EFFECTS OF SELF-IMITATION PRACTICE ON L2 PRONUNCIATION WITH THE USE OF GOLDEN SPEAKER BUILDER

Ewa Kusz, University of Rzeszów, Poland

One innovative method of L2 pronunciation improvement is self-imitation practice, which involves mirroring recordings of one's own voice after they have been resynthesized to match native speaker pronunciations of target sounds. The Golden Speaker Builder (GSB) (Ding et al., 2019) is a tool that allows users to generate such a personalised model voice, mirroring the learner's voice quality, but with a native accent. In this study we investigate the effects of using the GSB and to what extent it affects L2 learners' comprehensibility and fluency. Thirty-five participants in the study performed a three-week self-imitation task by repeating some of the sentences they had previously recorded. The participants took a pre-test, a post-test after three weeks of practice, and a delayed post-test. Each participant completed two qualitative questionnaires before and after the exercises, to guage their opinion about the tool used. The results show a significant improvement in pronunciation in terms of comprehensibility and fluency, but the feedback from the questionnaire indicates that GSB cannot replace the personalised comments received directly from a teacher during pronunciation training.

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

INTRODUCTION

Although the use of CAPT tools as a method of pronunciation training is still developing, the first work on imitation and self-imitation using voice synthesisers appeared in the late 1980s (e.g., Repp & Williams 1987). To date, we can find more and more studies that focus on the self-imitation method and its use in L2 pronunciation training (Bissiri et al., 2006; Bissiri & Pfitzinger, 2009; De Meo et al., 2012; Ding et al. 2019; Hirose et al., 2003; Hardison 2004; Peabody & Seneff, 2006; Pellegrino & Vigliano, 2015;). A pioneering study was conducted by Nagano and Ozawa (1990), who applied prosodic conversion to the pronunciation training of Japanese learners. Participants in the study were divided into two groups, one of which performed tasks involving imitation of the voice of an English native speaker, while the other imitated their own voices, previously synthesised to match the prosody of a native speaker of English. The results of this experiment showed that those who imitated the synthesised voice achieved better results (they sounded more native-like) in L2 pronunciation practice than those who practised with the voice of an English native speaker. Over a decade later, Probst, Ke and Eskenazi (2002) showed that a high level of similarity between students' and teachers' voices (i.e., 'golden speakers') can bring beneficial results in L2 pronunciation training. Peabody and Seneff (2006) changed the pitch contour of the recordings of L2 learners of Japanese whose L1 was English, revealing that the modified utterances were assessed as correct more frequently than unmodified recordings. Self-imitation practice was also shown to have beneficial effects in L2 pronunciation improvement in Bissiri et al. (2006). Felps et al. (2009) implemented an accent conversion method and with the use of PSOLA (a digital signal processing technique, which stands for Pitch Synchronous Overlap and Add) they adapted the learner's rate and pitch to the teacher's and reduced the participants' foreign accent. The

findings revealed that the techniques they used resulted in a significant reduction in foreign accent while not changing the voice quality of the L2 learner. The results also revealed a strong interdependence between accent and identity. De Meo et al. (2013) also highlighted the idea of a 'golden speaker' model in L2 pronunciation practice. In their study, they used a rhythmic-prosodic transplantation technique to evaluate the effectiveness of imitation and self-imitation pronunciation training of Chinese learners of Italian. The results indicated that post-training utterances of L2 learners who did self-imitation practice were significantly better than recordings of participants who did imitation tasks. All of these studies show that prosodic accent conversion may have a positive effect on L2 pronunciation improvement. Moreover, the studies reveal that self-imitation practice can bring satisfying results regardless of L1-L2 differences and similarities. Thus, this paper reports the results obtained from an experiment which drew inspiration from Ding et al.'s (2019) study, in which a program called Golden Speaker Builder (GSB) was used to investigate whether self-imitation practice helps in L2 comprehensibility and fluency improvement.

