams & Munro, 1978; Adams, 1979). Such studies suggest higher vocal effort along with narrow variability, interacting with high F0 minimum and lower speech rate when producing the target L2 point out to significantly high foreign-accented speech.

For the present research, the intensity variation coefficient (Varco-I, Figure 1i) presented higher vocal effort variability for the native accent when compared to the foreign accent productions, $F(1, 94) = 17.90$, $p < .001$, $R^2 = .62$.

He (2017) claims that the use of intensity metrics such as Varco-I, is a robust measure for speaker comparison. The studies still highlight that Varco-I seems to be more consistent than duration metrics in speaker recognition. He et al. (2015) yet address that Varco-I would have a higher probability and strength of improving speaker classification in ASR based on suprasegmental (prosodic) variability features.

3.1.5.3. Voice quality results

As one may observe from Figure 1j, there is higher variability of the HNR (Harmonic-to-noise ratio) parameter along with lower values produced by the native group, $F(1, 94) = 11.64$, $p < .001$, $R^2 = .58$. Such as Varco-I (see section 3.1.5b), this is as an indication of vocal effort. Such vocal effort is also stated by the LTAS (long-term average spectrum) slope of frequencies from 0 to 1 kHz and 4 to 8 kHz (Figure 1k), $F(1, 94) = 8.53$, $p = .004$, $R^2 = .17$, by the Local jitter, (Figure 1l), $F(1, 94) = 12.98$, $p < .001$, $R^2 = .21$, and the Local shimmer (Figure 1m), $F(1, 94) = 22.19$, $p < .001$, $R^2 = .57$.

By this sense, the higher amount of noise produced by the L1 speakers (more noise indicates lower HNR values) might be related to the social markers and other extralinguistic aspects as pointed out by Laver (1980, 1994). Regarding the LTAS, this parameter is straightly related to a sonorous, clear, and authentic voice (lower values presented by the Brazilian group), or a harsher, creakier, and leaky voices (higher values presented by the amount of creakiness produced in L1 speech) as attested by Esling et al., (2019) and Alcaraz (2023).

Farrús (2018) addresses that L2 voice quality figures as the class of acoustic parameters that could cause great difficulties to the lay listener at the task of speaker identification, especially in voice disguise context. According to Järvinen (2017) and

Järvinen et al. (2017), languages may differ in the type of phonation and vocal effort, however, L2 speech production seem to differ from L1 independent of the language. These researches still address that, changes from L1 to L2 resulted in a significant difference of the LTAS in frequencies between 50 and 1200 Hz, high F0 minimum values and relatively long F0 peak width in regions from zero to 300 Hz. These studies also reported that the perceptual differences between speaking L1 and L2, is that in the L2, there is a considerable diminishing performance of voice quality (more pressed, strenuous voice production, higher-pitched and fatigued voice) due to the input cognitive effort.

Acoustic parameters such as the ones mentioned in sections 3.1.5.1, 3.1.5.2 and 3.1.5.3 represent how much the native speaker perceives expressivity, tenacity, and enthusiasm in the target language. For the sake of expressivity or interpretation of good/bad characters in narrated stories, fairy tales, fables, films etc., it is common that voice quality gains more variability and different modulation in certain instances of the rendition for instance.

In the reading task, one might infer that the native group was able to interpret characters in such a way that L2 group could not. This is reflected in higher values of jitter and shimmer, lower and more variable HNR values, in addition to higher and more variable LTAS values between the 0–1 kHz and 4–8 kHz for the L1 group. According to Niebuhr et al. (2018), it means that the slope of the spectra during production is associated to changes in the mode of phonation (the degree of creakiness measured between 1–4 kHz, as well as the degree of breathiness measured between 5–8 kHz). Moreover, the results of HNR and LTAS for the present study corroborate Engelbert (2014), in which she evaluated both parameters in several spectral ranges cross-linguistically, involving Brazilian Portuguese (BP) speakers’ productions of English and Portuguese.

3.1.5.4. Duration results

The results presented in this section corroborate, to a certain extent, L2 speech studies, such as Munro (1999), Loukina et al. (2009), Fuchs (2016), Silva Jr. and Barbosa (2019, 2021, 2023), Teixeira and Lima Jr. (2021), inter alia.

Loukina et al. (2009), and Silva Jr. and Barbosa (2019), state that speech rate is one of the most reliable and consistent parameters for measuring L2 acoustic duration. Similar results were found in the present study (Figure 1n), $F(1, 94) = 79.02$, $p < .001$, $R^2 = .77$, as well as for the articulation rate (Figure 1o), $F(1, 94) = 36.23$, $p < .001$, $R^2 = .51$. On the other hand, Gut (2012) asserts that speech rate does not seem to be a reliable measure for L2 speech rhythm, for being a breaking point of sensitivity for other quantitative metrics, that is, the slower speech rate of non-native speech the more it might distort other measurements.

Along this section, it was brought up results and some discussion showing evidence that when reading in L2, a consistent cognitive vocal load is presented during speech planning and consequently, it highly affects melodic, voice quality, intensive and durational parameters. We may conclude, at least on a preliminary basis, that prosodic planning is somewhat neglected in several instances of speech, and one might think over and consider a revisitation on what (else) could aggregate, for instance, knowledge when speaking or training L2 pronunciation for different purposes as claimed by Krivokapic (2012), and Reed and Michaud (2015).

3.2. Experiment 2

For Experiment 2, The L1-English group will have speech chunks in authentic voice and the L2-English group, in disguised voice.

3.2.1. Participants

3.2.1.1. The L1-English group

For the second experiment, we remained only with the male L1-English speakers' samples from the first experiment, that is, samples of two speakers (see section 3.1.1). A novel male speaker was added forming the group for the second experiment. The novel speaker was 60 years old when the experiment was carried out. Participants were only male to guarantee the comparison with the L2 group. Participants' ages were in between 29 and 60 years (*M* = 44.6, *SD* = 25.1). The L1 group produce authentic voice only (all of the chunks from the story-reading task of experiment 1 + the interview chunks of the novel speaker). A total of three speakers participated in this experiment (2 speakers from Experiment 1 + 1 novel speaker = 3 L1-English speakers).

3.2.1.2. The L2-English group

For the L2-English, a male BP speaker of the first experiment participated (henceforth, 'BPS-DV', which stands for "BP Speaker with a Disguised Voice"). He is the first author of this paper and lived in Florida, U.S.A for two years. He also started to study English in Brazil at the age of eleven. When data were collected, he was 41 years old. The L2 group, i.e., the BPS-DV, produced disguised voices only (BPS-DV interview chunks + BPS-DV reading-imitation chunks). A total of 14 chunks was produced by the L2 group (4 chunks for the reading imitation + 10 chunks for the interview impersonation = 14 chunks).

3.2.2. Data collection

Corpus. The dataset for Experiment 2 consisted of:

a) *L1-English*: male chunks containing the same recordings of the fable-reading in authentic voice from Experiment 1, and chunks from an interview. The interview is referred to as the authentic voice either.
b) *L2-English* chunks containing recordings of the fable-reading imitation, as well as the chunks from an interview impersonation with a disguised voice by the BPS-DV.

The interview used for this experiment is from a voice-over artist (Redd Pepper) called "Meet the Epic Voice Behind Movie Trailers" (Great Big Story, 2018). The interview was extracted from YouTube streaming platform and it contains a total of 3 min 11 s.

Task. As well as for the story reading, the BPS-DV was previously shown the interview. Voice disguise was supposed to be performed as follows:

a) The BPS-DV had to memorize and reproduce the fable's words from the first experiment converging toward a native-like American English accent, and;
b) The BPS-DV had to listen to the voice-over artist interview, memorize and impersonate the whole rendition (see Appendix B for the transcript of the interview).

Recording procedures. BPS-DV was recorded from a studio room with appropriate acoustic conditions. Such as in experiment 1, recordings were held in a TASCAM DR-100 mkII Digital Recorder, and a unidirectional electromagnetic-isolated cardioid Shure SM7B dynamic microphone, at a sampling rate of 48 kHz and 16-bit quantization rate to ensure high quality and noise reduction in order to guarantee the preservation of the intensity and voice quality features used for later acoustic analysis. The interview was extracted from YouTube streaming platform using Audacity software, as well at a sampling rate of 48 kHz and 16-bit quantization rate to ensure high quality.

It is worthy highlighting that the interview impersonation did not have to be strictly with the same words of the original interview for the maintenance of (semi)spontaneous speech characteristics as pointed out by Masthoff (1996), and Schiller and Koster (1996) during the task of speaker comparison in the forensic field. The whole speech information of the interview was transcribed, for either memorization process or for labeling the segmentation in later acoustic analysis.

Differently from experiment 1, this experiment used Nolan's (1983/2009) protocol by means of what best fits forensic implication analyses. To do so, BPS-DV had all of his voice-disguise tasks recorded three times, so it would be able to check if the acoustic parameters would reveal a (non)significant variability within-speaker. The whole task took 5hr (2hr for the familiarization to the accent and voice modulations, 2hr for the edition of the interview, such as video-audio conversion and audio-cuts, and 1hr for the recording process) distributed in three days.

3.2.3 Acoustic analysis

For the acoustic analysis, we carried out the data segmentation procedures and the extraction of the prosodic parameters likewise we conducted in experiment 1 (see section 3.1.3a and 3.1.3b).

Data were segmented into 32 chunks:

{[story reading authentic = (4 chunks × 2 L1 participants = 8 chunks) + story reading disguised = (4 chunks × 1 BPS-DV = 4 chunks) + interview authentic = (10 chunks × 1 L1 novel participant = 10 chunks) + interview disguise = (10 chunks × 1 BPS-DV = 10 chunks)] = story reading authentic + story reading disguised + interview authentic + interview disguised = 32 chunks}.

