Language Learning & Technology
ISSN 1094-3501 CC BY-NC-ND

2022, Volume 26, Issue 1
pp. 1–21

ARTICLE

"A better me": Using acoustically modified learner voices as models

Alice J. Henderson, LIDILEM, University Grenoble–Alpes

Radek Skarnitzl, Institute of Phonetics, Charles University

Abstract

This paper presents the results of a brief mixed-methods intervention which sought to modify the production of prominence-related features in L2 English by four native French-speaking university lecturers, in read-aloud speech. Selected parts of participants' productions were acoustically modified and then used as the model in a Listen-and-Repeat protocol, where both quantitative (acoustic measures) and qualitative (free comments from discussion) data were collected. Acoustic measures were taken again from productions realized three months after the protocol, to trace longer term retention of modifications; expert listeners compared a selection of these productions to the original, diagnostic renditions, rating the degree of native-like rhythm and melody. Analysis of the quantitative and qualitative results confirm that imitating oneself can help individuals to modify prominence-related features of their pronunciation, that such changes can be retained over a 3-month period, but that people cannot reliably judge what they have modified. New potential is thus shown for Listen-and-Repeat, using one's own modified voice, as an effective technique in pronunciation instruction.

Keywords: *Pronunciation Teaching, L2 English, Prominence, PSOLA*

Language(s) Learned in This Study: *English*

APA Citation: 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

Introduction

Lecturers, researchers and administrative staff in European universities work with an increasingly diverse student body, one of the consequences of internationalization drives undertaken since the 1999 Bologna process. In this context, European bodies have tended to support Content-Based Instruction (CBI) both in secondary and tertiary education, as one way to facilitate the European ideal of integration and plurilingualism (Bonnet, 2012). In higher education, this has manifested itself in a trend towards English-medium instruction (EMI), resulting in field-specific content being taught in English by an increased number of non-native speakers of English. More generally, as student bodies have become more culturally and linguistically varied, more and more academic staff need to be able to communicate comfortably in English with a greater variety of speakers.

Some universities provide lecturers with excellent training to teach in English, including explicit work on producing speech which is intelligible (i.e., listeners can understand the message) and easily comprehensible (i.e., listeners do not perceive it as too difficult to decode someone's speech). Crucially, several studies show that the pronunciation of non-native teachers is important for students (e.g., Gorsuch, 2016; Kavas & Kavas, 2008), both in terms of performance and attitude. Research into the effectiveness of EMI-teacher training programs should therefore examine how speech can be modified and which features are likely to facilitate comprehension by students with enough English proficiency to study abroad in English, or to attend EMI courses in their own country.

In such contexts, aspects of prominence and rhythmic patterning may be regarded as high-value features.

We define prominence as relative acoustic and perceptual differences within a given unit of speech; for this paper, the distinction between word- and phrase-level prominence is not crucial (see Preparation of stimuli). The overall rhythmic patterning guides the segmentation of the speech stream in our native language (Cutler, 2012) and in other languages we may encounter (Levis, 2018). In addition, it has been shown that incorrect placement or phonetic realization of stress in English negatively affects intelligibility (Field, 2005). This applies not only to native listeners, but also to non-native ones, in international contexts (e.g., Lewis & Deterding, 2018). Importantly, many learners of English struggle with these aspects of speech, and French learners of English are no exception. Finally, prominence-related features of a foreign language are known to be teachable and learnable (Saito & Saito, 2017).

For these reasons, we devised a mixed-methods intervention as part of a Spoken English course for lecturers at a French university. Our objective was to improve their production of rhythmic patterning in read-aloud speech by using their own modified speech as auditory and auditory-visual feedback in a revamping of the traditional Listen-and-Repeat (LaR) paradigm.

In the following sub-sections, we summarize the prominence-related aspects of English and French, examine some key issues related to LaR, and explain the concept of the "Golden Speaker." We conclude this introductory section by presenting our research questions and hypotheses. The following sections describe the methodology behind this study and the results, respectively.

Prominence-Related Aspects of Speech

Prominence can be signaled by different acoustic cues – fundamental frequency (f_0), amplitude, duration, and vocalic formant structure – with the corresponding perceptual correlates being pitch, loudness, length, and vowel quality. However, languages differ as to the relative importance of these cues, and English and French belong among those where the difference is greatest (Frost, 2011), leading some researchers to talk about "stress deafness" in French learners (Peperkamp & Dupoux, 2002).

In French, stress does not have a distinctive function and it is fixed on the last syllable of the word, but the realization of stress depends on the position of the word within a phrase; hence, stress is regarded as a property of the accentual phrase (Frost, 2011; Jun & Fougeron, 2002). This phrase-level stress is manifested by longer duration and a salient pitch movement, which is typically rising when not at the end of an intonation phrase (Frost, 2011). In terms of its rhythmic patterning, French has been traditionally described as a syllable-timed language (e.g., Patel et al., 2006).

English is a language where lexical stress is not fixed to a specific syllable: although the majority of words in English are stressed on the first syllable (Cutler & Carter, 1987), English is typically described as having "free stress", where the rules for stress placement are rather complex. Lexically stressed syllables are characterized by a longer duration and a less steep spectral slope, as well as by a higher pitch level (Eriksson & Heldner, 2015). Concerning its rhythmic properties, English has been traditionally described as a stress-timed language (e.g., Patel et al., 2006). Although the concept of isochrony (whether syllable- or stress-based) is not without controversy, the fact remains that unstressed syllables tend to be temporally and spectrally compressed in English. As Low explains, "if learners aspire towards a globalist orientation, then stress-based timing should be taught" (2015, p. 133). Our EMI context is arguably more globalist than localist. Therefore, our study focuses on prominence-related aspects of English in native speakers of French. This includes the temporal patterning (i.e., the tendency for stressed syllables to be longer and for unstressed syllables to be shorter), as well as the melodic prominence of stressed syllables.

Listen-and-Repeat and Learning

Listen-and-Repeat (LaR) is a widespread technique in language teaching and was the dominant approach until the late 19th century. The technique exemplifies an Intuitive-Imitative Approach to teaching pronunciation (Celce-Murcia et al., 2010, p. 2).

Our study exploits the LaR technique in an innovative way, which was partly inspired by reviews of the challenges and advantages of Elicited Imitation (Jessop et al., 2007; Vinther, 2002). First, we imposed a

period of silence before the participants repeated the stimuli, in order to distinguish between mere repetition and mindful repetition. The latter involves *noticing* a feature, and noticing may have conscious and unconscious aspects, with both contributing to learning (for a review, see Ünlü, 2015). Therefore, we asked our subjects to comment on what they felt they had changed in their pronunciation; during this task, meta-phonological knowledge is being elicited, which may also facilitate better pronunciation. For example, Wrembel (2005) showed that meta-awareness raising interventions led to greater improvements in L2 pronunciation than instruction without metacognitive work. And more recently, Kennedy and Trofimovich (2010) found a positive correlation between the number of qualitative language awareness comments and higher pronunciation ratings. To summarize, the LaR technique was enhanced in this study by avoiding "blind" repetition, by urging the subjects to notice aspects of the model pronunciation, and by encouraging metalinguistic awareness comments.

Repeat After a "Golden Speaker" Model

As regards the model to be imitated, in the common classroom practice of LaR, the teacher typically serves as the model. This is problematic when teachers are not confident that they are a good model, and many teachers therefore use sound files as a substitute.

Inspired by the work of De Meo et al. (2013), Probst et al. (2002), and Wang and Lu (2011), we decided to have our participants imitate themselves. De Meo et al. (2013) summarize studies showing that the similarity of a teacher's and student's voice has an impact on pronunciation improvement, therefore arguing that "the most effective golden speaker to learn segmental and suprasegmental features of a second language is the learner's own voice with a native accent" (p. 90f).

The idea of using the learner's own voice as a model, with more target-like pronunciation, is not new. Felps et al. (2009) cite several earlier attempts at accent conversion, especially in the prosodic domain. Accent conversion refers to modifying those features of speech which contribute to foreign accentedness, while preserving the idiosyncratic voice quality of the speaker. A number of technologically advanced approaches have been proposed in recent years, some of them including not only the conversion of prosodic, but also segmental parameters: voice morphing (Aryal et al., 2013), FD-PSOLA (Felps et al., 2009), phonetic posteriorgrams (Zhao et al., 2018), or Sparse, Anchor-Based Representation (SABR; Ding et al., 2019).

