A Unified Framework to Improve Learners' Skills of Perception and Production Based on Speech Shadowing and Overlapping

Nobuaki Minematsu†, *Noriko Nakanishi*‡, *Yingxiang Gao*†, *Haitong Sun*†

†Graduate School of Engineering, The University of Tokyo, Japan
‡Faculty of Global Communication, Kobe Gakuin University, Japan
{mine,gyx,sunhaitong}@gavo.t.u-tokyo.ac.jp, nakanisi@gc.kobegakuin.ac.jp

Abstract

A unified framework to improve learners' skills in perceiving and producing L2 sounds is demonstrated based on speech shadowing and overlapping. Speech shadowing is a training method, where learners are asked to reproduce a given model speech (M) as immediately as possible, and it was proved to be effective in enhancing their L2 speech perception. After several trials of shadowing, the learners are provided with M's script to continue shadowing with no delay, called overlapping. By comparing the shadowing speech (S) and the script-shadowing speech (SS), shadowing breakdowns are measured sequentially, which can characterize listening breakdowns. By comparing M and SS, the prosodic and segmental gaps are analyzed sequentially and presented visually to learners along with imitation scores. All the tasks are implemented as interactive speech games, which help learners to become more proficient in L2 speech perception and production.

Index Terms: language learning, perception and production, speech shadowing, utterance comparison, speech game

1. Introduction

Learners of a new language often have difficulties in perceiving and producing L2 sounds, especially when the phonemic system of their L1 is remarkably different from that of the L2. Since perception and production are explained in speech science to be linked to each other, perception training is often implemented through production training. In this presentation, however, a framework is demonstrated to improve the individual skills in an explicit but unified way based on speech shadowing and overlapping. Given a model speech without its script, perception training is provided to learners, and after that, with the script, production training is provided. L2 speech perception and production training is connected seamlessly.

Speech shadowing was originally proposed in psycholinguistics [1] to analyze human listening processes. In the late 1990s, it was introduced to language education by a Japanese teacher of English [2] as a training method to enhance learners' word perception. In [2], the effects of shadowing was discussed based on the working memory theory [3]. By asking learners to vocalize the perceived message immediately, speech acoustics are converted into their abstract form called phonological representation, and it is stored firmly in working memory. The effectiveness of shadowing can be understood by comparing listening with reading. Although the past information in reading can be found visually on the text, that in listening exists only mentally in the brain. Because of this difference, listening is considered to invoke more frequent accesses to the mental representation stored in the brain. Speech shadowing can activate working memory and enhance its capacity.

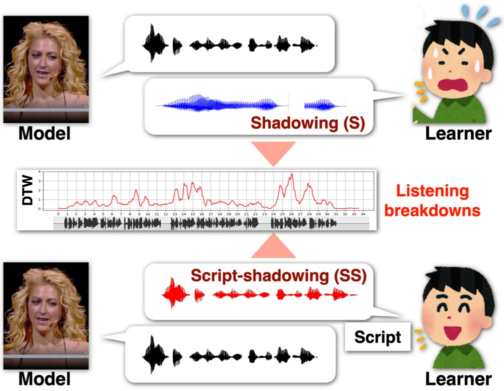


Figure 1: *Measurement of listening breakdowns*

As shown in Figure 1, a shadowing speech (S) and its corresponding script-shadowing speech (SS) are compared via. Dynamic Time Warping (DTW) after they are converted into Phonetic PosteriorGram (PPG) [4]. In shadowing, listening breakdowns often happen, which are observed acoustically as shadowing breakdowns. In script-shadowing, however, they never happen. PPT-DTW(S,SS) draws a curve of listening breakdowns, and [4] verified a high effectiveness of continuing shadowing drills over days to reduce the score of PPG-DTW(S,SS).

