Training of English prosody with acoustically modified voices

Radek Skarnitzl and Tomáš Bořil
Charles University

Prosodic aspects of speech are crucial for the comprehensibility of L2 speakers, but prosody is rarely targeted in English language lessons. This paper describes an innovative training of English phrasal prosody using participants' own speech as models in a modified listen-and-repeat paradigm, with their melodic and rhythmic patterns manipulated by means of PSOLA. The two-hour training was delivered individually to twelve intermediate native speakers of Czech. The comparison of a baseline recording of a read text before the training and two texts read six weeks after the training shows that using one's own PSOLA-modified voice for prosody training is beneficial: the participants were perceived as sounding significantly more competent in the after-training recordings, their phrasing corresponded more to text-based predictions, and their melodic variability was significantly greater. The contribution of targeted prosody modifications in the teaching of L2 pronunciation are discussed.

Keywords: prosody, PSOLA manipulation, intonation, visualization, English as a foreign language

1. Introduction

This study reports on an experiment in which modifications of melodic and rhythmical patterning of Czech learners' English as a foreign language (L2) were used in a two-stage prosody-focused training. The use of artificially modified prosodic patterns was combined with visualizations of these patterns and with a strong awareness-raising component. The following two subsections will discuss the role of prosody in L2 speech and previous applications of prosody manipulations.

https://doi.org/10.1075/jslp.24041.ska | Published online: 3 February 2025
Journal of Second Language Pronunciation 10:3 (2024), pp. 375–403. ISSN 2215-1931 | E-ISSN 2215-194X

1.1 Prosody in L2 speech

The prosody of speech concerns the way we use stress, intonation, phrasing and rhythm to communicate meanings. More specifically, speakers use prosodic means to divide the continuous stream of speech into shorter units (prosodic phrases) which can be easily processed (Chafe, 1988; Frazier, Carlson & Clifton, 2006); to signal the grammatical nature and structure of sentences (Croft, 1995); to mark the information structure of utterances by highlighting important words and suppressing others (Kügler & Calhoun, 2020); to express affective meanings (Scherer, 2003); or to indicate whether they are negotiating turn-taking in the interaction (Gravano & Hirschberg, 2011). Crucially, many of these functions of prosody are universal, but some are language-specific. For example, languages differ in whether and to what extent they make use of prosody to signal questions or to mark prominence at the level of prosodic phrases (Trouvain & Braun, 2020). Prosody thus provides signposts to listeners which help them to navigate to the speaker's meaning. However, speakers in their L2 may not be providing the signposts expected by listeners, whether native (L1) listeners of the target language or listeners with another L1.

Prosodic aspects have been shown to be essential for the *comprehensibility* of L2 speakers. In other words, good prosodic patterning helps speakers to be more effective communicators, while poor and unpredictable patterns make them more difficult to understand, with listeners forced to exert greater effort to decipher what is being said. Significant effects on comprehensibility have been reported for both melodic and rhythmical patterning. In research focused on the former, intonation has been conceptualized as a more general perspective of native-likeness of melodic patterns (Munro & Derwing, 1999) or as appropriate and varied *vs.* incorrect and monotonous use of melody (Bergeron & Trofimovich, 2017; Saito, Trofimovich & Isaacs, 2016), as well as individual dimensions of intonation like the adequacy of individual nuclear pitch patterns (Pickering, 2001) or phrasing (Phillips et al., 2022). Moreover, Saito et al. (2016) show that intonation affects the comprehensibility of learners at all levels of proficiency. The effect of speech rhythm on L2 speakers' comprehensibility is generally discussed by Derwing et al. (2014), and a highly significant effect of rhythm (operationalized as vowel reduction ratio) is reported by Crowther, Trofimovich & Isaacs (2016).

Apart from the clear effects of various prosodic dimensions on comprehensibility, Levis (2018) provides ample evidence that prosody is also important for the *intelligibility* (i.e., the actual understanding) of L2 speakers (see Chapters 6 and 7). For rhythmical patterning, Levis bases this claim on the close relationship of speech rhythm to lexical access (that is, the listeners' ability to identify and understand words in the stream of speech; see also Cutler, 2012). In the melodic domain, Levis specifically identifies intonation prominence (also referred to as phrasal or nuclear stress) as crucial for intelligibility: as shown by Hahn (2004), predictable prominence patterns of international teaching assistants resulted in listeners remembering more information, while misplaced prominence impaired intelligibility.

An essential aspect which needs to be emphasized is that prosodic patterns of an L2 can be successfully taught and learned (e.g., Anderson-Hsieh, 1990; Gordon & Darcy, 2016, 2022; Saito & Saito, 2017; see also the summaries in Chun and Levis, 2020 and O'Brien, 2022, or the more detailed discussion in the above-mentioned chapters in Levis, 2018). At the same time, however, prosody has been largely neglected in language lessons. To blame seems to be, at least in part, the belief that prosody overall, as well as its components are very difficult to teach (Rogerson-Revell, 2012). In addition, language teachers frequently report feeling under-prepared and lacking confidence when it comes to teaching prosody (Baker, 2011; O'Brien, 2022). Much therefore needs to be done to close the often-lamented cleft between research findings and teaching practice (Derwing, Levis & Sonsaat-Hegelheimer, 2022), certainly (but not only) in the Czech context.

Compared to many other L2 English varieties, informal observations and reports suggest that the English produced by L1 speakers of Czech is largely intelligible to both native and nonnative speakers of English. Nevertheless, there are extensive differences in the prosodic makeup of English and Czech. Impressionistically, Czech may be described as very monotonous. At the lexical level, stressed syllables (or vowels) are not manifested by systematic deviations in any acoustic domain: unlike in English (Eriksson & Heldner, 2015) they are not higher pitched, louder, longer, nor do they differ in vocalic quality (Skarnitzl & Eriksson, 2017). In the rhythmical domain, English is characterized by unstressed syllables being reduced in duration and vowel quality, whereas Czech has no systematic vowel reduction; Czech has thus been described as syllable-based (Dankovičová & Dellwo, 1999), despite its syllabic complexity (Šturm & Lukeš, 2017). Melodically, the perceived monotonousness of Czech seems to be due to significantly longer phrases than in English (Skarnitzl & Hledíková, 2022; Volín, 2019) and narrower pitch range (Volín, Poesová & Weingartová, 2015). Focusing on prosody – both melodic and rhythmic patterns – in pronunciation training is therefore expected to benefit especially the comprehensibility of L1 Czech learners.

1.2 Using manipulations of speech prosody

Manipulations of temporal or melodic properties of speech have been used in L2 speech research for decades. For example, Tajima, Port and Dalby (1997) time-warped English phrases pronounced by a Mandarin speaker to correspond in their temporal structure to that of a native speaker. The resynthesized phrases were played to listeners who performed a four-alternative forced-choice identification test. The results suggest that the intelligibility of temporally improved English improved by 15 to 25 percent. Conversely, the intelligibility of native English speech, manipulated to resemble the temporal structure of the Mandarin speaker, dropped by 15 percent. Kolly et al. (2017) transplanted durational patterns of English- and French-accented German onto native Swiss German speech and found that Swiss German listeners were able to identify the L2 (i.e., English or French) accent based only on the temporal changes. In a study focused on melodic manipulations, Volín and Poesová (2016) transplanted intonation contours from a native English speaker to a Czech-accented speaker of English and vice versa. In a word monitoring experiment, Czech listeners' reaction time to the target words differed significantly, with native English speech with unchanged intonation yielding shortest reactions (that is, the least cognitive effort, reflecting highest comprehensibility) and Czech-accented English with original Czech-sounding melody resulting in longest reactions. Some studies applied both temporal and melodic manipulations, transplanting native-like patterns onto the speech of L2 learners to probe the individual and combined effect on accentedness (Polyanskaya, Ordin & Busa, 2017) or accentedness and comprehensibility (Van Maastricht et al., 2021), revealing intricate relationships between the prosodic domains and dimensions of L2 speech.

The more recent studies typically use PSOLA for their prosodic manipulations (Pitch-Synchronous Overlap-and-Add;[1] Moulines & Charpentier, 1990), as does the current study. A brief explanation is therefore in order, although the technical knowledge is not strictly necessary, since PSOLA manipulations are implemented in Praat (Boersma & Weenink, 2024) as quite intuitive visual changes. As shown in Figure 1a, the first step consists in identifying the fundamental frequency (f_0) of the speech signal; hence the label pitch-synchronous.[2] The so-called pitch marks indicate centres of frames in each fundamental period of the waveform. Overlapping analysis segments called windows are then centred around each pitch mark (Figure 1b). The manipulation of f_0 consists in changing the degree of overlap of the individual windows (Figure 1c): a greater overlap shortens the fundamental period and raises f_0, while a smaller overlap leads to a longer period and lower f_0. The signal is lengthened by adding and shortened by removing individual periods of the signal, as also indicated in Figure 1c.