Our approach in the current study differs from those mentioned above in that prosodic conversion was performed locally, on selected words and phrases, to target specific aspects of the learners' speech. More importantly, we are convinced that our approach may be applied by ordinary teachers with minimal training necessary[1]. It would be especially useful in high-stakes one-on-one teaching contexts, for instance with international executives or, as is the case in this study, university professors or international teaching assistants who need to teach in English.

Beyond the task design issues discussed above, a crucial issue is choosing what type of feedback should be offered. Since many studies show that auditory-visual perceptual training is more effective than auditory-only training, and that the combination of different modalities also facilitates transfer into production (Inceoglu, 2016; Motohashi-Saigo & Hardison, 2009; Olson, 2014), the current study involves elicited productions of prominence-related aspects of L2 English based on auditory-only and combined auditory-visual feedback, where the visual component corresponds to the display of temporal and melodic patterning in target phrases.

Research Questions & Hypotheses

The objective of this study is to report on a short experiment in which prominence-related features of pronunciation were modified to create "golden speaker" stimuli for a LaR protocol. L2 English speakers' productions were compared from before the pronunciation-focused intervention, and from after receiving auditory-only and auditory-visual feedback (note that the auditory-visual feedback will be henceforth referred to only as "aud-vis" in the Figures). In addition, we analyzed the participants' comments, which may reveal traces of cognitive load and its influence on the participants' reflections about the task.

Based upon our experience of teaching pronunciation to a variety of learners, we formulated three research questions:

1. To what extent is it possible for learners to successfully modify prominence-related features of their English by imitating a model consisting of their own acoustically modified voice?
2. Do learners retain these modifications for three months without further intervention?
3. Is there a difference between what learners want to modify and what they actually modify? In other words, does the awareness of one's own speech correspond to the changes made?

To answer these questions, we gathered data which is both quantitative (acoustic measurements and perceptual assessment) and qualitative (participants' self-reflections), and from three moments in time: six months before the LaR protocol, during it, and three months after.

Our hypotheses are:

1. Learners can modify the prominence-related features of their pronunciation by imitating their own manipulated speech.
2. Some of the modifications can be retained for at least three months after an intervention.
3. Participants will not always be aware of what they have modified successfully.

Method

In this section, the participants are described, followed by an explanation of how the stimuli were prepared for the LaR protocol. The section concludes by explaining the two-step protocol.

Participants and Their Spoken English Course

Four native French university lecturers (three female, one male; ages 30, 35, 51 and 52) participated in this study; this article uses fictional names for them. They attended a Spoken English course (7 × 3 hours) over four months, followed by a three day intensive course on CLIL[2]-style pedagogy. Their objective was to be able to teach their subject – Cultural Studies, Psychology, Law, or History – in English, and they were motivated to improve their language and teaching skills to meet that challenge.

At the start of the course, DIALANG's Web version[3] was used to evaluate their CEFR-levels: C1 in reading and B2 in writing. In addition, 14 teachers of English assessed the lecturers' accentedness and overall spoken proficiency on scales from 1 to 7, with 7 corresponding to not accented at all and extremely high oral proficiency. The mean accentedness and oral proficiency scores, respectively, are as follows: Anne (3.64 and 5.33), Françoise (2.50 and 3.93), Jacques (4.14 and 5.17), and Sophie (3.93 and 4.50)

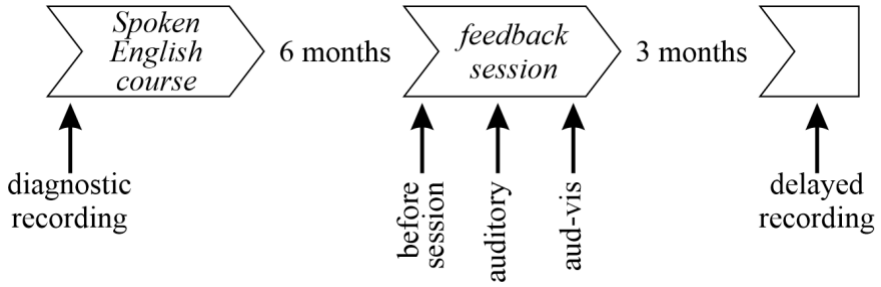
During the first class, they also recorded a diagnostic text from Dauer (1993) (see Appendix A). Then, based on that recording, they received feedback from the first author on the clarity and fluency of their pronunciation, as well as their phrasing, intonation, stress-related aspects and segmental features.

The course's guiding principle was Intelligibility (Levis, 2018). Each class was explicitly organized around the different levels of Gilbert's (2008) Prosody Pyramid, starting with pausing, chunking, and intonation, and then covering focus, stressed syllables, and peak vowels via pair and group work. Tips were also given related to articulatory setting when appropriate, using visuals, videos, gestures, and mirrors to compare French and English (Messum, 2017). Two participants did a presentation each week (5 to 15 minutes, about something from their field) and received immediate peer feedback as well as delayed feedback from the instructor on a variety of aspects, including gestures, eye contact, PPT slide design, lexis, grammar, pronunciation, and pragmatics.

The two-step intervention (described in the Listen-and-Repeat Protocol section), occurred six months after the Spoken English course had finished. Three months later, participants were asked to record the same text, to check whether there was any long-term retention of the temporal and melodic patterns. The timeline of the steps is shown in Figure 1, with each black arrow representing feedback and a recording being made.

Figure 1

A Schematic Timeline of the Study



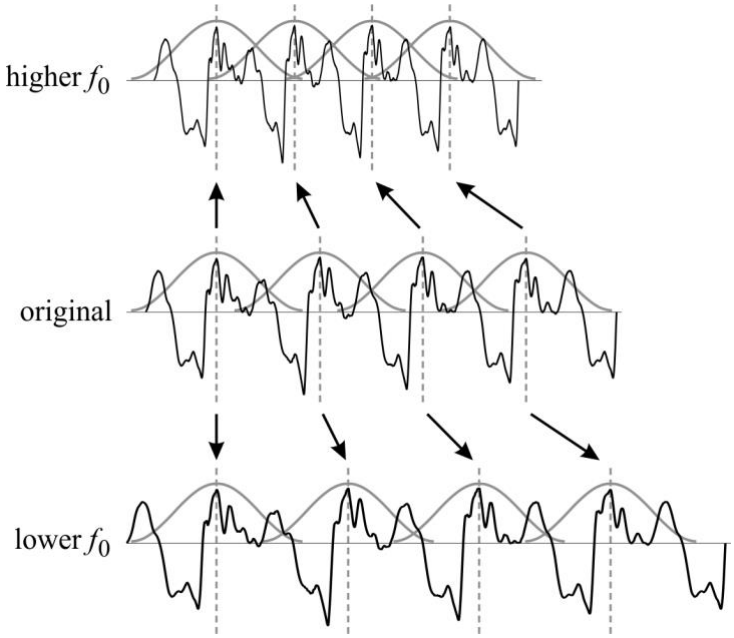
The first author taught the course and did the intervention sessions. Further details of the protocol are provided in the section Listen-and-Repeat Protocol, after the next section's explanation of how stimuli were prepared.

Preparation of Stimuli

To manipulate the prominence-related aspects of the participants' speech, we applied the TD-PSOLA (Time-Domain Pitch-Synchronous Overlap-and-Add) technique, as implemented in Praat (Boersma & Weenink, 2017). PSOLA is used for manipulating the f_0 of voiced (periodic) portions of the signal by overlapping short signal segments with different intervals (see Figure 2 for an example of raising and lowering f_0), or for manipulating duration by copying or deleting such segments and then overlapping them (Moulines & Charpentier, 1990).

