Topics in Linguistics (2022), 23(2), pp. 1-17

10.2478/topling-2022-0008

Effects of imitation and self-imitation practice on L2 pronunciation progress

Ewa Kusz*
University of Rzeszów, Poland

Abstract

The major aim of the current study is to verify whether an interdependence between self-imitation practice and L2 pronunciation improvement in the process of second-language acquisition is stronger than traditional imitation tasks. 35 Polish students of Applied Linguistics (at English level B2+) divided into two groups performed imitation and self-imitation exercises in order to improve their L2 pronunciation skills. Three acoustic parameters were considered, namely, articulation rate, speech rate and average syllable duration. The results of the research have revealed that there is a significant interdependence between L2 pronunciation improvement and self-imitation training in terms of speech rate. The outcomes of the research are in line with Ding et al. (2019), De Meo et al.'s (2013), and Felps et al.'s (2009) assertion that the better the match between learners' voices and their modified equivalents, the more positive impact there is on L2 pronunciation training.

Key words

imitation and self-imitation practice, L2 pronunciation improvement, L2 prosody

1. Introduction

Traditional methods of pronunciation teaching involve direct teacher-student interaction, and the teacher's immediate feedback on the student's speech production. This time-consuming and demanding approach is frequently neglected due to the insufficient training that teachers receive (Ding et al. 2019, Burgess and Spencer 2000) and their lack of self-confidence (Couper 2017). Moreover, there is a common belief among language teachers that the second-language (L2) pronunciation skill is of less importance than other skills and thus more suitable for self-study. Also, there is a common belief that an L2 learner's pronunciation improves by itself while acquiring a target language, thus it does not require teaching in the way that other skills do. This approach has developed from the main principles of communicative language teaching (CLT) which emphasize fluency over accuracy (Ding 2019; Levis and Sonsaat 2017).

There is a series of studies that support the view that second-language learners are more likely to achieve better results when they imitate a speaker with a voice similar to their own. Prior research also indicates that L2 self-imitation pronunciation training based on modification of students' voices and their adaptation to the prosody of an English native speaker's voice lead to beneficial results for L2 students who, consequently, can improve their foreign-language pronunciation skills. Starting from the 1990s, for the last few decades, the idea of automatic identification of foreign accents and the growing technological possibility to investigate the effects of students' self-imitation method using their own voices has attracted a large number of researchers who believe that prosodic (relating to the rhythm and intonation of language) and accent conversion can bring implicit, corrective and

* Address for correspondence: Ewa Kusz, University of Rzezsów, Kopisto 2B, Rzeszów, Poland. E-mail: ekusz@ur.edu.pl

encouraging feedback to L2 pronunciation learners (e.g. Repp and Williams 1987; Nagano and Ozawa 1990; Markham 1997; Munro and Derwing 1999; Eskenazi 1999; Bissiri and Pfitzinger 2009; De Meo et al. 2013; Pellegrino and Vigliano 2015; Ding et al. 2019 etc.).

According to Johansson (2008, p. 9), contrastive analysis is based on the comparison of two or more languages, with the aim of describing their similarities and differences. Being aware mainly of differences between L1 and L2 phonological systems is essential to avoid mistakes in pronunciation. Without such awareness, L2 learners tend to hear foreign-language structures in familiar ways, according to the categories with which they are familiar from their mother tongue.

Theoretically, the studies considering contrastive analysis can serve to specify how L2 learners accommodate variations in different phonological systems of L1 and L2, how practice of imitation/self-imitation methods enables L2 learners to acquire better L2 pronunciation skills, and whether (or to what extent) the effects of such methods differ for L2 pronunciation progress. Pedagogically, any effects of a prosodic-conversion technique (or any other technique that aims at improving L2 learners' pronunciation skill), such as L2 pronunciation progress in self-imitation practice, can serve to direct educational practice so as to promote this method in the L2 pronunciation learning process.

As a step towards the goal of investigating the considerable potential benefits of the use of computer-assisted training for L2 pronunciation improvement, the current paper describes the prosodic-conversion method used to modify and juxtapose accented utterances of Polish learners of L2 English to their prosodic-corrected equivalents. For the purposes of the current work, prosody is defined as "explicit characterisation of the length, pitch and loudness of the individual sounds" (see Hirst 2012, p. 15). In the present work, we investigate whether self-imitation practice enables L2 learners to improve their L2 pronunciation skills in terms of modified prosody of their recorded utterances.

Moreover, adjusting pitch curves to the model voice required modification of syllable durations. In modifying the phoneme durations and pitch contour of the participant's utterances to follow those of the model version, we considered three variables: speech rate, articulation rate and average syllable duration. The modified performances were obtained from the ProZed plugin implemented in Praat software, whereas the parameters were elicited from Praat algorithms. The significance of the current study lies in two facts. Firstly, it attempts to replicate to a certain degree research to corroborate or refute earlier findings, including Ding et al.'s (2019), De Meo et al.'s (2013), Felps et al.'s (2009) and Nagano and Ozawa's (1990). Secondly, it focuses on L2 English (stress-timed) and L1 Polish (a syllable-timed language) that, to the author's knowledge, has not yet been examined from the perspective of imitation and self-imitation tasks and their impact on L2 pronunciation skills. Finally, implementation of the ProZed plugin as a tool used for cloning the prosody of the model's utterance opens new paths to make the notoriously difficult evaluation of L2 prosody more objective.

The remaining sections of this article are organized as follows. Section 2 provides information about previous research on the discussed issue. In section 3 we present the purpose of our study and research questions and hypotheses. Methods, including participants, experimental sentences, procedures and data analyses are introduced in section 4. Results from this experiment are analysed in Section 5. The paper concludes with a discussion of the findings in section 6 and the theoretical and pedagogical implications of self-imitation practice in L2 pronunciation improvement in the final part, which discusses ideas for future research.

2. Previous research on imitation and self-imitation pronunciation practice

There are a handful of studies that have attempted to investigate the beneficial results of imitation and self-imitation of L2 pronunciation training as well as differentiated and evaluated traditional and modern (computer-assisted) methods. For instance, Repp and Williams (1987) proved that isolated L2 vowel production ([u]-[i] and [i]-[æ]) supported by the process of repeating students' own voices results in their better phonemic realizations than by using only a voice synthesizer.

On a larger scale, in their pioneering work Nagano and Ozawa (1990) used the prosodic-conversion method to evaluate the effectiveness of pronunciation training of Japanese learners. The method was based on the production of synthetic speech by converting such prosodic parameters as F0 contours and segmental duration of sonorants in subjects' speech into native English speakers' speech. The participants were divided into two groups among which the first one imitated an English native speaker's utterances, and the second repeated L2 structures produced by their own voices, previously changed so that they matched the prosody of the reference English speaker. The results were

subsequently assessed by English listeners. It turned out that the second group who imitated utterances recorded by their own voices achieved better results in terms of L2 pronunciation (they sounded more native-like).

Almost 12 years later Probst et al. (2002) showed the importance of the similarity of students' and teachers' voices in the process of L2 learning. Non-native speakers of English with various native languages, including Mandarin Chinese, German, French, Russian, Arabic, Thai, Bahasa Indonesian, and Spanish, were asked to choose and imitate their "golden speaker", that is, the one that matches their own voice. The results from their study revealed that the students' level of L2 pronunciation significantly improved if their teacher's voice ("golden speaker's voice") was well-matched to their own voices.

Also, Hardison (2004) proved that the use of a computer-assisted program helped L2 French learners to improve their pronunciation. In the pre-test–post-test designed experiment, 16 native English-speaking learners of French received three-week computer-assisted prosody training. Their pre- and post-training utterances were recorded and then categorized into filtered and unfiltered (unintelligible segmental information). Subsequently, they were presented to native speakers of French whose task was to rate the recordings in terms of prosody and segmental accuracy. The results revealed significant improvement of unfiltered utterances in prosody with generalization to segmental production and novel sentences. The main aim of the second part of the study was to analyse the relationship between prosody and the lexical content of utterances. By using filtered stimuli in a long-term memory task, the participants were asked to identify the exact lexical content of an average of 80% of the training sets based on prosodic cues that matched example-based learning models. The results of experiment 2 revealed that prosodic cues enabled the learners to facilitate the recall of the lexical content of the presented sentences.

