# Training English prosody with manipulated voices: Can both intermediate and advanced learners benefit?

*Radek Skarnitzl, Tomáš Bořil*

Institute of Phonetics, Faculty of Arts, Charles University
Prague, Czech Republic
{radek.skarnitzl|tomas.boril}@ff.cuni.cz

## Abstract

Prosodic aspects of speech such as intonation and rhythm are crucial for the comprehensibility and fluency of L2 speakers. However, prosody is rarely addressed in English pronunciation teaching. This paper reports on a short prosody-focused training, where participants' own PSOLA-modified speech was used as model in an adapted listen-and-repeat paradigm. The individualized training, comprising one audio-only and one audio-visual session, was delivered to six intermediate (B-level as per CEFR) and six advanced (C-level) Czech speakers of English. The comparison of before- and after-training reading performance shows that both intermediate and advanced speakers benefitted from the training: their phrasing was more in line with the linguistic structure of the texts, and particularly their melodic variability increased significantly. The implications of the study are discussed especially with respect to awareness-raising.

**Index Terms**: phrasing, melody, rhythm, prosody manipulations, learner proficiency

## 1. Introduction

Prosodic aspects of speech have been shown to be essential for successful communication in a second language (L2): good prosodic patterning results in better comprehensibility of L2 speakers [1], [2], [3] and also contributes to their increased fluency [4], [5]. At the same time, studies examining the effects of prosodic instruction have demonstrated that features like intonation, rhythm or lexical and phrasal stress are both teachable and learnable [6], [7], [8]. In this sense, the long-lamented cleft between the available results of empirical research on the one hand and their implementation into teaching practice on the other hand (see [9] for a recent example) does not seem to be shrinking: prosodic features tend to be neglected in pronunciation teaching, at least in part due to teachers' conviction that these phenomena are difficult if not impossible to teach [10], [11].

The context of Czech learners of English is not different: prosodic aspects are targeted only rarely in lessons of English as a foreign language, despite the considerable differences in the prosodic makeup of English and Czech. Specifically, Czechs have been shown to speak in significantly longer prosodic phrases in their L2 English than L1 English speakers do [12], [13]. Prominence marking at the lexical level is very weak or nearly nonexistent in Czech, with the stressed syllable not being higher, louder or longer, or manifesting more peripheral vocalic articulation [14], which tends to be transferred into their L2 English [15]. Moreover, pitch range in Czech speakers of English is very narrow [16], sometimes even more so than in native Czech speech [16].

The current study is conceptually based on our earlier study [17] who used short stretches of L1 French university teachers' L2 English speech with modified prosodic patterns as feedback in a training session; the study demonstrated the usefulness of training with the teachers' own voice, but with their word- and phrase-level prominence features modified to approximate those of the target language, English. The benefits of using learners' own voices for pronunciation training have been documented before [18] and seem to be grounded neurolinguistically [19]: perceiving familiar patterns strengthens neural response and results in more effective learning. In contrast with [17], however, this study focuses on Czech speakers of English, and the instruction concerned predominantly phrase-level prosody.

One of the main aims of the present research is to reveal whether both intermediate (i.e., B1- and B2-level, according to the Common European Framework of Reference, CEFR [20]) and advanced (C1- and C2-level) learners of English may benefit from a relatively short prosody-focused training where their own modified speech served as a model. On the one hand, research shows that prosody training can be successful even with learners who have been using the target language for a long time and may seem to be fossilized: Derwing and colleagues [21] showed that the comprehensibility of their Vietnamese speakers improved after these speakers had stayed in the L2 country (Canada) for an average of 19 years. It must be pointed out, however, that Derwing et al.'s speakers were not advanced (i.e., C-level) users of English; in fact, such highly advanced speakers are targeted only rarely in pronunciation training research. On the other hand, advanced L2 speakers, who use English daily in their professional lives, are likely to have progressed from explicit to implicit acquisition processes [22], which may render these speakers less susceptible to a brief intervention in the form of explicit pronunciation instruction. We therefore aim to compare the effectiveness of prosody-focused training in an intermediate and advanced group of participants, using their own modified voices.