Golden Speaker Builder

In 2019, Ding et al. introduced the GSB, an interactive tool for pronunciation training, in which they incorporated segmental accent-conversion into L2 speech synthesis (2019, p. 51). In their project, they introduced the overall system design, and the speech analysis process in which they mirrored L2 learners' voices and adjusted them to a native model (either female or male) so that the users could imitate their own voice but with an American English accent. Their study took three weeks and it included 15 Korean learners of English who imitated the utterances that were first synthesised with the voice of an English native speaker. The training session was preceded by a pre-test in which the L2 learners were asked to record 48 sentences. One week after the training, the participants of the study took an immediate post-test which included the same 48 sentences. They also took a delayed post-test three weeks after the training session. The sentences were then assessed by 95 native-English speaking undergraduate students who were divided into two groups, one for assessing comprehensibility and the other for fluency (Ding et al. 2019: 60).

The results of the user study revealed that there was a significant improvement in L2 learners' comprehensibility and fluency at post-test. At the delayed post-test, improvement was sustained for fluency, but not for comprehensibility. In terms of fluency, L2 learners highlighted the GSB training as being beneficial as they themselves noticed their improvement in L2 intonation, stress and connected speech (Ding et al. 2019, p. 64).

Research Questions

Ding et al.'s (2019) study was the inspiration for our experiment in which Polish students of English used a recent version of the GSB to practise L2 pronunciation. Thus, to a certain degree, this experiment replicated Ding et al.'s (2019) user study, and it was guided by the same research questions (Ding et al. 2019, p. 52). However, it is worth noting that in our study we used an updated version of the GSB but it only allowed participants to choose between a single male of female voice (rather than two voices from each in the original experiment).

- RQ1: What is the effect of using the GSB on learners' improvement of their comprehensibility and fluency?

• RQ2: What features of the GSB did learners find useful, and what did they find in need of improvement?

METHODS

Participants

There were two groups of participants. The first one included 35 Polish students of English. They were studying Applied Linguistics at the University of Rzeszów in Poland. Their ages were between 19-23 and their language level varied from B2+/C1 (CEFR). None of the students had spent more than several weeks in an English-speaking country prior to this research. All of the students attended English Phonetics and Phonology classes.

The second group of participants were 10 raters. They were Polish teachers of English who had graduated from the University of Rzeszów in the same year (2013). They all had at least nine years of experience in teaching English. The raters were part of two equal groups since comprehensibility (n=5) and fluency (n=5) were each rated by a separate group of raters. Similarly to Ding et al. (2019), both categories were measured using a 10-point Likert scale (0-9).

Procedure

The three-week practice sessions started in November 2021 and ended in December. Due to COVID-19 restrictions, all meetings were held online. Before building their own golden speakers, the participants were asked to record 48 sentences. Secondly, the participants were asked to build their own golden speakers by recording 24 of those 48 sentences, which were available in the web application. The same 48 sentences were used for the pre-tests and the post-tests. They are described in Ding et al.'s (2019) reading task (pp. 64-65).

After building their golden speaker models, the learners were also asked to answer several questions in an online survey. The main purpose of these questions was to examine if they had any issues in building a golden speaker model before the three-week training program.
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 24 sentences. There were nine 30-minute meetings in the three-week training programme, again replicating the methodology from Ding et al. (2019).

Training session

The participants listened to and repeated the resulting recordings of their own voices synthesised with the American English 'model voice'. Two types of activities were provided in the training process: say-listen-repeat and listen-repeat.

The participants were first urged to follow the instructions presented by the teacher before each session in week one. Carefully monitored by the researcher, the students worked individually in online breakout rooms to have the best opportunity to work at their own pace.

Following the three weeks of training, the participants took part in the immediate post-test, which was given one week after the training and included the same 48 sentences used in the pre-test. The

delayed post-test was administered four weeks (about a month) after the training and the students were again asked to record the same 48 sentences.