Over the next few years there were a growing number of studies that focused on imitation and self-imitation experiments. Peabody and Seneff (2006) presented a pattern that previously appeared in Nagano and Ozawa's (1990) study. In order to teach L2 pronunciation more efficiently, they focused on tonal (Mandarin) and non-tonal (English) languages. By changing the pitch contour and adjusting L2 utterances of English speakers learning Japanese, they have shown that the modified L2 structures were twice as likely to be classified correctly by a pattern classifier.

Again, similarly to Nagano and Ozawa (1990) and Peabody and Seneff (2006), Bissiri et al. (2006) came to the same conclusion by investigating the process of L2 German pronunciation training on Italian speakers. Participants' own voices with corrected prosody resulted in the improvement of L2 pronunciation. A year later Huckvale and Yanagisawa (2007) attempted to build a spoken language conversion system based on accent-morphing correction in Japanese and English. They concentrated on phonetic differences between L1 and L2 and their impact on intelligibility. The results of their research have shown that there are segmental and suprasegmental differences, as well as segmental-suprasegmental interactions that should be accommodated in a spoken language conversion (2007, p. 69). Moreover, they proved that phonetic details need to be matched to the prosodic context. This, as they claim, guarantees a significant improvement in pronunciation training.

Accordingly, Felps et al. (2009) focused their research on full (prosodic and segmental) foreign accent conversion in computer-assisted pronunciation training (CAPT) by using an accent conversion method involving the use of a spectral envelope vocoder. Moreover, they paid attention to accentedness and identity rather than intelligibility of the utterances produced by the participants in the research (native speakers of American English). The proposed method was evaluated through a series of perceptual experiments, including a foreign-accentedness experiment, an acoustic quality experiment, an identity experiment and an identity experiment with reversed speech. As they claim (2009, p. 921), "it would be beneficial for L2 students to be able to listen to their own voices producing utterances in a native accent". As a consequence, the most effective "golden speaker" from whom to learn segmental and suprasegmental features of a second language is the learner's own voice with a native accent. With the use of PSOLA, they adapted the learner's voice to the teacher's in terms of speaking rate and pitch so that they reduced the foreign accent and preserved the voice quality of the participants in the research. The results of their study revealed significant interdependence between accent and identity.

The idea of a "golden speaker" in the process of SLA was also taken into account in the research conducted by De Meo et al. (2013) whose major aim was to investigate the effects of imitation and self-imitation on the acquisition of L2 suprasegmental patterns. They took into consideration four variables: speech act identification, communication effectiveness, degree of foreign accent and improvements in

intelligibility. They also focused on four speech acts: request, order, granting and threat. By using a rhythmic-prosodic transplantation technique and computer-assisted learning methodology (with the use of Praat), they were able to transfer the acoustic parameters of L2 learners and adjust them to an Italian native speaker's voice. Moreover, similarly to Felps et al.'s (2009) study, they did not change the segmental sequence or identity of the synthesized voice (De Meo et al. 2013, p. 91). The results of the research revealed that participants who performed self-imitation exercises achieved better results than those who imitated a native speaker's voice. Gorjian et al. (2013, p. 34) also pointed out that "the learners that practiced stress and intonation through CALL [Computer Assisted Language Learning] approach were more successful than the students who were taught through traditional method". They came to this conclusion after research based on an experiment with Iranian university students who were encouraged to improve their L2 English pronunciation skill with the use of Praat software.

Pardo (2010) suggests that side-by-side synchronous reading allows for the convergence of rhythmic patterns in speech. In her other works (e.g. Pardo et al. 2017), she highlights that phonetic convergence should be considered as a general phenomenon, which appears not only with regard to speech rhythm but also with the formant structure of vowels, pitch, tempo, etc. By giving a full review of research on phonetic convergence in speech imitation tasks, Pardo et al. (2017) distinguish Howard Gile's communication accommodation theory (Giles et al. 1991, Shepard et al. 2001) and the similarity attraction hypothesis. The latter assumes that interlocutors attempt to become more similar to those to whom they are attracted in terms of their speech, including accentedness, speaking rate, phonological variants, etc. Furthermore, Pickering and Garrod (2013) emphasize that the so-called motor commands used in the process of language comprehension can lead to imitation in speech production. These and many other approaches presented in Pardo et al.'s (2017) work suggest that, despite differences, they all propose that speech perception involves the resolution of phonetic form from vocal tract activity (Pardo et al. 2017, p. 638).

In one of the most recent studies on entrainment and speech rhythm convergence, Cerda-Onate, Vega and Ordin (2021) tested the effect of co-presence on entrainment to speech rhythm by examining the differences in speech rhythm convergence in a dyadic reading task. Following other studies (e.g. Cummins 2002, 2009), they point out that English native speakers achieve better rhythmic convergence when they work with recordings obtained in synchronous conditions than with recordings of individual readers (Cerda-Onate et al. 2021, p. 2). Hence, one of the conditions introduced in their study was to use a live speaker rather than a recorded model speaker. In their experiment, they asked a model speaker and participants of the study to sit side-by-side and read two texts aloud at the same time. Each speaker was recorded separately. The next step was to ask participants of the study to read and attempt to synchronize with the recording obtained from the recurrent model speaker while side-by-side reading. Finally, they asked participants to synchronize their reading with a solo recording obtained from the same recurrent model speaker. The reasoning behind this procedure and the choice of a live speaker was the assumption that only a live speaker can modify and synchronize their speech online to achieve better convergence with their interlocutor. The findings of their study revealed that the participants achieved better results in terms of synchronized and more regular, rhythmic speech with the co-presence of a live speaker who had the strongest modulatory effect. The results of this study may lead to a discussion of whether, or rather to what extent, imitation practice might be different when choosing a live partner as a model speaker or recorded utterances resynthesized with the voice of the participants of the study. Further discussion on this matter requires more specific research. This, on the other hand, gives ample opportunities to explore the phenomenon of imitation practice and phonetic convergence.

Ding et al. (2019) introduce an interactive tool for pronunciation training, called the Golden Speaker Builder (GSB). By mirroring an L2 learner's voice and adjusting it to a native accent, they showed that using GSB leads to improved L2 fluency and comprehensibility. In their paper, they described the overall system design, as well as the speech analysis and algorithms used to resynthesize L2 learners' voices. In a three-week experiment, preceded by a pre-test, the participants in the study, who were 15 Korean learners of English, practised the pronunciation of sentences that were created from the synthesis of the students' own voices with that of an English native speaker. After three weeks of training, the learners were asked to take part in immediate and delayed post-tests. The recordings were then rated in terms of their comprehensibility and fluency by a separate group of 95 native-English speaking

undergraduate students. The study has shown that the raters evaluated post-test recordings significantly higher than those in the pre-test. Moreover, the improvement in fluency was maintained, which was shown in the delayed post-test.

Most of the above-mentioned studies reveal the interdependence between self-imitation practice and L2 pronunciation improvement by analysing the results obtained from native-speakers' assessment. Following that, the current research project focuses on L2 pronunciation evaluation from a different angle. It implements automatic measurement of suprasegmental parameters with the use of Praat and the ProZed plugin, hoping to enrich the current state of research with new tools for pronunciation assessment.

3. Purpose of the study

As Goronzy et al. (2006, p. 309) point out: "To date automatic speech recognition has not advanced to the point where it can be an alternative for human scorers on important evaluations of spoken language proficiency". Bearing the above in mind, this paper attempts to fill the gap between L2 pronunciation learner (and teacher) needs and CAPT instruments which may give valid and reliable feedback that is so crucial in the process of second-language acquisition. Incorporating prosodic-conversion capabilities by using CAPT tools allows L2 learners to resynthesize their own utterances with a native prosody. This can be done either manually (Martin 2004) or automatically (GenevaLogic 2007), using algorithms.