## 2. Method

### 2.1. Participants and training procedure

Six intermediate (B2/B1 as per the CEFR, four females and two males, mean age 26.7 years, SD = 7.3 years) and six advanced (C1/C2, four females and two males, mean age 25.8 years, SD = 7.4 years) speakers participated in the individualized prosody-focused training. The participants were recruited as volunteers who expressed their interest in improving their English spoken skills; participants in the advanced group were mostly university students of English and American Studies at the Faculty of Arts in Prague.


10.21437/ISAPh.2024-18

First, a baseline recording was acquired, consisting in reading a diagnostic text of 230 words after having a short time to prepare. Subsequently, portions of the recording were modified using PSOLA [23] as implemented in Praat [24], based on the judgement of the first author, who has been teaching the phonetics of English for over twenty years. Nine or ten such stimuli were used for each participant, with one stimulus comprising between one and three prosodic phrases. The features which were targeted during the manipulations include the following:

- phrasing and the corresponding phrase boundary signals (i.e., when a phrase boundary was added into the participants' original speech, phrase-final deceleration and a more marked nuclear pitch movement were generated typically in the last syllable)
- pitch range was increased
- temporal characteristics of syllables (specifically, shortening of unstressed syllables and weak grammatical words)
- prominence (when it was suitable to shift the nuclear-accented word from that originally produced by the participant).

The training itself consisted of two training sessions. The first session featured an awareness-raising introduction which covered information about prosody in general and the above-mentioned prosodic aspects of English. The importance of prosody in communication was explained, as well as the reasoning behind using the participants' own voice for individualized feedback. The prosodic phrase was emphasized as the central unit of connected speech – not only in terms of intonation [25], but also rhythm [26]. The differences between phrasing and pitch range in Czech and English were discussed, along with the potential consequences of speaking with intonation which is "too flat", and the principle of reducing unstressed syllables according to Dickerson's two-peak profile [26] was also explained.

In the training stage itself, the participants then listened repeatedly to each of the modified stimuli; they were instructed to listen as long as they wanted, so as to really "internalize" or "soak in" the new prosodic patterns. They would then stop the playback themselves and, after a 20-second silent period (*cf.* [17]), they were instructed to repeat the stimulus as closely to the manipulated version as possible.

The second training session, which took place on average ten days later and lasted approximately 50 minutes, was based on the same stimuli, with two additions. First, visual displays of the stimuli were provided (pitch contours and temporal information corresponding to portions which were lengthened and shortened, similarly to what appears in Praat's Manipulation window; see Figure 1 for an example of one such stimulus). Second, the respondents were encouraged to use any form of gesticulation to reinforce the prosodic patterns; gestures have been reported to facilitate learning the prosody of a second language (see [27] for an example). The benefits of applying multimodal aspects in pronunciation teaching, given learners' dominant memory types, have been emphasized at the beginning of this session.

A delayed post-test recording was obtained from the participants at least six weeks (48 days on average) after the second training session. During this last session, the participants were asked to record the initial diagnostic text, as well as a completely new text; they were encouraged to make any notes in the text that would help them with the reading. The analyses presented in the next section are based on the comparison of the first diagnostic recording and the two post-test recordings (referred to as "before text1", "after text1", and "after text2", respectively, and overall as "conditions"). Finally, the last session also featured a debriefing discussion with the experimenter in which the participants were asked about their overall experience with the training, its individual aspects and perceived usefulness.

