Gradual Improvements Observed in Learners’ Perception and Production of L2 Sounds Through Continuing Shadowing Practices on a Daily Basis

Takuya Kunihara1, Chuanbo Zhu1, Nobuaki Minematsu1, Noriko Nakanishi2

1Graduate School of Engineering, The University of Tokyo
2Faculty of Global Communication, Kobe Gakuin University

Abstract
Shadowing was proposed originally in psycholinguistics to investigate listeners’ process of perceiving speech, and now it is used widely in language education in Japan as a training method for learners to acquire a better skill in perceiving and producing L2 sounds. Recently, a shadowing-based assessment method was proposed to quantify learners’ skill in perception and production. By integrating the two methods, this study aims at tracking gradual improvements of learners’ skill during shadowing practices. We held a special program of 42-day Shadowing Marathon, where Japanese learners of English participated in shadowing practices every day for six weeks. Four new oral passages were presented daily, each of which was shadowed repeatedly. From the obtained data, we analyzed gradual improvements in learners’ perception and production of both segments and prosody through repeating shadowing within a day, and through continuing the shadowing practices over 42 days. The results of analysis showed that, while learners’ perception of segments and prosody improved significantly even with no explicit instructions, their production did not show significant improvement in a self-learning condition. The data also showed in what way learners depend on written input and auditory input when they learn perception and production of L2 sounds.

Index Terms: language learning, shadowing, perception and production, articulatory and prosodic control, assessment

1. Introduction
Shadowing is a task of listening to given utterances and reproducing them with as short delay as possible. It was originally proposed in psycholinguistics [1–4] and in [3], it was used to analyze human behaviors of perceiving spoken words and to model the process of accessing to the mental lexicon that was assumed to exist in mind. Later in [5], shadowing was introduced as a training method to simultaneous interpreters. Here, trainees are not asked to interpret given utterances into another language, but follow the raw speech image instead [14]. However, reproduction without referring to the L1 system may be difficult for na\’ive learners without any corrective feedback from teachers.

While the effects of shadowing on improving learners’ L2 performance are discussed in the literature of language learning and brain sciences [15–18], to the best of the authors’ knowledge, no study was conducted to investigate gradual changes of perception and production of L2 segments and prosody through shadowing practices. By shadowing the same model utterance repeatedly, learners’ shadowing behaviors will change gradually, and by continuing the shadowing practices over days, their performance will improve further. To clarify learners’ gradual changes in perception and production, we held a special program of 42-day Shadowing Marathon (SM), where Japanese learners of English attended shadowing practices every day. On a day, four new oral model passages of about 30 sec were presented to the learners, who shadowed each passage repeatedly. From the obtained data, we analyzed learners’ gradual changes through repeating shadowing within a day, and through continuing the shadowing practices for 42 days. The changes were automatically analyzed to be quantified. It should be noted that all the recordings were made at home due to COVID-19, meaning that shadowing and recording were made independently of teachers. In this sense, the paper reports how effective shadowing is in self-learning to improve L2 perception and production.

2. Quantitative observation of learners’ shadowing performance

2.1. Task adopted for Shadowing Marathon (SM)
SM consisted of 42 days and the task on a day is composed of four sessions. In a session, a new oral model passage (M) was presented to be shadowed, which was so long as about 30 sec. As shown in Figure 1, learners shadowed M three times (S1, S2, S3) and script-shadowed it once (SS). Script-shadowing is a...
special form of shadowing, where M is presented with its script, and it is equivalent to synchronized reading-aloud. Since all the words in the model passages were controlled to be known to the learners, SS can be reasonably regarded as shadowing with no listening breakdown. After SS, learners read aloud the script (R1) without listening to M. One week later, when learners were assumed not to bear in mind clear speech image of M, they read aloud the script again (R2).

In the following sections, learners’ performance of shadowing, i.e. perceiving and producing L2 sounds, were analyzed automatically to be quantified as scores with respect to the two aspects of speech, segments and prosody. Namely, automatic scoring was made for each case of (perception, production) × (segments, prosody). Another dimension is possible, which is analyzing gradual changes of the scores by shadowing repeatedly within a session, and by continuing the shadowing practices over days. Figure 2 overviews the analysis in this study.