1. More accurately, one should talk about TD-PSOLA (time-domain), since the manipulations are applied directly to the waveform (the time domain representation of the speech signal).

2. Strictly speaking, *pitch* refers to the subjective percept, while it is the *fundamental frequency* as the measurable quantity that the algorithm needs to identify.

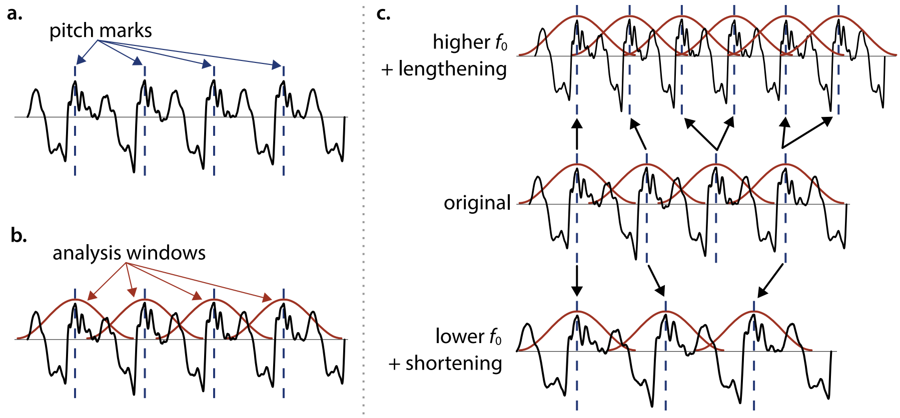

Figure 1. Block scheme of PSOLA manipulations: Identification of pitch marks (**a**); overlapping fundamental periods with windowing functions (**b**); and changing overlap to modify f_0 and adding or removing fundamental periods to modify duration (**c**)

As mentioned above, the Praat software features an intuitive environment in which PSOLA manipulations can be implemented. The process is illustrated in Figure 2 on the word *prosody*, generated in a British English female AI voice. Figure 2a shows Praat's entire Manipulation window, with its three panels labelled: the top panel shows the waveform along with the pitch marks; the middle panel shows the fundamental frequency (f_0) contour;[3] and the lower (empty) panel is designed for making temporal changes. Figure 2b shows the middle (melodic) panel, with stylization of f_0 contour applied. It is clear that the stylized curve is much simpler, allowing for easier subsequent manipulations; it also reflects much better than the original "rough" contour how people actually perceive pitch changes. Figure 2c shows specific manipulations: the intonation has been changed from a falling pattern into a falling-rising one, and the unstressed *schwa* [ə] vowel (and part of the closure of [d]) has been shortened.

Using PSOLA manipulations for pronunciation training has been attempted before (see Sundström, 1998 for one of the earliest attempts). Such pronunciation training is based on the findings that the more similar the learner's voice is to the modelling speaker's voice, the greater the pronunciation improvement (see De Meo et al., 2013). Repeating after oneself — with one's voice quality preserved but key characteristics of speech "improved" — should therefore yield best results, as confirmed by several studies (see Kusz, 2023 for a summary). Automatic voice conversion has been applied for this purpose (the Golden Speaker Builder

3. Note that some of the changes in the f_0 contour are not audible, such as the lower frequency at the beginning of [d].

by Ding et al., 2019 is one of the best-known tools; see also Felps, Bortfeld & Gutierrez-Osuna, 2009). However, the current study is based on manual prosody manipulations, primarily because direct input and feedback seems to be crucial for learners, even though automatic tools do result in better comprehensibility (Kusz, 2023).

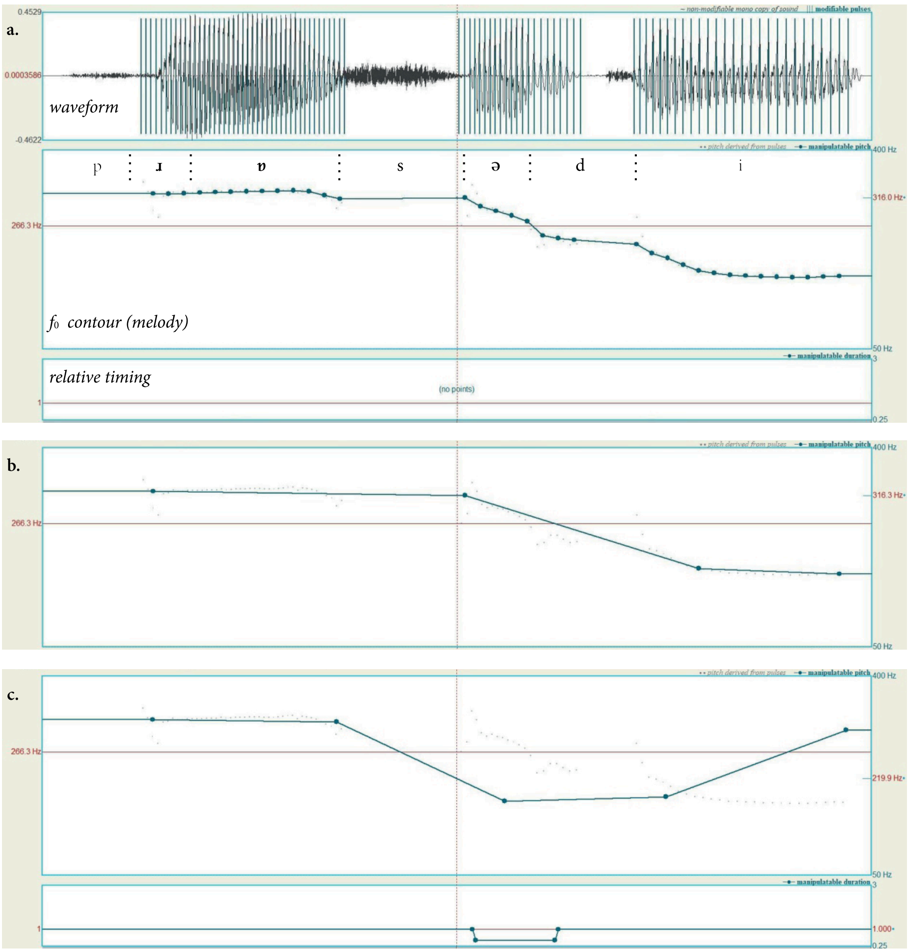

Figure 2. The implementation of PSOLA manipulations in Praat: The original Manipulation window with the main features labelled (**a**), with inserted speech sounds and boundaries between them; f_0 contour after stylization (**b**); and manipulated melody and duration (**c**). See text for details

1.3 Current study

This paper reports on an experiment which is conceptually based on a previous pilot study (Henderson & Skarnitzl, 2022), which used modified prosodic patterns of L1 French speakers of English. The modifications concerned predominantly word-level and partly also phrase-level prominence. The results indicated that the participants' temporal and melodic patterning did improve in the feedback session, during which they were asked to repeat the modified stimuli. Additionally, some carryover was reported in a delayed post-test three months after the feedback session.

The aim of the current study is not to compare the prosodic rendition of individual words or short phrases, but to reveal whether a relatively short but comprehensive prosody-focused training, centred on prosodic phrasing, leads to improved prosodic patterning overall. The training will target both melodic and temporal aspects of prosody, which has been relatively rare, as noted by Mennen and de Leeuw (2014) and reiterated by Trouvain and Braun (2020). The potential improvement will be examined using a perception test and acoustic measurements. In short, the intervention presented in the following section combines several aspects which have been recommended in literature – prosody as the crucial component for effective L2 communication; awareness about the main prosodic differences between English and Czech; the use of technology to modify learners' prosodic patterns; and a multimodal approach – to improve the prosodic patterning of L1 Czech speakers of L2 English.

2. Training with PSOLA-modified prosody

2.1 Speakers and material

Twelve intermediate native Czech speakers of English (eight females, four males; mean age 26.0 years, SD 7.3 years) participated individually in the training. They had previously registered as participants for psychological and linguistic experiments carried out at Charles University and responded to an email seeking participants for this study. To be included in the study, they had to have B1 or B2 level of English as per the CEFR; they were not present or past students of phonetics or an English-related university programme; they had not spent more than one month in an English-speaking country. A demographic questionnaire concerning their use of English and their self-assessed talent for languages and music, obtained at the beginning of the initial session along with an informed consent, shows that the group was fairly homogenous in these areas. During this initial session, each participant also recorded a diagnostic text of 230 words. They had time to go through the text before recording, which took place in a sound-treated studio of the Institute of Phonetics in Prague.

