BeatMaker: a computational system for foreign language pronunciation teaching based on speech prosody

Leônidas José da Silva Jr., Universidade Estadual da Paraíba (UEPB),
<leonidas.silvajr@gmail.com>
ORCID: https://orcid.org/0000-0002-3728-9851

Abstract:
This present research aims at presenting a system referred to as "BeatMaker" - a multilingual prosody-based program - to help researchers and teachers on foreign language pronunciation classes. For the purpose of testing the effectiveness of BeatMaker system, we conducted a pretest-posttest experiment where it was analyzed prosodic aspects, such as rhythm and intonation from the productions of ten Brazilians speakers of English as a foreign language in comparison to eight English native speakers. Data were collected once by the native group, and before and after the training with BeatMaker by the foreign-English speakers. Results showed a significant improvement for the intonational aspects after training, that is, the foreign-language group converged towards the native group in terms of intonation, nevertheless there was no significant correlation for the rhythmic aspects between groups. It is concluded on a preliminary basis that the use of technologies for teaching foreign-language prosody were of great help and practices involving these systems shall be considered for the sake of the importance of foreign-language prosody in oral communication.
Keywords: BeatMaker system. foreign language prosody. pronunciation teaching technologies.

BeatMaker: um sistema computacional de ensino de pronúncia de língua estrangeira baseado na prosódia da fala

Resumo:
A presente pesquisa tem como objetivo apresentar o sistema "BeatMaker" - um programa multilíngue e baseado na prosódia da fala - para auxiliar pesquisadores e professores em aulas de pronúncia de língua estrangeira. Com o propósito de testar a eficácia do sistema BeatMaker, realizamos um experimento pré-teste/pós-teste em que foram analisados ritmo e entonação das produções de dez brasileiros falantes de inglês como língua estrangeira em comparação com oito falantes nativos de inglês. Os dados foram coletados apenas uma vez pelo grupo de nativos, e antes e após o treinamento com o BeatMaker pelos falantes de inglês como língua estrangeira. Os resultados apontaram uma melhora significativa dos aspectos entoacionais após o treinamento, ou seja, o grupo de língua estrangeira convergiu em direção à prosódia entoacional do grupo nativo, no entanto não houve correlação significativa para os aspectos rítmicos entre os grupos. Conclui-se, preliminarmente, que o uso de tecnologias para o ensino da prosódia em língua estrangeira foi de grande ajuda, e práticas envolvendo esses sistemas devem ser consideradas em função da importância da prosódia da língua estrangeira na comunicação oral.
Palavras-chave: sistema BeatMaker. prosódia de língua estrangeira. tecnologias para o ensino de pronúncia.

1 Introduction

According to Derwing and Munro (2015), the use of speech technology by means of visualization and production in the pronunciation teaching context was built initially

to represent acoustic properties of speech on a printed page (in the 1980s). At that time, it was expected that visual depictions of speech might help learners to match their pronunciation in accordance to a native-speaker model. Speech technology has advanced to a certain point that waveforms and spectrograms can be generated by any user on nearly any kind of computer platform. Software and systems for these purposes became inexpensive or freely-available, such as Praat (Boersma and Weenink, 2021). Praat is a powerful tool capable to make phonetic analysis in different domains, such as in the segment level, in the prosodic level, in the statistical level among others.

As for the prosody-based pronunciation systems, Pyshkin et al. (2019) address that technology referred to as Computer-Assisted Prosody Teaching (CAPT) is a recently new topic of interest of software developers working together with language teaching communities. The authors developed an acoustically-based system that graphically compares student's pitch pattern output with a pronunciation pattern of a native speaker by the use of two pitch curves. Pyshkin et al.'s (2019) research is aligned with Chun's et al. (2008) study which state that certain types of visual speech analysis are known to provide effective feedback in the teaching of foreign language (L2) prosody. The authors claim that pitch information is, to a certain extent, a straightforward concept, and the display of pitch contours are relatively easy to be interpreted by learners by associating the rises and falls in visual patterns when pitch changes from one voice to another. Pedagogical effectiveness by using pitch visualization was provided by other studies (see Hardison, 2004).

As put forward and aligned to the abovementioned studies, this paper aims at presenting the "BeatMaker", a free system that helps researchers, teachers and L2 pronunciation practitioners on the work of speech comparison based on prosody. Hereon, the following research question is proposed:

- (1) Will the Brazilian speakers improve their L2 prosody after using a technology such as BeatMaker?

This article is divided into the following sections. Introduction, where it is presented some aspects about the importance of L2 prosody for communication, and some of the speech technologies based on L2 prosody, which includes BeatMaker; Theoretical framework, where it is presented studies that made effective use of several technologies in prosody-based pronunciation teaching in different contexts; Methods, where it is presented the experimental design and the participants of the present research, as well as the acoustic and statistical analyses used for comparison between the productions from the native group, and the productions before and after the use of BeatMaker from the foreign group; Results and discussion, where it is presented and discussed the pretest-posttest results based on the use of the system. Final remarks, where it is presented the summary of the research, limitations and some future directions for its continuation, as well as the references herein used.