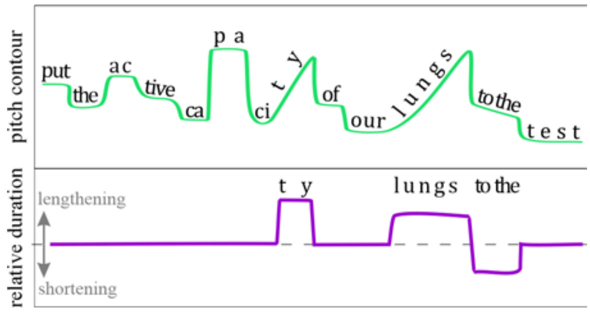


Figure 1: *An example of the visual display of one stimulus showing the melodic contour in green and the temporal changes in purple. The stimulus illustrates, among others, the addition of a phrase boundary behind the words "capacity" and "lungs", and shortening of the weak grammatical words "to the".*

### 2.2. Analyses

To compare the participants' prosodic patterning between the before- and after-training recordings, we analyzed phrasing and pitch range. The participants' phrasing was examined because the training was centred around the prosodic phrase. In both text1 and text2, the suitability of a prosodic boundary between any two neighbouring words was determined, based on the semantic and syntactic structure. Since virtually all boundaries marked as "obligatory" were observed by our participants, the phrasing score corresponds to the sum of those phrase boundaries which were placed incorrectly by the participants (in other words, where it disrupted the natural flow of the text). Note that the phrasing score is therefore "negative": the lower it is, the more were the phrase boundaries placed in line with the structure of the texts.

Fundamental frequency (henceforth marked $f_0$) was analyzed using default settings for filtered autocorrelation in Praat, only with $f_0$ ceiling changed to 500 Hz. The Pitch object in Praat was converted to PitchTier, and portions containing creaky voice were manually removed, so as not to negatively affect subsequent analyses. The rPraat package [28] was then used to convert $f_0$ to semitones (ST) relative to 50 Hz and to stylize the contour using either one point which corresponded to mean $f_0$ at the mean time point for vowels which were shorter than 100 ms, or two points (the first and last $f_0$ point within the vowel) fitted with linear regression for vowels which were longer than 100 ms. The aim of using stylized $f_0$ contours was to approximate the results to the way in which the human ear perceives pitch changes (see [29] for a more detailed justification of the choices made here). The stylized $f_0$ contours were subsequently used to compute the $f_0$ range (note that the stylization avoids reporting outlier values). This was done always for the entire prosodic phrase, but also separately for its nuclear and prenuclear portion (see the following section for a more detailed explanation).

## 3. Results and discussion

Phrasing results are presented in Figure 2 for both intermediate and advanced participants. It is clear that the advanced group had lower phrasing scores overall; this should not be surprising, as advanced language users are expected to be better able to segment a given text correctly (that is, in agreement with its syntactic and semantic structure). As for the comparison of before- and after-training recordings, most participants managed to phrase better in the after text1 condition (i.e., in the same text that they had read before the training, as well as at the end of the two training sessions); in other words, this improvement may also be ascribed to an effect of learning. Phrasing score in the after text2 condition was lower in some and higher in other participants; naturally, however, the two texts are not directly comparable in this respect.

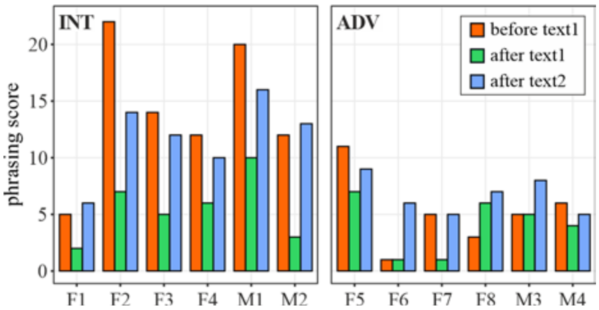


Figure 2: *Phrasing score (number of incorrectly placed phrase boundaries) of the intermediate (left) and advanced (right) participants in the three conditions.*

Turning to the results concerning melodic variability, we first ran a linear mixed effects model to see whether the effect on $f_0$ range of the prosodic training with participants' own modified voices was significant. The three texts (CONDITIONS) were used as the fixed effect, and the by-SPEAKER intercept and slope for the effect of CONDITION served as random effects, to account for the possibility that each participant behaved differently in terms of melodic range in each of the conditions.