Raters

As previously mentioned, two speech dimensions, comprehensibility and fluency, were rated by 10 teachers of English, 5 raters for each dimension. Both constructs were clarified to the raters before they started assessing the recordings. Comprehensibility was explained as the listener's judgment of how difficult it is to understand speech produced by an L2 learner (Derwing and Munro, 1998: 396). It was measured using a 10-point Likert scale (0-9) in which 0 indicated L2 utterances that were 'impossible to understand', whereas 9 meant there was 'no problem with understanding'. As for fluency, which refers to overall tempo and flow, smooth delivery and general proficiency (Derwing and Munro, 1998: 396), 0 represented 'an unnaturally slow tempo' and 9 a 'native-like tempo/flow'. All three tests for each student were included in the assessment. The procedure for both dimensions were exactly the same. The rating sessions were held online and there were ten meetings (one for each rater). Each meeting took approximately 60 minutes. All sentences were presented randomly and there was a pause after each sentence so that the raters had enough time to assess it. Zero represented a poor rating, whereas 9 meant an excellent rating for each dependent variable. Before the first rating session, the raters listened to a few sentences previously recorded by an experienced pronunciation teacher to help them to become familiar with the task. All in all, there were 5040 assessed utterances in each category which were rated on a 10-point Likert scale (from 0-9).

Data Analysis

For the analysis, we used the mean rating scores for each of the sentences produced by the 35 learners. The ratings were based on the 24 sentences that were used in the training session and 24 other sentences that were not practised by the students, but were recorded in the pre- and post-tests (overall 48 sentences). Students completed one pre-test before the training sessions and two post-tests (immediate and delayed). This resulted in 720 mean values of students' utterances for each category (48 sentences, 3 tests, 5 separate raters for each of the comprehensibility and fluency ratings).

A preliminary analysis showed that the mean ratings for comprehensibility and fluency that the 5 raters of each dimension gave to to each participant did not differ. In contrast, the mean values differ depending on the analysed tests (i.e., pre-, immediate post- and delayed-post-tests). Moreover, differences in mean values appear in both practised (sentences 1-24) and unpractised (25-48) sentences. To confirm the hypothesis that use of GSB led to statistically significant gains, ANOVA and MANOVA were used. Analyses were based on two factors: TRAINING (practised vs. unpractised) and TIME (pre-test, immediate post-test, delayed post-test).

RESULTS

Comprehensibility

Five raters assessed 1680 recorded sentences for comprehensibility, resulting in 8400 ratings. Each recording was assessed by each rater.

Comprehensibility ratings were normally distributed (p = 0.001). After checking the normality assumption by running the Shapiro-Wilk test for each dependent variable, ANOVA and MANOVA statistical methods were implemented.

Figure 1

Comprehensibility for practised sentences 1-24, pre-test, immediate post-test, delayed post-test