The manipulations were performed by the first author of this study. Nine or ten portions of the diagnostic recording were chosen based on their linguistic and technical suitability. As for the former, each portion had between six and ten words, forming a semantically self-standing text which may be prosodically realized as having between one and three prosodic phrases. From the technical perspective, portions containing creaky voice were avoided, because these usually render PSOLA manipulations unreliable.

2.2 Prosodic manipulations

The features which were targeted during the manipulations include several areas, based on the differences between English and Czech prosody (see Section 1.1). Most importantly, because the training was centred around the prosodic phrase, new phrase boundaries and boundary signals (a salient pitch movement and phrase-final deceleration) were introduced into the participants' original speech, or weakly realized boundaries were made more salient. Second, the pitch range was expanded. When phrase-level prominence was placed inadequately, a more suitable word (its stressed syllable) was prosodically highlighted as prominent and the original one suppressed. Finally, at the level of individual words, it was sometimes necessary to shorten unstressed syllables (including grammatical words) and lengthen those which were meant to be stressed. Figure 3 illustrates these types of manipulations (although these types naturally overlap) on one portion of speaker F1's recording. The phrase boundary after *question* was made more salient, pitch range was overall increased, prominence shifted to *first*, and weak grammatical words were shortened (and also suppressed melodically).

2.3 Procedure

The training consisted in two sessions, spaced ten days apart on average (with the break ranging between 7 and 16 days). The first session started with a short awareness-raising presentation which lasted approximately 20 minutes and covered the following areas:

- the basic distinction between segmental aspects and prosody
- prosodic dimensions and functions of prosody

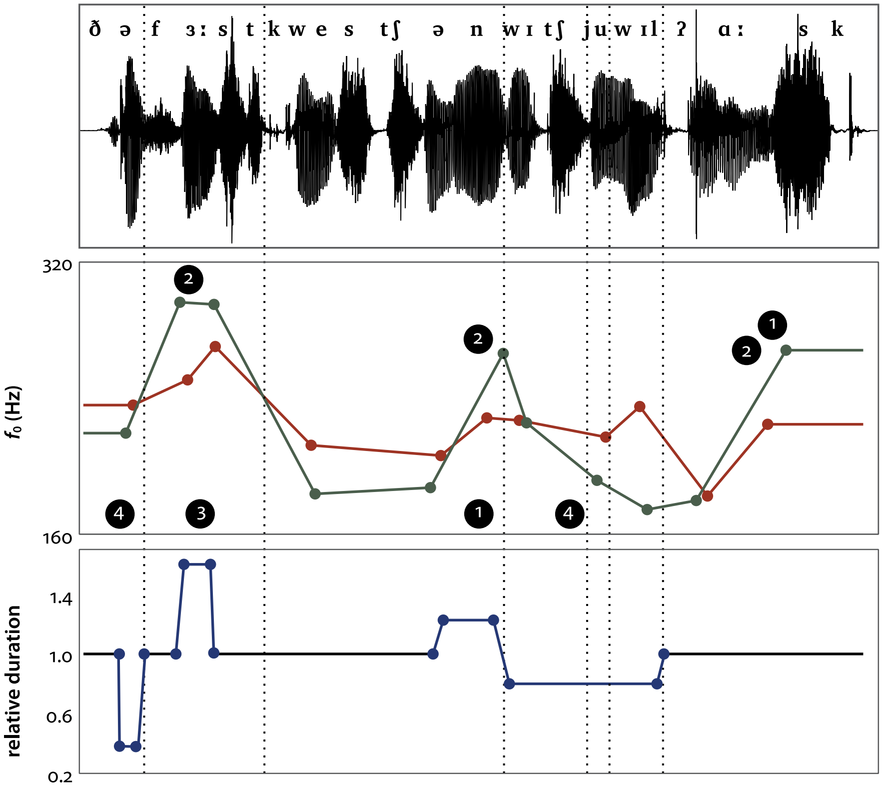

Figure 3. Examples of PSOLA manipulations on the portion *the first question which you will ask*: More salient phrasing signals (1), increased pitch range (2), prominence marking (3), shortening of unstressed syllables (4). The original melodic contour in red, the manipulated one in green

- the central role of the prosodic phrase in communicating meaning, language-specific length of phrases and its relation to comprehensibility, and the main tendencies of phrasing in English
- prominence and its role in signalling information structure (specifically the deaccenting of known information)
- speech rhythm in English (presented as a somewhat simplified version of the two-peak profile by Dickerson, 2019, according to which English rhythm is realized in the domain of prosodic phrases)
- intonation and specifically the much wider pitch range in English
- the justification of using one's own adapted voice for modelling (see above, as well as the concept of mirror neurons; Hickok, 2010)

During the presentation, the participants also practiced dividing a short text into phrases, shifting prominence and deaccenting, and realizing phrases only with two prominent words (peaks). The training *per se*, implemented as a PowerPoint presentation, could then begin. Participants listened repeatedly to each modified stimulus; they were instructed to listen as long as they wished, "to their ears' content". When the participants stopped playback, a 20-second silent period followed during which they were instructed to repeat the stimulus quietly and to savour the new prosodic patterns in their mind. This pause was introduced because we were interested in mindful repetition rather than "parroting" from short-term phonological memory. After the pause, the text of the stimulus reappeared, and they repeated it aloud as closely as possible. Each stimulus was then discussed with the experimenter: the participants would comment on what they perceived had been changed, check the phrasing or the prominent peaks etc. The aim was to reinforce the meta-awareness of English prosodic patterns, to facilitate the participants' noticing of prosodic features, and to combine this with the knowledge they had gained earlier. After the training, the participants were asked to read the diagnostic text again. Altogether, the entire first session lasted approximately 75 minutes.

Between the two training sessions, we prepared schematic visual displays of the manipulated prosodic patterns, because the objective of the second training session was to introduce multimodal aspects of pronunciation teaching. The visual displays resembled the contours in the Praat Manipulation window; however, they were usually somewhat simplified and made more intuitive (see Niebuhr et al., 2017 for a justification of using stylized pitch contours). An example is shown in Figure 4: the new melodic contour is shown in green (top panel), and the most significant temporal changes in purple (bottom panel).

The second session started with the explanation of the visual stimuli. Apart from incorporating the visual modality, participants were also encouraged to use gesticulation if it felt natural to them, to imitate the melodic movements or to indicate the rhythmic patterning; using gestures has been shown to facilitate learning L2 prosody (Li et al., 2023). In this training session, the participant would first only see the visual display and try to "read" it using the indicated prosodic patterns. Only then would the sound start playing, and they again listened repeatedly while following the visual display. A 20-second pause would again follow, during which they could also use the gestures along with the inner repetition; they then repeated the stimulus and discussed it with the experimenter, as in the first session. Finally, the participants were again asked to read the diagnostic text. Altogether, the second session lasted approximately 50 minutes.

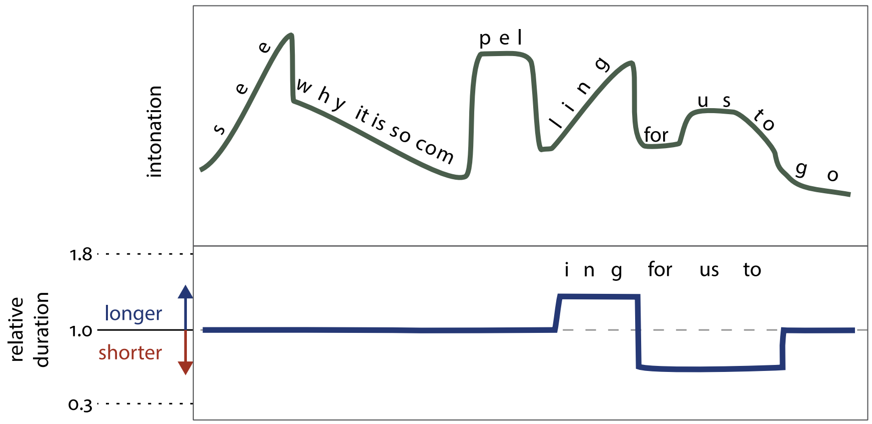

Figure 4. An example of the visual display of one stimulus (based on the recording of speaker F7)

A delayed post-test recording was obtained at least six weeks after the second training session, with the interval ranging between 42 and 58 days. The participants were asked to read the original diagnostic text, as well as a completely new text, which they could prepare in any way (e.g., by marking prosodic breaks or underlining prominent words). The analyses to be presented in the subsequent two sections will therefore be based on the comparison of three conditions: two correspond to the diagnostic recording (referred to as "text 1") obtained before and six weeks after the training, and this new text ("text 2") is the third condition. Both texts are provided in Appendix 1.