The propositions introduced here constitute a challenging test of implementing CAPT tools, which are admittedly of limited accuracy; however, the same accusation may be directed toward the traditional methods of L2 pronunciation assessment. Thus, the present study was designed to first examine the relevance of automatic pronunciation assessment, second the interdependence between imitation and self-imitation practice and L2 pronunciation progress, third to find the possible theoretical and pedagogical implications of the obtained results.

Also, the main purpose of this study was to see the effects of two methods: imitation and self-imitation practice, and check which is more effective. It is worth noting that in the experiment a delayed imitation practice was implemented. We focused on the differences between the utterances made by the participants of the study in the pre-test and post-test not post-test and the model. Additionally, due to the fact that the students who performed the imitation task used a British-English native speaker as a model (see subsection 4.3.), it was not possible to compare the model voice to the post-training effects with the use of Wilcoxon signed-rank test, as it is implemented to conduct a paired difference test of repeated measurements on a single sample.

Most importantly, the model value of golden speaker was used to determine the deviation of the utterances performed by the students from the model value before and after training. It is worth noting that the pre- and post- values do not refer to the results the subjects achieved, but to the deviations from the model value.

3.1. Research questions and hypotheses

The study was guided by the following research questions:

1. What is the effect (if any) of using self-imitation method on learners' improvement of their L2 pronunciation?
2. Which method of pronunciation practice (imitation or self-imitation) is more effective in terms of L2 pronunciation improvement?
3. Which of the given variables has the strongest correlation with L2 pronunciation progress?

We assume that the above-mentioned variables should produce more positive effects on L2 pronunciation improvement among the participants who performed a self-imitation task. We also predict positive effects for the group who did only an imitation exercise at the time of study; however, as has been observed in previous studies, we expect a stronger correlation between L2 pronunciation improvement and a self-imitation task.

4. Methods

4.1 Participants

Of the 40 students who initially took part in the experiment, 35 were taken into consideration for the final results. The reason for such elimination was either technical or acoustic issues which resulted in the poor quality of the recorded sounds, or students' errors. These forced us to reduce the number of subjects. Participants in the study were thus 35 Polish students (31 females and 4 males) attending English Phonetics and Phonology classes at B2+ level of English. The age of the participants varied from 19-21 and they were students of the first year of Applied Linguistics. They had all had to pass an English exam at upper-intermediate level when they finished school. None of the participants reported having spent more than a few days in an English-speaking country prior to the study. We chose to use Polish speakers due to both the segmental and suprasegmental difficulties that arise in the process of acquiring English as a second language. Additionally, Polish learners appear to make similar types of mistakes in terms of pronunciation, regardless of levels of English proficiency. It should also be pointed out that the research was conducted deliberately on both women and men, without singling out one gender or collecting a sample that is gender-balanced. The main reason for this was the choice of algorithms which made the participants' gender irrelevant (see subsection 4.4).

4.2 Research design

First, the participants were asked to perform the imitation task (pre-test) by repeating six carefully chosen sentences from *English Pronunciation in Use: Advanced*; however, they were slightly modified. The six experimental sentences appear in Table 1. The reasons for modifying the original statements were twofold: (a) to simplify long and complicated sentence structures and (b) to include lexical items that contain sounds that were problematic for Polish learners of English.

Table 1. List of sentences used in the pre-training task.

Sentence number	**Sentence text**
1.	Don't jump to conclusions.
2.	They're putting a brave face on it, Mark.
3.	He's had a change of heart.
4.	You can say that again, Eva.
5.	You may well ask.
6.	He took them in his stride.

The statements were pronounced by six British English native speakers (3 females and 3 males). The "native likeness" or, in other words, the model version/golden speaker was based on the mean value from six English native speakers (see section 4.3).

Students' recorded utterances (the so-called pre-test) were randomly divided into two groups. One group, consisting of 20 pre-training performances, was then modified by Praat algorithms so that the prosody of L2 students' recordings resembled the prosody of L1 model structures (see subsection 4.4). The next step was to invite both groups to attend the second part of the experiment, in which the participants were asked to perform an imitation (or self-imitation) task. This part of the experiment took place over three weeks. During each week, the students were invited to join a 30-minute online session three times a week in which they practised the production of the sentences from Table 1. All in all, there were nine 30-minute meetings in a three-week training programme.

As previously mentioned, one group performed an imitation task, which was based on the process of repeating British-English model voices, whereas the second group focused on a self-imitation task. The participants in this group listened and repeated after their own voices were synthesized with the "model voice", so that they imitated the "golden speaker" voice, i.e. their own voice with a native accent. Regardless of imitation and self-imitation practice, two types of activities were provided in the training process: say-listen-repeat and listen-repeat. In the first type, the students were asked to utter a sentence then listen to the model (or golden speaker's) voice, and subsequently, repeat the same

sentence once again. In the second type of exercise, the subjects listened to the utterance pronounced either by a model speaker or a golden speaker and then repeated what they have just heard.

The participants were first urged to follow the instruction presented by the teacher before each session in week one. After becoming comfortable with both types of tasks, students could move to the next sentence. The participants' work was monitored by the teacher who was always present during the meetings. The students worked individually in online breakout rooms to have the best opportunity to work at their own time and pace. Following the three weeks of training, the participants in the study took part in the immediate post-test, which was given one week after the training. Their utterances were recorded once again in order to compare their outcomes to the pre-test results. The pre- and post-tests included the same sentences (see Table 1).

The sentences that were used in the study were chosen because they focused on evident differences between the phonetic inventory of the two languages. They all included the most problematic sounds, which either do not exist in the Polish phonological system or the place and manner of articulation differ from those in English.

Prosodically, the production of those intonational phrases reveals that Polish and English differ in the ways that they use intonation and stress. For instance, in sentence 1 we may distinguish the diphthong /əʊ/, and the mid-central vowel sound /ə/ which does not exist in the Polish phonological system. Also, the consonants /dʒ/, /ʒ/, as well as vowels /a:/ and /u:/ have different vocal realizations in Polish and English. In English, the sound /dʒ/ is classified as a voiced, affricate, post-alveolar consonant, whereas its Polish equivalent is voiced, affricate, but alveolar. English /ʒ/ is a fricative, post-alveolar sound, whereas in Polish it is defined as fricative, alveolar. In addition, there are long and short vowel sounds in English (e.g. /u:/), but there is no such distinction in Polish. The pronunciation of the short high front vowel [ɪ], which appears in the word "putting", is another sound tested in this study. The sentences were also chosen in terms of their rhythm and melody so that speech rate, articulation rate and average syllable duration could be carefully examined.

According to Crystal (1985, p. 266-267), [rhythm] "refers to the perceived regularity of prominent units in speech. These regularities may be stated in terms of patterns of stressed vs. unstressed syllables, syllable length (long v. short) or pitch (high v. low), or some combination of these variables". Speech melody or, to be more precise, intonation, is the melody or music of a language (Crystal 1985). Following that, the chosen sentences, though short, are fixed phrases or idioms, in which each has only one speech unit (if we do not include the names used in sentences 2 and 4). The choice of sentences was dictated by the fact that in each of them the regularities mentioned by Crystal can be distinguished. Moreover, since the selected sentences are fixed phrases and idioms, they are usually said together in one speech unit and with two prominent syllables.

Also, there is a long high front vowel [i:], which was already mentioned as a sound whose pronunciation is challenging for Polish learners of English, who usually attempt to replace it with Polish [i]. Instead of the [ɪ] sound, Poles pronounce the Polish front centralised vowel [y], which is very close to English [ɪ].

The first research group, consisting of 15 subjects, received a model version of a recorded text pronounced by an English native speaker, whereas the second one (20 subjects) listened to and repeated their own utterances that had been previously recorded and modified in terms of pitch and the duration of utterance. After the three-week pronunciation training course, the participants' utterances were recorded once again in order to compare their initial performances to the post-training ones. Finally, both the pre- and post-training recordings were compared and analysed with the use of relevant statistical tools.

4.3 Data analysis

Since the prosodic model to be used for the imitation training was offered by six British English native speakers, it was of crucial importance to select the most communicatively pertinent and accurate native male and female voices. In order to choose the most suitable native subjects for the imitation training, we decided to carry out a pre-test in which students from the imitation group was asked to listen to the recordings of the six sentences made by six British English native speakers.