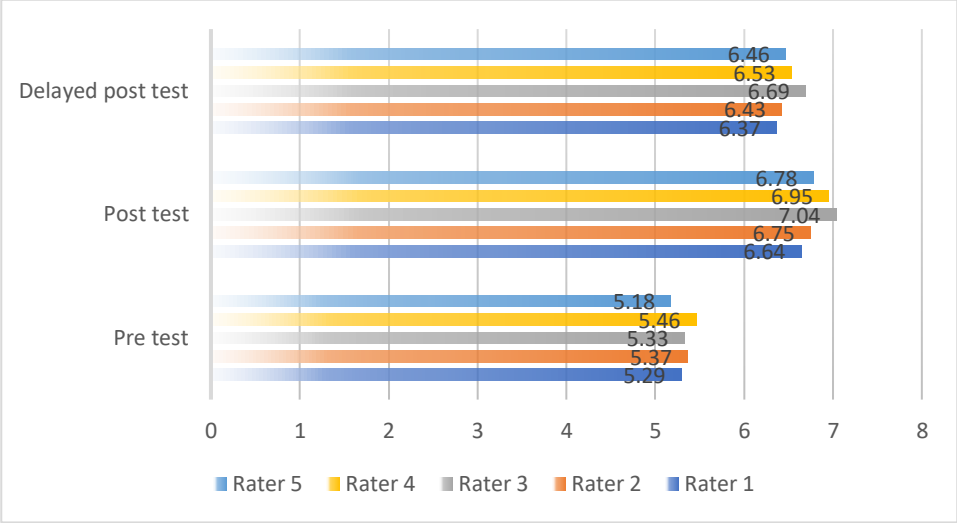
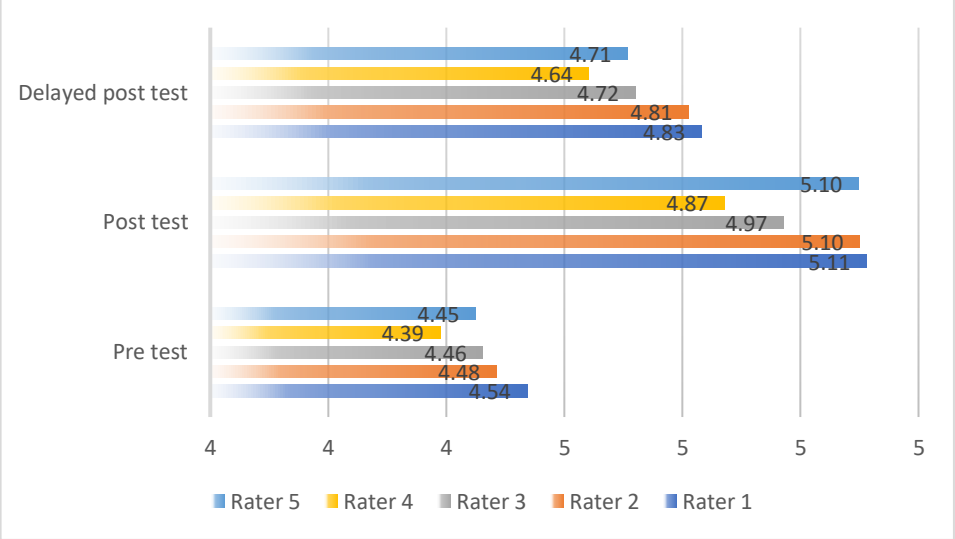


Figure 2

Comprehensibility for unpractised sentences (25-48), pre-test, immediate post-test, delayed post-test



One-way ANOVA was applied to see if TIME (pre- and post-tests) differentiates the assessment of the group of raters (p < 0.001). We can state that the differences between the results obtained from the pre-test, immediate post-test and delayed post-test are statistically significant. The mean values of the scores given by the group of raters (see Figure 1) clearly indicate progress between pre-test and post-test and that these differences were maintained at the delayed post-test for the practised sentences. A clear difference can also be noticed in the ratings of unpractised sentences

(see Figure 2), although there appears to be greater regression between the post-test and delayed post-test, relative to performance on the trained sentences.

Figure 3 and Table 1 show the analysis of TRAINING variable, in which the mean values of raters' scores for practised and unpractised sentences were considered. There is a significant difference between the two factors (p=0.000; F(1.718)=624.10) in terms of comprehensibility.

Figure 3

One-way ANOVA for comprehensibility practised (Yes) and unpractised (No) sentences.

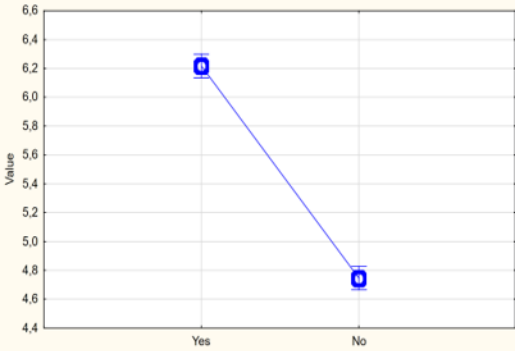


Table 1

One-way ANOVA results for comprehensibility, TRAINING variable

	SS	df	MS	F	p	Noncentrality parameter	Significance level (alpha=0.05)
intercept	21632.85	1	21632.85	34654.59	0.00	34654.59	1.000000
program using	389.59	1	389.59	624.10	0.00	624.10	1.000000
error	448.21	718	0.62				

The results obtained from one-way ANOVA show that there is a significant difference between the mean values of dependent variables in terms of comprehensibility.