The effect of CONDITION is significant for the intermediate ($\chi^2(2)$ = 20.01, $p$ < 0.001), as well as advanced group of participants ($\chi^2(2)$ = 6.55, $p$ < 0.05). To assess the significance of individual pairwise comparisons, we used Tukey-adjusted confidence intervals of the estimated marginal means (emmeans) with the alpha level of 0.05. Plots showing mean values of the competence score and their confidence intervals were created using the *emmeans* package [30] and plotted using the *ggplot2* package [31].

Figure 3 shows that $f_0$ range, extracted from the stylized $f_0$ contours, is markedly larger in the after-training conditions than it was before the training. The fact that the red arrows do not overlap means that the pairwise comparisons are significant: $f_0$ range was therefore significantly higher in the delayed post-test recording six weeks after the training than before the training, in both the intermediate and advanced group. The participants also spoke with significantly greater melodic variability in text2 than in text1 after the training; it should be noted, however, that the participants' prosodic performance in the two texts cannot be directly compared.

The emmean results presented in Fig. 3 suggest that there is very little difference between the two participant groups – both before and after the training. This is confirmed by the boxplots shown in Figure 4: the typical melodic range is very similar in the two groups before the training, and the increase in $f_0$ range after the training is also comparable.

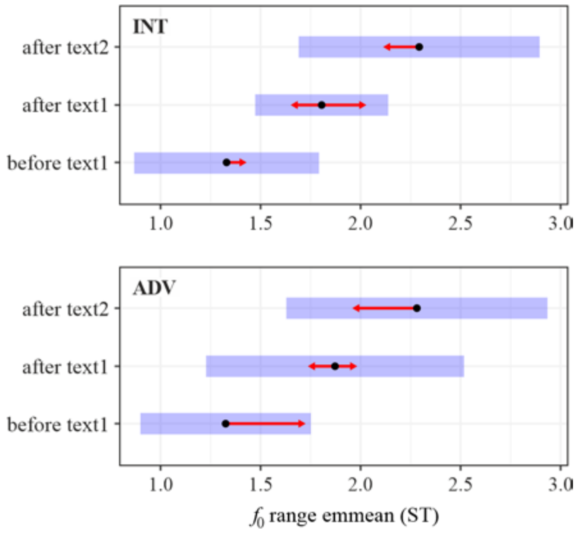


Figure 3: *Estimated marginal means of $f_0$ range illustrating the effect of condition in the intermediate (top) and advanced participants (bottom). Blue bars correspond to confidence intervals of the emmeans, and red arrows serve for comparisons among them.*

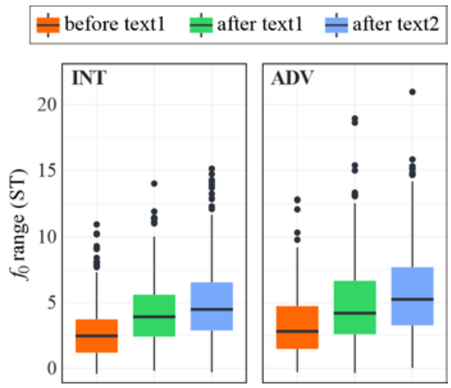


Figure 4: *$f_0$ range of the intermediate (left) and advanced (right) participants in the three conditions.*

In the next step, we wanted to examine the participants' melodic range separately in the nuclear portion of each prosodic phrase (i.e., in the final part of each phrase which carries the nuclear melodic pattern and most of the linguistic meaning of intonation) and in the prenuclear portion (the beginning of each phrase before the intonational nucleus). The justification for this is based only on an informal but repeated observation that Czech speakers of L2 English sometimes have very flat intonational nuclei but do manifest some melodic variability before the nuclei. The results in the form of boxplots are shown in Figure 5.