The pre-test was based on some of the procedures introduced by De Meo et. al (2013) who also focused on the results of imitation and self-imitation practice in terms of L2 pronunciation progress. Following that, the students from the imitation group were asked to evaluate natives' utterances in terms

of three categories: communication effectiveness on a five-point scale (1=min, 5=max); degree of standard British English accent (i.e. standard accent, mild regional accent, strong regional accent); and intelligibility on a three-point scale (poor, sufficient, good). The native speaker utterances were administered in random order to the listeners of group 1. Following the results of the pre-test, it was possible to match the voice of a British English native speaker with a particular participant of the study. Taking into account the above-mentioned categories, each participant, after obtaining the results of the pre-test, was given one model speaker, based on which he or she performed the self-imitation training session.

In order to evaluate the prosody of L2 speakers, the study focuses on such acoustic parameters as speech rate, articulation rate and average syllable duration. Speech rate, as one of the factors that determines native-like pronunciation (see Munro and Derwing 2015), measures the rate of the phones and the pauses of an utterance. It has been shown that L2 learners speak at a slower rate than native speakers of a particular language (see Kowal and O'Connell 1980, Levelt 1989, Clark 2002, Ginther et al. 2010). This may result in less effective communication as the L2 speaker may not provide a message at the expected time. Articulation rate measures the amount of time the speaker requires while uttering words, including filled pauses but excluding silence (see Kusz 2019, p. 69). Average Syllable Duration (ASD), associated with the rhythm of a language, is calculated by dividing speaking time by the number of syllables in a given utterance.

The significance of the following study lies in the fact that it concentrates on L2 English, which is a stress-timed language and L1 Polish, a syllable-timed language. Both types of languages have yet to be studied in terms of imitation and self-imitation tasks and their impact on L2 pronunciation skills. The distinction between stress-timed and syllable-timed languages can be noticed when durational variability and the proportion of vocalic speech material are compared. It has been shown that languages that are normally classified as stress-timed, including English, exhibit greater durational variability and a lower proportion of vocalic speech material than syllable-timed languages, e.g. Polish (Bunta and Ingram 2007; Gervain, Nespor, Mazuka, Horie and Mehler 2008; Grabe and Low 2002; Nazzi, Bertoncini and Mehler 1998; Ordin and Polyanaskaya 2019; Payne, Post, Astruc, Prieto, and Vanrell 2012; Ramus et al. 1999; White and Mattys 2007).

The differences between these types of languages also appear in terms of articulation rate, e.g., English is spoken at a slower rate (≈4.2–5.2 syl/sec) than French (a syllable-timed language) (Anderson-Hsieh and Venkatagiri 1994; Clopper and Smiljanik 2011; Dauer 1983; Ordin and Polyanaskaya 2019; Pellegrino, Coupe, and Marciso 2011; Quene 2005; 2007). Moreover, the research has shown that L2 speech is generally slower than L1 speech, however, speech rate and articulation rate increase with the higher level of L2 proficiency (Anderson-Hsieh and Venkatagiri 1994; Ordin and Polyanskaya 2014; 2015). Also, a faster articulation rate in L2 is perceived as less accented (Kang 2010; Kang, Rubin, Pickering 2010; Munro and Derwing 1998; 2001), although it does not influence the degree of perceived foreign accent as much as, for example, pitch range or mean length of silent pauses (Ordin and Poluanaskaya 2019).

The variables were obtained from the prosodic conversion technique by implementing ProZed: a speech prosody editor for linguists (Hirst 2000, 2012, 2015; Zhi et al. 2016), available with the use of Praat (Boersma and Weenink 2001). The obtained results were then compared statistically to see whether there was any significant difference between pre- and post-training utterances. Since the research data consisted of repeated measurements on two samples uttered by the same subject to assess whether their population means differ, the Wilcoxon signed-rank test was used. It is a nonparametric statistical hypothesis test used when comparing two paired groups. Its main aim is to determine whether two (or more) sets of pairs differ from one another in a statistically significant manner. This paired difference test meets the needs of this study as it calculates the repeated measures in the given samples and shows whether the population means differ.

4.4 Prosodic conversion with ProZed

Our prosodic conversion technique is based on an approach proposed by Daniel J. Hirst (2015; 2016) who presented an automatic procedure in which the normalized pitch and syllable durations of a native-speaker's utterance are cloned and resynthesized to the L2 learner's recording.

ProZed is software implemented as a freely downloadable plugin to Praat (Boersma and Weenink 2001), the aim of which is to manipulate the prosody of utterances using the MOMEL and INTSINT

algorithms. The MOMEL algorithm used in the Praat ProZed plugin factors out the microprosodic effects of individual speech segments on fundamental frequency (F0) and models the underlying macroprosodic contour as a smooth contour curve (Hirst 2016, p. 1039). One of the main aims of using the MOMEL algorithm was to neutralize speaker-specific, and, consequently, gender-specific effects of pitch. Thus, as Hirst (2016, p. 1039) explains, the MOMEL anchor points were pitch-normalized with the use of the OMe (Octave-Median) scale, known as a natural scale for speech prosody (De Looze and Hirst 2014). With the use of the OMe scale, pitch values are denoted as the deviation in octaves from the speaker's median pitch (Hirst 2016: 1039). Therefore, the application of the MOMEL algorithm enables its users to conduct research in which the group of respondents does not have to be gender-balanced to receive reliable results.

As explained above, MOMEL provides an automatic stylization of F0[1] curves, whereas INTSINT (International Transcription System for Intonation) represents an intonation pattern as a sequence of "tones". As Hirst (2012, p.13) explains, ProZed adjusts the rhythmic and tonal properties of the rhythm and melody of an utterance by means of symbolic representation to evaluate the appropriateness of different phonological models of pitch and rhythm. Prosodic features are modified with the use of a Praat TextGrid. Hirst (2015, p. 2) implemented a British-English model of rhythm in the ProZed environment. As he explains, in order to model the speech rhythm, "the rhythm units are delimited by word boundaries and by the onset of the (primary or secondary) lexically stressed syllable". Hence, they correspond to narrow rhythm units and anacruses[2] (Hirst 2005, p. 2). This enables linguists to manipulate the rhythm and melody of an utterance. Rhythm is manipulated by factoring segmental duration, which consists of three steps: intrinsic duration determined by phonemic identity, local modifications encoded on the rhythm tier and global variations of speech rate encoded on the intonation tier (Hirst 2012, p. 13). Similarly, melody is manipulated by tonal segments on the tonal and intonation tiers. Using global parameters of key and span enables users to change the pitch of an utterance. In order to modify the prosody of a given utterance three interval tiers are used, namely the rhythm tier, the tonal tier and the intonation tier. The first two (tiers) are responsible for controlling the short-term variability of prosody, whereas the intonation tier controls longer term parameters.

As Hirst (2016) carefully explains, the first step was to obtain phoneme and word durations from an automatic alignment of the orthographic transcription with the acoustic signal (Hirst 2016, p. 1039). It was possible due to the use of SPASS software (Speech Phonetization Alignment and Syllabification) (Bigi and Hirst 2012). The next step was to group the phoneme durations into syllables with the use of Praat script, in which the maximum-onset principle was applied in the word boundaries. Following Hirst's (2016, p. 1039) detailed explanation, the word boundaries were calculated as the maximum number of consonants which were grouped into the onset of a syllable.

In order to compare intonation patterns between the model speaker and the recordings made by the participants in the study, the MOMEL algorithm has been applied. By factoring the raw fundamental frequency curves into macro- and micromelodic components, we received continuous and smooth curves which could be juxtaposed and compared. The obtained pitch contours were subsequently resynthesized. This process was based on the normalization of anchor points and the time of the recorded utterances. Syllable boundaries were used as the unit of normalization, which means that there was a fixed duration for each syllable.