Figure 2

A Schematic Display of Overlapping Speech Signal Segments Using TD-PSOLA

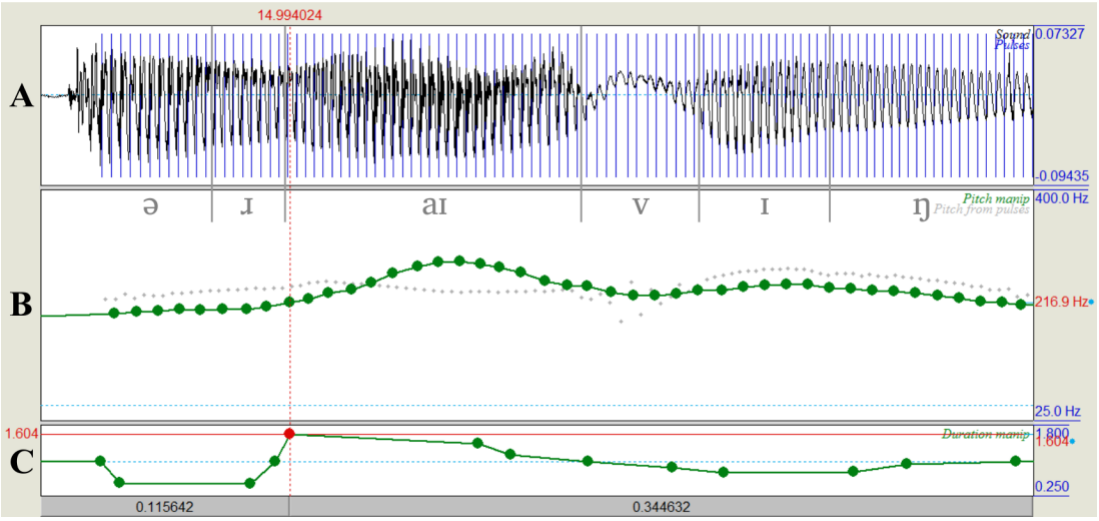


Note. The original waveform is shown in the middle panel, raised f_0 in the top panel, and lowered f_0 in the bottom panel. Identical points in the waveform (so-called pitch marks) are shown by the dashed grey lines.

Praat allows for easy manipulations of both f_0 and duration. However, relatively clean recordings are desirable for reliable manipulations; noisy recordings or recordings containing echo may render manipulations of especially f_0 difficult or even impossible. In addition, f_0 extraction may fail with creaky voices where periodicity is impaired or even absent. A screenshot of the Praat Manipulation environment is shown in Figure 3.

Figure 3

Screenshot of the Praat Manipulation Editor Window



Note. (A) = speech waveform of the word "arriving" in the top panel, (B) = f_0 manipulation panel in the middle, and (C) = duration panel at the bottom. Segment boundaries and labels have been superimposed.

Figure 3 shows that in this rendition of the word "arriving," manipulations have been performed in both the melodic and temporal domain. In panel B, the manipulated f_0 contour is shown in green, and the original contour is represented by the smaller grey dots: the rather flat melody with a minor melodic peak on the last syllable was thus replaced with a more prominent peak on the stressed syllable, also achieved by lowering of the neighboring unstressed syllables. In the manipulated version, the stressed vowel is nearly 6 semitones (ST) higher than the first vowel and nearly 3 ST higher than the last vowel.

In the duration panel C, the light blue dotted line corresponds to a relative duration of 1.0 (i.e., no temporal modification). Deviations above 1.0 indicate lengthening, those below 1.0 shortening. The figure shows the unstressed syllables considerably shortened (to about 0.5 in the first [ə] and 0.75 in the final [ɪŋ]) and the stressed vowel [aɪ] significantly lengthened. A closer look also reveals that the nucleus of the diphthong was lengthened more (the value of the point in red indicates a relative lengthening of 1.604) than its offglide, in line with the nature of English diphthongs.

From the diagnostic recordings, phrases containing between one and nine words were chosen by the second author for subsequent manipulations, based on two criteria. First, as mentioned above, PSOLA manipulation is rather sensitive to sound quality. Only such portions of the signal were therefore selected which allowed for successful modifications. Second, it was necessary to identify portions of the recordings where the L1 French effect on the L2 English prosodic realizations was obvious. The resulting manipulations had to be audibly different from the original renditions, so that the participants could pinpoint the differences during the intervention. In other words, the second criterion consisted in the potential benefit of the given phrase for the participants during the feedback sessions. It is thus a necessary feature of the design of this study that every speaker worked with a different set of phrases (although there was some overlap).

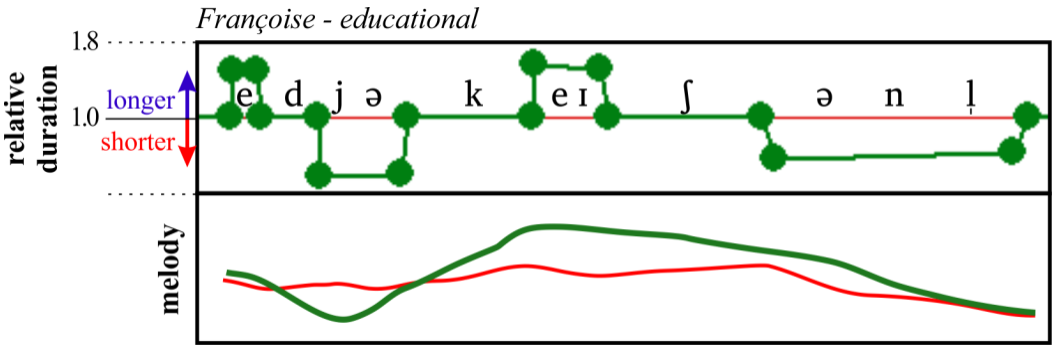
For three out of the four participants, seven phrases were chosen for manipulation, depending on the two

criteria mentioned above. For the last participant, it was only possible to select five target words or phrases. In total, therefore, 26 words or phrases were extracted for manipulations (see Appendix B).

The manipulations were performed by the second author with the aim of approximating native-like prosodic patterns (as shown in Figure 3). The naturalness of the modifications was then checked by the first author, a native English speaker and experienced teacher of L2 pronunciation. The final manipulated versions were subsequently used to create a PowerPoint presentation for each participant. Figure 4 shows an example with both temporal and melodic modifications.

Figure 4

Feedback Slide Example



Note. Original f_0 contour displayed in red, manipulated one in green.

The temporal panels were copied directly from the Manipulation Editor window, and the corresponding segment labels were inserted. To make the melodic contours more intuitive, these were created from smoothed and interpolated Pitch objects in Praat; smoothing serves to eliminate minor, imperceptible changes in f_0, while interpolation inserts f_0 values into voiceless portions of the signal, creating an uninterrupted contour, which better corresponds to how we perceive the pitch of the voice (see, e.g., Hermes, 2006).

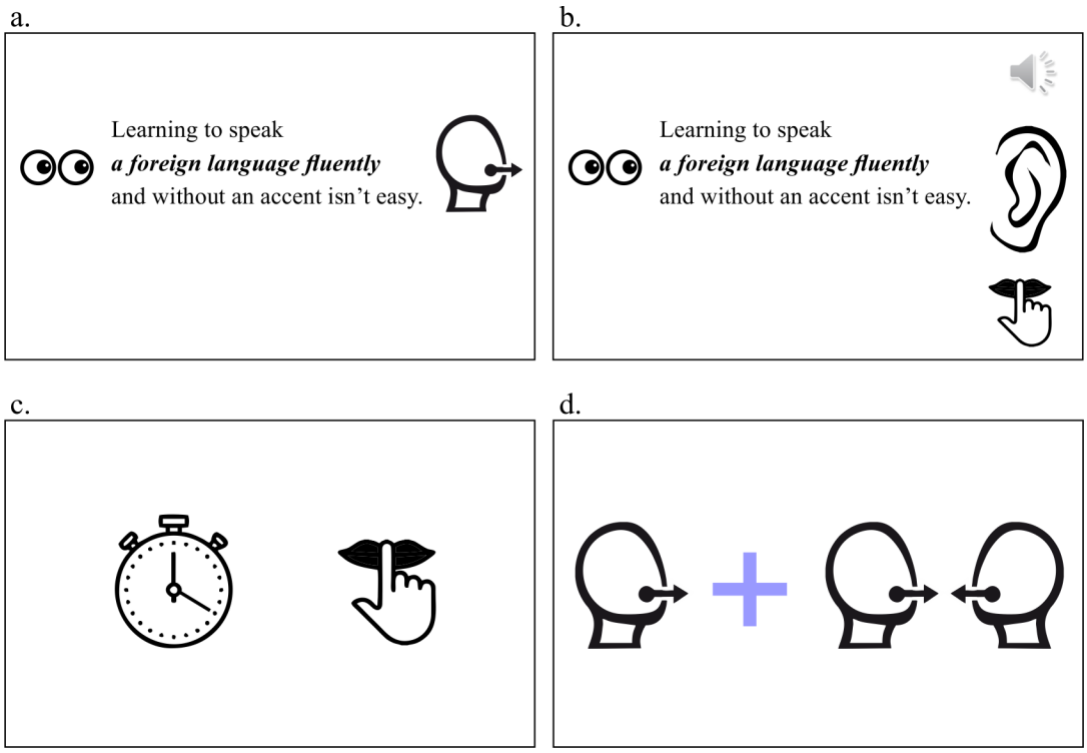
Listen-and-Repeat Protocol

The LaR protocol was delivered via a self-paced PowerPoint presentation on a laptop, with the participant wearing headphones for optimal sound signal quality. A portable digital recorder (MARANTZ PMD 561) was used to record the entirety of their imitations and free comments. Data collection sessions lasted between 15 to 60 minutes, depending on the amount participants talked.