3.2.4. Statistical analysis

Such as in section 3.1.4, we performed 1-way ANOVA statistics in order to assess the effect of the factor 'Accent' (native or foreign), on the prosodic parameters. The effect size for the factor was also determined by the adjusted R^2.

3.2.5. Results and discussion

This section presents the statistical results for the significant prosodic-acoustic parameters from Table 2 for each group. Figure 2 presents the performance of the groups.

As for the voice quality parameters in Table 2, the choice for the LTAS in two different levels of frequency bands (0–1:1–4 kHz and 0–1:4–8 kHz) are aligned to Niebuhr et al. (2018), Alcaraz (2023) among others, in studies for the comparison and the discrimination between disguised and authentic voice. For forensic speaker comparison, the choice for H1–H2 is aligned to San Segundo (2014, 2021).

In the case of Cepstral Prominence Peak (CPP), Procter (2019) considers CPP as being a measure that correlates to the harmonic structure and well-defined F0 parameters of the voice signal. Her study points out to CPP as being a consistent predictor of foreign accentedness.

The choice for CPP in the present research is for accent comparison purposes (native/foreign) once

		Native accent		Foreign accent				
Acoustic correlate	**Prosodic parameters**	*M*	*SD*	*M*	*SD*	*F*(1,30)	*p*	R^2
Intensity	Spectral emphasis	1.72	1.03	5.07	2.16	11.72		.26
Voice quality	LTAS (0–1:1–4 kHz)	–19.32	2.40	–11.40	3.07	24.34	***	.45
	LTAS (0–1:4–8 kHz)	–24.10	5.36	–14.74	7.06	6.82		.27
	H1–H2 amplitude difference	–0.83	5.03	–9.82	4.67	30.42	***	.42
	Cepstral Prominence Peak (CPP)	21.70	5.60	15.80	3.10	16.55	***	.27

Table 2. Means, Standard Deviations, and One-Way ANOVA measures: *F* values (degrees of freedom), and R^2 values for the effect size for the prosodic-acoustic parameters of intensity and voice quality (*** = $p < .001$).

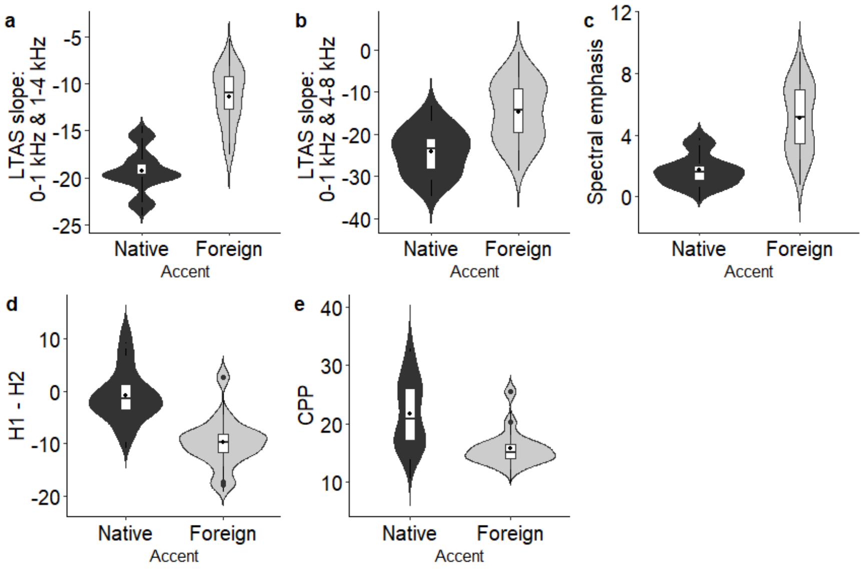


Figure 2. Violin plots (accompanied with boxplots, median and mean values) for the parameters that best discriminated between groups from Experiment 2 (L1 authentic voice and L2 disguised voice). Voice quality (2a, 2b, 2d and 2e) and intensity (2c).

this feature already compares different groups, such as dysphonic/normal speakers, female/male speakers, adolescents/children showing up some interesting degree of consistency as pointed out by Garret (2013). It seems to be a suitable parameter for checking the presence/absence of sustained vowels and syllables as addressed by Procter (2019).

As for the effect size that modeled the prosodic parameters observed in Table 2, all of the five prosodic-acoustic parameters extracted from the voice disguise task presented a weak-to-satisfactory effect size ($.21 \leq R^2 \leq .49$). These parameters include one of intensity (Spectral emphasis) and three of voice quality (LTAS in the frequencies 0–1:1–4 kHz, LTAS in the frequencies 0–1:4–8 kHz, H1–H2 and CPP). This means that, by the fact of these parameters are intrinsically distinct, they hardly might be impersonated as suggested by Leemann and Kolly (2015).

3.2.5.1. Intensity results

As for the intensity features, only the spectral emphasis in Figure 2c, $F(1, 30) = 11.72$, $p = .006$, $R^2 = .26$, presented a significant difference between the speeches for the voice disguise task. These results corroborate Modesto’s (2019) study where the author analyzed Spectral emphasis and its correlation to other kinds of vocal effort, such as relative intensity for BP speakers of English in production and perception of lexical stress. His study showed evidence that BP speakers of English find difficult to cope with intensive (besides durational and melodic) parameters.

Kapolowicz et al. (2016) used Spectral emphasis for foreign accent detection during speech recognition tasks and addressed its robustness when comparing Spectral emphasis to other acoustic parameters. Brouwer (2019) in a study of foreign accent familiarity in background speech-in-speech recognition points out that Spectral emphasis is a strong and reliable acoustic feature that the listener can rely on for foreign accent recognition in noisy background. The study highlights the strong correlation between Spectral emphasis and LTAS in different frequency bands.

In the ASR domain, Heldner (2001) addresses that spectral emphasis may be described as an acoustic feature reflecting the relative intensity in the higher frequency bands that seems to be more useful for the detection of accents than overall intensity and other intensive-based acoustic features. The author points out that Spectral emphasis was also found to outperform F0- and duration-based features in certain conditions, and it is argued to be inserted in ASR systems using a combination of acoustic features for automatic classification of prosodic categories and (foreign) accent.

3.2.5.2. Voice quality results

As it is presented in Figures 2a, 2b, 2d and 2e, voice quality parameters revealed significant differences between the authentic L1 voices and the BPS-DV. As one may observe in Table 2, as well as in Figure 2b, the foreign speech significantly steeped the LTAS slope in the medium and high frequencies when comparing to the authentic L1 voice, as well as the authentic L2 voice from Experiment 1 (Figure 1k, section 3.1.5) and to the native speech. In Figure 2a, the density curve in the plots illustrate a negatively-skewed distribution of the foreign data, which means that most of the speech productions of the disguised voice had higher values of frequencies in between 1 to 4 kHz resulting in a possible degree of creakiness as pointed out by Niebuhr et al. (2018).

As for the LTAS slope of frequencies in between 4–8 kHz, Figure 2b presents a density curve for the foreign speech data which reflects a “quasi” bimodal distribution, which means that there was a shift in the voice quality during the productions. One possibility for this performance is that there was cognitive and vocal load to maintain the disguised voice in the frequency range of 4 to 8 kHz, which suggests that the BPS-DV may have had some difficulty at keeping the disguised voice in terms of creakiness, breathiness and harsh for a longer period of time as attested by Maryn (2010); Tjaden et al. (2010); Hernandez (2012); Costa (2017); and Niebuhr et al. (2018).

Regarding the voice quality parameters in voice disguise context, results are aligned to other forensic L2 speech studies, such as Keating and Esposito (2007), Eriksson (2010), Fraile and Godino-Llorente (2014), Keating et al. (2015) and Alcaraz (2023). Such studies point out to a significant variability between the native and foreign accented speech in spectral modulation parameters such as the LTAS slope in frequencies of 0–1:1–4 kHz, $F(1, 30) = 24.34$, $p < .001$, $R^2 = .45$, and in frequencies of 0–1:4–8 kHz, $F(1, 30) = 6.82$, $p = .025$, $R^2 = .27$, as presented in Figures 2a and 2b, respectively, as well as the amplitude of H1–H2, $F(1, 30) = 30.42$, $p < .001$, $R^2 = .42$, and the CPP, $F(1, 30) = 16.55$, $p < .001$, $R^2 = .27$, respectively in Figures 2d and 2e.

For Garellek and Keating (2011), and Fraile and Godino-Llorente (2014), CPP is an acoustic measure of voice quality that has been qualified as the most promising and perhaps the most robust acoustic measure of breathy voices’ evaluation. Phonetic literature traditionally deals with CPP with applications to voice pathologies and/or disorders. As for H1–H2, Keating et al. (2015) highlight that this is a consistent parameter for measuring creaky and breathy voices, both in low and high frequency ranges. The authors yet address that H1–H2 figures as the best parameter to distinguishing creaky voice from other voice qualities, as well as different levels of creakiness. This measure generally reflects glottal constriction, with a lower value indicating greater constriction.

In the L2 speech domain, Duarte and Silva Jr. (2020) used H1–H2 to investigate differences in L1 and L2 English productions of glottal stops from both American and Brazilian speakers. The authors concluded that H1–H2 showed to be a reliable parameter for the determination of glottal stops and laryngeal gestures.