However, in [4], PPG-DTW(M,SS) was not reduced at all, indicating that the phonetic gaps (and the prosodic gaps) from SS to M were not reduced at all by continuing shadowing drills. This ineffectiveness has a clear reason that in all the shadowing training in [4], no corrective feedback in terms of the phonetic and/or prosodic gaps were provided, where learners cannot know which parts of their speech should be corrected. To solve this problem partially in [5], after SS, learners were asked to overlap the prosodic features (intensity, pitch, and tempo) of their imitative speech on M as precisely as possible, and after recording, the remaining prosodic gaps were instantly visualized for learners along with the prosodic imitation scores provided. This training effectively raised learners' attention to their prosodic features and how different they were from M. Although learners' prosody became effectively closer to M immediately after the overlapping drills, it became less similar again when their reading speech was recorded one week later [5]. Continuous training with corrective feedback will be necessary to acquire a good skill of prosodic control.

2. Task design and system development

2.1. Typical task of shadowing and overlapping

Being inspired by [4] and [5], we developed a system that connect regular shadowing drills and overlapping (no delay shadowing) drills seamlessly. The former aims at enhancing perception skills and the latter aims at enhancing production skills. For a model speech sample of M, S1–S2–S3–SS*–O*–R1*–R2* is assigned to learners as a typical task. Here, Sn is the n-th one-time recording of speech shadowing, and SS* means a recording of script-shadowing, where '$*$' means that re-recordings are allowed if vocalization errors are found. O is overlapping drills, where corrective feedback is provided for the learners. R1 is a recording of reading aloud immediately after the O drills, and R2 is another recording one week later. Re-recordings are allowed for O, R1, and R2 for the learners' best performance.

2.2. Interactive user interface for overlapping

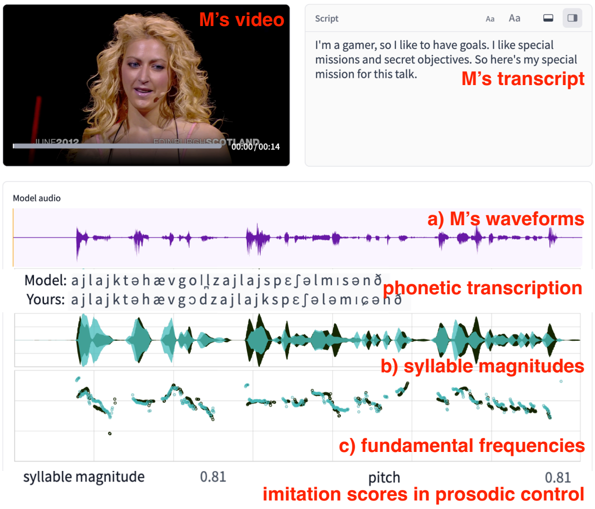
Figure 2 shows the interactive user interface in O*, where a model speech (M) is provided in the form of a video, and before recording, M's visualization is provided for learners in three ways of a) waveforms, b) normalized syllable magnitudes, and c) normalized logarithmic fundamental frequencies. In Figure 2, b) and c) are drawn in black. During recording, the indicator moves in M's waveforms from left to right, showing when the next phrase starts. This visual hint is very useful for successful overlapping. Instantly after recording of a learner's imitative speech, its visualization of b) normalized syllable magnitudes and c) normalized logarithmic fundamental frequencies is drawn in green, and correlation in b) and c) between M and the learner's production is calculated separately as similarity scores. When comparing a model speech with a learner's imitative speech, DTW is often conducted as preprocessing between them. In our system, however, automatic time warping is not performed and the learner tries to overlap his/her speech as precisely as possible on M. This induces the learner to acquire a good skill of duration control. English rhythm is composed of alternation of an stressed syllable and one or more unstressed syllables. The duration of a stressed syllable is longer than that of an unstressed syllable. A good durational control is very important to produce natural English rhythm.