2.2. Quantification of learners’ performance of production

In SM, learners were asked to reproduce a model speaker’s oral passage by imitating both the speaker’s pronunciation and his/her prosodic features, i.e. both articulatory control for segments and prosodic control for intensity, pitch, and duration.

Learners’ performance of imitating the model articulatory control was measured as quasi-phonetic distance from Sn, SS, and Rn to M [19, 20]. It was calculated automatically by Phonetic PosteriorGram-based Dynamic Time Warping, and PPG-DTW(X,M), where X = Sn, SS, and Rn, was used as quantitative metric for the imitation performance. Spectrogram is an acoustic representation of an utterance with extra-linguistic factors such as age and gender explicitly represented. PPG is, however, a quasi-phonetic representation with the extra-linguistic factors well suppressed [19, 21, 22]. A speech frame at time $t$ is converted to its posterior probability distribution over the entire set of context-dependent phones (CDP). The size of CDP is generally so large as some hundreds up to a few thousands. If we regard a discrete probability distribution as a multi-dimensional vector, we can say that a speech frame of a spectrogram is converted to its probability vector with extra-linguistic factors suppressed. This conversion is made by using the front-end of WSJ-KALDI recipe [23], a Deep Neural Network (DNN)-based speech recognizer. Phoneticians transcribe a speech sample as sequence of phones, but PPG represents it as sequence of probabilistic distributions of the phonetic classes.

PPG-DTW of two utterances, $x$ and $y$, PPG-DTW($x$, $y$) provides us with the average phonetic gap between $x$ and $y$ as well as the temporal alignment between them. Based on the alignment, learners’ performance of imitating the model control of intensity and pitch were calculated. The intensity was measured as the log of the speech signal energy, and the pitch was as the log of the fundamental frequency. After phonemic boundary detection via forced alignment using the DNN acoustic models, for each node on the DTW alignment path, it was examined whether both speech frame on $x$ and that on $y$ belonged to vowels. By using only the nodes whose $x$-frame and $y$-frame are in vowels, correlation in intensity control and that in pitch control were calculated between $x$ and $y$. Correlation in duration control was measured differently. The number of the vowel classes used in the DNN models was 45, where each class is stress-dependent, i.e. primary, secondary, and no stress. For each of $x$ and $y$, the averaged duration was calculated for each vowel class, and between $x$ and $y$, correlation of the averaged vowel durations was calculated over the entire vowel classes.

2.3. Quantification of learners’ performance of perception

How to observe a learner’s performance of perception? Listening skill is often quantified by asking him/her to conduct manual dictation of given utterances. The word-based correct dictation rate is used as objective score for his/her listening skill [24]. However, this approach is inadequate when a learner wants to know his/her online listening skill. Manual dictation of an utterance generally takes a longer time than the utterance; hence the learner may rephrase what s/he actually heard by guessing while manual dictation. Further, due to memory capacity, the presented utterance has to be short enough. These problems can be solved by changing the task from manual dictation to oral dictation, that is shadowing [25–27]. Sn is oral dictation often accompanied by listening breakdown, while SS is oral dictation without it as explained in Section 2.1. Then in this study, a learner’s listening skill is quantified as averaged score of PPG-DTW(Sn, SS). Shadowing is always online and real-time, and a model utterance so long as 1 min can be presented to learners.

In [26], by using learners’ Sn and SS, and the manual transcriptions of the Sn and SS, it was verified that PPG-DTW(Sn, SS) can approximate well the gap between the two transcripts, which clearly indicates that PPG-DTW(Sn, SS) of a learner can be regarded as a valid score of his/her listening skill. As explained in Section 2.2, while PPG-DTW($x$, $y$) can quantify deviation in articulatory control between $x$ and $y$, it can also provide us with correlation in intensity control, pitch control, and duration control between $x$ and $y$. In the following sections, we apply this approach to Sn and SS. The correlations are thought to represent learners’ performance of perceiving prosodic features that are present in a given model utterance.