With regard to the use of the CPP, Procter (2019) conducted an experiment with French and Spanish L2 speakers of English, and a control (L1 speakers) group from the U.S.A. In her study, she evaluated the influence of CPP measures on foreign accent degree rated in a perceptual analysis. As mentioned in this section, Procter (2019) showed that CPP was a great predictor of accentedness. Moreover, when the factor Language is controlled for the factor Gender, results are still more consistent.

In summary, this section presented reliable results when using LTAS, H1–H2 and CPP for the determination of breathiness and creakiness in disguised tasks of foreign-accented voices. We conclude on a preliminary basis, that the parameters presented in this section (as well as Spectral emphasis in section 3.2.5a) were of the utmost importance for modeling the disguised voice samples herein presented.

It is worth noting that the disguised tasks presented here were produced by BPS-DV in a short period of time, i.e., production of memorized chunks that did not take more than 20-to-30 seconds of duration. Neuhauser (2008) suggests that it is necessary to test if acoustic features that succeeded in the disguise task could be maintained over a longer period of time.

3.3. Experiment 3

For Experiment 3, lay L1-English listeners rated both the native accent and the foreign accent samples produced by the L1-L2 English groups.

3.3.1. Participants

For the perceptual experiment, data were collected from a L1-English-speaking group (ten Americans who lived in Brazil for about two years when the experiment was run). The American participants from the first and second experiments were not included as part of the perceptual experiment. This group consisted of 50% female/male participants with ages between 24 and 56 years (*M* = 37.5, *SD* = 12.8).

3.3.2. Data collection

For this third experiment, Participants were asked to rate speech chunks by the degree of foreign accent.

Corpus. 60 randomly-distributed speech chunks of approximately twenty seconds, organized in the following configuration:

a) Thirty chunks from Experiment 1: (story reading in authentic voice for both L1 and L2 groups. Fifteen chunks per group);
b) Thirty chunks from Experiment 2: (L1 chunks: the story reading and the interview in authentic voice; L2 chunks: story reading in authentic voice; story reading and interview imitation in disguised voice. Fifteen chunks per group).

3.3.3. Perceptual analysis

For the perceptual analysis, Multiple Forced Choice (MFC) listening experiment was carried out in Praat software (Boersma & Weenink, 1992–2021).

Task. Listeners had to rate the foreign accent degree of the 60 randomly-distributed chunks through a 7-point Likert scale (the higher the score the higher the foreign accent degree. The anchor points were: ‘No foreign accent’ = 0; ‘Very low foreign accent’ = 1; ‘Low foreign accent’ = 2; ‘Neutral foreign accent’ = 3; ‘High foreign accent’ = 4; ‘Very high foreign accent’ = 5, and ‘Extreme foreign accent’ = 6). As mentioned in section 3.3.2, this experiment contained 60 speech chunks that were about twenty seconds long. The whole experiment lasted around twenty minutes per participant. The whole task took a total of 3hr 20 min distributed in four days.

It is important to highlight that using a 7-point scale for grading foreign accent, is neither norm-referenced nor a commonsense, and therefore, may be interpreted only in a relative sense as pointed out by Munro and Derwing (1995, 2020). Munro (2018) reported that listeners consistently spread over the 9-point scale since 19 out of the 21 listeners in his experiment used at least eight of the nine points. Derwing and Munro (2009) also present successful use of a 9-point interval scale. On the other hand, Busch and Turner (1993) mentions that 5-point interval scaling is used by L2 researchers to measure learners' characteristics, accent attitudes and opinions for validity of the research aims.

Southwood and Flege (1999) made a comparison between direct magnitude estimation (DME) and a 7-point interval scale for measuring perception of foreign accent. Their study analyzed (via ANOVA) the effect of the methods (DME vs. interval scaling) accounting for the score frequency of the intra/inter-judge rates. Results of posterior regression analyses indicated that there was a significant correlation between both of the methods, which suggest that accentedness may be a metathetic continuum at least for L1-Italian speakers of English listened by L1-English-speaking listeners. Southwood and Flege (1999) also draw attention to the "ceiling effect" (mode = 7 for one of the listeners' groups) that might be caused due to the insufficient number of scale intervals.

The study suggests that, although frequently used, a 7-point scale may not be sensitive enough for a number of listeners to discriminate among speech chunks. They also suggest that 9- or 11-point scales might improve listener's sensitivity when scaling degree of foreign accent.

Once in Southwood and Flege's (1999) study, there was no difference between DME continuum and 7-point interval scale methods, in the present study, we considered a 7-point scale for being the mean value between a 5-point scale, suggested by Busch and Turner (1993) and a 9-point scale, suggested by Munro and Derwing (1995, 2020), Derwing and Munro (2009), and Munro (2018).

3.3.4. Statistical analysis

For this experiment, we performed a Kruskal-Wallis test in order to evaluate the effect of the factor 'Voice style' (authentic or disguised), on the rating scores for both native and foreign speech.

3.3.5 Results and discussion

As one may see in Table 3 and Figure 3, we detailed the results from the Experiment 3.

	Native accent		Foreign accent				
Voice style	*n*	%	*n*	%	$\chi^2(1)$	*p*	η^2
Authentic (L1-L2 groups)	15	25.0	15	25.0	31.08	***	.82
Authentic (L1) / Disguised (BPS-DV)	15	25.0	15	25.0	4.37		.26

Table 3. Number of samples (*n*), proportional values (%), Kruskal-Wallis χ^2 values (degrees of freedom) [$\chi^2(1)$] and the eta effect size (η2) for the factor 'Voice style' (authentic and disguised voices), on the rating scores of both native/foreign speech (*** = $p < .001$).

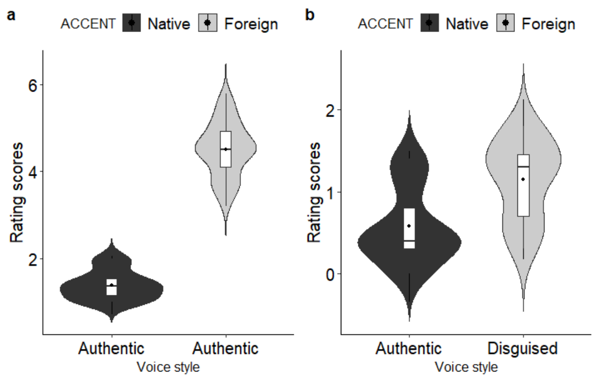


Figure 3. Violin plots (accompanied with boxplots, median and mean values) for the rating scores between the speech chunks of 'a' (authentic voices) and 'b' (L1 authentic voice / L2 disguised voice).

Figure 3a and 3b present how the listeners performed when rating between authentic and voice-disguised chunks of speech. Results significantly differed when the task to be rated was the story reading (Figure 3a, both groups using authentic voice) $\chi^2(1, N = 60) = 31.08$, $\eta^2 = .82$, $p < .001$. A strong effect size (η^2) explains 82% of the variation between groups. The score mean values for the native group was 1.38 and for the foreign group, 4.51. The most extreme scaling scores in the comparison of authentic voices were '0' for the native group, and '6' for the foreign group.

On the other hand, results were statistically conflicting and inconclusive (Greenland, 2019; Amrhein & Greenland, 2022) when the rated task was to differentiate between the L1 authentic speech and the L2 disguised speech (Figure 3b), $\chi^2(1, N = 60) = 4.37$, $\eta^2 = .26$, $p = .056$, where a weak-to-satisfactory effect size explained 26% of the variation between both native and foreign-accented speeches. The scores mean values for the native group was 0.57 and for the foreign group, 1.15. The most extreme scaling score in the comparison authentic/disguised voices was '2' for native speech and '3', for the foreign speech.

In terms of native speech, similar findings were attested by Munro and Derwing (1995, 2020) where the raters scored one of the native speeches worse, at 2.4, given the variability in speech rate and voice quality parameters. The study yet state that these parameters also affect foreign accent comprehensibility.

4. General discussion

In this section, we will answer the research questions put forward in the Introduction, as well as promote a general discussion around the three experiments presented. It will also be presented some implications to the forensic field concerning voice disguise and foreign accent in the forensic domain, as well as some further insights for L2 pronunciation teaching directions.

The first research question was:

I. *As a (Brazilian) foreign speaker of English converges toward the prosodic structure of the L2 when voice disguising, will it be more difficult to the lay listener to recognize this L2-English accent?*

Answer-I: Yes. It seems to be more difficult to the lay listener to recognize the L2-English speaker's degree of accent in a scalar rating task when he converges toward a native-like accent. The results of the Experiments 1 and 2 also provided evidence that imitation might foster that prosodic-acoustic features, at least in the experiments herein conducted, performed differently for both authentic and disguised voices.

Eriksson (2010) comments that the target of the imitation is the vocal behavior of a specific individual. There is a chance that such behavior might be related to the individual linguistic aptitude. According to Wen (2019), L2 (phonological/prosodic) aptitude is basically related to two models: the acquisitional and long-term developmental aspects of L2 knowledge phonemes (and prosody), and the processing that regulates and coordinates attentional resources implicated in L2 comprehension and production.

For Wen (2019), the phonological/prosodic components are conceived as a key construct of language aptitude, which can be further demarcated into a phonological (and prosodic) short-term store, as well as an articulatory rehearsal mechanism. Such components have been claimed to play an instrumental role in acquiring novel phonological (and prosodic) forms which facilitates the chunking process of linguistic sequences of different levels ranging from phonemes, words, and phrases to morphosyntactic constructions.

It seems to be the case that was done by the BPS-DV, when designated for the tasks of memorizing and imitating the chunks of the fable, and of the interview (using the prosodic short-term store and the articulatory rehearsal mechanism).