2 Theoretical framework

In this section, it will be presented the importance of L2 prosody for language teaching and some studies that used technological systems for the teaching of L2 prosody, as well as the functionality and use of BeatMaker system.

2.1 The importance of L2 prosody in language teaching

According to Lengeris (2012), pronunciation accuracy in L2 requires mastering production of both segmental (consonants and vowels) and prosodic features of speech (features that extend to lexical/phrasal stress, pitch accent, rhythm, intonation and voice quality). Therefore, the teaching of prosody is traditionally neglected in language

classrooms. Levis (2018) addresses that the prioritization of prosodic aspects during L2 pronunciation classes is an attempt to minimize pronunciation deviations and enhance conversational intelligibility in oral communication instances. Such intelligibility might reflect a fluid L2 speaker-listener accommodation that go beyond the segment level.

In a broad sense, Fletcher (2010), and Jackson and O'Brien (2011) pose that (L2) prosody is synonymous of variations in the suprasegmental parameters of the paralinguistic domain such as duration, F0 and intensity that contribute to the production and perception of stress, rhythm and tempo, lexical tone and intonation besides voice quality. On this point of view, the utterance cannot be reduced to individual consonants and vowels, but to syllabic and higher units. The authors highlight that deviations or inadequate production of L2 prosody can lead to misunderstandings on both semantic and pragmatic domains. Such misunderstandings operate on the word, the sentence, and the discourse levels. Derwing and Rossiter (2003) yet state that if the goal of pronunciation teaching is to help students become more understandable, a stronger emphasis on prosody shall be considered.

For Chun (2022), one of the main reasons that L2 prosody has been understudied and under-taught is the sheer complexity of prosodic systems in all languages. Linguists have investigated L2 prosody and its acoustic features applied to language teaching for decades (see Adams, 1979). Gussenhoven and Chen (2020) pose that L2 teachers (and researchers) are often not as familiar with prosody, much less are acquainted of how to teach it. Chun (2022) yet addresses the importance of acquisition of L2 prosody and/or how it can be taught. It is essential to focus both on perception and production, and furthermore, in authentic communicative situations. Learners must be trained to perceive prosodic markings such as, stress, rhythm and intonation that signal meaning in authentic speech and must have opportunities to practice and produce these markings throughout the utterances.

Silva Jr. (2021) suggests that the teaching of pronunciation must adopt an integrated approach that prioritizes L2 prosody such as, the inclusion of acoustic features related to rhythm, intonation and voice quality. Celce-Murcia, et al. (2010) point out that learners from a syllable-based language background will present, at least to some extent, difficulties in assigning greater length to the stressed syllables of content words within the sentence or discourse level. The authors also emphasize L2 prosody teaching pointing out that taking classes on stress-based rhythm helps to improve L2-English speech fluency of learners whose L1 is syllable-based promoting reduction on foreign accent degree, and consequently, enhancing communication.

In order to help L2 learners acquire L2 prosody, Chun and Levis (2020), and Chun (2022) propose different types of instruction that target awareness, perception, and production of different prosodic features based on the use of technology referred to as computer-assisted pronunciation training (CAPT) applications and programs. These technologies shall provide effectiveness of training for perception (based on subjective human ratings), production (computer-based acoustic analyses), or both.

2.2 The use of speech technology for prosodic-based pronunciation teaching

Chun et al. (2008), Derwing and Munro (2015), and Pyshkin et al. (2019) advocate that visualization of pitch contours are useful for sentence-level or discourse-level chunks of a certain language, but there are screen limitations on how much is visible at one time. In the same vein, spectrographic displays for prosodic information are easily created by most current software programs; however, they are not as easily interpretable as pitch contours by non-specialists, particularly L2 learners. Moreover, Jenkins (2007) addresses the need for empirically established phonological norms for pronunciation models in

foreign language teaching and also stresses intelligibility as the main point. In addition to specific segmental items, such as assimilation processes in connected speech, the author claims for the appropriate use of prosody, such as the direction of pitch movements to signal attitude, the location of word (and phrasal) stress, which consequently, leads to a higher performance of [± stress/syllable]-timed rhythm.

Chun et al. (2008) highlight some difficulties faced by learners to access intonation via visualizations of pitch changes in computer systems, due to technical limitations on the representations of intonation, and a lack of pedagogical input related to those visualizations. The authors yet address that a further restriction was the usual focus on sentence-level intonation for contrasting (syntactic) sentence types, such as declarative statements, *yes-no* questions, *wh*-questions, and exclamations. The study suggests that technologies should provide learners systems for visualization of their intonational patterns to help them improve semantic interpretation during their speech perception and production in the metacognitive domain (see also Pennington and Ellis, 2000; Krivokapic, 2012; Reed and Michaud, 2015; Ran et al. 2020 for similar suggestions).