2.3. Normalized syllable magnitudes

Alternation of an stressed syllable and one or more unstressed syllables in M is visualized by using its waveform envelopes. After full-wave rectification, the moving average by 100 msec is calculated to derive the waveform envelope pattern of M. The same algorithm is applied to a learner's speech to draw its waveform envelope pattern. In Figure 2, the syllable magnitudes of M are drawn in black and those of the learner are drawn in green. Both are normalized to have 1.0 as the maximum magnitude. The correlation between the two is calculated to be used as an imitation score in syllable magnitude control.

2.4. Normalized logarithmic fundamental frequencies

Fundamental frequencies are extracted from M and the learner's overlapping, and their logarithmic values are calculated. The log average over an utterance is calculated for each of the two, and by subtracting the log average from each, normalized logarithmic fundamental frequencies are obtained for M and the learner's overlapping. In Figure 2, the two normalized patterns are drawn in black and green. To calculate the imitation score in pitch control, the weighted correlation is adopted. If we denote the original correlation as $cor(p_M(t), p_L(t))$, the weighted correlation is obtained as $cor(m_M(t)p_M(t), m_L(t)p_L(t))$. Here, $p_X(t)$ is the logarithmic fundamental frequency of X at time t, and $m_X(t)$ is the syllable magnitude. X is M (model) or L (learner). The weighted correlation emphasizes pitch gaps with larger magnitudes and suppresses those with lower magnitudes, which is considered to accord with subjective assessment.



Script

I'm a gamer, so I like to have goals. I like special missions and secret objectives. So here's my special mission for this talk.

Model audio

Model: a j l a j k t ə h æ v g o l l̪ z a j l a j s p ɛ ʃ ə l m ɪ s ə n ð
Yours: a j l a j k t ə h æ v g ɔ d z a j l a j k s p ɛ ʃ ə l ə m ɪ ɕ ə n ð

syllable magnitude 0.81 pitch 0.81

Figure 2: *Interactive user interface for overlapping*

2.5. Universal phone recognition to generate phonetic gaps

In addition to the prosodic gaps between M and a learner's overlapping, the phonetic gaps between the two are provided for the learner. Allosaurus [6], which is a multi-language phone recognizer, is used to represent the two speech samples using quasi-phonetic symbols. While the prosodic gaps are illustrated with figures, the phonetic gaps are presented with symbols.

2.6. Voices from actual users of our system

Our interface works instantly and interactively with learners, who enjoy shadowing and overlapping drills like speech games. The video in Figure 2 can be replayed after replacing the original audio track with the learner's imitative voice. According to questionnaires, some learners enjoyed our drills until midnight.

3. Conclusions

In this presentation, a unified framework to improve learners' skill of perception and production is demonstrated based on seamless connection of shadowing and overlapping. Our interface is language-independent, and it can be applied directly to perception and production training in any language.

4. References

[1] W. D. Marslen-Wilson, "Speech shadowing and speech comprehension," *Speech Communication*, vol. 4, pp. 55–73, 1985.

[2] K. Tamai, "The effect of shadowing on listening comprehension," in *The Study of Current English*, vol. 36, 1997, pp. 105–116.

[3] A. Baddeley, *Working memory*. Oxford University Press, 1986.

[4] T. Kunihara, C. Zhu, N. Minematsu, and N. Nakanishi, "Gradual improvements observed in learners' perception and production of L2 sounds through continuing shadowing practices on a daily basis," in *Proc. INTERSPEECH*, 2022, pp. 1303–1307.

[5] N. Nakanishi and N. Minematsu, "Effects of karaoke shadowing on EFL learners' segmental and prosodic features," in *Proc. CamTESOL*, 2023.

[6] X. Li, S. Dalmia, J. Li, M. Lee, P. Littell, J. Yao, A. Anastasopoulos, D. R. Mortensen, G. Neubig, A. W. Black, and F. Metze, "Universal phone recognition with a multilingual allophone system," in *Proc. ICASSP*, 2020, pp. 8249–8253.