First, the overall data do not confirm the informal hypothesis that $f_0$ range would be higher in prenuclear portions of prosodic phrases. However, the prenuclear and nuclear range were very similar – in both intermediate and advanced speakers – before the training, with $f_0$ range in most prosodic phrases not extending 5 ST, as shown by the orange boxes. It may be that, since we expect most melodic movement to occur in the intonational nucleus, we perceive the relative lack of melodic variability here as greater variability in the prenuclear portions. Second, the increase in $f_0$ range between the before- and after-training conditions is greater in the nuclear than in the prenuclear portions. That is not surprising, of course; what is interesting, however, is that melodic variability also increased in the prenuclear portions of prosodic phrases, again in both groups. The participants' speech thus became livelier overall after the training with their own voices, not only in the intonational nuclei. Finally, melodic range increased – again in both participant groups – especially in the nuclear portions of text2.

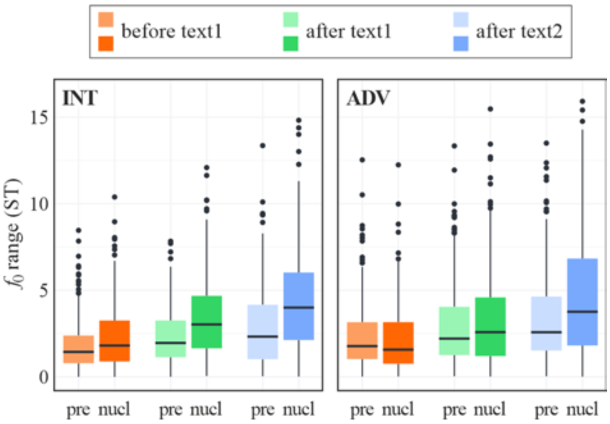


Figure 5: *$f_0$ range of the intermediate (left) and advanced (right) participants in the three conditions. Prenuclear (marked pre) data are always shown in a lighter shade on the left, nuclear (marked nucl) data in a darker shade on the right.*

The final aspect of the analysis to be presented here focuses on individual speakers, since it was to be expected that the previous displays would conceal a lot of variability and that some participants would change their prosodic behaviour more than others. The results are presented in Figure 6; for the sake of easier comparison, the figure shows $f_0$ range in entire prosodic phrases (that is, regardless of their prenuclear and nuclear portions).

It is obvious at first glance that $f_0$ range increased in all participants between the before- and after-training recordings, with the exception of the advanced speaker F5; note, however, that this participant's $f_0$ range was already relatively wide before the training, with the mean value around 5 ST. This speaker has virtually a native-like American accent in her English, and her melodic patterning did not change much anymore after the training.

It is worth pointing out two participants whose delivery was extremely flat before the training: the intermediate speaker M2 and the advanced speaker F8. While M2 did not increase his melodic range dramatically in absolute terms – his post-training $f_0$ range only rarely exceeds 5 ST – it still is a marked change, and certainly one which is clearly audible. On the other hand, the $f_0$ range increase in participant F8 is considerable, especially in text2. As in the previous group analyses, there are no consistent differences between the intermediate and the advanced participants: it cannot be specified which speakers benefitted more from the training.