In terms of methodological approaches to the use of technology for L2 prosody teaching, the phonetic literature has brought forward the use of isolated scripted sentences/phrases in developing prosodic awareness. The main advantage of this approach is that it provides the teacher or researcher with content control, the focusing of learner attention as attested by Chun et al. (2008), Krivokapic (2012), Derwing and Munro, (2015), and Pyshkin et al. (2019).

As far as online applications are concerned, Polushkina and Tareva (2021) conducted a study using Google tools (Google translator, YouTube, G-board) to teach L2 prosody due to the Covid-19 pandemic. Their findings showed that the suggested training had an important effect on L2 prosody acquisition by the students as well as it generated a more autonomously-guided way of studying these L2 prosodic aspects.

2.3 The BeatMaker system

BeatMaker is a multilingual system that aims at helping professionals devoted to L2 pronunciation teaching (teachers, instructors, practitioners and researchers) on the work of speech comparison based on prosodic elements such as fundamental frequency (F0) and duration. The program is a script for Praat software (Boersma and Weenink, 2021), and it receives two audio files to be compared (a reference audio file and a comparison audio file).

The audio files' content may be a speech chunk containing different prosodic levels[1] as either: i) a single prosodic word (ω), such as [girl]ω or [danced]ω, as in "the [GIRL]ω danced well" or "the girl [DANCED]ω well"; ii) a prosodic phrase (φ), such as [the girl]φ [danced well]φ, or even a higher speech unit, as iii) an intonational phrase (I), such as [[the girl danced well]φ [because she had practiced]φ [for months]φ]I. Along with the speech material, a plaintext file containing the linguistic information of the reference audio file should be provided. BeatMaker, then, realizes an automatic forced-alignment of the text based on the reference audio file, and returns a two-tier TextGrid file (a word and a phrase tier).

[1] For a better performance of the system, the suggested prosodic levels that compose the chunks of speech might be, i) the *prosodic words*, which according to Nespor and Vogel (2007, [1986]), represent individual words that have their own pitch and rhythm patterns within a phrase; ii) the *prosodic phrases*, which represent groups of (prosodic) words that are pronounced with a particular intonation pattern, and iii) the *intonational phrases*, which represent a grouping of prosodic words or phrases that forms a coherent unit marked by a change in pitch direction (falling or rising pitch contour) in the phrase boundary.

The phrase-level alignment is based on intensity features, and the word-level alignment is run from a built-in command in Praat. The system also returns a multi-colored plot which contains both of the F0 contours and the aligned TextGrid. From the audio files, the program extracts the delexicalized prosodic information such as, the timing (beat) and pitch and then, creates new sound files based on prosody.

BeatMaker brings up to ten different languages such as, English (U.S. and U.K.), Portuguese (Brazil and Portugal), Spanish (Latin America and Spain), French, Japanese, Russian and Interlingua. The user can also choose the speaker's gender for a more precise extraction of the F0 contours and, for didactic purposes as greatly suggested Chun et al. (2008), and Derwing and Munro (2015), the user can choose one out of six different colors for plotting the F0 contours of both audio files.

Figure 1 presents the workflow of system with input and output objects' details.

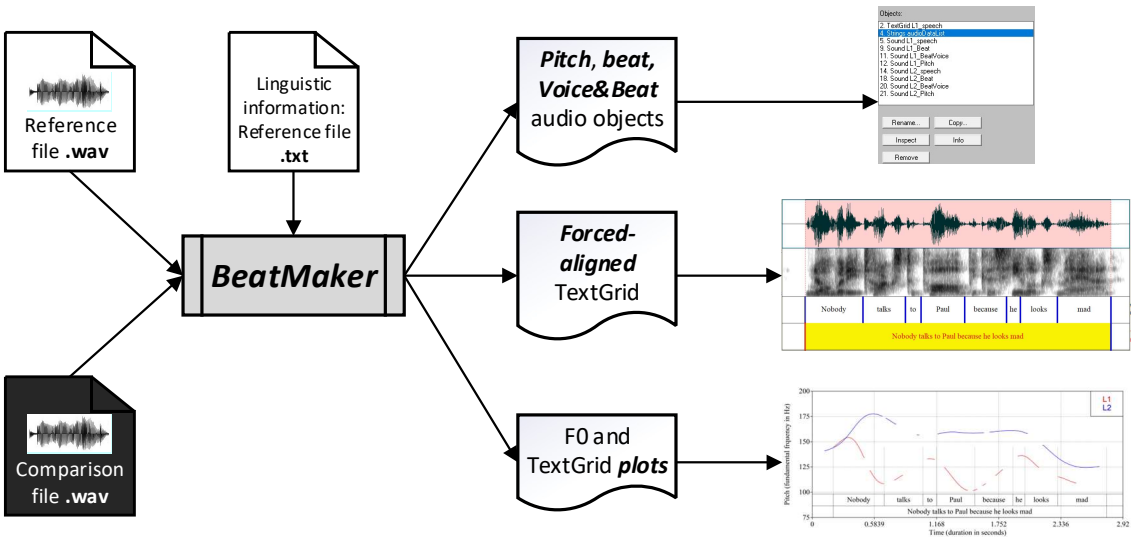

Figure 1 - The workflow of BeatMaker. Input audio files are: i) an L1-English production (reference file), and ii) an L2-English production (comparison file). Output files are screenshots of: iii) the audio objects, iv) a TextGrid object, and v) F0 contour plot and aligned TextGrid for the phrase "Nobody talks to Paul because he looks mad"
Source: personal collection