3. Perceptual assessment of competence

First, we were interested in seeing whether the training with their own manipulated voices would result in improved perceptual ratings of the participants. When listening to unfamiliar speakers, we form a range of rapid impressions: apart from comprehensibility, which has been researched extensively in the L2 context, studies have examined ratings of various personality dimensions, typically subsumed under the headings of "status" and "solidarity" (Lambert et al., 1960). In this study, we opted for a status-related dimension of competence; the extent to which a speaker is perceived as competent may have a large impact on their social interactions and success in life. We hypothesized that the speakers would receive higher competence scores in extracts from the recordings obtained six weeks after the training (whether text 1 or text 2) than in extracts from before the training.

3.1 Method

When identifying stimuli for the perception test, it was necessary to proceed as objectively as possible, without skewing the selection. That is why at least eight extracts of approximately four seconds were chosen as potential items for the perception test from each of the three recordings. Our aim was to include identical extracts from the before- and after-training recordings from text 1; that is why the criterion was that none of the potential extracts should contain major dysfluencies. In addition, the extracts had to correspond to syntactically and semantically meaningful stretches of text. Of the eight or more extracts per recording, the one to be used in the perception test was then chosen randomly. The perception test thus contained 36 target items (12 speakers × 3 conditions). Moreover, we included 45 filler items of similar nature and duration from other Czech speakers of English. The 81 items were divided into two blocks, with the target items assigned so that the extract of one speaker's before-training recording of text 1 always appeared in one block and the after-training of text 1 in the other block. Another four extracts were used as training items.

The test was designed in the Praat MFCExperiment environment. The respondents were instructed to imagine that they were choosing their personal spokesman and use a five-point scale to rate how competent each speaker sounded to them. They were allowed to repeat each item once. The perception test was administered to 39 first-year students of English and American Studies (29 females, 10 males, mean age 20.5 years, C1-level as per CEFR) at Charles University, before they received any instruction about the prosody of English. The respondents completed the perception test during a lesson in a quiet classroom, using their own laptop but identical high-quality headphones (Koss UR/20).

To assess the statistical significance of the effect of the prosodic training on perceived competence, we constructed linear mixed-effects (LME) models using R (R Core Team, 2017) and the *lme4* package (Bates et al., 2015). Competence score served as the dependent variable, CONDITION (text) as the fixed effect, and by-SPEAKER slope for the effect of CONDITION and LISTENER intercepts as random effects. We then used Tukey-adjusted confidence intervals of the estimated marginal means (emmeans) for pairwise comparisons 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 (Lenth, 2023); all plots were generated using the *ggplot2* package (Wickham, 2016). To assess inter-rater reliability, we computed the Intraclass Correlation Coefficient (ICC) using the *rptR* package (Stoffel, Nakagawa & Schielzeth, 2017).

3.2 Results and discussion

First, inter-rater reliability based on the ICC is excellent at 0.908, showing that the concept of perceived competence was intuitive for the respondents. According to the LME analysis (see Appendix 2 for detailed results), the effect of the prosody-focused training with participants' own modified voices on perceived competence is significant: $\chi^2(2) = 9.44$, $p = 0.009$. Figure 5, which shows the global results using emmeans, confirms our hypothesis that speakers in the extracts from the after-training conditions would be perceived by the respondents as significantly more competent than in those from the before-training condition: this is indicated by the nonoverlapping arrows. At the same time, there is no significant difference between the two after-training conditions. This result indicates that even a short training with their prosodic patterns manipulated using PSOLA can significantly improve the perception of the speakers.

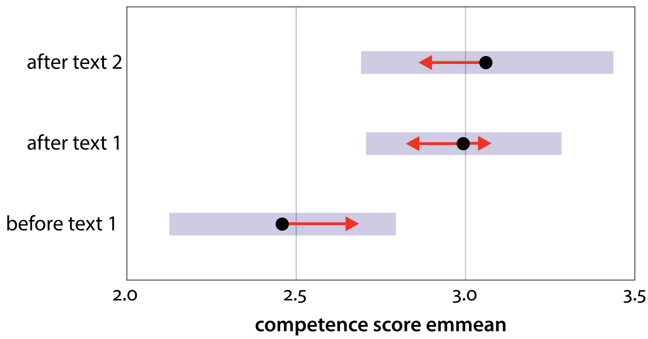

Figure 5. Estimated marginal means of perceived competence scores illustrating the effect of condition. Bars correspond to confidence intervals of the emmeans, and arrows serve for comparisons among them

It is to be expected that there will be differences in how individual participants are evaluated. The more detailed perspective is provided in Figure 6, where each speaker's perceived competence score is shown on the vertical axis and the error bars correspond to 95% confidence intervals. The individual results reveal a nearly categorical speaker gender difference: except for speaker F2, the female speakers received higher competence scores in at least one of the post-training conditions (and usually in both), while this is only the case in the male speaker M3. It will be interesting to compare whether a similar gender difference will be observed in the acoustic results (Section 5). Of course, one must keep in mind that a wider range of factors including voice quality are likely to affect the evaluation of speaker characteristics.

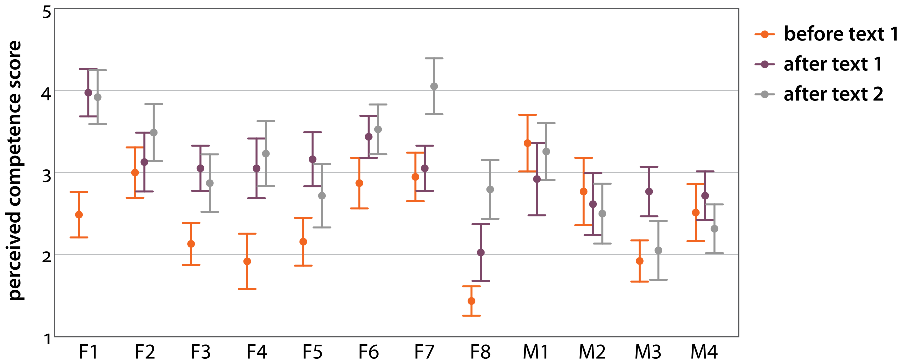

Figure 6. Perceived competence score of individual speakers in three conditions

4. Analysis of phrasing

Since the PSOLA-based training was centred around the concept of the prosodic phrase, our first aim as regards the participants' production was to examine their phrasing (that is, the placement of prosodic boundaries) and to see whether it is possible to build a model that would use phrasing information to predict the speaker competence scores obtained from listeners' assessments, as discussed in Section 3.

4.1 Method

First, we determined the optimal placement of phrase breaks from the texts; this was done by two experienced teachers of English, and the very few disagreements were discussed to arrive at the final version. Initially, four possibilities were distinguished:

- an obligatory break at sentence or clause ends
- a prosodic break realized in a place where this would never happen in fluent speech, for example between an article and the following noun or adjective (referred to as an "impossible" break)
- a prosodic break realized in an untypical but not impossible place, where this resulted in an impression of speech which is a little too emphatic (referred to an "untypical" break)
- a prosodic break which was possible but not obligatory in the given place

For instance, a prosodic break would be obligatory after the verb "*be*" in the sequence "*my answer must at once be, it is no use*"; a break would be possible after "*answer*"; and untypical and slightly emphatic breaks might be placed around "*at* *once*". As the obligatory breaks turned out to be realized nearly always by all the speakers, the modelling presented below will be based only on the impossible and untypical breaks.

The identification of prosodic phrases in the participants' speech was conducted by means of careful auditory analysis, observing ToBI conventions for analyzing prosodic break indices (Beckman & Ayers Elam, 1997). Major (break index 4, BI4) and minor prosodic phrases (BI3) were marked, but they were not distinguished in the subsequent analyses in this or the next section.