Figure 4 and Table 2 show that in self-imitation L2 pronunciation practice, there is a significant difference between pre-, post-, and delayed post-test (p=0.000; F(2.717)=69.21) for comprehensibility. The results reveal that regardless of the TRAINING variable (practised or unpractised sentences), students' level of comprehensibility improved. The largest difference can be noticed between the mean scores of pre- and immediate post-test (pre-test: 4.894; post-test: 5.930), but the mean scores of delayed post-test also reveal a higher level of L2 comprehensibility than in pre-test recordings.

Figure 4
One-way ANOVA for comprehensibility pre-test (Pre), post-test (Post) and delayed post-test (Delayed) results

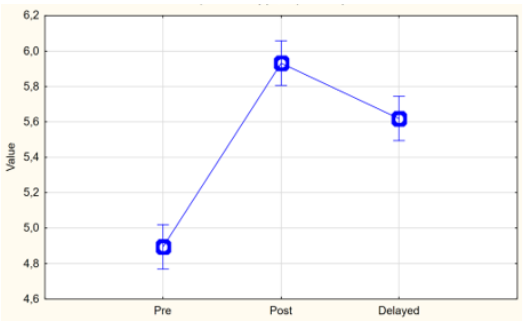


Table 2
One-way ANOVA for comprehensibility, TIME variable

	SS	df	MS	F	p	Noncentrality parameter	Significance level (alpha=0.05)
intercept	21632.85	1	21632.85	22087.93	0.00	22087.93	1.000000
category	135.57	2	67.79	69.21	0.00	138.42	1.000000
error	702.23	717	0.98				

To identify the level of significance and differences in means between two dependent variables: TIME and TRAINING, we conducted a one-way MANOVA (Table 3). The results show that there was a statistically significant difference between TIME and TRAINING in terms of comprehensibility (p=0.000, F(1. 714) = 997.02).

Table 3
One-way MANOVA for comprehensibility

	SS	df	MS	F	p	Noncentrality parameter	Significance level (alpha=0.05)
intercept	21632.85	1	21632.85	55361.64	0.00	55361.64	1.000000
category	135.57	2	67.79	173.47	0.00	346.94	1.000000
program using	389.59	1	389.59	997.02	0.00	997.02	1.000000
cat*pr.us	33.64	2	16.82	43.04	0.00	86.08	1.000000
error df	279.00	714	0.39				

Fluency

As for fluency, the mean values of the scores given by the raters (see Figure 5) also indicate progress between pre-test, post-test and delayed post-test for practised sentences. There is a noticeable difference in the ratings of unpractised sentences (see Figure 6). The largest difference occurred between the mean values of the scores for the pre-test and post-test results. The mean values of the scores for the delayed post-test are also higher than those of the pre-test, but, similarly to comprehensibility results, the means were slightly lower than for the post-test.

Figure 5

Fluency for practised sentences 1-24, pre-test, immediate post-test, delayed post-test