BeatMaker can be used for speech comparison based on the prosodic dimensions of stress, rhythm, intonation and voice quality as well as supporting foreign language teachers at reducing prosodic errors, such as pitch accent and stress shifts that compromise effective communication by promoting prosody audible and visualized information as suggested by Chun et al. (2008), and Munro and Derwing (2015). For a better use of BeatMaker, it is recommended that the user tags the first two characters of the file names as: "V1 - V2" ("Voice 1/2"), or "RV - CV" ("Reference/Comparison Voice"), or "L1 - L2" / "NS - FS" ("Native/Foreign Speaker). The BeatMaker system, the user's manual[2], and the audio and text files for tests can be freely downloaded from: <https://github.com/leonidasjr/ProsodyCode>.

[2] For a more detailed description of the functionality, as well as for setting up the system's input parameters, consult the user's manual from <https://github.com/leonidasjr/ProsodyCode/blob/main/BeatMaker_UserManual.pdf>.

3 Methods

For the experimental design, we adopted Hardison's (2004) protocol (cf. p. 39-40). A pretest-posttest design was used to measure the effects of two weeks of training (ten sessions of about 50 minutes each) in English prosody using visual displays of pitch contours and delexicalized audio information provided by BeatMaker. As for the description along this section, it is detailed the participants, the speech corpus used for the study, data collection before and after the use of BeatMaker, the training with the system, as well as the acoustic and statistical analyses carried out in the present research.

3.1 Participants

We collected audio data from two groups; one of L1-English speakers and the other one of L2-English speakers. Both the L1- and the L2-English groups contained participants who were 50% female/male.

The L1-English group consisted of eight graduate Americans from the United States who occupied different job positions such as, farmer, dentist, missionary, CEOs *inter alia*. All of the American speakers lived in Brazil for about two years when the experiment was run. They were also fluent in Brazilian Portuguese. The group were aged between 26 and 60 years old (mean, M = 39.4; standard deviation, SD = 14.3).

The L2-English group consisted of ten undergraduate Brazilian students from the state of Paraíba). The group had participants with ages between 19 and 25 years (M = 21.5; SD = 2.2). The group was submitted to the Oxford Online Placement Test (OOPT) for proficiency level purposes. Speakers were qualified at a transition in between B2→C1 (mean score = 73; see Pollitt, 2019, p. 9, and Polushkina and Tareva, 2021, p. 40, for details on the use of mean scores applied to proficiency level transition), according to the Common European Framework of Reference (CEFR, Council of Europe, 2001).

3.2 Speech corpus and data collection

Both groups read eight speech chunks extracted from the story "The Simple Joys of Life", available in: <https://github.com/leonidasjr/L2SpeechCorpus>. The chunks (CH) were compound of at least two syntactic clauses which contained a minimum of one pause in between the phrases for the maintenance of rhythmic and intonational patterns during speech planning (see Krivokapic, 2012; Reed and Michaud, 2015; Silva Jr. and Barbosa, 2021 for further details and applications). Speech chunks are as described in Table 1:

Table 1 - Extracted speech chunk *number* (CH#), and the linguistic information of each chunk produced by both L1- and L2-Ensglish groups.

CH#	Linguistic information for each extracted speech chunk
CH1	I want to stay at home, but I need to go to a library
CH2	He was celebrating because he was approved
CH3	I wanted to text you, but I don't have your cell phone number
CH4	I go to the mall every week, because I love window shopping
CH5	The virus cannot live in immunized individuals, nor in nature
CH6	Nobody talks to Paul because he looks mad
CH7	I always take a book to read, yet I never seem to turn a single page.
CH8	She is very old but still attractive

Source: Adapted from Silva Jr. and Barbosa (2021).

Each participant was previously shown the text, so that they could be able to be familiar with the words and syntax for inputting their prosody during recordings. The participants could read the text as much as they wanted before recording process begins. Each participant recorded the text three times before training and three times after so that

they could feel more comfortable with the arranged syntactic sequences, nevertheless it was considered only one (normalized) recording per speaker. The chunks were extracted from each participant‘s recording with the use of Praat software (Boersma and Weenink, 2021). A total of 224 chunk tokens were used for the experiment as presented in Formula 1:

Formula 1: $[(8_{chunks} * 8_{L1\ participants} = 64_{L1\ tokens}) + (8_{chunks} * 10_{L2\ participants} = 80_{L2\ tokens\ BEFORE\ training}) + (8_{chunks} * 10_{L2\ participants} = 80_{L2\ tokens\ AFTER\ training}) = \mathbf{224_{tokenss}}]$.