4.2 Phrasing score

The aim of the modelling was to arrive at a phrasing score which would best reflect the score of perceived competence. The phrasing score will be determined as the linear combination of the number of impossible and untypical phrase breaks, where an impossible break is assigned the coefficient of 1:

$$\textit{phrasing score} = c_1\ \textit{n\_untypical}+\textit{n\_impossible}$$

and an optimum value of the untypical break's weight c_1 will be sought using cross-validation, so that the mean squared error (MSE) of residuals of the linear fit between phrasing score and mean perceived competence rating is the lowest. To tune the model, we used the participants' phrasing in both conditions (before and after training) of text 1. Naturally, after the optimum weights have been identified, the behaviour will also be presented for text 2. The optimum value of the c_1 coefficient was sought in the interval $<0;\ 8>$ with a step of 0.025, using three repeats of a 10-fold cross-validation. The individual folds were generated using the *caret* package in R (Kuhn, 2008).

The relationship between MSE and the value of c_1 is shown in Figure 7. Using the above-described procedure results in the optimum value of $c_1 = 0.725$. It is clear from the very small range of the vertical axis, however, that the resulting scores and residuals are not very sensitive to the precise specification of the weights. This can also be explained by a relatively high correlation between the occurrence of both types of breaks: the Pearson's correlation of the number of impossible and untypical breaks in individual speakers is 0.78 for text 1 (note that both before- and after- training recordings are included), 0.47 for text 2 (on which the model was not trained), and 0.7 for all texts (conditions) together.

As mentioned above, this procedure allowed us to calculate a slightly more precise phrasing score for each participant in the three conditions, as depicted in Figure 8; the phrasing score is negative in the sense that the higher it is, the more phrase breaks were placed in incorrect or untypical positions. It can be seen that

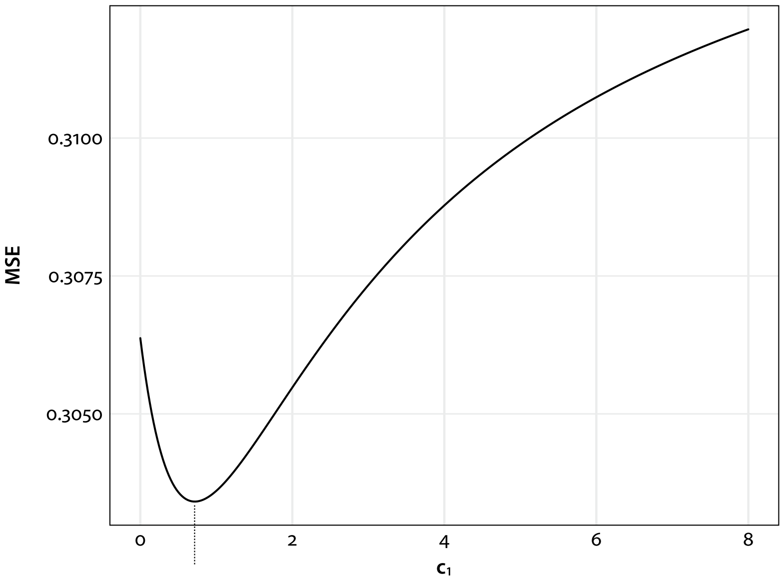

Figure 7. Relationship between the mean square error (MSE) and the value of coefficient c_1; the lowest value of c_1 is indicated using the dotted line

in most speakers — more specifically, with the exception of speaker M1 — phrasing score dropped quite markedly in text 1 between the pre- and post-training recording. The relationship between the pre-training phrasing score and the post-training one in text 2 is less clear-cut, with an improved quality of phrasing in some speakers (e.g., speakers F4, F6 or M2), very little difference in phrasing score in others (e.g., F1, F2 or M4), and a higher phrasing score in two speakers (F3 and F7). However, we may tentatively conclude that the training procedure, using PSOLA-manipulated phrasing, may indeed result in improved segmentation of speech into prosodic phrases.

Returning to the phrasing score, it should be pointed out that although the analysis yielded 0.725 as the untypical break coefficient, Figure 7 shows that the residual error remains robust for other weightings as well, not least due to the above-mentioned strong correlation between the occurrence of the two error types. This means that, for practical purposes, it would also be possible to use a simple sum of these two error types to obtain the phrasing score.

The Pearson correlation coefficient between the phrasing score and the perceived competence score is relatively weak (−0.37 for both conditions of text 1, −0.39 for text 2, and −0.32 for all conditions together). Of course, the phrasing score constitutes only a very rough indicator of the participants' prosodic behaviour; that is why the following section will focus on more detailed analyses in the melodic domain.

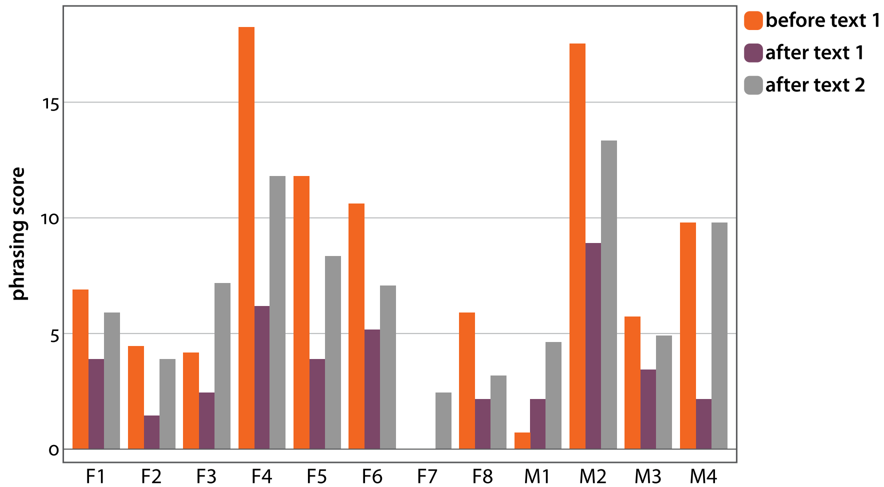

Figure 8. Phrasing score of individual speakers in three conditions

5. Production analysis

In the final type of analysis, we investigate how phrasing was signalled melodically before and after the training. Given the focus of the training, we were particularly interested in various measures of melodic variability. We hypothesized that the participants' melody would be more variable in the after-training conditions than in the pre-training condition.

5.1 Method

For the purpose of acoustic analysis, we only included "complete" phrases; in other words, phrases which included hesitations or other disfluencies were excluded from the analyses. For each prosodic phrase (see Section 4.1), we extracted a stylized fundamental frequency (f_0) contour. First, we used filtered autocorrelation in Praat with the ceiling of 500 Hz to generate the Pitch object, converted it to a PitchTier and then converted the values to semitones (ST) re 50 Hz using the rPraat package (Bořil & Skarnitzl, 2016). The stylization itself, which was performed so that the results correspond better to human perception (Hermes, 2006), depends on the duration of vowels. For vowels shorter than 100 ms, one f_0 value is identified as the mean f_0 of the entire vowel, placed in its temporal midpoint. For vowels longer than 100 ms, all f_0 values within the vowel were fitted with a linear regression model, which was used to predict two values, one at the vowel's onset and one at the offset.

Using these stylized melodic contours, we computed f_0 range (note that the stylization is quite robust with respect to the frequent problem with errors in f_0 extraction) and the so-called cumulative slope index (CSI; Hruška & Bořil, 2017). The CSI is a more sophisticated measure of melodic variability in that it corresponds to the sum of absolute frequency differences between subsequent f_0 points divided by the number of syllables; it therefore takes into account potential multiple melodic movements within a prosodic phrase.

To assess the statistical significance of the effect of the prosodic training on melodic variability, we proceeded in a manner similar to that described in Section 3.1: we constructed LME models with the melodic measures as the dependent variable, CONDITION (text) as the fixed effect, and by-SPEAKER slope for the effect of CONDITION as the random effect. Tukey-adjusted confidence intervals of emmeans were then used for pairwise comparisons with the alpha level of 0.05.

5.2 Results and discussion

LME analysis confirms our hypothesis, with both f_0 range ($\chi^2(2) = 24.77$, $p < 0.001$) and CSI ($\chi^2(2) = 25.42$, $p < 0.001$) significantly higher in the post-training conditions. Figure 9 shows the global results using emmeans for both measures of melodic variability. It can also be seen that both range and CSI are significantly higher (as shown by the nonoverlapping arrows) in text 2 (i.e., the new text) than in text 1; note, however, that the two texts are not directly comparable. It may be concluded that the training, in which pitch range was expanded using PSOLA and the participants were instructed to repeat these manipulated prosodic patterns, led to a significantly higher pitch range in the delayed post-test conditions.