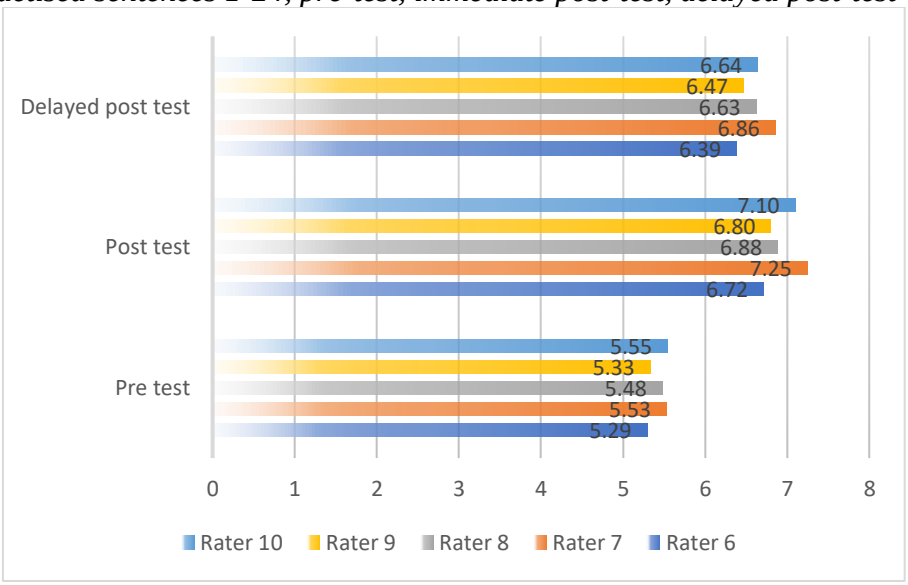


Figure 6

Fluency for unpractised sentences (25-48), pre-test, immediate post-test, delayed post-test

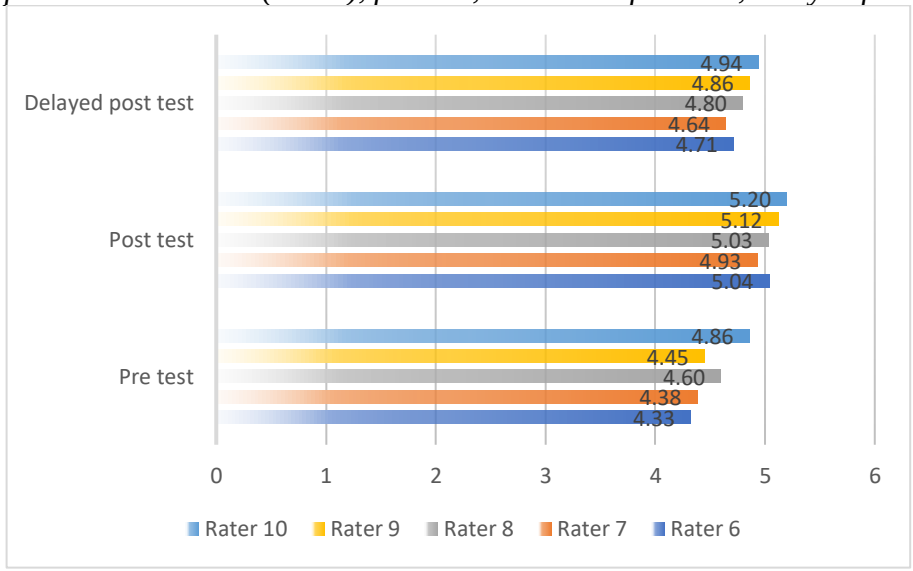


Figure 7

One-way ANOVA for fluency practised (Yes) and unpractised (No) sentences for fluency

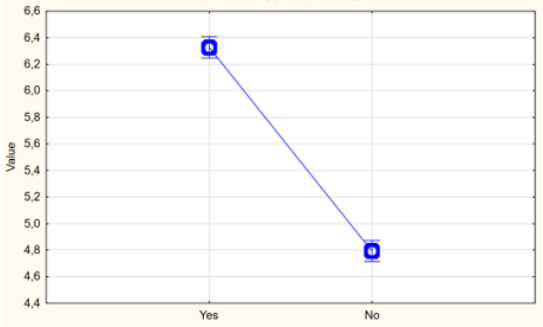


Table 4

One-way ANOVA for fluency, TRAINING variable

	SS	df	MS	F	p	Noncentrality parameter	Significance level (alpha=0.05)
intercept	22258.75	1	22258.75	38185.14	0.00	38185.14	1.000000
program using	424.38	1	424.38	728.04	0.00	728.04	1.000000
error	418.53	718	0.58				

Figure 8

One-way ANOVA for fluency pre-test (Pre), post-test (Post), and delayed post-test (Delayed) results for fluency