Data collection was performed in a quiet room from a Zoom H1 Handy PCM Recorder using a unidirectional on-board Zoom H1 microphone, at a response frequency from 30 to 16 kHz, a sampling frequency of 44.1 kHz, and a 16-bit quantization rate. Signal-to-noise ratio was higher than 30 dB to ensure greater data quality and fidelity. The described settings could guarantee a better capture of the F0 values.

3.3 Acoustic and statistical analyses

Acoustic analysis was conducted in Praat (Boersma and Weenink, 2021) and data were segmented and labeled into: i) vowel onset to the next vowel onset (V-V) units, and ii) chunk (CH) units based on Table 1. After the segmentation and the labeling processes, acoustic data of duration and F0 were extracted (the raw absolute values) and normalized into two steps.

Firstly, we performed Bark Difference Metric normalization, where duration and F0 of the same chunk for each of the three productions were normalized within-speaker. The Bark difference normalization is associated with how the auditory system processes the F0 contours of each speaker’s production, that is, the pitch per speaker as suggested by Smith et al. (2019) in a study for L2 tense and lax vowels when accounting for intra-speaker variability of both duration and F0. Secondly, we performed Lobanov‘s (1971) z-score normalization of duration and F0 values from each speech chunk between speaker’s production. According to Barbosa and Madureira (2015), Lobanov normalization is used to minimize microprosodic effects in duration, such as number of phones and perceptually-related peaks in the syllables, and in the F0 of vowels (and consonants) that do not have prosodic-linguistic function in addition to smoothing differences between female/male voices.

The normalized values were automatically extracted from the script for Praat “SpeechRhythmExtractor” (Silva Jr. and Barbosa, 2023b). This procedure was successfully used by Silva Jr. and Barbosa (2023a) for L2 prosodic analysis and comparison of foreign accent degree, which is one of BeatMaker’s functions.

As for the statistical analysis, we performed one-way Analysis of Variance (ANOVA) test statistics to check the normalized duration and F0 values controlled for the factor Group of L1 speech (L1speech), L2 speech before the use of BeatMaker (L2speech_before), and L2 speech after the use of BeatMaker (L2speech_after) and checked for the effect size from the coefficient of determination (R^2), which provides the variance explained by the group levels for duration and F0.

For group pair differences (L1speech vs. L2speech_before; L1speech vs. L2speech_after; L2speech_before vs. L2speech_after), we performed the TukeyHSD statistics that account for a pairwise mutiple comparison of the means (M) as well as the standard deviations (SD).

3.4 Training

After runing BeatMaker, the L2 group listened to each prosodic audio and were asked to mimicry the beat and melody as they were listening. Next, since one of the audio files returned by the system provides both prosodic and linguistic information simultaneously, they listented and were asked to mimicry both prosodic and linguistic information for metacognitive aprehension as suggested by Krivokapic (2012), and Reed and Michaud (2015). In the sequence, students were shown the F0 and TextGrid plots for visualization of linguistic and prosodic information. As long as they saw the plots, we played the audio showing the upwards and downwards of the F0 countors on both L1 and L2 pronuncaiation.

After the training sessions, students were asked to repeat as much as they wanted in order to feel confortable for the recordings. As mentioned in section 3, the training with BeatMaker was held in ten sessions of about 50 minutes each for the period of two weeks.

4 Results and discussion

In Table 2 and Figure 2, we present the summary of the results and the plots respectively for the duration and for the F0:

Table 2 – The mean values from the acoustic features of duration and F0, the three-level speech factor, the F-statistics, the P-value and the coefficient of determination (R^2) produced by both L1- and L2-Ensglish groups.

Acoustic features	Speech			F(2, 221)	P-value	R^2
	L1	*L2 after*	*L2 before*			
Duration	-0.47	-0.09	-0.28	3.49	=.048	.17
F0	0.87	0.96	1.02	6.01	=.008	.44