The PPT slides were in the same format for each participant, even though each participant worked on a slightly different set of words and phrases. During the protocol, they could ask for clarification in French or English at any point. At the very start of the protocol, subjects were told that they would hear manipulated versions of their own voice. At the start of Steps 1 and 2, the four slides of Figure 5 and Figure 6 were used to show participants exactly what to expect.

Figure 5

Presentation of Stimuli in Step 1 (four example slides)

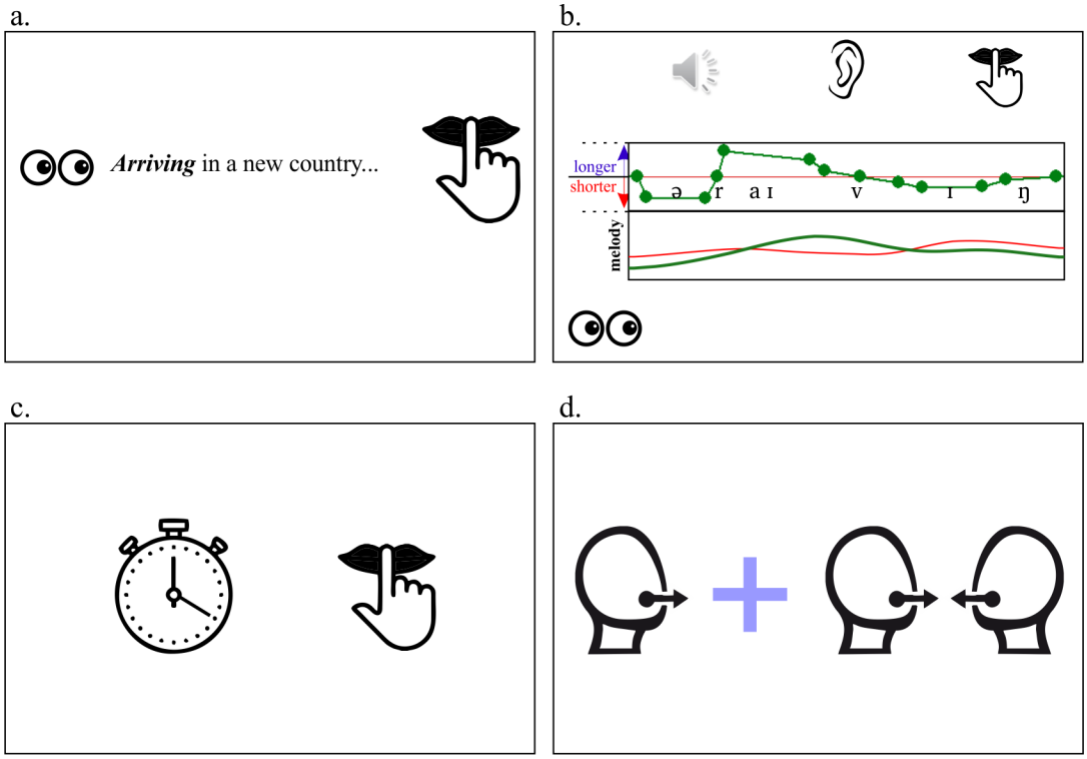


The schematic slides in Figure 5 show how the stimuli were presented in Step 1. First, the participant saw several lines from the initial diagnostic text and had to read aloud the highlighted section of that text (slide a; this part of the recording is referred to as "Before session" in the analyses). In the next slide (b), they said nothing but heard their manipulated voice reading aloud that section. They could listen as many times as they wished. Slide c indicated that they had to remain silent for 20 seconds[4] until the final slide (d), where they said the target word or phrase aloud again. Finally, they explained what they felt they had modified and what they had intended to modify; these explanations are the basis for the qualitative data.

In Step 2, visual information was added and briefly explained, and participants could ask for clarification. The slides in Figure 6 show how the stimuli were presented in Step 2. First, the participant remained silent and merely saw several lines from the text (a). In the next slide (b), they saw the Praat-generated visual and heard their manipulated voice reading aloud the highlighted section of that text; they could listen as many times as desired. Then they remained silent for 20 seconds (slide c), until the final slide (d), where they said the target word or phrase aloud once more. Finally, they again explained what they felt they had modified and what they had intended to modify.

Figure 6

Presentation of Stimuli in Step 2 (four example slides)



Results and Discussion

This section includes quantitative, qualitative and mixed data. First, we present the quantitative results of a perceptual evaluation of the diagnostic and post-feedback phrases by expert listeners. This is followed by acoustic and auditory analyses of participants' productions and comparisons of these realizations with their intentions (as expressed after each stimulus), in order to show speakers' immediate meta-phonological awareness of their modified speech. The section concludes with an analysis of participants' free comments from the end of Steps 1 and 2.

Perceptual Comparison

Ten expert raters, teachers of English pronunciation or phonetics, were asked to choose which recording had more native-like rhythm and melody, the diagnostic or the post-feedback recording; this was regarded as a reasonable request given the listeners' expertise. Included was the original recording and that of the two feedback (audio-only or auditory-visual) recordings. A self-paced 2AFC perceptual test was designed in the Praat ExperimentMFC tool, with the order of original vs. feedback recordings distributed randomly. The test was administered individually, using high-quality headphones. The respondents could replay the stimulus pair at will. The listening test thus comprised 26 comparisons and took approximately 10 minutes.

The results of the perceptual comparison are presented in Table 1 and in Figure 7. Table 1 shows that a post-feedback rendition was preferred more than the original rendition, although the preferences for the four speakers differed.

Table 1

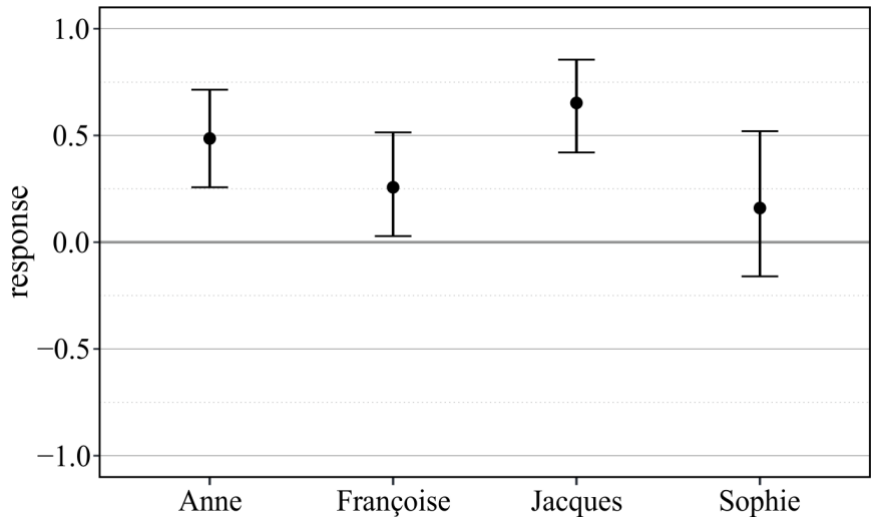
Preferences for Post-Feedback Renditions Versus Original Renditions by Speaker for One Utterance

	Preference for Post-Feedback	**Preference for Original**
Anne	52	18
Françoise	44	26
Jacques	57	13
Sophie	29	21

To assess the statistical significance of the perceptual comparisons, we employed the bootstrap method which is suitable for estimating the confidence interval of the mean value of a relatively small number of binary responses which are not normally distributed. The responses of post-feedback preference were recoded as 1, those of original preference as –1.