The second research question was:

II. *Which prosodic parameters, from fundamental frequency, intensity, duration and voice quality, best correlate with listeners' judgments of L2 speakers' degree of foreign accent?*

Answer-II: For this study, since listeners were acquainted with Brazilians' foreign accent of English, all of the four acoustic correlates, i.e., F0, intensity, duration and voice quality, seem to be related to their perception when comparing foreign accent degree, nevertheless, it would highly depend on the voice style, once, in the present study, F0 and duration were significantly consistent at differentiating groups in authentic voices, and intensity and voice quality were consistent at differentiating both authentic and disguised voices on the experiments presented here.

Besides, voice style did play a role on the determination of foreign accent degree, but as one could see from the results presented in section 3.3.5, the L1-English-speaking lay listeners were able to judge, even from a (statistically) conflicting and inconclusive form, between the L1 authentic speech and L2 disguised speech. One plausible explanation to be inferred is that cognitive vocal load highly occurs:

a) When producing prosodic features in a certain language that is not one's mother tongue (in authentic or) in disguised voice style;
b) In the cerebral insular cortex, which coordinates higher-order cognitive aspects of foreign speech and language processing, such as L2 prosody as proposed by Golestani and Pallier, 2007, and Hernandez (2012).

When speaking the L2 prosody, the brain area presents greater asymmetry in the left insula/prefrontal cortex when compared to one's L1 causing deficit in memory, attention, and emotion which is reflected in the speaker's phonetic performance (Golestani & Pallier, 2007; Costa, 2017). Costa (2017) yet highlights that L2 prosodic-acoustic parameters related to F0, amplitude/intensity and duration are greatly affected and very difficult to be impersonated.

As posed in section 2, Andrews (2019) explains that L2 acquisition is a dynamic process with periods of more intensive acquisition and loss of different language levels, and that perception-production process occurs simultaneously. We might infer that both the individual phonetically varies lifelong, especially if we think over the perception models for L2 (Best & Tyler, 2007). This is where the studies concerning within-speaker variability comes to scene.

A number of studies in forensic phonetics have tried to account for within-speaker variability of acoustic features. Hollien and Majewksi (1977) reported that voice disguise can lead to high within-speaker variability, which affects various acoustic parameters in long-term under distinct speaking conditions. For Endres et al. (1971), Eriksson and Wretling (1997), Leemann and Kolly (2015) among others, it is frequent that in long-term condition, F0, intensity and formant parameters are the most affected features during disguise context. These studies still address that high within-speaker variability in disguised voices makes it difficult for forensic experts to draw conclusions about the speakers' identity. Eriksson (2010) pinpoints that the type of voice disguise (through impersonation, for instance) is of particular interest in forensics, namely whether the voice disguise can produce an accurate-like copy of a specific native speech.

4.1. Some implications to the forensic field

As for the implications of foreign accent in forensic research, Eriksson (2005) addresses that the definition of what really sounds as foreign accent has laid down much on accent unfamiliarity rather than having a none-to-extremely strong degree of foreign-accented speech. The author still claims that shorter duration of speech samples is more difficult to be recognized on both native and foreign accent especially foreign accented voices. Eriksson's (2005) study concludes that results seem to be somewhat ambiguous, since there is a tendency for foreign accent to be less well recognized, although the difference is usually non-significant. He yet highlights that it is highly likely that experienced professionals, like linguists, perform better at recognizing foreign-accented voices than lay listeners.

Results from the rating scores of our study are (somewhat inconclusively) aligned, to some extent, to Schiller and Koster (1996), Rojczyk (2010) and Fernández-Trinidad's (2022). These studies pose that voice recognition is just as equally easy/hard to be done for foreign and native voices, since it will depend a great deal on the listener's language background. It might have been the case that the raters, who are L2-BP speakers, were influenced by the BPS-DV English proficiency (C2) on the productions of the L2 prosodic features.

Yet Procter (2019) points out that native speakers of a language may perceive a voice as normal if its accent is commonly found within the community and upholds the expected and anticipated variations of prosodic rate, fluency and voice quality features. The fact is that, there seems to be a consensus that even at determining foreign accent degree, an accurate accent identification, at least by means of rating, is rare when one is not familiar with the foreign accent in agenda. In the forensic field, this is the case of 'similarity' and 'typicality'.

A questioned voice can hold several kinds of prosodic parameters similar to (some of) the ones of the reference voice in the target L2, and one might mask his/her foreign accent degree to a certain extent (which is the case of BPS-DV in our study). On the other hand, forensic speech scientists/experts need to assess not only the similarity between the voices, but also, crucially, the typicality of features in the wider population as pointed out by Hughes and Wormald (2017), and thoroughly discussed by Brescancini and Gonçalves (2020) as part of the weighting model for sociophonetic evidence in the task of speaker comparison. Besides similarity and typicality, Brescancini and Gonçalves (2020) propose a third level for the analysis between reference and questioned voice; the 'individuality'.

Individuality might have been played a crucial role for the raters when judging BPS-DV chunks with a higher degree of foreign accent when compared to the L1-English group. Features of vocal effort, such as 'Spectral emphasis', 'H1–H2', 'CPP' and 'LTAS' in different frequencies seemed to outperform at identifying BPS-DV's speech chunks. A possible cause for this inference is that BPS-DV's chunks might have kept little within-speaker variability from authentic to disguised voice although this needs to be investigated in future studies. Foulkes et al. (2010) and Thomas (2011) claim that these are fine phonetic details very early apprehended by production, perception and cognitive processing sociophonetically produced by the individual in a number of social contexts.

As far as the LTAS is concerned, contrarily to what Alcaraz (2023) suggests, in our study the LTAS seems to figure as a reliable parameter for the forensic field, either in calculations or for speaker identification and comparison. Along our research, we could find evidence that goes on the opposite direction of Alcaraz (2023) in relation to the LTAS although the author states that his general impression of the study is that it is not yet conclusive and requires further in-depth, consistent research, as well as it is not prudent to rule outright in favor (or not) of the usefulness of LTAS in the forensic field. Besides, the reliability of LTAS measures will directly depend on (background) noise of the environment.

In terms of a (possible) real forensic scenario, what if there is a considerable amount of background noise? How would the listener judge a certain voice?

It is not a novel practice that researchers claim and debate about the quality (and quantity) of samples in different realistic forensic scenarios. According to Hollien and Majewksi (1977), Nolan, (1983/2009), Rose (2002), Eriksson (2005), ENFSI (2021), Fernández-Trinidad (2022), in forensic practice, we start from the fact that the information available is commonly scarce and has been collected by limiting telephone/microphone channels, sometimes in noisy environments, which results in low-quality sound samples. This forensic reality questions the applicability of acoustic and statistical analysis regardless of the voice quality parameters (Hollien & Majewksi, 1977; Fernández-Trinidad, 2022) herein used such as, LTAS, jitter, shimmer among others, that could discriminate between authentic and disguised voices in this research.

In the way to mitigate the number of problems for the low quality of the samples for forensic phonetics, Fernández-Trinidad (2022) suggests a multiparametric study (the present research was conducted under a number of parameters) including melodic, durational and glottic parameters which are less sensitive to non-optimum quality from the audio material collected in the forensic scenarios from the experts.

Moreover, Fernández-Trinidad (2022) draws attention to voice quality parameters of long-term laryngeal configurations as having a high discriminant power and better resisting to the attempts at imposition, camouflage, or dissimulation. The author yet poses that laryngeal functioning is more difficult to modify or impost, since it does not seem that we have such accurate and conscious control of our vocal system, compared to the one we execute on the articulatory system. It seems the data presented from experiment 2 (Table 2 and Figure 2) corroborates Fernández-Trinidad (2022) inferences for the forensic field.

In a perspective of speech recognition via perception, Brouwer (2019) attests that speech communication, in different situations, rarely takes place under quiet listening conditions, and interlocutors are in noisy environments in which they need to segregate the target signal from background noise (that could be either speech). Her study points out that being familiar and proficient in the background language plays a role in speech recognition.

As mentioned in this section, Fernández-Trinidad's (2022) study reiterates that the degree of familiarity with the background language might only be of influence when participants are highly proficient in that language, which was somewhat the case of our listeners. The American listeners of this study had lived in Brazil for two years when the experiment was conducted, and most of them were fairly nice fluent in BP language.

4.2. Further insights

Besides forensic implications, the results of the experiments herein presented have implications for foreign language instruction, more specifically to pronunciation teaching (Munro & Derwing, 1995, Derwing & Munro, 2009; Munro, 2003; McCullough, 2013; Grosjean & Li, 2013; De Marco, 2020; Silva Jr. & Barbosa, 2021). For L2 learners of English who express accent reduction as a priority, language instructors would be wise to focus attention on acoustic details of the speech signal that contribute most to the perception of the lay listener of the target language.

Furthermore, finding out what prosodic-acoustic (and segmental) parameters could best model foreign accent degree of English in the research field, it would provide useful information for the development of tools and protocols for proficiency level assessment, other than for the development of pronunciation applications, such as the 'BeatMaker', a computational software for L2 prosody teaching (Silva Jr., 2023).

5. Conclusions

By general means, the present research explored the effect of authentic and disguised voice styles on a series of prosodic-acoustic parameters (speech production experiments), and listeners' score ratings (perceptual experiment) for both L1- and L2-English. For the authentic voice style, 15 acoustic features proved to be significant based on *p*-valued descriptions with 47% of the features presenting a strong effect size, 33%, a weak-to-satisfactory effect size and 20%, a weak effect size. For the disguised voice style, five acoustic features showed significance, nevertheless presenting a weak-to-satisfactory effect size. Four out of the five parameters were of voice quality and one, of intensity.