Source: personal collection

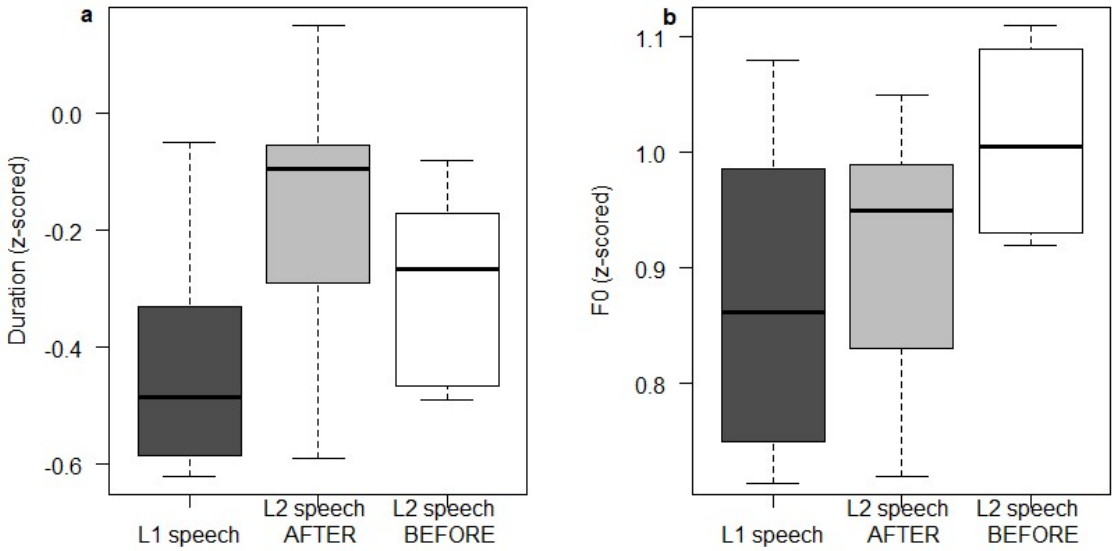

Figure 2 - Boxplots of Lobanov-normalized duration (panel 'a'), and F0 (panel 'b') for the native speakers of English (L1 speech), and the Brazilian speakers before (L2 speech BEFORE) and after (L2 speech AFTER) the prosody training session with BeatMaker.
Source: personal collection

As one may see in Figure 2a, our results showed significant differences when pointing the global results for the normalized duration before and after the use of prosody training with BeatMaker system, [$F(2, 221) = 3.49$, $p = .049$, $R^2 = .17$]. Tukey's post hoc

results pointed out to a significant difference on the productions between both L1- and L2-English levels, i.e., before (M = .137, SD = 3.2, p = .038), and after (M = .268, SD = 4.8, p = .038) training, but there is no significant difference between the L2 production levels, (M = .131, SD = -2.9, p = .416).

As for the normalized F0 values, Figure 2b presents global results with significant differences between L1 and L2 productions only before the training with BeatMaker [F(2, 221) = 6.01, p = .008, R^2 = .44]. Tukey's post hoc results pointed out to: i) a significant difference on the productions between L1 and L2 before training, (M = -0.716, SD = -1.43, p = .008), ii) an inconclusive (to some extent) difference for both L2 levels, i.e., before and after training, (M = -0.271, SD = -4.78, p = .059), and iii) there was no significant difference between the L1 and L2 productions for the after-training level, (M = -0.981, SD = -3.70, p = .632).

From the results herein presented, it is possible, at least to some extent, to infer about the importance of L2 prosody training when teaching pronunciation and, consequently, enhance oral communication. The explained variance of the F0-related results (the model explains 44% of variation) showed to be more robust in comparison to the duration-related ones (only 17%). This indicates that, from these results, it seems that the speaker retains more attention to the melodic aspects of the L2. The F0 contour visualization and its slow manipulation during the training is likely to present more promising results. Durational aspects are harder to be retain, especially when the speaker's L1 (Brazilian Portuguese in the examples here presented) is rhythmically-based on duration (see Barbosa, 2006 for a thoroughly detailed explanation). Yet, Gut (2012) asserts that duration does not seem to be well apprehended during L2 speech rhythm training or practice, for being an intrinsically-based correlate able to distort other prosodic measurements. The findings for the z-scored duration of the V-V units are aligned to Li, et al. (2018) study on Mandarin speakers of L2-English.

On the one hand, aspects of pronunciation related to rhythm (long and irregular duration of V-V units due to L2 cognitive load) and intonation (little variation in the F0 contour for the productions before training) indicate difficulties for Brazilian Portuguese speakers on the effective use of L2 prosody. This fact can compromise the speaker's intelligibility in semantic aspects (the shift of lexical/phrasal stress and pitch accent may affect prosodic boundaries towards a morphosyntactic boundary, for example) leading to a pragmatic dimension of the discourse (misunderstandings during conversational turns). On the other hand, the after-training level provided, to a certain extent, correlation between L1-L2 groups on the F0 domain. These findings are also aligned with Derwing et al. (1998) study, which provided evidence that learners who had received instruction on features such as, intonation, showed significant improvement when they produced the corpus after the training sessions.

5 Final remarks

In this work, we aimed to present "BeatMaker", a free system to help researchers, and L2 teachers and practitioners devoted to prosodic aspects and visualization during L2 pronunciation teaching. We presented the system and its features, as well as how its functionality may be worked in classroom. We also ran a pretest-posttest experiment with the use of BeatMaker, which provided to some extent, significant improvement on the L2 intonational acquisition by both listening directly to prosodic information and visualizing the related prosodic aspects.