Figure 7

Confidence Interval Estimation of the Mean Responses for the Four Speakers



Note. (1 = post-feedback preference, –1 = original preference).

Figure 7 shows that, at the alpha level of 0.05, for three speakers (Anne, Françoise, and Jacques) the responses significantly favoured the post-feedback renditions; only for Sophie, whose confidence interval intersects the zero value, is the preference not statistically significant. It is important to realize that for Sophie, it was only possible to manipulate the temporal aspects of her speech due to the lower quality of the recording, particularly the presence of background noise (*cf.* the Preparation of Stimuli section). It therefore appears that her improvements were not noticeable to the listeners.

Acoustic Analyses and Speaker Awareness

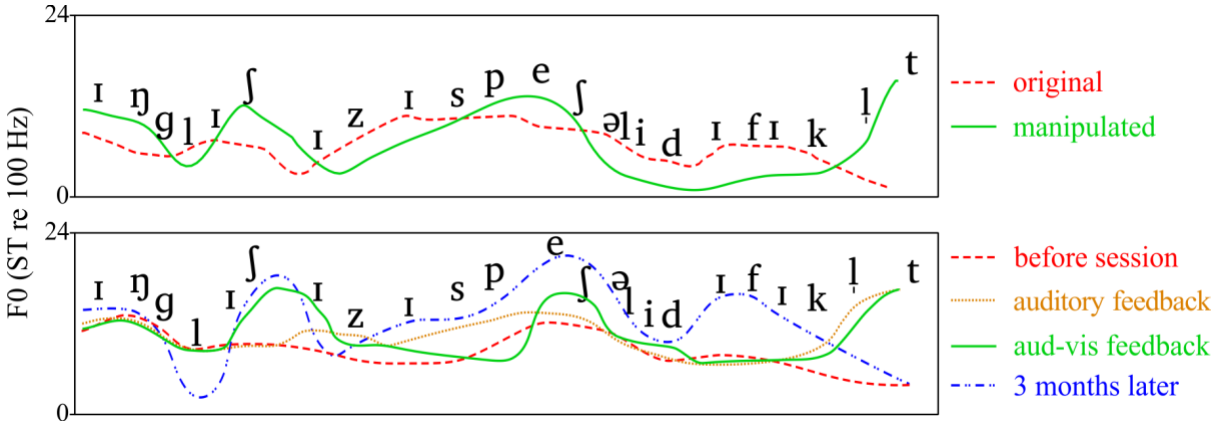
This part compares the individual renditions of selected words and phrases. The nature of the study does not allow us to present "hard" experimental data; however, the results nonetheless reveal interesting intra-individual modifications. As may be expected with this short intervention, long-term improvement (three months after the session) is less obvious, but we will provide examples suggesting that some generalization has been achieved. Each speaker's realizations will also be compared to what they felt they had modified,

to show their degree of awareness.

First, we provide examples of improvements in the melodic domain. Figure 8 shows the melodic realizations of Françoise's phrase "English is especially difficult".

Figure 8

Melodic Patterns in Françoise's Renditions of the Phrase "English is especially difficult"



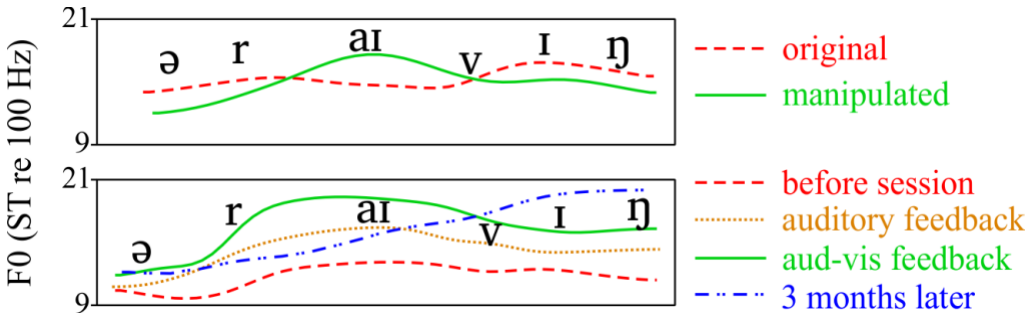
The top panel shows, in red, the original (i.e., diagnostic) realization with relatively flat intonation; the correct syllables were stressed in the lexical words. Fundamental frequency was manipulated (shown in green) so as to feature a falling-rising intonation on the word "English" in a separate prosodic phrase, and then another fall-rise on "(e)specially difficult" (since the original sentence continued with another clause), with a clear expansion of pitch range. In the lower panel, the red line shows Françoise's melodic pattern before the feedback session started, again relatively flat. The orange and green lines correspond to the renditions after the auditory-only and auditory-visual feedback, respectively. After audio feedback, the speaker realized a falling-rising melodic pattern on the latter part of the phrase, with a marked expansion of pitch range. The melodic realization after auditory-visual (indicated as "aud-vis" in all figures) feedback very much resembles the PSOLA-manipulated version, with clear imitation of both fall-rises and pitch range expansion even greater. The blue line shows how Françoise read the phrase three months after the feedback session; it is clearly visible that the stressed syllables are realized with clear melodic prominences and that the subject part of the phrase ("English") is realized with a very clear falling-rising pattern. The final tone was realized as falling.

In her recorded comments, Françoise indicated uncertainty about what she had modified: "I really don't know. Maybe the pronunciation of "especially". Maybe I accentuate the end of this word "difficult", but I'm not really sure." She is correct about "especially", in that she modified it in both feedback conditions. However, she did the exact opposite with "difficult", with analysis showing a (correct) lack of accentuation both at the end of the word and of the entire phrase.

Speaker Anne's pronunciation of a single word, "arriving", is shown in the melodic renditions of Figure 9. The original flat contour, where the last syllable had highest f_0, was manipulated so that the stressed syllable contained a clear melodic peak (see the top panel).

Figure 9

Melodic Patterns in Anne's Renditions of the Word "Arriving"

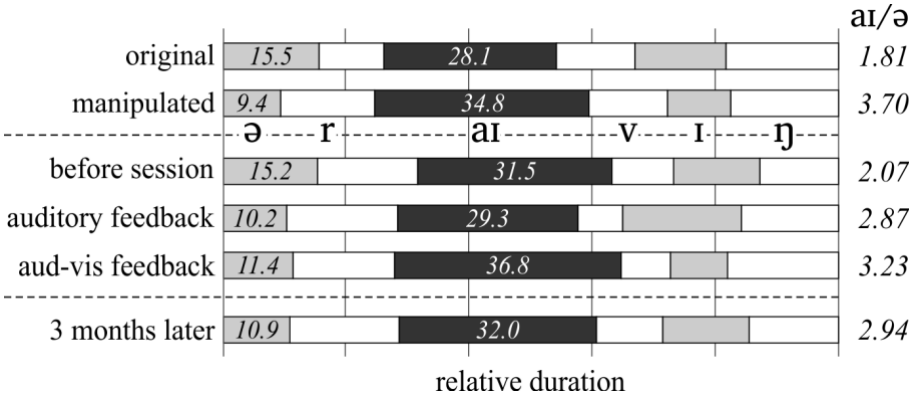


Before the feedback session, Anne did realize a melodic peak on the stressed syllable (red line in the bottom panel) but still the difference relative to the neighboring syllables was quite small (2.4 ST compared to the first and 0.7 ST compared to the third syllable). The difference is much greater with both auditory-only (5.5 and 2.2 ST) and auditory-visual feedback (7.8 and 3.1 ST). The blue line manifests a gradually rising intonation (as part of the phrase "arriving in a new country"), but the perceptual prominence of the stressed syllable is clear; the difference in f_0 between the central portions of [ə] and [aɪ] is approximately 3 ST.