By creating a Manipulation object in Praat, each syllable uttered by the participants in the study was modified by placing two points on the duration tier, the first 5 ms after the beginning and the second 5 ms before the end of each syllable. As Hirst (2016, p. 1040) lucidly explains, the value of the two duration points was the quotient of dur1/dur2, where dur 1 was the duration of L1 (model speaker), whereas dur2, the duration of L2 (the participant of the study). MOMEL anchor points were modified in a similar way, i.e. the F0 value was manipulated on the pitch tier to F01 * median1/median2. F01 is the F0 of the anchor point for the model speaker, whereas median1 and median2 are the F0 medians for both, the model speaker and the participant of the study. After such modifications, L2 utterances corresponded to L1 recordings.

[1] F0 (fundamental frequency) is the course of the lowest frequency in a harmonic vibration.

[2] Narrow rhythm units have a tendency to be equal in length, whereas anacrusis is one or more unstressed syllables at the beginning of a verse.

To visualize the process of resynthesis, the acoustic example of the speech signals are presented on spectrograms generated in Praat. The first figure shows the original recording made by a Polish native speaker. The second is the model voice, whereas the third figure depicts the recording after the resynthesis procedure, and this is the golden speaker's voice.

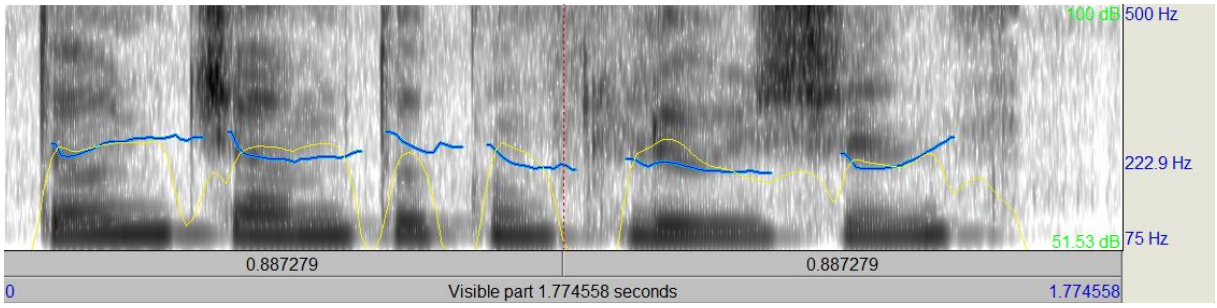


Figure 1. Recording of the sentence "Don't jump to conclusions" made by one of the participants in the study, whose mother tongue is Polish.

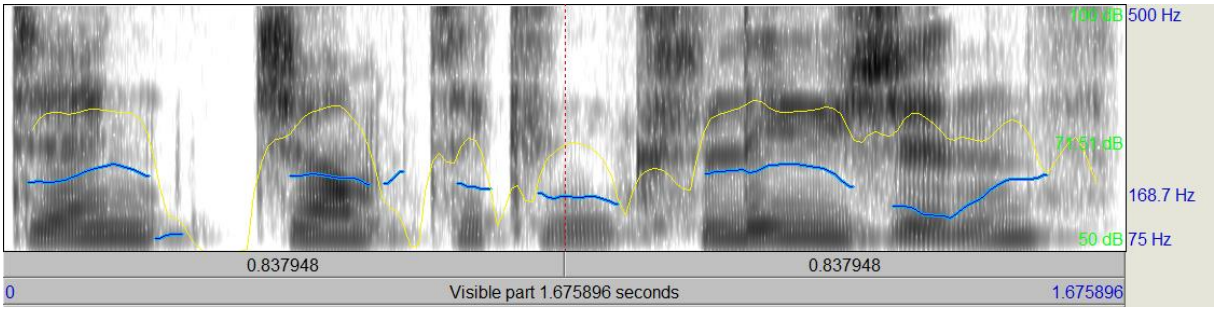


Figure 2. Model voice generated as the mean values of six British English native speakers. Sentence produced: "Don't jump to conclusions".

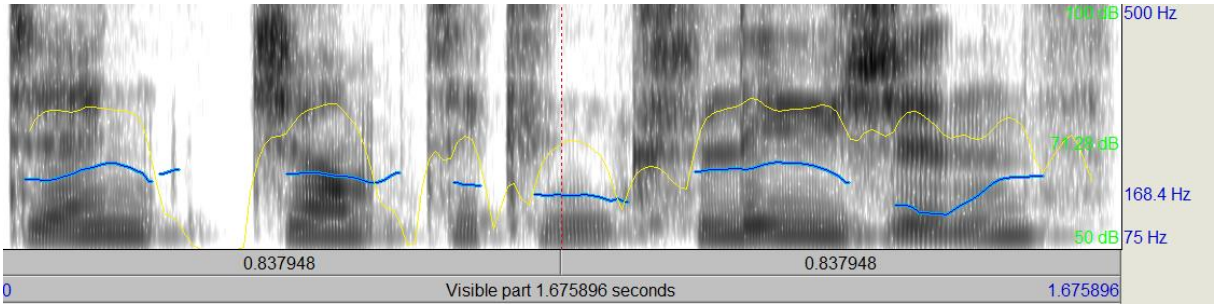


Figure 3. Polish synthesized voice, the "golden speaker's voice". Blue line: pitch contour; yellow line – intensity.

As can be seen in Figures 1, 2 and 3, the Polish original voice (Figure 1) received the same segmental duration, pitch movement and intensity contour of the English "golden speaker's voice" (Figure 2) resulting from the proposed method of rhythmic-prosodic transfer. The duration of the whole intonational phrase, as well as their segments in the "golden speaker's voice", strongly correspond to the English model utterance. By comparing spectrograms, we can easily notice that the pitch contour was perfectly matched to the model intonation.

5. Results

A total of 420 recorded utterances were obtained from 35 speakers who participated in pre- and post-tests. Among these, 120 recordings were resynthesized with the use of the ProZed plugin implemented in Praat. The obtained results reveal that there are significant differences in deviations from the model

value, especially for the Wilcoxon signed-rank test results. It should be noted that the numbers presented in Tables 2 and 3 show the difference between the golden speaker's voice and students' utterances. The smaller the value, the better result was achieved. In terms of imitation practice, the mean values were calculated as the difference between the model's voice and the student's utterances, whereas for self-imitation practice the difference between golden speaker's voice (resynthesized voice) and student's recorded utterances were calculated. As for Tables 4 and 5, we compare pre- and post-training results. The larger the difference between the given parameters, the better the result in L2 pronunciation improvement achieved.

Speech rate is measured in the total duration of an intonational phrase including pauses. Articulation rate is calculated as the quotient of the number of perceptually fluent syllables in each utterance and the duration (seconds) of the utterance, removing all disfluencies, such as repetitions, self-correcting, false starts and pauses greater than 250 milliseconds. Average syllable duration is the amount of speaking time divided by the number of syllables produced in an intonational phrase (Kusz 2019, p. 110).

5.1. Descriptive statistics for recorded sentences

Table 2 and Table 3 present the results based on the recorded sentences. Speech rate mean values differ from the self-imitation group (mean 0.55 and 0.98 respectively) with standard deviation (0.33 and 0.55 respectively) and the imitation one (mean 0.66; 1.04; SD 0.50; 0.56 respectively). The mean values for the articulation rate and average syllable duration are slightly higher for the self-imitation group (mean 0.47; 0.04 respectively) than for the imitation group (mean 0.43; 0.03 respectively). No interdependence can be noticed between the results obtained from the utterances given by the students who belonged to the imitation group. There is a small decrease in ASD values between pre- and post-training utterances (ASD 0.033; 0.038); however, the difference between these two variables is too insignificant to indicate any correlation.

Table 2. Descriptive statistics for pre- and post-training results of speech rate, articulation rate and average syllable duration (ASD) for the imitation group.

Variable	**Descriptive statistics – imitation task**				
	Number of participants	**Mean**	**Minimum**	**Maximum**	**SD**
Speech rate pre-training	15	0.66400	0.040000	1.64000	0.50015
Articulation rate pre-training	15	0.43067	0.050000	1.09000	0.37393
ASD pre-training	15	0.03367	0.004000	0.07400	0.02551
Speech rate post-training	15	1.04467	0.320000	2.26000	0.56982
Articulation rate post-training	15	0.54800	0.010000	1.71000	0.51490
ASD post-training	15	0.03867	0.000000	0.10200	0.03116

Table 3. Descriptive statistics for pre- and post-training results of speech rate, articulation rate and average syllable duration (ASD) for the self-imitation group.