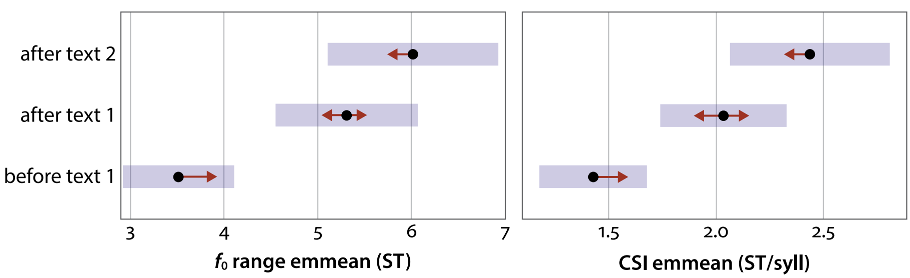

Figure 9. Estimated marginal means of f_0 range (left) and CSI (right) illustrating the effect of condition. Bars correspond to confidence intervals of the emmeans, and arrows serve for comparisons among them

Concerning the results of individual speakers, it was to be expected that some participants would change their prosodic behaviour more than others. Since the results are very similar for f_0 range and CSI, Figure 10 only shows the results of CSI. The boxplots confirm the overall tendencies, with melodic variability captured by CSI higher in most speakers in at least one of the after-training conditions and slightly higher in after-training text 2 than text 1, but the difference varies with speaker, and condition. Very little change in melodic variability may be observed in participant F3; it is interesting to note, however that her post-training recordings – while being similar in terms of melodic patterns to the pre-training recording – subjectively sounded considerably more fluent. There is therefore a marked improvement in her spoken performance, and it appears that listeners did react to that (*cf.* Figure 6). As for notable improvements in melodic variability, it is worth examining speaker M4 in more detail: he had extremely flat intonation before the training, with negligible intonation patterns even in nuclear contexts. The increase in CSI is not huge in absolute terms – it was not likely that he would completely overhaul the way he speaks – but relatively to his speaking habits we believe the change is remarkable. The greatest increase in melodic variability can be seen in speaker F1; notice that this is the participant in whom listeners perceived the greatest increase in competence (*cf.* Figure 6). A further comparison of perceived competence and melodic variability reveals similarities in most female participants. This was not the case with the male speakers, but we should point out that in three of them, the perceived competence score was relatively high already in the pre-training recording; there may thus have been a saturation effect, where the objective melodic improvement was no longer reflected in higher competence ratings.

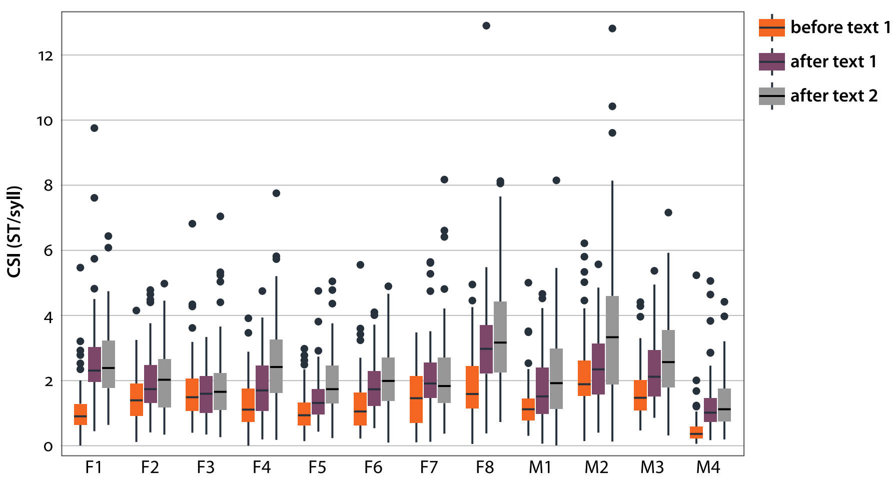

Figure 10. Cumulative slope index (CSI) of individual speakers in three conditions

6. General discussion

This paper has described a study centred on the training of English prosody, which was delivered individually to 12 native speakers of Czech. The cornerstone of the training consisted in using the participants' own voices, modified in their melodic and rhythmical properties using PSOLA, and repeating each stimulus after having listened to it multiple times. An awareness-raising session, focused on the phrase as the crucial unit of English prosody and on key differences in the prosodic patterns of the two languages, was an important component of the training. The comparison of the participants' reading before the training and six weeks after it documents its effectiveness: the participants were perceived as sounding significantly more competent (Section 3), the placement of prosodic breaks in their reading corresponded more with text-based predictions (Section 4), and their melodic variability was significantly greater (Section 5) in the delayed post-test recordings. Given the short duration of the training – approximately two hours spread over two sessions (see Section 2.3) – we believe that this convergence of positive results across three analysis domains is noteworthy.

Based on the results of the present study, we are convinced that targeted manipulations of learners' melodic and rhythmical patterns and their subsequent use for pronunciation training have great potential. Manipulations with PSOLA are implemented in an intuitive way in Praat (see Section 1.2 and Figure 2), which allows users to adjust the melodic and durational points carefully and listen to the result repeatedly to arrive at the intended prosodic representation. A detailed step-by-step video tutorial[4] was prepared by the first author of this study, so that language teachers can become familiar with the procedure. Such prosodic manipulations may be put to beneficial use in one-on-one teaching contexts, where the teacher would modify the learner's speech during lesson preparation and base pronunciation training on it in a future lesson. However, manipulating prosody may be instrumental even in classroom contexts: the teacher could take a student's homework recording, for example, and use it to illustrate key characteristics of L2 prosody by comparing it with the original rendition. Of course, this is based on the presumption that teachers are able to perform the manipulations, not only in the technical sense. Metalinguistic knowledge is also required; more specifically, they must be able to imagine what a given phrase should sound like: where should one insert a prosodic break, which word should receive nuclear prominence, which other syllables will be stressed, and which syllables need to be shortened. One of our future goals will therefore be to attempt to train teacher trainees to perform some elementary speech prosody manipulations.

4. https://fonetika.ff.cuni.cz/en/research/from-our-research/psola-modification/

Naturally, our study has its limitations. Most importantly, no control group was used to compare the results presented here with another training condition. Recruiting participants and only asking them to read text 1 and then both text 1 and text 2 six weeks later was not considered effective; it is extremely unlikely that the prosodic behaviour of such control group speakers would change in any meaningful way. We are convinced that the changes reported for our participants can only be due to the entire training procedure. In the debriefing discussions (which took place after the two training sessions and after the delayed post-test session), they reported that training with their own modified voice was a very positive and useful experience (*cf.* Henderson & Skarnitzl, 2022, where one participant's experience differed). Many of our speakers felt that they would not have improved as much if they had listened to and repeated after the voice of another speaker, and existing research does support the effectiveness of using one's own voice (see Section 1.2).

Another limitation is that we only examined the participants' performance in a reading task. It should be noted, however, that reading aloud, let alone in a foreign language, is not an activity that most of our participants would engage in on a regular basis; on the other hand, it would be unrealistic to expect major improvements in their prosodic behaviour in spontaneous speech based on such a short intervention. As mentioned in Section 2.3, the participants were instructed to make any notes they liked to prepare the new text (i.e., text 2). In the final debriefing session, after they had read both texts, most of the participants actually mentioned this ability to prepare a spoken performance in English (a presentation they may have to deliver, for instance) as the greatest benefit of the training. They believed that they now "knew what to focus on during the preparation" or "knew how to make a sentence sound the way it is supposed to" because of the training.

A crucial question in pronunciation pedagogy is the extent to which the gains observed in pronunciation training can be long-lasting (French, Gagné & Collins, 2020) and whether they can transfer to real-life situations. It is our hope that the participants' belief in their acquired ability to prepare the delivery of a presentation in English will stand the test of real-life communication contexts.

To conclude, this study has shown that using PSOLA to acoustically modify learners' melodic and rhythmical patterning resulted in improved prosodic performance of the participants, as well as perceptual assessment by listeners, demonstrating the potential contribution of targeted modifications of speech prosody in the teaching of L2 pronunciation.

Funding

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).

References

Anderson-Hsieh, J. (1990). Teaching suprasegmentals to international teaching assistants using field-specific materials. *English for Specific Purposes*, *9*, 195–214.

Baker, A.A. (2011). Discourse prosody and teachers' stated beliefs and practices. *TESOL Journal*, *2*, 263–292.

Beckman, M.E., & Ayers Elam, G. (1997). *Guidelines for ToBI labelling, version 3*. The Ohio State University Research Foundation.