Figure 10 turns to the temporal aspects but continues with Anne's word "arriving". The duration of the individual sounds, relative to the duration of the entire word, is indicated by the length of the rectangles. The absolute durations in milliseconds of the first two vowels, [ə] and [aɪ], are given in italics, and the temporal ratio of the stressed vs. unstressed vowel is shown on the right.

Figure 10

Relative Duration of Sounds in Anne's Renditions of the Word "Arriving" and Temporal Ratio of the Stressed and First Unstressed Vowel



The results confirm a tendency to approximate the manipulated version after feedback. While the stressed vowel is only approximately twice as long in the original diagnostic recording, as well as in the recording just before the feedback session, its relative lengthening is obvious in both feedback conditions, and in the retention recording three months later.

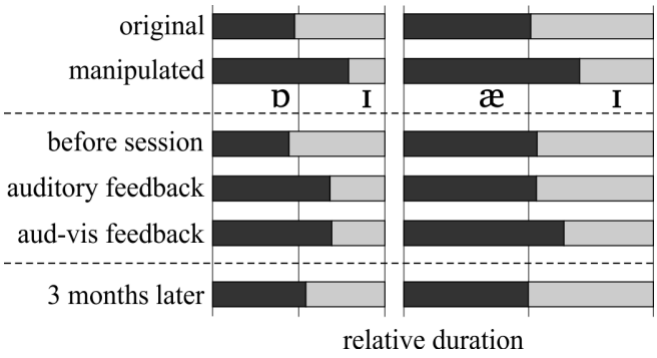
Anne's intentions are difficult to identify, as she states that she had "modified the start a bit, the same aspiration where the vowel starts"; her comments thus do not indicate an explicit awareness of what she had actually modified.

In Figure 11, only the relative vowel durations in Jacques' phrase "foreign language" are shown, because

the large differences in consonantal durations would have obscured the picture. As in the previous example, the manipulations consisted of lengthening the stressed vowels and shortening the unstressed ones.

Figure 11

Relative Duration of Vowels in Jacques' Renditions of the Phrase "Foreign Language"



The results show that in the word "foreign", there is considerable improvement in the temporal ratio in both feedback sessions, while only the auditory-visual feedback led to clear improvement in the word "language". However, a more native-like temporal patterning is not manifested in the delayed recording.

Jacques explained his intentions in French. Depending on how his adjectives are translated, his comments could be interpreted as showing awareness of the changes in vowel duration. He felt he had modified

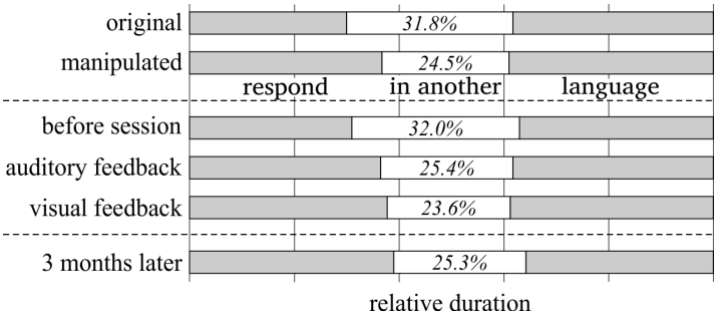
> …not necessarily much, but I think that the end of my word was very (aigue) sharp/high-pitched, lanGUAGE. I tried to be (plus clair) clearer, to articulate more the second time. But I don't know.

His reference to "language" in his comments makes the second syllable longer (indicated above in capital letters), and in the auditory feedback condition ("the second time") he did not manage to lengthen the stressed vowel: "clearer" might be referring to this non-lengthening, given his recurring insistence that it is correct for everything to be clearly articulated. This might also explain why he reverted to his original vowel durations in the retention recording three months later: the desire to speak "clearer" was deeply ingrained and required him to avoid lengthening or shortening.

Sometimes the temporal relationships are not particularly illustrative at the level of individual sounds, and it was more useful to take into account individual words; for example, Sophie's renditions of the phrase "respond in another language" is shown in Figure 12.

Figure 12

Relative Duration of Lexical and Function Words in Sophie's Renditions of the Phrase "Respond in Another Language"



Note. Percentage of the duration of the function words is indicated out of 100%.

The temporal manipulation concerned a relative lengthening of the first word, "respond" (especially the rhyme of its stressed syllable, [ɒn]), and a relative shortening of the function words "in another". The relative duration of the two function words with respect to the duration of the entire phrase is given as a percentage, showing that Sophie managed to modify both these aspects in the feedback sessions, as well as three months later.

It is noteworthy that Sophie noticed her modified pronunciation of "another", explaining that:

> The change is on the music of the rhythm. 'Another' was pronounced more rapidly than I did the first time. It's more musical, with ups and downs.

Participants' Free Comments

At the end of Steps 1 and 2, the participants were asked: "Try to explain what you tried to change or do differently." Table 2 shows the five categories of speech features which emerged from their comments.

Table 2

Number of Mentions, per Participant, of the Five Thematic Categories

Category of Speech Feature	Anne	Sophie	Françoise	Jacques
Accentuation, insistence	11	4	5	3
Duration, rapidity	4	7	8	4
Separation, linking	2	0	2	10
Melody, pitch, intonation	2	4	6	3
Rhythm	0	1	0	0

Rhythm was the prosodic feature which drew the least attention, mentioned only once by Sophie: "The change is on the rhythm."

Verbatim example comments from the other categories are provided in Table 3.

Table 3

Examples of Participants' Comments for Four of the Categories

Category	Example Comments
Accentuation, insistence	In French the accent is not like English should be. I accentuated it better. I increase the voice to accentuate it. I make the syllable more noticeable.
Duration, rapidity	It's shorter. I said it faster.
Separation, linking	It's more connected. It's not chopped up enough. I thought I had articulated more clearly. I tried to break it up into parts. I made it more linked "inanother" with no pause.
Melody, pitch, intonation	My voice goes up & down. My voice is more monotone. It's more musical (hand gesture).

The comments reveal that each participant tended to focus their attention on certain prosodic features, resulting in a sort of individual profile. Anne focused on "accentuating" certain syllables and words, to differentiate them noticeably from non-prominent ones. Jacques aimed for the opposite; he constantly wanted to make every syllable more separate and of similar clarity, which to him made them more identifiable. His overall, laudable goal was to make it easier for listeners to decipher the speech stream, to

eliminate "words that are too compressed" because they are pronounced too rapidly. Sophie and Françoise are similar in that they are more balanced, as the comparative range in the number of their comments is narrower than for the others. These two also commented more on duration and melody than Anne and Jacques[5].

In order to answer our research questions, four pre-determined themes were used to further analyze the participants' comments from the end of Steps 1 and 2:

1. Metacognition about the task
2. Ideas about "normal" or "correct" English
3. Perceived usefulness of the schemas
4. Attitude toward self-imitation: "a better me"

The comments will be discussed by referring to thematic categories (1, 2, 3, or 4) and speakers.

Category 1 comments about task metacognition show, among other things, that the participants are not unanimously in favor of having the visual information:

> It's rather difficult to change. It would be better if we could repeat with the picture right away. (Sophie)
>
> The second part was harder, maybe because I was concentrating more? (Françoise)
>
> The more words there are, the more complicated it is to read the visual. (Anne)
>
> It would have been easier to be able to repeat right away but that was forbidden. (Sophie)

A plethora of comments in this category belong to Jacques, who commented repeatedly about the overall nature of the task, expressing great discomfort and frustration:

> I'm not very convinced by what I hear. At the start, what I was hearing didn't match what I thought I had said, it was annoying. […] It's unsettling, we don't know if our way of pronouncing things is the right way, it really throws you. […] You think what you're doing is right for years and years and then, suddenly, well, you don't know if you should continue to do it that way or not. That's what's bugging me.

He was frustrated because the task instructions required him to contradict the way he had approached English for over 20 years; he seemed to assume that "proper" English required one to avoid connected speech phenomena and vowel reduction. He spent 9 minutes of an hour commenting solely on this aspect, quite obviously puzzled and almost angry until his cognitive dissonance was later resolved.