Variable	**Descriptive statistics – imitation task**				
	Number of participants	**Mean**	**Minimum**	**Maximum**	**SD**
Speech rate pre-training	20	0.55750	0.00000	1.3300	0.33460
Articulation rate pre-training	20	0.47450	0.11000	1.2300	0.28445
ASD pre-training	20	0.04120	0.01100	0.0840	0.02236
Speech rate post-training	20	0.98800	0.11000	2.1900	0.55776
Articulation rate post-training	20	0.57900	0.05000	1.6400	0.47719
ASD post-training	20	0.04510	0.00400	0.1560	0.03861

5.2. Wilcoxon signed-rank test results

The main aim of the use of Wilcoxon signed-rank test is to compare two sets of scores that come from the same participants from one time point to another. The purpose of this study was to investigate changes in the results obtained from pre- and post-tests, thus we chose to analyse the given data using a Wilcoxon signed-rank test. Moreover, the normal approximation of the Wilcoxon signed-rank test tests the hypothesis that the distribution of differences has a median of zero (Riffenburgh 2006: 282-283). There were seven steps in performing the Wilcoxon signed-rank test. First, we calculated the differences between pre- and post-training pairs of recordings made by each participant of the study. Second, we ranked the magnitudes, which were the differences without signs. Third, we reattached the signs to the ranks and subsequently totalled up the positive and negative ranks. The unsigned value of the smaller sum of ranks was denoted by T, whereas n stands for sample size (i.e., number of ranks). The sixth step was to calculate in order to obtain the p-value:

$$\mu = n(n + 1)/4,\ \sigma 2 = (2n + 1)\mu/6, \text{ and then } Z = (T - \mu)/\sigma$$

After a detailed analysis of the data for individual sentences, there were no statistically significant differences in the obtained results. Due to space limitation, we decided to present tables which show average results for particular indicators. Hence, Tables 2 and 3 (see subsection 5.1.) reflect the values of the indicators for all sentences before and after a three-week training session.

The results of the Wilcoxon signed-rank test revealed an interdependence between self-imitation practice and L2 pronunciation improvement, due to the larger differences between pre- and post-training measures of the recorded utterances, but only in terms of speech rate. Table 5 shows that there are significant changes between all variables, speech rate (T=48.5; z=2.109; p=0.03), articulation rate (T=88; z=0.63; p=0.52) and average syllable duration (T=104.5; z=0.01; p=0.98); however, it should be noted that for both self-imitation (Table 5) and imitation (Table 4) groups, only the data for speech rate value is statistically significant. The result obtained from the speech rate analysis indicates a rejection of the null hypothesis, which assumes that the medians do not differ significantly. For the other two variables we cannot reject the null hypothesis. The outcomes suggest that speech rate may be one of the most important variables that should be taken into consideration while implementing the self-imitation and imitation technique, however, considering the obtained results, it may be more beneficial for self-imitation practice.

Table 4. The Wilcoxon signed-rank test for imitation practice before and after the experiment.

	Wilcoxon signed-rank test results for imitation task p <0.05			
Variables	**N**	**T**	**Z**	**p**
Speech rate pre- & Speech rate post-training	15	24.00000	2.044663	0.040889
Articulation rate pre- & Articulation rate post-training	15	48.00000	0.681554	0.495521
ASD pre- & ASD post-training	15	53.00000	0.397573	0.690945

Table 5. The Wilcoxon signed-rank test for self-imitation practice before and after the experiment.

	Wilcoxon signed-rank test results for self-imitation task p <0.05			
Variables	**N**	**T**	**Z**	**p**
Speech rate pre- & Speech rate post-training	20	48.5000	2.109296	0.034920
Articulation rate pre- & Articulation rate post-training	20	88.0000	0.634655	0.525654
ASD pre- & ASD post-training	20	104.5000	0.018666	0.985107

The results of the Wilcoxon signed-rank test cannot confirm that there is an interdependence between self-imitation practice and L2 pronunciation improvement with regard to articulation rate and average syllable duration. The only potential we can notice in self-imitation practice and the prosodic conversion technique in order to improve L2 pronunciation level is in terms of speech rate.

6. Discussion

The present study described an experiment in the application of the prosodic conversion technique with the use of the ProZed plugin implemented in Praat software. It was used to provide information about a potential positive correlation between self-imitation practice and L2 pronunciation improvement. According to this, changing the prosodic features of an utterance and adjusting them to a golden speaker's voice can have positive effects on the L2 pronunciation process for Polish learners of English who take part in a delayed self-imitation practice. Less significant effects for Polish learners of English were expected for those who used the traditional method of imitating a native speaker's voice (model voice). The results of the study only partially support these predictions and allow us to conclude that self-imitation practice may be more beneficial than imitation practice for speech rate only. Future research with a larger number of subjects and additional training tools should be conducted to confirm or refute the given outcomes.

Although the results of the Wilcoxon signed-rank test indicated that there is an interdependence between self-imitation practice and L2 pronunciation improvement in terms of speech rate, descriptive statistics of the same data do not confirm such outcomes between speech rate difference. In line with that, the results could be attributed to several factors, including the production of sounds (/əʊ/, /dʒ/, /ə/, /ʒ/) that do not exist in Polish or have a different realization in the Polish phonological system (see Research Design section). Secondly, it is the speed/rate of speech and intonation which are phonetically relevant and decide to what extent a structure was pronounced in a native-like manner.

Additionally, showing that there are no significant differences between the results obtained from individual sentences and providing average results of particular indicators do not reject the conclusion that speech rate may be one of the most important variables in self-imitation and imitation practice. As the Wilcoxon signed-rank test has shown, the result obtained from the speech rate analysis indicates a rejection of the null hypothesis, which assumes that the medians do not differ significantly. Considering the obtained results, we may suggest that self-imitation practice may be more beneficial than imitation practice in terms of speech rate, however, due to the limitations of this study, including the small number of participants and training of relatively short sentences, future research should be conducted to confirm or refute the given outcomes.

Responding to the research questions, we should especially pay attention to the interaction between speech rate and L2 pronunciation improvement, which is the only variable that reveals a significant difference between pre- and post-test results. Focussing on speech rate as the main variable can be especially beneficial for Polish learners of English due to the fact that Polish is a syllable-timed language. The theoretical and pedagogical implications of these findings, which turned out to be consistent with our predictions, are discussed in the following paragraphs.

From a theoretical standpoint, the present outcomes provide evidence in favour of the presented assumption, in which the prosodic conversion technique using Praat algorithms and the ProZed plugin was the main subject. When L2 learners are encouraged to repeat their own modified utterances, they tend to incorporate changes in speech rate, thanks to which this aspect of pronunciation improves.

Although the obtained results are, of course, extremely preliminary, it is encouraging to follow this type of analysis in the hope that more efficient metrics can be obtained automatically in future research. As Hirst (2016) explains, the process of cloning prosody with the use of the ProZed plugin can provide an objective evaluation of one's L2 pronunciation improvement. All steps described in this study can be carried out fully automatically. This can be used to compare automatic pronunciation assessment with an expert's/teacher's evaluation. An attempt to compare the resynthesized voice with the utterances after a training session might also be encouraging. In the present study, we compared the results obtained before and after a training session, however, a comparative analysis of the model voice or resynthesized voice with the participants' final utterances would also provide an opportunity for further data analysis.