Bergeron, A., & Trofimovich, P. (2017). Linguistic dimensions of accentedness and comprehensibility: Exploring task and listener effects in second language French. *Foreign Language Annals*, *50*(3), 547–566.

Boersma, P., & Weenink, D. (2024). *Praat: Doing phonetics by computer (Version 6.4)*. Retrieved from www.praat.org.

Bořil, T., & Skarnitzl, R. (2016). Tools rPraat and mPraat: Interfacing phonetic analyses with signal processing. In: P. Sojka, A. Horák, I. Kopeček & K. Pala (Eds.), *Proceedings of the 19th International Conference on Text, Speech and Dialogue* (pp. 367–374). Springer International Publishing.

Chafe, W.L. (1988). Linking intonation units in spoken English. In: J. Haiman & S.A. Thompson (Eds.), *Clause combining in grammar and discourse* (pp. 1–27). John Benjamins.

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

Croft, W. (1995). Intonation units and grammatical structure. *Linguistics*, *33*, 839–882.

Crowther, D., Trofimovich, P., & Isaacs, T. (2016). Linguistic dimensions of second language accent and comprehensibility: Nonnative listeners' perspectives. *Journal of Second Language Pronunciation*, *2*(2), 160–182.

Cutler, A. (2012). *Native listening*. MIT Press.

Dankovičová, J., & Dellwo, V. (1999). Czech speech rhythm and the rhythm class hypothesis. In: J.J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A.C. Bailey (Eds.), *Proceedings of the 16th International Congress of Phonetic Sciences* (pp. 1241–1244, San Francisco, August 1–7, 1999.

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

doi Derwing, T. M., Levis, J. M., Sonsaat-Hegeheimer, S. (2022). Bridging the research-practice gap in L2 pronunciation. In: J. M. Levis, T. M. Derwing, & S. Sonsaat-Hegelheimer (Eds.), *Second language pronunciation: Bridging the gap between research and teaching* (pp. 1–18). Wiley Blackwell.

doi Derwing, T. M., Munro, M. J., Foote, J. A., Waugh, E., & Fleming, F. (2014). Opening the window on comprehensible pronunciation after 19 years: A workplace training study. *Language Learning*, *64*(3), 526–548.

Dickerson, W.. (2019). The ripples of rhythm: Implications for ESL instruction. J. Levis, C. Nagle & E. Todey (Eds.), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching conference* (pp. 36–54). Iowa State University, September 2018.

doi Ding, S., Liberatore, C., Sonsaat, S., Lučić, I., Silpachai, A., Zhao, G., Chukharev-Hudilainen, E., Levis J., & Gutierrez-Osuna, R. (2019). Golden speaker builder — An interactive tool for pronunciation training. *Speech Communication, 115*, 51–66.

doi Eriksson, A. & Heldner, M. (2015). The acoustics of word stress in English as a function of stress level and speaking style. In: S. Möller, H. Ney, B. Möbius, E. Nöth & S. Steidl (Eds.), *Proceedings of Interspeech 2015* (pp. 41–45). Dresden, September 6–10, 2015.

doi Felps, D., Bortfeld, H., & Gutierrez-Osuna, R. (2009). Foreign accent conversion in computer assisted pronunciation training. *Speech Communication, 51*, 920–932.

doi Frazier, L., Carlson, K., & Clifton, C. Jr. (2006). Prosodic phrasing is central to language comprehension. *Trends in Cognitive Sciences*, *10*(6), 244–249.

doi French, L. M., Gagné, N., & Collins, L. (2020). Long-term effects of intensive instruction on fluency, comprehensibility and accentedness. *Journal of Second Language Pronunciation*, *6*(3), 380–401.

doi Gordon, J., & Darcy, I. (2016). The development of comprehensible speech in L2 learners. *Journal of Second Language Pronunciation*, *2*, 56–92.

doi Gordon, J., & Darcy, I. (2022). Teaching segmentals and suprasegmentals: Effects of explicit pronunciation instruction on comprehensibility, fluency, and accentedness. *Journal of Second Language Pronunciation*, *8*(2), 168–195.

doi Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech and Language, 25*, 601–634.

Henderson, A. J., & Skarnitzl, R. (2022). "A better me": Using acoustically modified learner voices as models. *Language Learning & Technology*, *26*(1), 1–21. http://hdl.handle.net/10125/73462

doi Hermes, D. (2006). Stylization of pitch contours. In: S. Sudhoff et al. (Eds.), *Methods in empirical prosody research* (pp. 29–62). De Gruyter.

doi Hickok, G. (2010) The role of mirror neurons in speech perception and action word semantics. *Language and Cognitive Processes*, *25*(6), 749–776.

doi Hruška, R., & Bořil, T. (2017). Temporal variability of fundamental frequency contours. *Acta Universitatis Carolinae — Philologica, 3*, 35–44.

doi Kolly, M.-J., Boula de Mareüil, P., Leemann, A., & Dellwo, V. (2017). Listeners use temporal information to identify French- and English-accented speech. *Speech Communication, 86*, 121–134.

doi Kügler, F., & Calhoun, S. (2020). Prosodic encoding of information structure: A typological perspective. In: C. Gussenhoven & A. Chen (Eds.), *Oxford handbook of language prosody* (pp. 454–467). Oxford University Press.

doi Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26.

doi Kusz, E. (2023). Effects of self-imitation practice on L2 pronunciation with the use of Golden Speaker Builder. In: R.I. Thomson, T.M. Derwing, J.M. Levis & K. Hiebert (Eds.), *Proceedings of the 13th Pronunciation in Second Language Learning and Teaching conference*, Brock University, June 2022.

doi Lambert, W.E., Hodgson, R.C., Gardner, R.C., & Fillenbaum, S. (1960). Evaluational reactions to spoken languages. *Journal of Abnormal and Social Psychology*, *60*(1), 44–51.

Lenth, R.V. (2023). *emmeans: Estimated Marginal Means, aka Least-Squares Means, v. 1.8.8*. Retrieved from https://CRAN.R-project.org/package=emmeans

doi Levis, J.M. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press.

doi Li, P., Baills, F., Alazard-Guiu, C., Baqué, L., & Prieto, P. (2023). A pedagogical note on teaching L2 prosody and speech sounds using hand gestures. *Journal of Second Language Pronunciation*, *9*(3), 340–349.

doi Mennen, I., & de Leeuw, E. (2014). Beyond segments: Prosody in SLA. *Studies in Second Language Acquisition, 36*, 183–194.

doi Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*(5–6), 453–467.

doi Munro, M.J., & Derwing, T.M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *49*, 285–310.

doi Niebuhr, O., Alm, M., Schümchen, N., & Fischer, K. (2017). Comparing visualization techniques for learning second language prosody: First results. *International Journal of Learner Corpus Research*, *3*(2), 250–277.

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

doi Phillips, S., Aguilar Perez, A., Alt, H., & Darcy, I. (2022). Pause for thought (groups): non-native pausing behavior and ease of processing of L2 speech. In: J. Levis & A. Guskaroska (Eds.), *Proceedings of the 12th Pronunciation in Second Language Learning and Teaching conference*, Brock University, June 2021.

doi Pickering, L. (2001). The role of tone choice in improving ITA communication in the classroom. *TESOL Quarterly*, *35*(2), 233–255.

doi Polyanskaya, L., Ordin, M., & Busa, M.G. (2017). Relative salience of speech rhythm and speech rate on perceived foreign accent in a second language. *Language and Speech*, *60*(3), 333–355.

Rogerson-Revell, P. (2012). Can or should we teach intonation? *IATEFL Pronunciation SIG Newsletter*, *47*, 16–20.

doi Saito, Y., & Saito, K. (2017). 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*, *21*(5), 589–608.

Saito, K., Trofimovich, P., & Isaacs, T. (2016). Second language speech production: Investigating linguistic correlates of comprehensibility and accentedness for learners at different ability levels. *Applied Psycholinguistics, 37*, 217–240.

Scherer, K.R. (2003). Vocal communication of emotion: A review of research paradigms. *Speech Communication, 40*, 227–256.

Skarnitzl, R., & Eriksson, A. (2017). The acoustics of word stress in Czech as a function of speaking style. In: F. Lacerda, D. House, M. Heldner, J. Gustafson, S. Strömbergsson & M. Włodarczak (Eds.), *Proceedings of Interspeech 2017* (pp. 3221–3225, Stockholm, August 20–24, 2017.