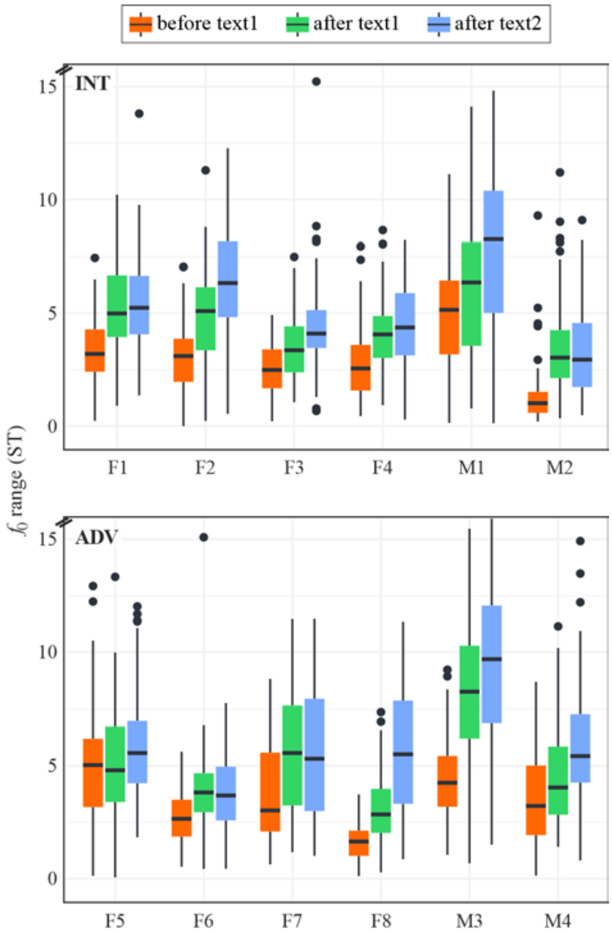


Figure 6: *$f_0$ range of individual intermediate (top) and advanced (bottom) participants in the three conditions.*

## 4. General discussion

This study examined the effects of training of English prosody, which revolved around using the participants' own voices as models, modified in their melodic and rhythmical patterns to approximate native-like English speech. The results indicate that this short (approximately two hours in total) prosody-focused training, comprising an awareness-raising session and audio-only and audio-visual (see Fig. 1) feedback sessions with the speakers' voices modified using PSOLA, yields marked improvements in the melodic variability of prosodic phrases and modest improvements in the structuring of the text, even in a delayed post-test recording obtained six weeks after the training.

The main research question of this study was whether the potential benefit of the training would apply to intermediate (B2- and B1-level), as well as advanced (C1- and C2-level) participants. As advanced language users are assumed to rely more on implicit than explicit acquisition processes [22], we wanted to see whether explicit prosodic instruction would result in comparable increase in melodic variability. The results clearly confirm that both intermediate and advanced participants benefitted from our prosody training. Importantly, the improvement is very similar in both the intermediate and advanced speakers: it cannot be determined in which group the improvement is greater. Moreover, there appears to be an effect of condition (text), with text1 – which was familiar to the participants (having been asked to read it also at the end of each training session) – manifesting a lower $f_0$ range than the new text2, which they were asked to read during the post-test session for the first time. Interestingly, the between-participant variability that we expected to find was not borne out to a great extent: with the exception of speaker F5, an increase in $f_0$ range was observed in all our participants (see Fig. 6).

In addition to the melodic variability data, the results also suggest that participants' after-training rendition of the texts was more in line with the linguistic structure of the text; in other words, there were fewer incorrectly placed prosodic breaks than before the training. In this area, the analysis did reveal a clear effect of speaker proficiency, with advanced participants violating the text structure overall to a much lesser extent (*cf.* Fig. 2).

Based on the results of this study, we are convinced that using targeted modifications of learners' prosodic patterns for pronunciation training has great potential, regardless of learner proficiency. From the participants' feedback, gathered during the final debrief within the delayed post-test recording session, we may conclude that the use of their own voice for training was highly beneficial: it allowed them to realize what exactly they could sound like in their L2 English. Naturally, it would be unrealistic to expect that the participants would adopt the new prosodic patterns in their everyday speech. However, they claimed that – if they were required, for instance, to deliver a presentation in English – they knew what they needed to focus on when preparing their performance. A conscious application of the newly learned patterns in prepared presentations could undoubtedly serve as the first step towards more natural, English-sounding speech in general.