3. Experiments

3.1. Participants in Shadowing Marathon

SM was designed as a special program to improve learners’ English oral skills, and it was introduced as optional assignment at home in the summer vacation. The period of the marathon is exactly the same as that of the summer vacation. 35 freshmen or sophomores majoring in Global Communication participated in SM. Their L1 was Japanese. Since SM was assigned as optional, 20 students attended SM every day while the others were absent from SM only on a few days. Before the special program, they took an online test, Versant Test [28], and their
English oral proficiency was assessed to be A1 or A2 on the CEFR scale [29]. All the participants had participated in another short shadowing program in May. Although this prior program consisted only of four shadowing sessions, all the students had gotten accustomed to using our web-based system designed to collect shadowing and reading utterances. Figure 3 shows the interface for SS, R1, and R2. The script and the waveforms of the model passages used on Day-01 were used again on Day-23 with about three-week separation, and those on Day-20 were used again on Day-42. We will depict shortly after entering at Shin-Kobe, so please be ready to get off before the train stops.

**3.2. Model utterances used for Shadowing Marathon**

Since shadowing is a task of high cognitive load, the model passages were extracted from the listening part of EIKEN Grade-2 tests [30], which are made for English learners at the level of CEFR A2–B1. The listening part of the tests uses about 30-sec monologue passages. For SM, we extracted them from the EIKEN Grade-2 tests of 2010 to 2020. Four passages were presented every day and in total, 168 passages were used for SM.

For the beginning two weeks, the speaking rate of the model passages reduced by a factor of 0.8 with SoX software [31]. For the next two weeks, it changed by a factor of 0.9, and for the last two weeks, the original passages were used.

It was highly expected that the participants’ shadowing performance would depend heavily on the semantic content of the individual passages [32], but in the current study, as all the model passages were from the EIKEN Grade-2 listening tests, the variance of semantic difficulty was considered to be suppressed to some degree. For fair and strict comparison of the participants’ performance before and after shadowing practices, the model passages used on Day-01 were used again on Day-23 with about three-week separation, and those on Day-20 were used again on Day-42.

As already shown in Figure 2, for each of the two kinds of graduate changes, within a session and over sessions, each of the two aspects of speech training, i.e. perception and production, is analyzed in the following sections. Further for each aspect, both segments and prosody are taken into account.

**4. Results and discussion**

**4.1. Gradual changes in perceiving segments and prosody**

Figure 4 shows an example of PPG-DTW(S1,SS), where segmental gaps or articulatory gaps between S1 and SS are sequentially quantified for a given model passage. As explained in Section 2.3, these sequential gaps are attributed reasonably to listening breakdown that takes place while listening. In this section, we discuss the averaged PPG-DTW(X,SS), where X

![Figure 3: The web interface for SS, R1, and R2](image)

![Figure 4: Listening breakdown quantified via DTW(S1,SS)](image)

| Table 1: Averaged segmental and prosodic gaps from X to SS |
|-----------------|-----------------|-----------------|-----------------|-----------------|
|                 | Seg \( \downarrow \) | Int \( \uparrow \) | Pit \( \uparrow \) | Dur \( \downarrow \) |
| D-01             | 1.14            | 0.71            | 0.72            | 0.70            |
| D-23             | 0.72            | 0.72            | 0.72            | 0.72            |
| D-42             | 1.09            | 1.10            | 1.05            | 0.91            |
|                  |                |                |                |                |
|                  | 0.70            | 0.72            | 0.72            | 0.72            |
|                  | 0.70            | 0.74            | 0.75            | 0.70            |
|                  | 0.70            | 0.74            | 0.74            | 0.63            |
|                  | 0.70            | 0.74            | 0.74            | 0.63            |
|                  | 0.70            | 0.74            | 0.74            | 0.63            |
|                  | 0.70            | 0.74            | 0.75            | 0.72            |