For the present research, all of the four classes (F0, intensity, duration and voice quality) of prosodic-acoustic parameters conducted throughout the experiments performed consistently when comparing different foreign accent degree for the authentic voice. Intensive and voice quality parameters performed consistently for both authentic and disguised voice styles. From our findings and to a certain extent, intensity and voice quality seemed to present reliable parameters for accent classification in the forensic scenarios once they maintained small within-speaker variability.

Besides the parameters mentioned along this study, foreign accent provides important social information, such as speaker's origin, education, language proficiency, and further background. All of these aspects shall be taken into account for a robust methodological design when speaker identification and/or comparison comes to scene in forensic studies.

For the perceptual level, on the one hand, the assessing rates for the authentic voice proved to be consistently robust, from the *p* value and the effect-size descriptions. On the other hand, the evaluated rates for the disguised voices presented somewhat significantly (but conflicting and inconclusive) different results as well as a weak effect size value. This means that listeners were able to judge the degree of foreign accent of the disguised samples, at least to a certain extent, as being disguised and non-authentic. More investigation needs to be carried out on this topic.

6. Limitations and future directions

At this point, it is important to mention the limitations of this study and some highlighting points for directing its continuation. These findings bring a forward necessity to apply this protocol into more data even though the dataset herein presented was able to make preliminary and, somewhat, reliable inferences about the mean difference between the groups on the speaking styles presented. The authors of the present research are aware to increase the number of samples for future work for the sake of reliability and consistency of the results based on other prosodic correlates, such as F0.

Since F0 retains individually-related features, its maintenance/variation would be better explained with more samples (from different individuals) in voice disguise contexts. Accounting for more individuals in voice-disguising scenarios would bring, at least to a certain extent, more reliability when describing results based on prosodic features of F0, once the F0 measurements rely on both individual and general characteristics.

We also state that more acoustic features should be analyzed in voice-disguising context, especially for forensic applications. In addition, for the continuation of this research, it is intended to move forward in the following directions that could have not been explored in the present study:

a) Analyze how consistently the robust prosodic-acoustic parameters (other than other parameters) could be maintained over a longer period of time in the voice disguise task;
b) Include prosodic-acoustic metrics (largely) studied in the phonetic literature for the study of (L2) speech rhythm;
c) In the forensic domain, check the consistent prosodic parameters in voice lineups for speaker identification (in progress);
d) Apply prosodic-based models in the development of automatic foreign accent speech recognition systems (in progress);
e) Apply the robust acoustic models in L2 pronunciation teaching tools (in progress).

Acknowledgments

This study was supported by the National Council for Scientific and Technological Development – CNPq, grant 307010/2022-8 for the first author, and grant 302194/2019-3 for the second author. The authors also acknowledge the reviewers for

critically reading the manuscript indicating significant specific suggestions for its improvement. Last but not least, the authors thank the participants of this research for kindly collaborating with valuable contributions.

References

Adams, C. (1979). *English Speech Rhythm and the Foreign Learner*. De Gruyter Mouton. https://doi.org/10.1515/9783110879247

Adams, C., & Munro, R. (1978). In search of the acoustic correlates of stress: Fundamental frequency, amplitude and duration in the connected utterance of some native and non-native speakers of English. *Phonetica*, *35*(3), 125–156. https://doi.org/10.1159/000259926

Alcaraz, J. (2023). The long-term average spectrum in forensic phonetics: From collation to discrimination of speakers. *Estudios de Fonética Experimental*, *32*, 87–110. https://doi.org/10.1344/efe-2023-32-87-110

Amrhein, V., & Greenland, S. (2022). Rewriting results in the language of compatibility. *Trends in Ecology & Evolution*, *37*(7), 567–568. https://doi.org/10.1016/j.tree.2022.02.001

Andrews, E. (2019). Cognitive Neuroscience and Multilingualism. In J. W. Schwieter, & M. Paradis (Eds.), *The Handbook of the Neuroscience of Multilingualism* (pp. 19–47). Wiley Blackwell. https://doi.org/10.1002/9781119387725.ch2

Barbosa, P. (2006). *Incursões em torno do ritmo da fala*. Pontes Editora.

Barbosa, P. (2020). *ProsodyDescriptorExtrator* (Version 2.0) [Praat script]. GitHub. https://github.com/pabarbosa/prosody-scripts

Best, C. (1995). A direct realist view of cross-language speech perception. In W. Strange (Ed), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 171–204). New York Press.

Best, C., & Tyler, M. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-S. Bohn, & M. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). John Benjamins. https://doi.org/10.1075/lllt.17

Boersma, P., & Weenink, D. (1992–2021). Praat: Doing phonetics by computer (Version 6.1.38) [Computer program]. https://www.praat.org/

Brescancini, C., & Gonçalves, C. (2020). O peso da evidência sociofonética na perícia da comparação de locutor. In P. Barbosa (Ed.), *Análise fonético-forense: Em tarefa de comparação de locutor* (pp. 67–87). Millennium Editora.

Brouwer, S. (2019). The role of foreign accent and short-term exposure in speech-in-speech recognition. *Attention, Perception, & Psychophysics*, *81*, 2053–2062. https://doi.org/10.3758/s13414-019-01767-8

Busch, M., & Turner, J. (1993). Using Likert scales in L2 research. *TESOL Quarterly*, *27*(4), 736–739. https://doi.org/10.2307/3587408

Carroll, D. W. (1994). *Psychology of language* (2nd ed.). Brooks/Cole Pub.

Chin, W. W. (1998). Issues and opinion on structural equation modeling. *MIS Quarterly*, *22*(1), vii–xvi. http://www.jstor.org/stable/249674

Clark, J., & Foulkes, P. (2007). Identification of voices in electronically disguised speech. *The International Journal of Speech, Language and the Law*, *14*(2), 195–221. https://doi.org/10.1558/ijsll.v14i2.195

Cohen, J. (1992). A Power Primer. *Psychological Bulletin*, *112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

Costa, A. (2017). *El cerebro bilingüe: La neurociencia del lenguaje*. Penguin Randon House.

Das, A., Zao, G., Levis, J., Chukharev-Hudilainen, E., & Gutierrez-Osuna, R. (2020). Understanding the effect of voice quality and accent on talker similarity. In H. Meng, B. Xu, & T. Zheng (Eds.), *Proceedings of Interspeech 2020*, 1763–1767. International Speech Communication Association. https://doi.org/10.21437/Interspeech.2020-2910

De Marco, A. (2020). Teaching the prosody of emotive communication in a second language.

In C. Savvidou (Ed.), *Second language acquisition: Pedagogies, practices and perspectives* (pp. 1–18). IntechOpen. http://doi.org/10.5772/intechopen.87210

Derwing, T., & Munro, M. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, *42*(4), 476–490. https://doi.org/10.1017/S026144480800551X

Duarte, M., & Silva Jr., L. (2020). A oclusiva glotal e outros gestos laríngeos na produção de falantes de inglês como L1 e L2. *Palimpsesto*, *19*(34), 241–265. https://doi.org/10.12957/palimpsesto.2020.54117

Endres, W., Bambach, W., & Flösser, G. (1971). Voice spectrograms as a function of age, voice disguise and voice imitation. *The Journal of the Acoustical Society of America*, *49*(6B), 1842–1848. https://doi.org/10.1121/1.1912589

ENFSI [European Network of Forensic Science Institutes]. (2021). *Best practice manual for the methodology of forensic speaker comparison* (ENFSI-BPM-FSC-01, vs. 01). https://enfsi.eu/wp-content/uploads/2021/07/2021-07-07-final-draft-BPM-SPEAKER-COMPARISON.pdf.

Engelbert, A. P. P. F. (2014). Cross-linguistic effects on voice quality: A study on Brazilians' production of Portuguese and English. *Concordia Working Papers in Applied Linguistics*, *5 [Proceedings of the International Symposium on the Acquisition of Second Language Speech]*, 157–170. URL: http://doe.concordia.ca/copal/documents/13_Engelbert_Vol5.pdf

Eriksson, A. (2005, September 4-8). *Tutorial on forensic speech science. Part I: Forensic phonetics* [Conference presentation]. Interspeech 2005 – Eurospeech (9th European Conference on Speech Communication and Technology), Lisbon, Portugal. International Speech Communication Association.