Perhaps the most important implication of this study concerns awareness. Many of the participants were amazed by the extensive differences between prosody in Czech and English and wondered why they had not been told about them during the many years of English instruction at various stages of their education. Given that $f_0$ range was very similar before the training in the intermediate and advanced participants (*cf.* especially Figs. 3 and 4), our results point to the profound lack of awareness about English prosody even in very advanced speakers. Since – as one of the participants put it – "this is no rocket science", awareness-raising appears to be a labour- and cost-effective solution, and this is the logical direction in which our future efforts should be heading.

## 5. Acknowledgements

This study was supported by the European Regional Development Fund project entitled "Beyond security: Role of conflict in resilience-building" (registration no.: CZ.02.01.01/00/22_008/0004595).

We are also grateful to all our participants for their dedication.

## 6. References

[1] A. Lengeris, "Prosody and second language teaching: Lessons from L2 speech perception and production research," in *Pragmatics and prosody in English language teaching*, J. Romero-Trillo, Ed. Dordrecht: Springer, pp. 25–40, 2012.

[2] K. Saito, P. Trofimovich, and T. Isaacs, "Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels," *Applied Psycholinguistics*, vol. 37, pp. 217–240, 2016.

[3] D. M. Chun, and J. M. Levis, "Prosody in L2 teaching: Methodologies and effectiveness," in *Oxford handbook of language prosody*, C. Gussenhoven and A. Chen, Eds. Oxford: Oxford University Press, pp. 619–630, 2020.

[4] S. L. Dowhower, "Speaking of prosody: Fluency's unattended bedfellow," *Theory Into Practice*, vol. 30, pp. 165–175.

[5] P. Trofimovich, and W. Baker, "Learning prosody and fluency characteristics of second language speech: The effect of experience on child learners' acquisition of five suprasegmentals," *Applied Psycholinguistics*, vol. 28, pp. 251–276, 2007.

[6] T. M. Derwing, and M. Rossiter, "The effects of pronunciation instruction on the accuracy, fluency and complexity of L2 accented speech," *Applied Language Learning*, vol. 13, pp. 1–17, 2003.

[7] Y. Saito, and K. Saito, "Differential effects of instruction on the development of second language comprehensibility, word stress, rhythm, and intonation: The case of inexperienced Japanese EFL learners," *Language Teaching Research*, vol. 21, pp. 589–608, 2017.

[8] J. Gordon, and I. Darcy, I., "Teaching segmentals and suprasegmentals: Effects of explicit pronunciation instruction on comprehensibility, fluency, and accentedness," *Journal of Second Language Pronunciation*, vol. 8, pp. 168–195, 2022.

[9] T. M. Derwing, J. Levis, and S. Sonsaat-Hegelheimer, "Bridging the research-practice gap in L2 pronunciation," in *Second language pronunciation: Bridging the gap between research and teaching*, T. M. Derwing, J. Levis and S. Sonsaat-Hegelheimer, Eds. Oxford: Wiley Blackwell, pp. 1–18, 2022.

[10] P. Rogerson-Revell, "Can or should we teach intonation?" *IATEFL Pronunciation SIG Newsletter*, 47, pp. 16–20, 2012.

[11] M. G. O'Brien, "Making the teaching of suprasegmentals accessible," in *Second language pronunciation: Bridging the gap between research and teaching*, T. M. Derwing, J. Levis and S. Sonsaat-Hegelheimer, Eds. Oxford: Wiley Blackwell, pp. 85–106), 2022.

[12] J. Volín, "The size of prosodic phrases in native and foreign-accented read-out monologues," *Acta Universitatis Carolinae – Philologica* 2/2019, pp. 145–158, 2019.

[13] R. Skarnitzl, and H. Hledíková, "Prosodic phrasing of good speakers in English and Czech," *Frontiers in Psychology*, vol. 13, paper 857647, 2022.