In relation to the research question put forward in section 1: *Will the Brazilian speakers improve their L2 prosody after using a technology such as BeatMaker?*

- Yes, to a certain extent. Intonational aspects succeeded after the training sessions. One may infer that both intonational perception and visualization are promptly to help L2 learners acquiring such prosodic aspects of the target language. On the other hand, rhythmic aspects were hard to be applied once rhythm is considered to be a difficult feature to be changed in L2 prosodic production because of its intrinsically-based durational characteristics. During training or practice, one may distort other prosodic measurements when trying to impersonate the L2 rhythm. Besides, we are aware that problems exist when comparing findings across studies because of differences in tools, instructions, time for instructions, amount of used content, students' feedback, number of samples, testing and training conditions and procedures, as well as other types of limitations in knowledge or technology. Longitudinal-like studies are lacking on examining long-range effects of training. We applied ten moments of training of 50 minutes each. We are also aware that limitations extend to the domain of pedagogical applications because of the time needed to train instructors, and to determine how much training would be necessary to be applied to the learners.

5.1 Limitations and Future directions

Considering we presented a system that, besides F0 and the delexicalized information, returns automatically-aligned linguistic information, we are yet aware that effective systems to align especially foreign speech is a difficult task that poses a number of challenges. These challenges include: speech variability (intra- and inter-speaker), different voices, accents, styles, contexts, as well as speech rates), recognition units (words and phrases, syllables, phonemes, diphones and triphones), language complexity (vocabulary size and difficulty), ambiguity (homophones, word boundaries, syntactic and semantic ambiguity), and environmental conditions (e.g., background noise, several people speaking simultaneously, etc.) as described by Levis and Suvorov (2020).

In accordance to previous sections (2.1 and 3.4), some of the future directions, as an attempt to minimize prosodic problems and optimize the teaching of L2 prosodic dimension in pronunciation classes, we may suggest (which is already in progress) an integration between multiple components of speech prosody produced by both L1 and L2 speakers and the use of metacognitive strategies involving L2 stress, rhythm and intonation were used in order to obtain from learners a more accurate, intelligible and understandable pronunciation, taking into account processes that occur in the suprasegmental domain.

7 Acknowledgments

We gratefully acknowledge the grant from the National Council for Scientific and Technological Development (CNPq), (grant nº 307010/2022-8), and the participants of this research.

8 References

ADAMS, C. **English Speech Rhythm and the Foreign Learner**. The Hague: Mouton Publishers, 1979.

BARBOSA, P. **Incursões em torno do ritmo da fala**. Campinas: Ed. Pontes, 2006.

BARBOSA, P.; MADUREIRA, S. **Manual de Fonética Acústica Experimental: aplicações a dados do português**. São Paulo: Cortez, 2015.

BOERSMA, P., WEENINK, D. **Praat: doing phonetics by computer**. Available in: https://www.fon.hum.uva.nl/praat/, 2021. Accessed on Sept, 23, 2021.

CELCE-MURCIA, M.; BRINTON, D.; GOODWIN, J. **Teaching Pronunciation: A course book and reference guide**, 2 ed. New York, Cambridge University Press, 2010.

CHUN, D. M. **L2 Prosody acquisition**. In: Verbetes LBASS. <http://www.letras.ufmg.br/lbass/>, 2022.

CHUN, D. M., HARDISON, D. M., PENNINGTON, M. C. Technologies for prosody in context: Past and future of L2 research practice. In: EDWARDS, J. G., **Phonology of Second Language Acquisition**. Philadelphia: John Benjamins Publishing, p. 323-346, 2008.

CHUN, D. M.; LEVIS, J. (2020). Prosody in L2 teaching: Methodologies and effectiveness. In: GUSSENHOVEN, C.; CHEN, A. (Eds.). **The Oxford Handbook of Language Prosody**. Oxford: Oxford University Press, p. 619-630, 2020.

COUNCIL OF EUROPE. **Common European Framework of Reference for Languages: learning, teaching**. London: Assessment. Press Syndicate of the University of Cambridge, 2001.

DERWING, T., MUNRO, M. **Pronunciation Fundamentals: Evidence-based Perspectives for L2 Teaching and Research**. New York: John Benjamins Publishing, 2015.

DERWING, T.; ROSSITER, M. The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. **Applied Language Learning**, v. 13, p. 1-17, 2003.

DERWING, T.; MUNRO, M.; WIEBE, G. Evidence in favor of a broad framework for pronunciation instruction. **Language Learning**, v. 48, n. 3, p. 393–410, 1998.

FLETCHER, J. The Prosody of Speech: Timing and Rhythm. In: HARDCASTLE, W.; LAVER, J.; GIBBON, F. (Eds.). **The Handbook of Phonetic Sciences**. 2 ed. Oxford: Wiley-Blackwell, p. 523-602, 2010.

GUSSENHOVEN, C.; CHEN, A. (Eds.). **The Oxford Handbook of Language Prosody**. Oxford: Oxford University Press, 2020.