Eriksson, A. (2010). The disguised voice: Imitating accents or speech styles and impersonating individuals. In C. Llamas, & D. Watt (Eds.), *Language and identities* (pp. 86–96). Edinburgh University Press. https://doi.org/10.1515/9780748635788

Eriksson, A., & Wretling, P. (1997). How flexible is the human voice? A case study of mimicry. In G. Kokkinakis (Ed.), *Proceedings of the 5th European Conference on Speech Communication and Technology (Eurospeech 1997)* (pp. 1043–1046). International Speech Communication Association. https://doi.org/10.21437/Eurospeech.1997-363

Esling, J., Moisik, S., Benner, A., & Crevier-Buchman, L. (2019). *Voice quality: The laryngeal articulator model*. Cambridge University Press. https://doi.org/10.1017/9781108696555

Farrús, M. (2018). Voice Disguise in Automatic Speaker Recognition. *ACM Computing Surveys*, *51*(4), Article 68. https://doi.org/10.1145/3195832

Farrús, M., Wagner, M., Anguita, J., & Hernando, J. (2008). Robustness of prosodic features to voice imitation. In D. Burnham (Ed.), *Proceedings of Interspeech 2008* (pp. 613–616). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2008-196

Fernández-Trinidad, M. (2022). Hacia la aplicabilidad de la cualidad de la voz en fonética judicial. *Loquens*, *9*(1–2), Article e093. https://doi.org/10.3989/loquens.2022.e093

Fletcher, J. (2010). The prosody of speech timing and rhythm. In W. Hardcastle, J. Laver, & F. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (2nd ed., pp. 523–602). Wiley Blackwell. https://doi.org/10.1002/9781444317251.ch1

Foulkes, P., Scobbie, J., & Watt, D. (2010). Sociophonetics. In W. Hardcastle, J. Laver, & F. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (2nd ed, pp. 703–759). Blackwell. https://doi.org/10.1002/9781444317251.ch19

Fraile, R., & Godino-Llorente, J. (2014). Cepstral peak prominence: A comprehensive analysis. *Biomedical Signal Processing and Control*, *14*, 42–54. https://doi.org/10.1016/j.bspc.2014.07.001

Fuchs, R. (2016). *Speech rhythm in varieties of English: Evidence from educated Indian English and British English*. Springer Science Business Media. https://doi.org/10.1007/978-3-662-47818-9

Garellek, M., & Keating, P. (2011). The acoustic consequences of phonation and tone interactions in Jalapa Mazatec. *Journal of the International Phonetic Association*, *41*(2), 185–205. https://doi.org/10.1017/S0025100311000193

Garrett, R. (2013). *Cepstral- and spectral- based acoustic measures of normal voices* [Master dissertation, The University of Wisconsin]. UWM Digital Commons. https://dc.uwm.edu/etd/217

Golestani, N., & Pallier, C (2007). Anatomical correlates of foreign speech sound production. *Cerebral Cortex*, *17*(4), 929–934. https://doi.org/10.1093/cercor/bhl003

Gonzales, A. R., Ishihara, S., & Tsurutani, C. (2013). Perception modeling of native and foreign-accented Japanese speech based on prosodic features of pitch accent [Meeting abstract]. *The Journal of the Acoustical Society of America*, 133(5_Supplement), 3572. https://doi.org/10.1121/1.4806547

Great Big Story (2018, August 2). *Meet the epic voice behind movie trailers* [Video file]. https://www.youtube.com/watch?v=6N5l0sgPP5k

Greenland, S. (2019). Valid p-values behave exactly as they should: some misleading criticisms of p-values and their resolution with s-values. *The American Statistician*, *73*(Sup1), 106–114. https://doi.org/10.1080/00031305.2018.1529625

Grosjean, F., & Li, P. (2013). *The Psycholinguistics of Bilingualism*. Wiley Blackwell.

Gut, U. (2012). Rhythm in L2 speech. *Speech and Language Technology*, *14–15*, 83–94.

Harrington, J. (2010). *The Phonetic Analysis of Speech Corpora*. Wiley Blackwell.

He, L. (2017). *Speaker idiosyncratic intensity variability in the speech signal* [Doctoral dissertation, University of Zurich]. Green Open Access. https://doi.org/10.5167/uzh-145283

He, L., Ulrike, G., & Volker, D. (2015). Comparisons of speaker recognition strengths using suprasegmental duration and intensity variability: An artificial neural networks approach. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences* (Article 395). The International Phonetic Association.

Heldner, M. (2001). Spectral Emphasis as an Additional Source of Information in Accent Detection. *Proceedings of ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding* (Paper 10). International Speech Communication Association.

Henseler, J., Ringle, C., & Sinkovics, R. (2009). The use of partial least squares path modeling in international marketing. In R. Sinkovics, & P. Ghauri (Eds.), *New challenges to international marketing* (pp. 277–320). Emerald Publishing. https://doi.org/10.1108/S1474-7979(2009)0000020014

Hernandez, A. (2012) *The Bilingual Brain*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199828111.001.0001

Hollien, H., & Majewksi, W. (1977). Speaker identification by long-term spectra under normal and distorted speech conditions. *The Journal of the Acoustical Society of America*, *62*(4), 975–980. https://doi.org/10.1121/1.381592

Hughes, V., & Wormald, J. (2017) Assessing typicality in forensic voice comparison: How can sociophonetics help? (and how can sociophonetics benefit?) [Conference presentation]. Innovative Methods in Sociophonetics II Workshop, 4th Workshop on Sound Change, Edinburgh, United Kingdom.

Idemaru, K., Wei, P., & Gubbins, L. (2018). Acoustic sources of accent in second language Japanese speech. *Language and Speech*, *62*(2), 333–357. https://doi.org/10.1177/0023830918773118

Jackson, C. N., & O'Brien, M. G. (2011). The interaction between prosody and meaning in second language speech production. *Die Unterrichtspraxis / Teaching German*, *44*(1), 1–11. https://doi.org/10.1111/j.1756-1221.2011.00087.x

Järvinen, K. (2017). *Voice characteristics in speaking a foreign language: A study of voice in Finnish and English as L1 and L2* (Acta Electronica Universitatis Tamperensis; No. 1776). Tampere University

Press. http://urn.fi/URN:ISBN:978-952-03-0424-9

Järvinen, K., & Laukkanen, A.-M. (2015). Vocal loading in speaking a foreign language. *Folia Phoniatrica et Logopaedica*, *67*(1), 1–7. https://doi.org/10.1159/000381183

Järvinen, K., Laukkanen, A.-M., & Geneid, A. (2017). Voice quality in native and foreign languages investigated by inverse filtering and perceptual analyses. *Journal of Voice*, *31*(2), 260–267. https://doi.org/10.1016/j.jvoice.2016.05.003

Kapolowicz, M. R., Montazeri, V., & Assmann, P. F. (2016). The role of spectral resolution in foreign-accented speech perception. In N. Morgan (Ed.), *Proceedings of Interspeech 2016*, 3289–3293. International Speech Communication Association. http://doi.org/10.21437/Interspeech.2016-1585

Kaur, H., & Kaur, R. (2021). Identification and comparison of disguised voices with the genuine voices under various circumstances using spectrographically analysis: A review study. *International Journal of Advanced Trends in Computer Applications (IJATCA)*, *8*(1), 54–58.

Keating, P., & Esposito, C. (2007). Linguistic Voice Quality. *UCLA Working Papers in Phonetics*, *105*, 85–91.

Keating, P., Garellek, M., & Kreiman, J. (2015). Acoustic properties of different kinds of creaky voice. In The Scottish Consortium for ICPhS 2015 (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences* (Article 821). The International Phonetic Association.

Krivokapic, J. (2012). Prosodic planning in speech production. In S. Fuchs, M. Weihrich, D. Pape, & P. Perrier (Eds.), *Speech planning and dynamics* (pp. 157–190). Peter Lang. https://doi.org/10.3726/978-3-653-01438-9

Ladefoged, P., & Maddieson, I. (2008). *The Sounds of the World's Languages*. Wiley Blackwell. (Original work published 1996).

Laver, J. (1980). *The phonetic description of voice quality*. Cambridge University Press.

Laver, J. (1994). *Principles of Phonetics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139166621

Leemann, A., & Kolly, M. (2015). Speaker-invariant suprasegmental temporal features in normal and disguised speech. *Speech Communication*, *75*, 97–122. https://doi.org/10.1016/j.specom.2015.10.002

Levis, J. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press. https://doi.org/10.1017/9781108241564

Loukina, A., Kochanski, G., Shih, C., Keane, E., & Watson, I. (2009). Rhythm measures with language-independent segmentation. In R. Moore (Ed.), *Proceedings of the Interspeech 2009*, 1531–1534. International Speech Communication Association. https://doi.org/10.21437/Interspeech.2009-464

Magen, H. (1998). The perception of foreign-accented speech. *Journal of Phonetics*, *26*, 381–400. https://doi.org/10.1006/jpho.1998.0081

Maryn, Y. (2010). *Acoustic measurement of overall voice quality in sustained vowels and continuous speech* [Doctoral dissertation, Ghent University]. Ghent University. http://hdl.handle.net/1854/LU-888156

Masthoff, H. (1996). A report on a voice disguise experiment. *The International Journal of Speech, Language and the Law*, *3*(1), 160–167. https://doi.org/10.1558/ijsll.v3i1.160

McCullough, E. (2013). *Acoustic correlates of perceived foreign accent in non-native English* (Document No. osu1374052897) [Doctoral dissertation, The Ohio State University]. OhioLINK and Electronic Theses & Dissertations Center. http://rave.ohiolink.edu/etdc/view?acc_num=osu1374052897

Mennen, I., Schaeffer, F., & Docherty, G. (2011). Cross-language difference in f0 range: A comparative study of English and German. *The Journal of the Acoustical Society of America*, *131*(3), 2249–2260. https://doi.org/10.1121/1.3681950

Modesto, F. (2019). *Acoustic analysis of lexical stress in English by Brazilian Portuguese speakers, and inferences of production and perception* [Master dissertation, Universidade Estadual de Campinas]. Repositório da produção científica e intelectual da UNICAMP.

https://doi.org/10.47749/T/UNICAMP.2019.1083121

Moksony, F. (1999). Small is beautiful. The use and interpretation of R^2 in social research. *Szociológiai Szemle, Special issue*, 130–138.

Munro, M. (1999). The role of speaking rate in the perception of L2 speech. *The Journal of the Acoustical Society of America* [Meeting abstract], *105*, 1032. https://doi.org/10.1121/1.424931

Munro, M. (2003). A primer on accent discrimination in the Canadian Context. *TESL Canadian Journal*, *20*(2), 38–51. https://doi.org/10.18806/tesl.v20i2.947

Munro, M. (2018). Dimensions of pronunciation. In O. Kang, R., Thomson, & J. Murphy. *The Routledge Handbook of Contemporary English Pronunciation* (pp. 413–431). Routledge.