Skarnitzl, R., & Hledíková, H. (2022). Prosodic phrasing of good speakers in English and Czech. *Frontiers in Psychology*, *13*, 857647.

Stoffel, M.A., Nakagawa, S., & Schielzeth, H. (2017). rptR: repeatability estimation and variance decomposition by generalized linear mixed-effects models. *Methods in Ecology and Evolution*, *8*(11), 1639–1644.

Sundström, A. (1998). Automatic prosody modification as a means for foreign language pronunciation training. In: *Proceedings of ETRW on Speech Technology in Language Learning (STiLL)* (pp. 49–52, Marholmen, May 25–27, 1998. Retrieved from https://www.isca-archive.org/still_1998/sundstrom98_still.html

Šturm, P., & Lukeš, D. (2017). Fonotaktická analýza obsahu slabik na okrajích českých slov v mluvené a psané řeči [A phonotactic analysis of the content of syllables on word boundaries in spoken and written Czech texts]. *Slovo a slovesnost*, *78*(2), 99–118.

Trouvain, J., & Braun, B. (2020). Sentence prosody in a second language. In: C. Gussenhoven & A. Chen (Eds.), *Oxford handbook of language prosody* (pp. 605–618). Oxford University Press.

Van Maastricht, L., Zee, T., Krahmer, E., & Swerts, M. (2021). The interplay of prosodic cues in the L2: How intonation, rhythm, and speech rate in speech by Spanish learners of Dutch contribute to L1 Dutch perceptions of accentedness and comprehensibility. *Speech Communication*, *133*, 81–90.

Volín, J. (2019). The size of prosodic phrases in native and foreign-accented read-out monologues. *Acta Universitatis Carolinae — Philologica* 2/2019, 145–158.

Volín, J., & Poesová, K. (2016). Perceptual impact of speech melody hybridization: English and Czech English. *Research in Language*, *14*(1), 31–41.

Volín, J., Poesová, K., & Weingartová, L. (2015). Speech melody properties in English, Czech and Czech English: Reference and interference. *Research in Language*, *13*, 107–123.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. Available at: https://ggplot2.tidyverse.org/.

Appendix 1

A. Text 1: a slightly adapted extract from the famous quote of George Leigh Mallory

The first question which you will ask, and which I must try to answer, is the following: 'What is the use of climbing Mount Everest?' and my answer must at once be, It is no use. There is not the slightest prospect of any profit whatsoever. Oh, we may learn a little about the behaviour of the human body at high altitudes, we willingly put the active capacity of our lungs to the test, and possibly medical men may use our observations for the purposes of aviation. But otherwise nothing will come of it. Neither positive, nor negative. We shall not bring back a single bit of gold or silver, nor any coal or iron. We shall not find a single foot of earth that can be planted with crops. So it is no use, and you have to live with that. Nevertheless, if you cannot understand that there is something in man which responds to the challenge of this mountain, longs for it, and goes out to meet it, then you won't see why it is so compelling for us to go. What we get from this adventure is just sheer happiness. And happiness, after all, is the end of life. We don't live to eat and make money. We eat and make money to be able to live. That is what life means and what life is for.

B. Text 2: an extract from https://ideas.time.com/2013/12/30/new-years-resolutions-are-bad-for-you/, adapted by ChatGPT to a B1 level

Everyone has a mental list of habits they would like to change, and the New Year seems like a perfect time to start. "New Year, new you" is a phrase you will see repeated everywhere. However, just because it sounds right to your ear does not mean that it contains any meaningful truth. Just the act of making a resolution can make you feel better, but this tends to be only temporary. The statistics are depressing: only 8% of people who make New Year's resolutions stick to them, and those who don't usually abandon them after just one week! Unrealistic resolutions are doomed to fail. And it is unrealistic to think that you can immediately overcome a habit which you have spent years forming. But is this necessarily harmful? Actually, there's a good chance that it is. If your New Year's resolution is to eat less, but you have no plan in place – or even if you do have a plan and you fail – you will do damage to your sense of self-worth. Here's a better idea. Instead of listing an abstract goal like "lose weight", think of specific small steps you can take, every day, that will have the same result. If you fail at any of these small steps – which you inevitably will – brush it off, and realize that failure and recovery is part of any process.

Appendix 2. LME results related to perceived competence

Linear mixed model fit by REML ['lmerMod']
Formula: response ~ condition + (1 + condition | speaker) + (1 | ID)
REML criterion at convergence: 3928.7

Scaled residuals:

Min	1Q	Median	3Q	Max
-3.3860	-0.7112	-0.0536	0.7230	2.9364

Random effects:

Groups	Name	Variance	Std.Dev.	Corr	
ID	(Intercept)	0.1195	0.3457		
speaker	(Intercept)	0.2898	0.5383		
	conditionpost1	0.2862	0.5350	-0.65	
	conditionpost3	0.3353	0.5790	-0.41	0.68
Residual		0.8942		0.9456	

Number of obs: 1385, groups: ID, 39; speaker, 12
Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	2.4591	0.1707	14.404
conditionpost1	0.5345	0.1665	3.210
conditionpost3	0.6013	0.1783	3.371

Correlation of Fixed Effects:

	(Intr)	cndtn1
conditnpst1	-0.620	
conditnpst3	-0.415	0.653

Models:

```
fit_condition0: response ~ (1 + condition | speaker) + (1 | ID)
fit: response ~ condition + (1 + condition | speaker) + (1 | ID)
```

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
fit_condition0	9	3949.9	3997	-1965.9	3931.9			
fit	11	3944.4	4002	-1961.2	3922.4	9.4377	2	0.008925

Appendix 3. LME results related to melodic variability

A. Cumulative slope index (CSI)

Linear mixed model fit by REML ['lmerMod']
Formula: CSI ~ condition + (1 + condition | speaker)
REML criterion at convergence: 7514.5

Scaled residuals:

Min	1Q	Median	3Q	Max
-2.883	-0.609	-0.177	0.430	8.473

Random effects:

Groups	Name	Variance	Std.Dev.	Corr	
speaker	(Intercept)	0.1756	0.4190		
	conditionpost1	0.1367	0.3697	-0.20	
	conditionpost3	0.1124	0.3353	0.44	0.73
Residual		1.3586	1.1656		

Number of obs: 2368, groups: speaker, 12
Fixed effects:

	Estimate	Std. Error	t value
(Intercept)	1.4290	0.1280	11.167
conditionpost1	0.6050	0.1219	4.964
conditionpost3	1.0067	0.1133	8.885

Correlation of Fixed Effects:

	(Intr)	cndtn1
conditnpst1	-0.281	
conditnpst3	0.231	0.674

B. f_0 range

Linear mixed model fit by REML ['lmerMod']
Formula: max_min ~ condition + (1 + condition | speaker)
Data: tab
REML criterion at convergence: 10900.4

Scaled residuals:

```
Min     1Q      Median  3Q     Max
-3.1331 -0.6409 -0.0535 0.5861 4.5948
```

Random effects:

```
Groups   Name            Variance Std.Dev. Corr
speaker  (Intercept)     1.0201   1.0100
         conditionpost1 0.7837   0.8853   -0.05
         conditionpost3 0.8222   0.9068   0.36  0.81
Residual                 5.6618   2.3794
```

Number of obs: 2368, groups: speaker, 12
Fixed effects:

```
               Estimate Std. Error  t value
(Intercept)    3.5103       0.3038 11.555
conditionpost1 1.7956       0.2824 6.359
conditionpost3 2.5032       0.2881 8.689
```

Correlation of Fixed Effects:

```
            (Intr) cndtn1
conditnpst1 -0.129
conditnpst3 0.231  0.758
```

Models:

```
m1: max_min ~ (1 + condition | speaker)
full: max_min ~ condition + (1 + condition | speaker)
     npar AIC   BIC   logLik  deviance Chisq  Df Pr(>Chisq)
m1   8    10938 10984 -5461.0 10922
full 10   10917 10975 -5448.6 10897    24.767 2  4.186e-06  ***
```

Address for correspondence

Radek Skarnitzl
Institute of Phonetics
Faculty of Arts, Charles University
nam. Jana Palacha 1/2
116 38 Praha 1
Czech Republic
radek.skarnitzl@ff.cuni.cz
https://orcid.org/0000-0002-5682-1962

Co-author information

Tomáš Bořil
Institute of Phonetics
Faculty of Arts, Charles University
tomas.boril@ff.cuni.cz
https://orcid.org/0000-0002-5839-313X

Publication history

Date received: 28 August 2024
Date accepted: 16 December 2024
Published online: 3 February 2025