GUT, U. Rhythm in L2 speech. **Speech and Language Technology**, v. 14, n. 15, p. 83-94, 2012.

HARDISON, D. Generalization of computer-assisted prosody training: quantitative and qualitative findings. **Language Learning & Technology**, v. 8, n. 1, p. 34-52, 2004.

JACKSON, C.; O'BRIEN, M. The interaction between prosody and meaning in second language speech production. **Die Unterrichtspraxis:Teaching German**, v. 44, n. 1, p. 1-9, 2011.

JENKINS, J. **English as a Lingua Franca: Attitude and Identity**. Oxford: Oxford University Press, 2007.

KRIVOKAPIC, J. Prosodic planning in speech production. In: FUCHS, S.; WEIRICH, M.; PAPE, D.; PERRIER, P. (Eds.). **Speech Planning and Dynamics**. Frankfurt: Peter Lang, p. 157-190, 2012.

LENGERIS, A. Prosody and Second Language Teaching: Lessons from L2 Speech Perception and Production Research. In: ROMERO-TRILLO, J. (Ed). **Pragmatics and Prosody in English Language Teaching**. Dordrecht: Springer, p.25-40, 2012.

LEVIS, J. **Intelligibility, Oral Communication and the Teaching of Pronunciation**. Cambridge: Cambridge University Press, 2018.

LEVIS, J.; SUVOROV, R. Automatic speech recognition. In. CHAPELLE, C. **Concise Encyclopedia of Applied Linguistics**, New York: Wiley Blackwell, 2020.

LI, V.; LUO, X.; MOK, P. **L1 and L2 phonetic reduction in quiet and noisy environments**. In. 9th International Conference on Speech Prosody 2018, p. 13-16, Poznań: 2018.

LOBANOV, B. Classification of Russian Vowels Spoken by Different Speakers. The **Journal of the Acoustical Society of America**, v. 49, n. 2B, p. 606-609, 1971.

NESPOR, M.; VOGEL, I. **Prosodic Phonology: With a New Foreword**. Berlin: De Gruyter, 2007 [1986].

PENNINGTON, M.; ELLIS, N. Cantonese speakers' memory for English sentences with prosodic cues. **Modern Language Journal**, v. 84, n. 3, p. 372–389, 2000.

POLITT, A. **The Oxford Online Placement Test: The Meaning of OOPT Scores**. Available in: oxfordenglishtesting.com, 2019. Accessed on July, 3, 2022.

POLUSHKINA, T.; TAREVA, E. Developing L2 prosodic competence online: implications of the emergency remote teaching. **XLinguae**, v 14, n. 1, p. 38-48, 2021

PYSHKIN, E., BLAKE, J., LAMTEV, A., LEZHENIN I., ZHUIKOV, A., BOGACH, N. Prosody Training Mobile Application: Early Design Assessment and Lessons Learned. In: IEEE Ukraine Section I&M (ed**). The 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications**, IDAACS 2019. Metz, France, September, 18-21, 2019, Proceedings, IEEE, p. 1-7, 2019.

RAN, Y.; VAN DE WEIJER, J.; SLOOS, M. Intonation in Hong Kong English and Guangzhou Cantonese-accented English: A Phonetic Comparison. **Journal of Language Teaching and Research**, v. 11, n. 5, p. 724-738, 2020.

REED, M. MICHAUD, C. Intonation in Research and Practice: The Importance of Metacognition. In: REED, M.; LEVIS, J. (Orgs.). **The Handbook of English Pronunciation**. West Sussex: John Wiley & Sons, p. 454-470, 2015.

SILVA Jr., L. **L2 Prosody**. In: Verbetes LBASS. <http://www.letras.ufmg.br/lbass/>, 2021.

SILVA Jr., L.; BARBOSA, P. Foreign Accent and L2 Speech Rhythm of English a pilot study based on metric and prosodic parameters. In. **Proceedings of the II Brazilian Conference on Prosody (Anais II Congresso Brasileiro de Prosódia)**, v. 1, p. 41-50, 2023a.

SILVA Jr., L.; BARBOSA, P. **SpeechRhythmExtracto**r. Computer Program for Praat, version 1.1. Available in: https://github.com/leonidasjr/SpeechRhythmCode, 2023b.

SILVA Jr, L.; BARBOSA, P. Efeitos da Prosódia de L2 no Ensino de Pronúncia e na Comunicação Oral. **Prolíngua**, v. 16, n. 1, p. 126-141, 2021.

SILVA Jr., L; BARBOSA, P. A. Speech Rhythm of English as L2: an investigation of prosodic variables on the production of Brazilian Portuguese speakers. **Journal of Speech Sciences**, v. 8, n. 2, p. 37-57, 2019.

SMITH, B.; JOHNSON, E; HAYES-HARB, R. ESL learners' intra-speaker variability in producing American English tense and lax vowels. **Journal of Second Language Pronunciation**, v. 5, n. 1, p. 139-164, 2019.