![Figure 5: Gradual reduction of listening breakdown in S1](image)
Table 2: Averaged segmental and prosodic gaps from $X$ to $M$

A = audio and T = text, presented when recording $X$.

| Seg = segments, Int = intensity, Pit = pitch, and Dur = duration. Numbers in bold are discussed in the paper. | A = audio and T = text, presented when recording $X$. |
|---|---|---|---|---|---|---|---|
| SS = segmental | SS = prosodic | SS = segmental | SS = prosodic | SS = segmental | SS = prosodic |
| Segmental gap | Proximal gap | Segmental gap | Proximal gap | Segmental gap | Proximal gap |
| $|S1−M|$ | $|S2−M|$ | $|S3−M|$ | $|S4−M|$ | $|R1−M|$ | $|R2−M|$ |
| A | A | A | A | A+T | T | T | T |
| D-01 | 1.64 | 1.76 | 1.75 | 1.53 | 1.56 | 1.59 |
| D-23 | 1.64 | 1.62 | 1.62 | 1.53 | 1.55 | 1.53 |
| D-20 | 1.62 | 1.76 | 1.75 | 1.51 | 1.50 | 1.52 |
| D-42 | 1.69 | 1.68 | 1.65 | 1.50 | 1.50 | 1.50 |
| Int† | D-01 | 0.61 | 0.62 | 0.60 | 0.62 | 0.61 | 0.60 |
| D-23 | 0.63 | 0.63 | 0.63 | 0.64 | 0.62 | 0.62 |
| D-20 | 0.56 | 0.56 | 0.57 | 0.59 | 0.58 | 0.55 |
| D-42 | 0.58 | 0.58 | 0.58 | 0.61 | 0.56 | 0.56 |
| Pit† | D-01 | 0.52 | 0.51 | 0.53 | 0.51 | 0.51 | 0.31 |
| D-23 | 0.64 | 0.60 | 0.61 | 0.60 | 0.41 | 0.35 |
| D-20 | 0.61 | 0.62 | 0.61 | 0.62 | 0.44 | 0.39 |
| D-42 | 0.65 | 0.65 | 0.66 | 0.64 | 0.47 | — |
| Dur† | D-01 | 0.56 | 0.60 | 0.58 | 0.65 | 0.61 | 0.59 |
| D-23 | 0.61 | 0.65 | 0.64 | 0.65 | 0.61 | 0.64 |
| D-20 | 0.62 | 0.62 | 0.64 | 0.67 | 0.60 | 0.65 |
| D-42 | 0.64 | 0.65 | 0.68 | 0.70 | 0.64 | — |

As discussed above, Figure 5 shows gradual reduction of the averaged $|S1−SS|$ over 42 days, which was plotted using the data from the 20 participants who attended SM every day. After the first two weeks, the speaking rate of M was modified by a factor of 0.8, and for the next two weeks, it was decreased by 0.9. For the first two weeks, similar enhancement was observed, but the score of D-23 was much lower than D-01 while that of D-42 is also much lower than D-20. In this study, shadowing practices made by S-1 were independent of teachers, but they were found to be very effective in improving the skill in L2 perception.

As for prosody, $|S3−SS|$ is larger than $|S1−SS|$ ($p<.05$) except for D-23 with Pit and for D-01 and D-23 with Dur. By shadowing repeatedly within a session, the produced prosody tends to get closer to the prosody produced in SS. However, Pit correlation is very different in [R1−SS], compared to Int and Dur. Only in Pit, [R1−SS] is significantly lower ($p<.01$) than $|S3−SS|$ on every day. R1 was recorded only after text recording SS, which is with text and audio. It seems difficult for learners to produce a model-like pitch pattern without audio. Over sessions, improvements are significant in [S1−SS] ($p<.05$) except for D-20 to D-42 of $|S1−SS|$ with Pit. It is interesting that, on D-23 and D-42, the correlations are very close to each other irrespective of $n$ in Sn and the kind of prosody.