The question that arises after statistical analysis is what aspects of speech production are critical for the interpretation of results and what should be considered as only individual variations without any interpretative significance (Hirst 2015). As Hirst (2015, p. 11) points out, the results obtained from the statistical and/or automatic analysis, enable us, for example, to compare the time and frequency values of the pitch points which define the Momel curve (see subsection 4.4). Hence, despite the objectivity of statistical results, statistical comparisons require a number of assumptions in order to demonstrate what aspects can be put together. Despite the increasing number of studies on language prosody that rely on statistics, there is no universal rule on what values should be compared. Following Hirst (2015, p. 11), "there is [...] no general consensus on what statistical measurements best reflect the prosodic similarity between two versions of an utterance". Therefore, despite the many limitations of this study, including the small number of participants and training of only six utterances repeated regularly for 30 minutes a week for three weeks, the general idea of self-imitation practice as a method of L2 pronunciation improvement should still be taken into consideration, even though the obtained results might not be satisfactory.

An important but unresolved issue of the current study is its practical use on a larger scale, which may bring tangible effects for students with different levels of English. Additionally, the choice of structures recorded by the participants in this study make it impossible to carry out exact comparisons of our findings to results obtained from research conducted in different countries. This may serve as motivation for future research ideas (see Future Research section).

7. Future research

A logical next step in this area of research is to expand the number of participants to be able to achieve results on a larger scale. A continuous study on this issue would allow us to provide relevant information on whether the speech rate of the examined utterances is the most important parameter positively influencing L2 pronunciation improvement. Moreover, the outcomes of the research have revealed the importance of using and implementing CAPT tools in L2 pronunciation practice. The fact that Praat software and the ProZed plugin are available for free, enabling L2 learners to have unlimited access, encourages us to promote these tools, which can be used either in the classroom or in a self-study environment (see Beňuš, 2021).

In addition to examining the effects of self-imitation practice and automatic pronunciation assessment, future research should expand to include a comparison of the results obtained from the automatic pronunciation assessment and L2 utterances assessed by English native speakers and/or experienced teachers of English phonetics and phonology. Taking into account the limited accuracy of ASR (automated speech recognition), this idea respects the instruments used in previous studies (Ding et al. 2019; De Meo et al. 2013; Felps et al. 2009; or Nagano and Ozawa 1990) in which native-speakers' assessment was the main tool to evaluate students' L2 pronunciation improvement. Additionally, speech rate variables may be helpful in the process of L2 pronunciation progress for syllable-timed languages, including Polish and other languages. Lastly, given that the research on imitation and self-imitation practice and their effects on L2 pronunciation improvement has not been thoroughly examined, there is still room for extending this area of research. This would help to clarify whether L2 learners benefit from self-imitation practice and the prosodic conversion technique, including pitch and time (syllable) modifications. A study on this issue would allow us to compare our findings with the results obtained from other studies; however, the performed utterances must be first unified in terms of the recorded structures.

Addressing these suggestions in future studies may contribute to the improvement of automatic pronunciation assessment, as well as the process of learning L2 sounds at different levels of L2

acquisition. We believe that prosodic conversion and self-imitation practice could play a significant role in L2 pronunciation improvement. By reducing foreign accent and, at the same time, preserving the voice quality of the L2 learner, it would be possible to achieve satisfying results.

References

Anderson-Hsieh, J. and Ventakagiri, H., 1994. Syllable duration and pausing in the speech of Chinese ESL speakers. *TESOL Quarterly*, vol. 28, no. 4, pp. 807-812.

Beňuš, S., 2021. *Investigating spoken English. A practical guide to phonetics and phonology using Praat.* Springer International Publishing.

Bigi, B. and Hirst, D.J., 2012. SPeech Phonetization Alignment and Syllabification (SPPAS): a tool for the automatic analysis of speech prosody. In: *Proceedings of the 6th International Conference on Speech Prosody*, Shanghai, May 2012.

Bissiri, M.P., Pfitzinger, H.R. and Tillman, H.G., 2006. Lexical Stress Training of German Compounds for Italian Speakers by means of Resynthesis and Emphasis. In: *Proceedings of the 11th Australian International Conference on Speech Science & Technology*. University of Auckland, New Zealand, pp. 24-29.

Bissiri, M.P. and Pfitzinger, H.R., 2009. Italian speakers learn lexical stress of German morphologically complex words. *Speech Communication*, vol. 51, no. 10, pp.933 -947.

Boersma, P. and Weenink, D., 2001. Praat: doing phonetics by computer [Computer program]. Version 6.2.01, retrieved 17 November 2020 from http://www.praat.org/

Bunta, F. and Ingram, D., 2007. The acquisition of speech rhythm by bilingual Spanish- and English-speaking four-and five-year-old children. *Journal of Speech, Language, and Hearing Research 50*, pp. 999-1014.

Burgess, J. and Spencer, S., 2000. Phonology and pronunciation in integrated language teaching and teacher education. *System*, vol. 28, no. 2, pp. 191-215.

Cerda-Onate, K., Vega, T.G. and Ordin, M., 2021. Speech rhythm convergence in a dyadic reading task. *Speech Communication*, vol. 131, pp. 1-12.

Clark, H.H., 2002. Speaking in time. *Speech Communication 36, pp.* 5-13.

Clopper, C.G. and Smiljanic, R., 2011. Effects of gender and regional dialect on prosodic patterns in American English. *Journal of Phonetics 9*, pp. 237-245.

Couper, G., 2017. Teacher cognition of pronunciation teaching: teachers' Concerns and Issues. *TESOL Q.*, vol. 51, no. 4, pp. 820-843.

Crystal, D. (ed.), 1985. *Dictionary of Linguistics and Phonetics*, 2nd ed. Oxford: Blackwell.

Cummins, F., 2002. On synchronous speech. *Acoustics Research Letters Online*, vol. 3, no. 1, pp. 7-11.

Cummins, F., 2009. Rhythm as entrainment: the case of synchronous speech. *Journal of Phonetics*, vol. 37, no. 1, pp. 16-28.

Dauer, R., 1983. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics 11*, pp. 51-62.

De Looze, C. and Hirst, D.J., 2014. The OMe (Octave-Median) scale: a natural scale for speech prosody. In: N. Campbell, D. Gibbon and D. J. Hirst, eds. *Proceedings of the 7th International Conference on Speech Prosody* (SP7). Trinity College, Dublin, Ireland, May 2014.

De Meo, A. Vitale, M. Pettorino, M. Cutugno, F. and Origlia, A., 2013. Imitation/self-imitation in computer-assisted prosody training for Chinese learners of L2 Italian. In: J. Levis and K. LeVelle, eds. *Proceedings of the 4th Pronunciation in Second Language Learning and Teaching Conference*. Aug. 2012, 90-100, Ames, IA: Iowa State University.

Ding, H., Hirst, D. and Hoffmann, R., 2015. Cross-linguistic prosodic comparison with OMProDat database. In: *2015 International Conference Oriental COCOSDA held jointly with 2015 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, Shanghai, pp. 212-215.

Ding, S., Liberatore, C., Sonsaat, S., Lučić, I., Silpachai, A., Zhao, G., Chukharev-Hudilainen, E., Levis, J. and Gutierrez-Osuna, R., 2019. Golden speaker builder - An interactive tool for pronunciation training. In: *Speech Communication*, vol. 115, pp. 51-66.

Eskenazi, M., 1999. Using automatic speech processing for foreign language pronunciation tutoring: some issues and a prototype. *Language Learning & Technology*, vol. 2, no. 2, pp. 62-76.

Felps, D., Bortfeld, H. and Gutierrez-Osuna, R., 2009. Foreign accent conversion in computer assisted pronunciation training. *Speech Communication*, vol. 51, no. 10, pp. 920-932.

GenevaLogic. 2007. *SpeedLingua*. http://www.speedlingua.com

Gervain, J. Nespor, M., Mazuka, R., Horie, R. and Mehler J., 2008. Bootstrapping word order in prelexical infants: a Japanese-Italian cross-linguistic study. *Cognitive Psychology*, vol. 57, no. 1, pp. 56-74.

Giles, H., Coupland, J. and Coupland, N., 1991. Accommodation theory: Communication, context, and consequence. In: H. Giles, J. Coupland, N. Coupland, eds. *Contexts of accommodation: Developments in applied sociolinguistics*. Cambridge, UK: Cambridge University Press, pp. 1-68.

Ginther, A., Dimova, S. and Yang, R., 2010. Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, vol. 27, no. 3, pp. 379-399.