Munro, M., & Derwing, T. (1995). Foreign accent, comprehensibility and intelligibility in the speech of second language learners. *Language Learning*, *45*(1), 73–97. https://doi.org/10.1111/j.1467-1770.1995.tb00963.x

Munro, M., & Derwing, T. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech. *Studies in Second Language Acquisition*, *23*(4), 451–468. https://doi.org/10.1017/S0272263101004016

Munro, M., & Derwing, T. (2020). Foreign accent, comprehensibility and intelligibility, redux. *Journal of Second Language Pronunciation*, *6*(3), 283–309. https://doi.org/10.1075/jslp.20038.mun

Neuhauser, S. (2008). Voice disguise using a foreign accent: Phonetic and linguistic variation. *The International Journal of Speech, Language and the Law*, *15*(2), 131–159. https://doi.org/10.1558/ijsll.v15i2.131

Neuhauser, S. (2011). Variation of glottal activity in French accent imitation produced by native Germans. *The International Journal of Speech, Language and the Law*, *18*(2), 207–231. https://doi.org/10.1558/ijsll.v18i2.207

Neuhauser, S., & Simpson, A. P. (2007). Imitated or authentic? Listeners' judgments of foreign accents. In W. Barry (Ed.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1805–1808). International Phonetic Association.

Niebuhr, O., Skarnitzl, R., & Tylečková, L. (2018). The acoustic fingerprint of a charismatic voice: Initial evidence from correlations between long-term spectral features and listener ratings. In K. Klessa, J. Bachan, A. Wagner, M. Karpiński, & D. Śledziński (Eds.), *Proceedings of Speech Prosody 2018* (pp. 359–363). International Speech Communication Association. https://doi.org/10.21437/SpeechProsody.2018-73

Nolan, F. (2009). *The phonetic bases of speaker recognition*. Cambridge University Press. (Original work published 1983)

Nooteboom, S. (1997). The prosody of speech: Melody and rhythm. In W. Hardcastle and J. Laver (Eds.), *The Handbook of Phonetic Sciences* (1st ed., pp. 640–673). Wiley-Blackwell.

Ortega-Llebaria, M., Silva Jr., L., & Nagao, J. (2023). Macro and micro-rhythm in L2 English: Exploration and refinement of measures. In R. Skarnitzl, & J. Volín (Eds.), *Proceedings of the 20th International Congress of Phonetic Sciences* (pp. 1582–1586). Guarant International.

Pellegrino, E., He, L., & Dellwo, V. (2021). Age-related rhythmic variations: The role of syllable intensity variability. *Travaux Neuchâtelois de Linguistique*, *74*, 167–185. https://doi.org/10.26034/tranel.2021.2924

Perrot, P., Aversano, G., & Chollet, G., (2007). Voice disguise and automatic detection: Review and perspectives. In Y. Stylianou, M. Faúndez-Zanuy, & A. Esposito (Eds.), *Progress in Nonlinear Speech Processing*. (pp. 101–117). Springer. https://doi.org/10.1007/978-3-540-71505-4_7

Procter, T. (2019). *Acoustic analysis of the voice in native and non-native English speakers* [Master dissertation, University of Houston]. Electronic Theses and Dissertations. https://hdl.handle.net/10657/4651

Purpura, J. (2009). *The Oxford Online Placement Test: What does it measure and how?* Oxford University Press. https://www.oxfordenglishtesting.com/

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(1), 265–292. https://www.sciencedirect.com/science/article/pii/S0010027700001013

Reed, M., & Michaud, C. (2015). Intonation in research and practice: The importance of metacognition. In M. Reed, & J. Levis (Eds.), *The Handbook of English Pronunciation* (pp. 454–470). Wiley. https://doi.org/10.1002/9781118346952.ch25

Rojczyk, A. (2010). *Temporal and spectral parameters in perception of the voicing contrast in English and Polish*. Wydawnictwo Uniwersytetu Śląskiego.

Rojczyk. A. (2015). Using FL Accent imitation in L1 in foreign-language speech research. In E. Waniek-Klimczak, & M. Pawlak (Eds.). *Teaching and researching the pronunciation of English* (pp. 223–233). Springer. https://doi.org/10.1007/978-3-319-11092-9_14

Rose, P. (2002). *Forensic Speaker Identification*. Taylor & Francis. https://doi.org/10.1201/9780203166369

San Segundo, E. (2014). *Forensic speaker comparison of Spanish twins and non-twin siblings: A phonetic-acoustic analysis of formant trajectories in vocalic sequences, glottal source parameters and cepstral characteristics* [Doctoral dissertation, Universidad Internacional Menéndez Pelayo & Centro Superior de Investigaciones Científicas]. Biblioteca Virtual Miguel de Cervantes. https://www.cervantesvirtual.com/nd/ark:/59851/bmcm9293

San Segundo, E. (2021). International survey on voice quality: Forensic practitioners versus voice therapists. *Estudios de Fonética Experimental*, *30*, 9–34. https://raco.cat/index.php/EFE/article/view/396210

San Segundo, E., Univaso, P., & Gurlekian, J. (2019). Sistema multiparamétrico para la comparación forense de hablantes. *Estudios de Fonética Experimental*, *28*, 13–45. https://raco.cat/index.php/EFE/article/view/365395

Schiller, N., & Koster, O. (1996). Evaluation of a foreign speaker in forensic phonetics: A report. *The International Journal of Speech, Language and the Law*, *3*(1), 176–185. https://doi.org/10.1558/ijsll.v3i1.176

Silva Jr., L. (2023). BeatMaker: A computational system for foreign language pronunciation teaching based on speech prosody. *Revista Novas Tecnologias Na Educação*, *21*(1), 341–352. https://doi.org/10.22456/1679-1916.134363

Silva, C., & Arantes, P. (2021). Quantitative analysis of fundamental frequency in Spanish (L2) and Brazilian Portuguese (L1): Evidence of learning and language attrition. *Journal of Speech Sciences*, *10*, Article e021003. https://doi.org/10.20396/joss.v10i00.15779

Silva Jr., L., & Barbosa, P. (2019). Speech rhythm of English as L2: An investigation of prosodic variables on the production of Brazilian Portuguese speakers. *Journal of Speech Sciences*, *8*(2), 37–57. https://doi.org/10.20396/joss.v8i2.14996

Silva Jr., L., & Barbosa, P. (2021). Efeitos da prosódia de L2 no ensino de pronúncia e na comunicação oral. *Prolíngua*, *16*(1), 126–141. https://doi.org/10.22478/ufpb.1983-9979.2021v16n1.58725

Silva Jr., L., & Barbosa, P. (2022). Foreign accent and L2 speech rhythm of English: A pilot study based on metric and prosodic parameters. *Anais do Congresso Brasileiro de Prosódia*, *2*, 41–50.

Southwood, M., & Flege, J. (1999). Scaling foreign accent: Direct magnitude estimation versus interval scaling. *Clinical Linguistics & Phonetics*, *13*(5), 335–349. http://doi.org/0.1080/026992099299013

Teixeira, L., & Lima Jr., R. (2021). Análise do desenvolvimento do ritmo do inglês-L2 por brasileiros por meio de três métricas rítmicas. *Revista X*, *16*(5), 1258–1292. http://doi.org/10.5380/rvx.v16i5.81413

Thomas, E. (2011). *Sociophonetics: An introduction*. Palgrave Macmillan.

Tjaden, K., Sussman, J., Liu, G., & Wilding, G. (2010). Long-Term Average Spectral (LTAS) measures of dysarthria and their relationship to perceived severity. *Journal of Medical Speech Language Pathology*, *18*(4), 125–132.

Tremblay, A., Broersma, M., Coughlin, C., & Choi, J. (2016). Effects of the native language on the learning of fundamental frequency in second-language speech segmentation. *Frontiers in Psychology*, *29*(7), Article 985. https://doi.org/10.3389/fpsyg.2016.00985

Urbani, M. (2012). Pitch range in L1/L2 English: An analysis of F0 using LTD and linguistic measures. In M. Busà, & A. Stella (Eds.), *Methodological Perspectives on L2 Prosody* (pp. 79–83). Cooperativa Libraria Editrice Università di Padova.

Wen, Z. (2019). Working memory as language aptitude: The phonological/executive model. In Z. Wen, P. Skehan, A. Biedroń, S. Li, & R. L. Sparks (Eds.), *Language aptitude: Advancing theory, testing, research and practice* (pp. 187–214). Routledge.

Wrembel, M. (2007). The impact of voice quality resetting on the perception of a foreign accent in third language acquisition. In A. S. Rauber, M. A. Watkins, & B. O. Baptista (Eds.), *New Sounds 2007: Proceedings of the Fifth International Symposium on the Acquisition of Second Language Speech* (pp. 481–491). Federal University of Santa Catarina.

Zheng, L., Li, J., Sun, M., Zhang, X., & Zheng, T. (2020). When automatic voice disguise meets automatic speaker verification. *IEEE Transactions on Information Forensics and Security*, 16, 824–837. https://doi.org/10.1109/TIFS.2020.3023818

Appendices

Appendix A. Text applied in the ‘story reading’ task.

Chunks for the Lion and the Mouse (adapted Aesop’s fable)

#	Text
1	Once when a lion, the king of the jungle, was asleep, a little mouse began running up and down on him. This soon awakened the lion, who placed his huge paw on the mouse, and opened his big jaws to swallow him.
2	—Pardon, O King! cried the little mouse. Forgive me this time. I shall never repeat it and I shall never forget your kindness. And who knows, I may be able to do you a good turn one of these days! The lion was so tickled by the idea of the mouse being able to help him that he lifted his paw and let him go.
3	Sometime later, a few hunters captured the lion, and tied him to a tree. After that they went in search of a wagon, to take him to the zoo.
4	Just then the little mouse happened to pass by. On seeing the lion’s trouble, he ran up to him and bit away the ropes that bound him, the king of the jungle. —Was I not right? said the little mouse, very happy to help the lion.