Evaluative comments about what is "normal" or "correct" English are grouped together in Category 2:

> My voice is more monotone normally, but I tried to make more of a difference. (Françoise)
>
> In English you have this noticeable "highlighting" which we don't have in French. (Anne)
>
> I wanted to make the end of the sentence more correct, more clear, so that you could hear THEM better (in the phrase "can't understand them"). (Jacques)
>
> It needs to be clearer, more "broken into parts" than in the modified voice. (Jacques)
>
> I heard blrrrrrr. Ooh la la, my pronunciation is so awful! (Jacques)
>
> There is almost no E pronounced. I think I have to eat the E (of "respond"). (Françoise)

The comments reveal some awareness of connected speech phenomena and include one comment about intonation, but none related to stress or rhythm. Only Françoise's comment about "eating the E" could be interpreted as referring to rhythm. At the very least, we can infer that she understands this is part of what spoken English involves: sounds and syllables can be modified or even disappear. This aspect is precisely what Jacques dislikes, referring to the blurry, acoustically modified stimulus created from his voice as "so awful!" Understandably, hearing a recording of one's own voice can provoke a powerful affective reaction,

and this can then influence one's ability – and sometimes even willingness – to do a speaking task.

Category 3 comments about the perceived usefulness of the Praat-generated schemas range from an enthusiastic "Oh yes, it helps!", to a less positive "not really, they were too strange" and even "I didn't really look at the melody line":

> Oh yes, it helps!! Especially to see where to shorten or insist. (Sophie)
>
> Not each time but yes, especially for the little words and to see which syllables to shorten. (Anne)
>
> Not really, they were too strange. Trying to connect my eyes and ears was hard. (Françoise)
>
> I didn't really look at the melody line. (Jacques)
>
> Not the melody line but definitely the first schema, longer-shorter, really helped, like a mnemonic technique. (Jacques)

One of the most intriguing comments is "Trying to connect my eyes and ears was hard." which could be considered as evidence of cognitive overload. However, the duration-related comments were quite positive.

Category 4 comments refer to the self-imitation aspect of the task and express some marked affective reactions, for example:

> Like a mirror, makes a good accent seem do-able, attainable. (Anne)
>
> Oh, it's horrible!! (Sophie)
>
> I sing so I'm used to hearing recordings of my voice but here, the texture has been "tampered with." (Jacques)
>
> After a bit we no longer realize that it's our own voice. I'm surprised because I really didn't recognize myself. (Françoise)

Françoise was surprised she could not recognize her own voice, Sophie found her own voice "horrible," while Anne described the experience of imitating herself positively:

> It's interesting to see the changes we could make, what 'a better me' would sound like, with a good accent.

This is what De Meo et al. (2013) refer to as the "golden speaker": an attainable model for imitation which spurs learners on, facilitating changes in pronunciation.

Conclusion

Our study reports on a teaching intervention which aimed at modifying the prominence-related aspects of four French lecturers' English pronunciation, using a LaR protocol. An important aspect of the study is that the participants' own speech, locally modified using PSOLA to approximate native English prosodic patterns, was used in the feedback intervention.

The nature of the intervention did not allow for an experimental study design. However, the results show that the feedback session did elicit more native-like pronunciation: the participants' production of temporal and melodic patterning related to English prominence did improve during the feedback session. This is confirmed by the expert evaluators (see Table 1 and Figure 7) and is illustrated with specific examples in Figures 8 to 12, thus supporting Hypothesis 1 (see Research Question and Hypotheses). It was to be expected, however, that in some cases participants' productions fell short of the manipulated goal. In the case of Sophie, the lack of significance can be attributed to the fact that only temporal (and not melodic) properties could be manipulated. It seems that melodic improvements are much more noticeable to listeners.

Despite the short duration of the LaR intervention, the results indicate some degree of carry-over three months later: learners seemed able to retain some of the modifications, lending support to Hypothesis 2.

The strong motivation of these learners is probably a key factor in their improvement. One could argue that since the diagnostic recordings were recorded before the Spoken English course, the improvements might come from the combination of both the course and the intervention. However, the six-month period between the course and the intervention is considered long enough to diminish short- or long-term learning effects.

The last hypothesis concerned the participants' metacognitive awareness of their pronunciation changes. The results show that the modifications they made in their pronunciation did not always match their post-hoc reflections. In other words, they did not always accurately identify what they had modified, confirming Hypothesis 3. Nonetheless, their comments yielded valuable and diverse insights, highlighting the fact that each learner is quite different. It would be interesting to explore whether more learners share the powerful wish, expressed by Jacques, to separate elements in spoken English to counter the reality of connected speech, which Cauldwell (2013) refers to as the "jungle" of authentic speech. The mismatch between the lecturers' metacognitive comments and their productions also highlights the role well-trained teachers should play in guiding learners to notice what is already good, what needs improvement, and how to make the necessary adjustments to achieve even more success. To that end, in a truly learner-centered instructional perspective with adults, part of that improvement process should include trying to elicit self-reflective or metalinguistic comments from learners. For example, anecdotal evidence shows that adult learners often feel that connected speech processes and rhythm-induced vowel reductions are somehow incorrect or sloppy. This clearly needs to be explicitly addressed if it prevents them from accepting these key features; Jacques' strong affective reaction led to a long discussion at the end of which he understood – and accepted – what adjustments he needed to make. Therefore, we have shown one way that reacting to or commenting upon one's own production can be a useful technique in pronunciation instruction, similar to using this technique in writing instruction.

The data obtained do not allow us to empirically compare the effectiveness of the auditory and auditory-visual feedback, but the results presented in Figures 8 to 12 do suggest a more faithful approximation of the models in the auditory-visual condition as compared to the audio-only condition. However, this may simply be due to more learning, rather than more effective learning, as auditory-visual feedback was always provided after auditory-only feedback. Just as importantly, the participants did not react in the same way to the different feedback types. Praat-generated visualizations of relative duration and pitch were appreciated differently. However, this would probably be the case with any visual, such as the animated sagittal vocal tract views on the Sounds of Speech app. Assessing the relative value of feedback modalities would require an experimental paradigm, with the order of feedback modalities counterbalanced and with the phonological and syntactic complexity of the stimuli controlled for; that was not viable in the current teaching intervention.

The intervention was based on a LaR protocol designed especially for this study, combining auditory and auditory-visual feedback based on the participants' own modified speech. Performing local manipulations of the speech signal to create the stimuli is relatively time-consuming and has limitations. However, given the results of this study, we are convinced that in an instructed setting with motivated, adult, non-native speakers who need to prepare for high-stakes situations, creating such a "golden speaker" model could be quite effective. Examples of such learners include lecturers teaching in English, but also businesspeople preparing for negotiations, airplane pilots and air traffic controllers, and even medical personnel joining international teams. Regarding language teachers, we feel that it would be empowering for them to create their own such models; the instructional videos recently developed should bring this within their reach.

To conclude, "a better me," as one participant termed it and which inspired the title of this article, not only has the potential to increase learners' motivation to improve their L2 pronunciation, it can also help to improve their prosodic performance.

Acknowledgements

The second author was supported from European Regional Development Fund-Project "Creativity and

Adaptability as Conditions of the Success of Europe in an Interrelated World" (No. CZ.02.1.01/0.0/0.0/16 019/0000734). We would also like to thank the anonymous reviewers for helping us to improve this paper.

Notes

1. To address the training issue, easy-to-understand instructional videos can be found at https://fonetika.ff.cuni.cz/en/research/from-our-research/psola-modification/.
2. CLIL stands for Content Language Integrated Learning and refers to learning situations where content is taught through a non-maternal language.
3. https://dialangweb.lancaster.ac.uk/
4. This imposed 20-second period of silence was meant to go beyond the influence of echoic or auditory memory (limited to 4 seconds) and give the participants enough time to move the sounds into short-term memory (limited to approx. 20 minutes). They were given 20 seconds to "play" with each utterance – articulating or rehearsing sounds or phrases (be it silent or voiced) – before repeating it aloud. Twenty seconds was chosen because each participant had to work through 5-7 target words or phrases per Step, allowing 2-3 minutes total to work through the different slides associated with each utterance. It was important to finish each entire Step within the 20-minute limit of short-term memory, so that they could remember enough to be able to comment meaningfully in the final slide.
5. The content of the PPT-slide prompts does not explain this difference.