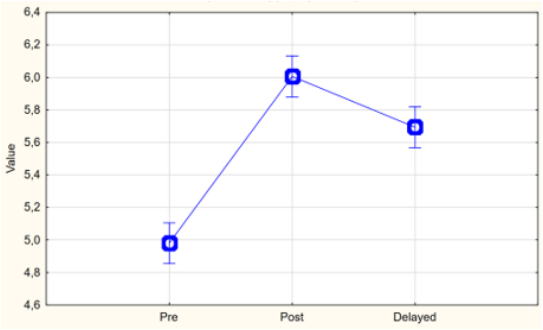


Table 5
One-way ANOVA for fluency, TIME variable

	SS	df	MS	F	p	Noncentrality parameter	Significance level (alpha=0.05)
intercept	22258.75	1	22258.75	22478.46	0.00	22478.46	1.000000
category	132.93	2	66.46	67.12	0.00	134.24	1.000000
error	709.99	717	0.99				

One-way ANOVA results reveal that there is a significant difference between the mean fluency ratings for practiced and upracticed sentences (p=0.000; F (1. 718)=728.04). See Figure 7 and Table 4.

Figure 8 and Table 5 show that there is a significant difference between pre- post- and delayed post-test (p=0.000; F(2. 717)=67.119) for fluency.
The results of one-way MANOVA (see Table 6) show a statistically significant difference between TIME and TRAINING for ratings of fluency (p=0.000 F(2.714)=189.25).

Table 6
One-way MANOVA results for fluency

	SS	df	MS	F	p	Noncentrality parameter	Significance level (alpha=0.05)
intercept	22258.75	1	22258.75	63381.32	0.00	63381.32	1.000000
category	132.93	2	66.46	189.25	0.00	378.50	1.000000
program using	424.38	1	424.38	1208.42	0.00	1208.42	1.000000
cat*pr.us	34.86	2	17.43	49.63	0.00	99.26	1.000000
error	250.75	714	0.35				

Polish students' GSB experience

To answer the second research question, two surveys were conducted, the first immediately after the students had built their own golden speaker models, and the second after the immediate post-test. The first survey focused mainly on the issues that might have appeared in building a golden speaker model, whereas the second survey concentrated on the impression the whole experiment made on the participants.

Among the 35 participants who joined the experiment and trained regularly, 21 students (60%) stated that they had no problems in building their golden speakers. The analysis of the comments

of students who had problems revealed two main issues. If it took too long to synthesise the sentences, meaning that it was necessary to try again to build the golden speaker model. Secondly, 7 students (20%) indicated that the synthesised sentences did not sound like they expected, describing the final result as 'machine-like' or 'robot-like'. However, 4 students (11%) admitted that this was the result of a weak internet connection, and/or poor sound quality of their microphone, and after re-recording the sentence or refreshing the website, this problem either disappeared or the synthesised utterances appeared to sound more natural.

In the second survey the students were asked if they had any issues while practising with the GSB. Sixteen students (46%) indicated that they had no problems; however, 19 participants said that in some situations, the GSB either did not record or played back high-pitched sounds (2 students). In addition, 17 students (48%) admitted that at the end of the training sessions or after playing the same utterance several times in a row, the programme required a page refresh; otherwise either the quality of the synthesised sentences was poor or it was not possible to play them because the website crashed.

However, when asked if they found this method advantageous, all participants recommended that self-imitation training with the use of the GSB should be implemented in regular L2 pronunciation practice. All participants confirmed that they noticed that their pronunciation improved.

When the students were asked if there were any drawbacks in the project, they highlighted that they lacked immediate feedback (14 students) from a teacher. Another point that students considered a disadvantage, despite the clear rules set out at the beginning of the experiment, was the monotony in repeating the same sentences (11 students).

DISCUSSION

The results obtained from this study confirm that, despite having a different L1 from the participants in previous research (i.e., Ding et al., 2019), the Polish students of English improved their L2 comprehensibility and fluency. These findings lend support to earlier studies which highlighted the benefits of self-imitation practice (Hirose et al., 2003; Hardison 2004; Peabody & Seneff, 2006; Bissiri & Pfitzinger, 2009; Bissiri et al., 2006; De Meo et al., 2012; Pellegrino & Vigliano, 2015; Ding et al. 2019).