Appendix B. Transcript of the interview.

Chunks for Meet the Epic Voice Behind Movie Trailers

#	Interview transcript
1	—My name is Redd Pepper, I'm a voiceover artist. In the UK, I voice hundreds of movie trailers, "Men in Black," saving the Earth from the scum of the universe.
2	"Blair Witch Project," "Armageddon," "Space Jam," "Mr. Bean's Holiday," so many, I forget half of them, to be honest with you. I started doing television adverts, animations, audio books, I do a lot of video games as well and a lot of them are sound effects, a goose in the background. Mr. Bean.
3	When I first started doing movie trailers, that was fun: one man, coming soon to a cinema near you. Sometimes I do romantic movies . . . in a sleepy town; sometimes I'm doing horror . . . don't answer the door!
4	You've got to use your voice; you've got to raise it sometimes; and you've gotta take it to the depths. Very occasionally I get recognized, but generally, no, but the time I do get recognized, the phone goes off on the train, guaranteed.
5	Hello, and people look up from behind the newspapers. The way I got into being a voice artist was kind of strange, I used to drive trains on the London Underground; one morning I was making my announcements, all stations to Harrow, mind the doors.
6	A television executive was a passenger on my train; he got off at the next stop, ran up to my cab, we exchanged details; and the rest is history.
7	I've had some strange experiences as a voice artist, I was doing a trailer for "Jurassic Park, The Lost World," Steven Spielberg movie and, kindly, they chose me to do the voice in the UK.
8	Something is coming, something big! And as I said that, a voice in my headphones said, "Wow, that's a great voice!"
9	And I didn't recognize it was Steven Spielberg, he was listening in from the States into London. Well, I swore, I said, who the (expletive) is that? Everybody went crazy in the studio, "Shh, shh, no, it's Spielberg on the set!" I appreciate what I do;
10	I'm still meticulous about what I do; I'm still proud of what I do. I really don't look at it as a job. I'm having fun. It really is a cool job; it's gotta be up there with one of the coolest jobs on the planet. Right, you got it? Cool, I'm out of here. Oh, that's a wrap.

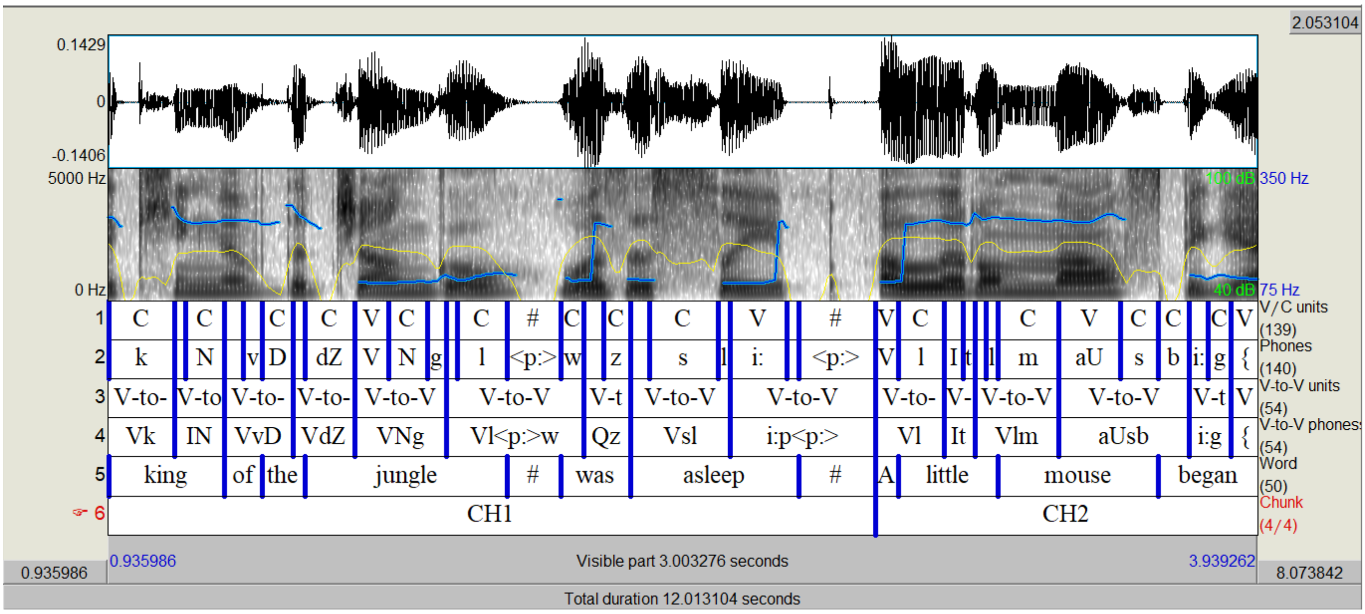
Appendix C. Table for all of the prosodic-acoustic parameters used in this research.

[Acoustic correlate] Class of the acoustic parameters. [Acoustic parameter] Total of the prosodic-acoustic features used in this research extracted by the algorithm 'ProsodyDescriptorExtractor' (LTAS = Long-Term Average Spectrum). [Unit] Unit of measurement of each related parameter (st = semitones; σ/s = syllables per second; σ/(s–pauses) = syllables per [second minus pauses]). [Description] Brief description of the parameter's function.

Acoustic correlate	Acoustic parameter	Unit	Description
F0	Minimum	st	F0 minimum value.
	Semi-amplitude between quartiles	st	Calculates the non-parametric F0 variability interquartile.
	SD	st	Calculates the F0 variability value.
	Negative slope	st	Calculates the Downward F0 value.
	positive slope SD	st	Calculates the upward F0 variability value.
	negative slope SD	st	Calculates the downward F0 variability value.
	total slope SD	st	Calculates the upward and downward F0 variability.
	peaks SD	st	Calculates the F0 peak variability.
Duration	Speech rate	σ/s	Calculates the velocity of speech production.
	Articulation rate	σ/(s–pau)	Calculates the velocity of articulatory production.
Intensity	Variation coefficient	%	Calculates the low-high-related intensity change during speech production.
	Spectral emphasis	dB	Calculates the vocal effort.
Voice quality	HNR	dB	Harmonics-to-Noise Ratio. Calculates the relation between the amount of noise and harmonics produced in speech.
	LTAS slope: 0–1:1–4 kHz	Hz/bin	A composite signal representing the spectrum of the glottal source and the resonant characteristics of the vocal tract for the detection of breathy, creaky or laryngealized voices (0–1 kHz and 1–4 kHz). Highly correlated to vocal effort.
	LTAS slope: 0–1:4–8 kHz	Hz/bin	A composite signal representing the spectrum of the glottal source and the resonant characteristics of the vocal tract for the detection of the degree of breathiness of the speech signal (4–8 kHz). Highly correlated to vocal effort.
	Jitter (local)	%	Calculates the sound wave amplitude irregularity.
	Shimmer (local)	%	Calculates the vocal cycle irregularity.
	H1–H2	dB	Calculates the difference between the first (the F0), and second (first multiple of the F0) harmonics. It is highly correlated with the degree of glottal constriction for voice quality determination.
	CPP	dB	Cepstral Peak Prominence. A measure that distinguishes breathy voice from other voice qualities such as hoarse, creaky or modal.

Appendix D. Praat's screen containing the segmentation arrangement used in the present research.

Partial waveform, broadband spectrogram with F0 (blue) and Intensity (yellow) contours, and six tiers respectively segmented and labeled as: 1) vocalic (V), consonantal (C), and pause (#) units; 2) vocalic and consonantal phones; 3) onset-to-onset units of vowels (V-to-V); 4) V-to-V phones; 5) Words of two different speech chunks; 6) higher level units (chunks – CH) produced by a female native speaker of English. Chunk 1: "(…) *king of the jungle was asleep*." Chunk 2: "*a little mouse began* (…)".


2.053104
0.1429
0
-0.1406
5000 Hz
0 Hz
100 dB
40 dB
350 Hz
75 Hz
1 C C C C V C C # C C C V # V C C V C C C V
2 k N v D dZ V N g l <p:> w z s l i: <p:> V l I t l m aU s b i: g {
3 V-to- V-to V-to- V-to- V-to-V V-to-V V-t V-to-V V-to-V V-to- V- V-to-V V-to-V V-t V
4 Vk IN VvD VdZ VNg Vl<p:>w Qz Vsl i:p<p:> Vl It Vlm aUsb i:g {
5 king of the jungle # was asleep # A little mouse began
6 CH1 CH2
V/C units (139)
Phones (140)
V-to-V units (54)
V-to-V phones (54)
Word (50)
Chunk (4/4)
0.935986
0.935986
Visible part 3.003276 seconds
3.939262
8.073842
Total duration 12.013104 seconds

Appendix E. Functional magnetic resonance imaging (fMRI) for the brain areas used for (non)skilled L2 speakers when producing the L2 target accent.

fMRI results of the meta-analysis in the brain cortex areas for highly competent (panels a, b and c) and less competent (panels d, e and f) L2 speakers. *Z* scores (z) measures the amount of the brain activity area. The blue line intercepts indicate higher brain signal activation. It should be noted that, by convention, the right part of each brain image corresponds to the left hemisphere (Costa, 2017, p. 174).