Gorjian, B., Hayati, A. and Pourkhoni, P., 2013. Using Praat Software in Teaching Prosodic Features To EFL Learners. *Procedia - Social and Behavioral Sciences*, vol. 84, pp. 34-40.

Goronzy, S., Tomokiyo, L.M., Barnard, E. and Davel, M., 2006. Other challenges: Non-native speech, dialects, accents, and local interfaces. In: T. Schultz and K. Kirchhoff, eds. *Multilingual speech processing*, San Diego and London: Academic Press, pp. 273–316.

Grabe, E. and Low, L., 2002. Acoustic correlates of rhythm class. In: C. Gussenhoven and N. Warner, eds. *Laboratory Phonology 7*. New York: Mouton de Gruyter, pp. 515-546.

Hardison, D.M., 2004. Generalization of computer-assisted prosody training: Quantitative and qualitative findings. *Language Learning & Technology*, vol. 8, pp. 34-52.

Hewings, M., 2017. *English Pronunciation in Use: Advanced. Book with Answers and Downloadable Audio*. Cambridge University Press.

Hirst, D., 2000. ProZed: A multilingual prosody editor for speech synthesis. In: *Proceedings of IEE Colloquium. State-of-the-Art in Speech Synthesis*. London.

Hirst, D., 2012. ProZed: A speech prosody analysis-by-synthesis tool for linguists. *Speech Prosody*, pp. 15-18.

Hirst, D., 2015. ProZed: A Speech Prosody Editor for Linguists, Using Analysis-by-Synthesis. In: K. Hirose and J. Tao, eds. *Speech Prosody in Speech Synthesis: Modeling and generation of prosody for high quality and flexible speech synthesis. Prosody, Phonology and Phonetics*. Berlin, Heidelberg: Springer.

Hirst, D., 2016. On the automatic comparison and cloning of native and non-native speech prosody. *Speech Prosody*, pp. 1038-1042.

Huckvale, M. and Yanagisawa, K., 2007. Spoken language conversion for text-to-speech synthesis. In: *Proc. ISCA Speech Synthesis Workshop*, Bonn, Germany, pp. 64-70.

Johansson, S., 2008. *Contrastive analysis and learner language: A corpus-based approach.* Oslo: University of Oslo.

Kang, O., 2010. Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System 38*, pp. 301-315.

Kang, O., Rubin, D. and Pickering, L., 2010. Suprasegmental measures of accentedness and judgements of language learner proficiency in oral English. *The Modern Language Journal*, vol. 94, no. 4, pp. 554-566.

Kowal, S. and O'Connell, D., 1980. Pausological research at Saint Louis University. In: H. W. Dechert and M. Raupach, eds. *Temporal Variables in Speech: Studies in Honour of Frieda Goldan-Eisler*. The Hague: Mouton, pp. 61-68.

Kusz, E., 2019. *Musical Intelligence and Its Impact on English Pronunciation Skills in the Process of Second Language Acquisition*. Berlin: Peter Lang.

Levelt, W.J.M., 1989. *Speaking: From Intention to Articulation.* Massachusetts: MIT Press.

Levis, J. and Sonsaat, S., 2017. Pronunciation in the CLT era. In: *The Routledge handbook of English pronunciation*, pp. 267-283.

Martin, P., 2004. WinPitch LTL II, a Multimodal Pronunciation Software. ISCA.

Markham, D., 1997. Phonetic Imitation, Accent, and the Learner. In: *Travaux de l'Institut de linguistique de Lund*. Lund: Lund University Press, p. 269.

Munro, M.J. and Derwing, T.M., 1998. The effects of speaking rate on listener evaluations of native and foreign-accented speech. *Language Learning 48*, pp. 159-182.

Munro, M.J. and Derwing, T.M. 1999. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, 49 (supplement 1), pp. 285-310.

Munro, M.J. and Derwing, T.M., 2001. Modeling Perceptions of the accentedness and comprehensibility of L2 speech. The role of speaking rate. *Studies in Second Language Acquisition 23*, pp. 451-468.

Munro, M.J. and Derwing, T.M., 2015. A prospectus for pronunciation research in the 21st century: A point of view. *Journal of Second Language Pronunciation*, vol. 1, no. 1, pp. 11-42.

Nagano, K. and Ozawa, K., 1990. English speech training using voice conversion. In: *1st International Conference on Spoken Language Processing* (ICSLP 90), Kobe, Japan, pp. 1169-1172.

Nazzi, T. and Ramus, F., 2003. Perception and acquisition of linguistic rhythm by infants. *Speech Communication 41*, pp. 233-243.

Ordin, M. and Polyanskaya, L., 2014. Development of timing patterns in first and second languages. *System 42*, pp. 244-257.

Ordin, M. and Polyanskaya, L., 2015. Acquisition of speech rhythm in second language by learners of English with rhythmically different native languages. *Journal of Acoustical Society of America*, vol. 138, no. 2, pp. 533-545.

Ordin, M. and Polyanskaya, L., 2019. The effect of speech rhythm and speaking rate on assessment of pronunciation in a second language. *Applied Psycholinguistics*, vol. 40, no. 3, pp. 1-25.

Pardo, J.S., 2010. Expressing oneself in conversational interaction. In: Morsella, E., ed. *Expressing Oneself/Expressing One's Self: Communication, cognition, language, and Identity*. Hove: Psychology Press/Taylor & Francis,pp. 183–196.

Pardo, J.S., Urmanche, A., Wilman, S. and Wiener, J., 2017. Phonetic convergence across multiple measures and model talkers. *Attention, Perception, & Psychophysics*, vol. 79, no. 2, pp. 637-659.

Payne, E., Post, B., Astruc, L., Prieto, P. and del Mar Varnell, M., 2012. Measuring child speech. *Language and Speech*, vol. 55, no. 2, pp. 203-229.

Peabody, M. and Seneff, S., 2006. Towards automatic tone correction in non-native Mandarin. In *Chinese Spoken Language Process*, pp. 602-613.

Pellegrino, F., Coupé, Ch. and Marsico, E., 2011. Across-Language Perspective on Speech Information Rate. *Language*, vol. 87, no. 3, pp. 539-558.

Pellegrino, E. and Vigliano, D., 2015. Self-imitation in prosody training: a study on Japanese learners of Italian. *SLaTE*, pp. 53-57.

Pickering, M. J. and Garrod, S., 2013. An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, vol. 36, pp. 49–64.

Probst, K., Ke, Y. and Eskenazi, M., 2002. Enhancing foreign language tutors – in search of the golden speaker. *Speech Communication*, vol. 6, no. 1, pp. 1-14.

Quené, H., 2005. Modeling of between-speaker and within-speaker variation in spontaneous speech tempo. In: *Proceedings of the ninth European conference on Speech Communication and Technology (InterSpeech)*, Lisbon, Portugal, 4-8 September, pp. 2457–2460.

Quene, H., 2007. On the just noticeable difference for tempo in speech. *Journal of Phonetics 35*, pp. 353-362.

Ramus, F. and Mehler, J., 1999. Language identification with suprasegmental cues: a study based on speech resynthesis. *Journal of Acoustical Society of America 105*, pp. 512-521.

Repp, B.H. and Williams, D.R., 1987. Categorical tendencies in imitating self-produced isolated vowels. *Speech Communication*, vol. 6, no. 1, pp. 1-14.

Riffenburgh, R.H., 2006. Tests on Ranked Data. In: Robert H. Riffenburgh, ed. *Statistics in Medicine, 2nd ed*. Academic Press, pp. 281-303.

Shepard, C.A., Giles, H. and Le Poire, B.A., 2001. Communication accommodation theory. In W. P. Robinson and H. Giles, eds. *The new handbook of language and social psychology*. New York, NY: Wiley, pp. 33-56.

White, L. and Mattys, S., 2007. Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, vol. 35, no. 4, pp. 501-522.

Zhi, N., Hirst, D., Bertinetto, P.M., Li, A. and Jia, Y., 2016. An analysis-by-synthesis study of Mandarin speech prosody. In *Proceedings of Speech Prosody*, pp. 104-108.