However, we mostly compare our findings to the results of Ding et al.'s (2019) study, as it was their tool that was used to help L2 learners improve their pronunciation. As mentioned above, in both studies, immediate post-test results indicate significant improvements in L2 learners' fluency and comprehensibility. The main difference between Ding et al. and this study is that L2 pronunciation improvement was sustained at the delayed post-test for fluency in both experiments, whereas for comprehensibility only for this study. Bearing in mind that the experiments were not identical and the group of participants assessing the students' utterances differed significantly, it would be worthwhile to continue the research. It would also be justifiable to investigate how the GSB might assist in improving L2 pronunciation of those whose mother tongue is neither Polish nor Korean.

An important limitation in this study is that no control group was used, so that we were not able to compare our results to a group in which different model voices were used. There could be, as Ding

et al. (2019, p. 63) suggest, a synthesized voice created with two native voices or the imitation of a native speaker's voice without synthesizing it with other voices.

Apart from the quality of the synthesised voices and the limited number of sentences for practice, there is another issue that should be taken into account in further research. The GSB is aimed at people who want to work on American English. However, there is no such tool for British English or other accents.

Golden Speaker Builder is an interactive tool for pronunciation training that undoubtedly deserves to become better known amongst L2 pronunciation teachers and students. However, most of the participants admitted that the method lacked direct feedback from a teacher to highlight individual problems with L2 pronunciation. Such comments are important in the light of L2 pronunciation teaching, as they reassure teachers that despite the developing technology, CAPT tools and apps for L2 pronunciation practice, direct feedback from experts is still the most desirable method.

ACKNOWLEDGMENTS

I wish to express my deepest gratitude to all the BA and MA students of Applied Linguistics from the University of Rzeszów who participated in this experiment.

ABOUT THE AUTHOR

Ewa Kusz is an Assistant Professor of Applied Linguistics at the University of Rzeszów in Poland. Her main areas of interest include phonetics and pronunciation pedagogy, methodology of teaching phonetics to adults in EFL context, and CAPT tools.

REFERENCES

Bissiri, M., Pfitzinger, H., & Tillman, H. (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* (pp. 24-29). University of Auckland, New Zealand.

Bissiri, M. P. & Pfitzinger, H. R. (2009). Italian speakers learn lexical stress of German morphologically complex words. In *Speech Communication*, 51(10), pp. 933-947.

De Meo, A. Vitale, M. Pettorino, 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*. Aug. 2012, 90-100, Ames, IA: Iowa State University.

Derwing, T., & Munro, M. (1998). Evidence in favour of a board framework for pronunciation instruction. *Language Learning,* 48(3), 393-410.

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. *Speech Communication*, 115, 51-66.

Felps, D., Bortfeld, H., & Gutierrez-Osuna, R. (2009). Foreign accent conversion in computer assisted pronunciation training. In *Speech Communication* 51(10), 920-932.

Hardison, D. (2004). Generalization of computer-assisted prosody training: Quantitative and qualitative findings. In *Language Learning & Technology*, 8, 34–52.

Hirose, K., Gendrin, F., & Minematsu, N., (2003). A pronunciation training system for Japanese lexical accents with corrective feedback in learner's voice. In: Proceedings of the Annual Conference of the International Speech Communication Association (INTERSPEECH). doi: 10.21437/Eurospeech.2003-787.

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

Peabody, M., & Seneff, S. (2006). Towards automatic tone correction in non-native Mandarin. In *International Symposium on Chinese Spoken Language Process*, 602-613. DOIhttps://doi.org/10.1007/11939993_62

Pellegrino, E. & Vigliano, D. (2015). Self-imitation in prosody training: a study on Japanese learners of Italian. In *Speech and Language Technology in Education*, 53-57.

Probst, K., Ke, Y., & Eskenazi, M. (2002). Enhancing foreign language tutors – in search of the golden speaker. *Speech Communication* 6(1), 1-14.

Repp, B., & Williams, D. (1987). Categorical tendencies in imitating self-produced isolated vowels. *Speech Communication 6*(1), 1-14.