[14] R. Skarnitzl, "Fonetická realizace slovního přízvuku u delších slov v češtině," *Slovo a slovesnost*, vol. 79, pp. 199–216, 2018.

[15] R. Skarnitzl, and J. Rumlová, "Phonetic aspects of strongly-accented Czech speakers of English," *Acta Universitatis Carolinae – Philologica*, pp. 109–128, 2019.

[16] J. Volín, K. Poesová, and L. Weingartová, "Speech melody properties in English, Czech and Czech English: Reference and interference," *Research in Language*, vol. 13, pp. 107–123, 2015.

[17] A. J. Henderson, and R. Skarnitzl, "'A better me': Using acoustically modified learner voices as models," *Language Learning & Technology*, vol. 26, pp. 1–21, 2022.

[18] A. De Meo, M. Vitale, M. Pattorino, F. Cutugno, and A. Origlia, "Imitation/self-imitation in computer-assisted prosody training for Chinese learners of L2 Italian," in *Proc. of the 4th Pronunciation in Second Language Learning and Teaching conference*, J. Levis and K. LeVelle, Eds., Iowa State University, 2013, pp. 90-100.

[19] G. Hickok, “The role of mirror neurons in speech perception and action word semantics,” *Language and Cognitive Processes*, vol. 25, pp. 749–776, 2010.
[20] Council of Europe (2001). *Common European framework of reference for language learning, teaching, assessment*. Cambridge: Cambridge University Press.
[21] T. M. Derwing, M. J. Munro, J. A. Foote, E. Waugh, and J. Fleming, “Opening the window on comprehensible pronunciation after 19 years: A workplace training study,” *Language Learning*, vol. 64, pp. 526–548, 2014.
[22] R. DeKeyser, “Skill Acquisition Theory,” in *Theories in second language acquisition: An introduction*, B. VanPatten and J. Williams, Eds. Lawrence Erlbaum Associates Publishers, 2007, pp. 97–113.
[23] E. Moulines, and F. Charpentier, “Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones,” *Speech Communication*, vol. 9, pp. 453–467, 1990.
[24] P. Boersma, and D. Weenink, *Praat: Doing phonetics by computer*, v. 6.4.06 Retrieved from http://www.praat.org/
[25] J. C. Wells, *English intonation: An introduction*. Cambridge: Cambridge University Press, 2006.
[26] W. Dickerson, “The ripples of rhythm: Implications for ESL instruction,” in Proceedings of the 10th Pronunciation in Second Language Learning and Teaching conference, J. Levis, C. Nagle and E. Todey, Eds., 2019, pp. 36–54.
[27] P. Li, F. Baills, C. Alazard-Guiu, L. Baqué, and P. Prieto, “A pedagogical note on teaching L2 prosody and speech sounds using hand gestures,” *Journal of Second Language Pronunciation*, vol. 9, pp. 340–349, 2023.
[28] T. Bořil, and R. Skarnitzl, “Tools rPraat and mPraat: Interfacing phonetic analyses with signal processing,” in *Proc. of the 19th International Conference on Text, Speech and Dialogue*, P. Sojka, A. Horák, I. Kopeček and K. Pala, Eds. Cham: Springer International Publishing, pp. 367–374, 2016.
[29] D. J. Hermes, “Stylization of pitch contours,” in *Methods in Empirical Prosody Research*, S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzk, I. Mleinek, N. Richter, J. Schließer, Eds. Berlin: De Gruyter, pp. 29–62, 2006.
[30] R. V. Lenth, *emmeans: Estimated Marginal Means, aka Least-Squares Means*, v. 1.8.8. Retrieved from https://CRAN.R-project.org/package=emmeans, 2023.
[31] H. Wickham, “*ggplot2: Elegant graphics for data analysis*”, *Journal of Statistical Software*, vol. 5, pp. 1–23, 2016.