References

Aryal, S., Felps, D., & Gutierrez-Osuna, R. (2013). Foreign accent conversion through voice morphing. In *Proceedings of Interspeech 2013* (pp. 3077–3081). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2013-191

Boersma, P., & Weenink, D. (2017). *Praat: Doing phonetics by computer* (Version 6.0.36) [Computer software]. www.praat.org

Bonnet, A. (2012). Towards an evidence base for CLIL: How to integrate qualitative and quantitative as well as process, product and participant perspectives in CLIL research. *International CLIL Research Journal*, *1*(4), 66–78.

Cauldwell, R. (2013). *Phonology for listening*. Speech in Action.

Celce-Murcia, M., Brinton, D., Goodwin, J. M., & Griner, B. (2010). *Teaching pronunciation: A course book and reference guide* (2nd ed.). Cambridge University Press.

Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words.* MIT Press.

Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, *2*(3–4), 133–142. https://doi.org/10.1016/0885-2308(87)90004-0

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, *11*(1), 51–62.

Dauer, R. M. (1993). *Accurate English: A complete course in pronunciation*. Pearson Education ESL.

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.

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. https://doi.org/10.1016/j.specom.2019.10.005

Eriksson, A., & Heldner, M. (2015). The acoustics of word stress in English as a function of stress level and speaking style. In *Proceedings of Interspeech 2015* (pp. 41–45). International Speech Communication Association. https://doi.org/10.21437/Interspeech.2015-9

Felps, D., Bortfeld, H., & Gutierrez-Osuna, R. (2009). Foreign accent conversion in computer assisted pronunciation training. *Speech Communication, 51*(10), 920–932. https://doi.org/10.1016/j.specom.2008.11.004

Field, J. (2005). *Listening in the language classroom*. Cambridge University Press.

Frost, D. (2011). Stress and cues to relative prominence in English and French: A perceptual study. *Journal of the International Phonetic Association, 41*(1), 67–84. https://doi.org/10.1017/S0025100310000253

Gilbert, J. B. (2008). *Teaching pronunciation: Using the prosody pyramid*. Cambridge University Press. https://www.tesol.org/docs/default-source/new-resource-library/teaching-pronunciation-using-the-prosody-pyramid.pdf

Gorsuch, G. (2016). International teaching assistants at universities: A research agenda. *Language Teaching, 49*(2), 275–290. https://doi.org/10.1017/S0261444815000452

Hermes, D. J. (2006). Stylization of pitch contours. In S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, & J. Schließer (Eds.), *Methods in empirical prosody research* (pp. 29–62). De Gruyter. https://doi.org/10.1515/9783110914641.29

Inceoglu, S. (2016). Effects of perceptual training on second language vowel perception and production. *Applied Psycholinguistics, 37*(5), 1175–1199. https://doi.org/10.1017/S0142716415000533

Jessop, L., Suzuki, W., & Tomita, Y. (2007). Elicited imitation in second language acquisition research. *Canadian Modern Language Review, 64*(1), 215–238.

Jun, S. A., & Fougeron, C. (2002). Realizations of accentual phrase in French intonation. *Probus, 14*(1), 147–172. https://doi.org/10.1515/prbs.2002.002

Kavas, A., & Kavas, A. (2008). An exploratory study of undergraduate college students' perceptions and attitudes toward foreign accented faculty. *College Student Journal, 42*(3), 879–890.

Kennedy, S., & Trofimovich, P. (2010). Language awareness and second language pronunciation: A classroom study. *Language Awareness, 19*(3), 171–185. https://doi.org/10.1080/09658416.2010.486439

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

Lewis, C., & Deterding, D. (2018). Word stress and pronunciation teaching in English as a Lingua Franca contexts. *CATESOL Journal, 30*(1), 161–176.

Low, E.-L. (2015). The rhythmic patterning of English: Implications for pronunciation teaching. In M. Reed, & J. Levis (Eds.), *The handbook of English pronunciation* (pp. 125–138). Wiley Blackwell.

Messum, P. (2017). Bringing the English articulatory setting into the classroom: (1) The tongue. *Speak Out!, 57*, 29–39.

Motohashi-Saigo, M., & Hardison, D. M. (2009). Acquisition of L2 Japanese geminates: Training with waveform displays. *Language Learning & Technology, 13*(2), 29–47. http://dx.doi.org/10125/44179

Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, *9*(5–6), 453–467. https://doi.org/10.1016/0167-6393(90)90021-Z

Olson, D. J. (2014). Benefits of visual feedback on segmental production in the L2 classroom. *Language Learning & Technology*, *18*(3), 173–192. http://dx.doi.org/10125/44389

Patel, A. D., Iversen, J. R., & Rosenberg, J. C. (2006). Comparing the rhythm and melody of speech and music: The case of British English and French. *Journal of the Acoustical Society of America*, *119*(5), 3034–3047. https://doi.org/10.1121/1.2179657

Peperkamp, S., & Dupoux, E. (2002). A typological study of stress deafness. In C. Gussenhoven, & N. Warner, N. (Eds.), *Laboratory phonology 7* (pp. 203–240). Mouton de Gruyter. https://doi.org/10.1515/9783110197105.1.203

Probst, K., Ke, Y., & Eskenazi, M. (2002). Enhancing foreign language tutors – In search of the golden speaker. *Speech Communication*, *37*(3–4), 161–173. https://doi.org/10.1016/S0167-6393(01)00009-7

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. https://doi.org/10.1177/1362168816643111

Ünlü, A. (2015). How alert should I be to learn a language? The noticing hypothesis and its implications for language teaching. *Procedia – Social and Behavioral Sciences*, *199*(3), 261–267. https://doi.org/10.1016/j.sbspro.2015.07.515

Vinther, T. (2002), Elicited imitation: A brief overview. *International Journal of Applied Linguistics, 12*(1), 54–73. https://doi.org/10.1111/1473-4192.00024

Wang, R., & Lu, J. (2011). Investigation of golden speakers for second language learners from imitation preference perspective by voice modification. *Speech Communication*, *53*(2), 175–184. https://doi.org/10.1016/j.specom.2010.08.015

Wrembel, M. (2005). *Phonological metacompetence in the acquisition of second language phonetics* [Unpublished doctoral dissertation]. Adam Mickiewicz University, Poznań.

Zhao, G., Sonsaat, S., Levis, J., Chukharev-Hudilainen, E., & Gutierrez-Osuna, R. (2018). Accent conversion using phonetic posteriograms. In *Proceedings of ICASSP 2018* (5314–5318). IEEE.

Appendix A. Diagnostic Text

Diagnostic Speech Sample, from Dauer, R. (1993)

Learning to speak a foreign language fluently and without an accent isn't easy. In most educational systems, students spend many years studying grammatical rules, but they don't get much of a chance to speak. Arriving in a new country can be a frustrating experience. Although they may be able to read and write very well, they often find that they can't understand what people say to them. English is especially difficult because the pronunciation of words is not clearly shown by how they're written. But the major problem is being able to listen, think, and respond in another language at a natural speed. This takes time and practice.

Appendix B. Phrases and Words Used as Feedback Material for Individual Lecturers

Speaker	Phrases and Words
Anne (7 phrases)	Foreign language fluently Educational Much of a chance to speak And respond in another language Arriving Especially difficult This takes time and practice
Jacques (7 phrases)	Foreign language Arriving Understand what people say to them Natural They can't understand English is especially difficult And respond in another language
Françoise (7 phrases)	Foreign language fluently Much of a chance to speak Being able to listen And respond Educational Can't understand what people say to them English is especially difficult
Sophie (5 phrases)	Educational Arriving English is especially difficult But the major problem is being able to listen And respond in another language

About the Authors

Alice Henderson is an Associate Professor at University Grenoble–Alpes in France, where she teaches English for Specific Purposes to STEM students. Her research focuses on the learning and teaching of L2 English pronunciation, the perception of foreign-accented speech, and English Medium Instruction (EMI).

E-mail: alice.henderson@univ-grenoble-alpes.fr

Radek Skarnitzl is an Associate Professor at Charles University in Prague. His research focuses on the impact of various pronunciation features on the socio-psychological evaluation of a speaker. He is also interested in the teaching of L2 pronunciation of a foreign language and in issues related to speaker identification.

E-mail: radek.skarnitzl@ff.cuni.cz