4.2. Gradual changes in producing segments and prosody

Table 2 shows the averaged segmental gaps and the averaged prosodic correlations, calculated via PPG-DTW(0,M) on the four special days, where $X = S1, S2, S3, SS, R1,$ and $R2$. To conduct two-way ANOVA post hoc multiple comparisons test, we chose $|S1−M|, |S3−M|, |SS−M|,$ and $|R1−M|$. By comparing $|S1−M|$ with $|S3−M|$ within a session for Seg, we can say that learners’ articulatory control for segments became significantly closer ($p<.01$) to the model control (M) except for D-23. However, $|S3−M|$ and $|R1−M|$ are significantly smaller ($p<.001$) than $|S3−SS|$ on every day. R1 was recorded only after text recording SS, which is with text and audio. It seems difficult for learners to acquire how to control their vocal organs with no instruction.

By comparing $|S1−M|$ on D-01 with that on D-23, and $|S1−M|$ on D-20 with that on D-42, we can say that articulatory control in S1 became significantly closer ($p<.001$) to M through 23-day training. In $|SS−M|$ and $|R1−M|$, however, both between D-01 and D-23 and between D-20 and D-42, no differences were observed, indicating that 23-day training did not change learners’ articulatory control when text cue is available. Their pronunciation seems to be fossilized and stuck to the visual form of words, and to promote learners to improve their pronunciation, explicit instructions will be necessary.

Learning difficulty is found again in the prosody correlations. This is reasonable because learners’ production heavily depends on text cue, which does not help learners to learn prosodic control. By comparing $|S1−M|$ with $|S3−M|$ within a session for Int, Pit, and Dur, significant improvements were not observed in any case. Even with text cue in SS, significant improvements from S3 were observed only in a few cases, such as D-20 and D-42 with Int ($p<.001$) and D-01 with Dur ($p<.01$). With text cue, learners may be at least able to detect the rhythmical structure for the text, but not the intonational structure (Pit). Further, in almost all the cases, the scores tend significantly ($p<.001$) to decrease from $|SS−M|$ to $|R1−M|$. R1 was recorded immediately after SS, but it was difficult for learners to duplicate the prosodic control of M when recording R1. Especially, the decrease from $|SS−M|$ to $|R1−M|$ is remarkable with Pit, indicating that duplicating the pitch control of M is extremely difficult even after shadowing M four times. In [33], visual presentation of pitch control was found to be more effective to Chinese learners of Japanese than audio presentation. This may also be the case with Japanese learners of English.

As for changes over sessions, while the scores of $|S1−M|$ and $|S3−M|$ for all the three prosodic features were increased significantly ($p<.05$) from D-01 to D-23, no differences were found in either case from D-20 to D-42. In $|SS−M|$ and $|R1−M|$, significant improvements ($p<.05$) are observed in Int and Pit, but only from D-01 to D-23. We can say that the learners’ performance of imitating the model prosodic control reaches a plateau in the beginning half of the marathon.

To wrap up, we claim that SM was remarkably effective to improve L2 perception, but it was difficult to improve L2 production with no explicit instructions. If we prepare explicit instructions for the next SM, we will visualize articulatory gaps and prosodic gaps observed between learners and the model.

5. Conclusions

This paper reported what we found in our intensive training of 42-day Shadowing Marathon. In shadowing, both L2 perception and L2 production are involved, and by introducing script-shadowing, the gradual changes of perception and production were observed by using PPG-DTW. The obtained data were analyzed for each case of (perception, production) × (segments, prosody) × (within a session, over sessions). Results of analysis showed that shadowing can improve learners’ performance of L2 perception significantly without any explicit instructions from teachers. However, shadowing in a self-learning scenario was also found not to be effective for L2 production. To make learners aware of articulatory gaps and prosody gaps to model speakers, explicit and visual instructions will be necessary, which will be provided to learners in our future work.

Every model passage was so long as about 30 sec, and it is surely impossible for learners to store the entire pitch movement over the passage in memory. Assuming that the pitch control in learners’ production was made based on their internal model for pitch control, we can say that they do not have a good pitch control model for reading aloud.
6. References


[31] SoX SourceForge.net.
