WithYou: An Interactive Shadowing Coach with Speech Recognition

Xinlei Zhang
Interfaculty Initiative in Information Studies,
The University of Tokyo
Bunkyo-ku, Tokyo, Japan
xinleizhang@g.ecc.u-tokyo.ac.jp

Takashi Miyaki
Interfaculty Initiative in Information Studies,
The University of Tokyo
Bunkyo-ku, Tokyo, Japan
miyaki@acm.org

Jun Rekimoto
Interfaculty Initiative in Information Studies,
The University of Tokyo / Sony CSL
Bunkyo-ku, Tokyo, Japan
rekimoto@acm.org

ABSTRACT
Speech shadowing, in which the subject listens to native narration sound and tries to repeat it immediately while listening, is a proven way of practicing speaking skills when learning foreign languages. However, since the narration is independent of user's speech, the playback cannot make an adjustment when the learner fails to catch up, and this makes shadowing difficult. We propose WithYou, a system based on Automated Speech Recognition (ASR) that is able to adjust narration playback during a live shadowing speech. WithYou compares the student's live speech with the narration playback to detect shadowing mistakes. In addition, WithYou is able to handle pauses and recognize repetitive phrases in shadowing practice. A user study shows that practicing shadowing with WithYou is easier and more effective compared with conventional methods.

Author Keywords
Computer-Assisted Language Learning; Speaking Support; Speech Recognition; Shadowing; Voice Interface

ACM Classification Keywords
H.5.2. Information Interfaces and Presentation: Voice I/O

INTRODUCTION
Mastering speaking skill in a second language is essential to communicate with others in international situations.

A typical way for students to practice speaking skill is listening to native speaker's narrations and then repeat word by word while listening – an approach that is commonly referred as shadowing. By doing this, it has been reported that second language (L2) learners can expect great improvement in their general language ability, especially in listening [2] and speaking skills [5].

UIST'16 Adjunct, October 16-19, 2016, Tokyo, Japan
ACM 978-1-4503-4531-6/16/10.
http://dx.doi.org/10.1145/2984751.2985704

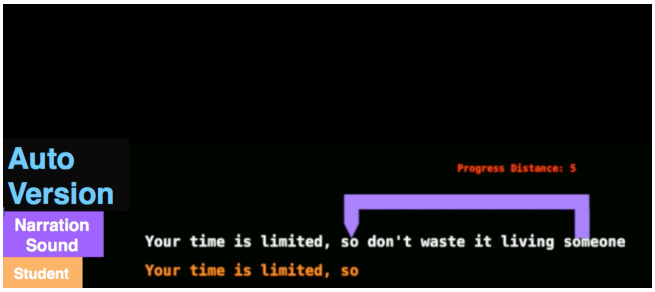

Figure 1. Demonstration of WithYou. When the difference of these two progress indicators cross the pre-defined threshold, the narration will be rewinded to the last punctuation.

However, in shadowing practice, it is common that students fail in catching up with the narration because of mispronunciation or misarticulation. In this situation, students can do nothing but restart the playback from the top because the narration is not aware of the student's performance, and thus cannot make playback adjustments when they fail.

We propose WithYou, a shadowing support system which is capable of adjusting the narration playback based on user's speech progress in real-time. WithYou tracks user's speech on-the-go and compares it with the word that is being played. If there is a huge distance (over the transcription) between the two, which indicates mispronunciation or user's failing behind the narration, then the playback is automatically rewound to the last punctuation that is near to the user's mistake.

In addition, WithYou is capable of dealing with pauses and recognizing repetitive phrases in a learner's speech. This has made the recognition robust, and can ensure that the feedback it provides is timely. As a result, WithYou can improve shadowing efficiency by eliminating unnecessary user operation and providing fast recovery from shadowing mistakes.

SYSTEM DESCRIPTION
To practice shadowing using WithYou, learner prepares a narration sound file and its transcription as the template for learning.

As demonstrated in Figure 2, in WithYou, speech recognition is used to track user's speech. To do that, 1st-pass recogni-

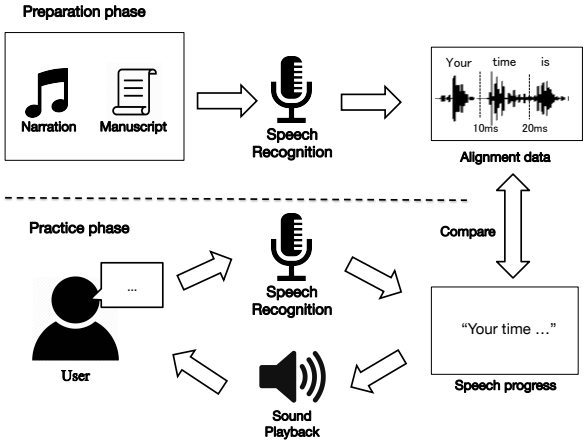

Figure 2. System Architecture.

tion in "Julius" [6] is adopted. The transcription is converted to a grammar file to restrict the recognizer to recognize only the words contained in the transcription. To ensure recognition robustness, an acoustic model that represents "silence" in speech is added to the grammar file at places that correspond to punctuations in the transcription. As the result, WithYou is able to recognize live shadowing speech which contains repeating phrases and pauses with high recognition rate in real-time.

On the other hand, WithYou monitors playback progress as the "correct answer" to refer to. To do that, the system first aligns the transcribed text with the narration sound, identifying which time segments in the speech data correspond to particular words in the transcription data. Then, during practice, system use this "word-timing" mapping to control narration playback on a word basis.

In WithYou, distance (over the transcription) between the user's real-time speech and the narration playback is used to determine shadowing mistakes. We set a distance threshold to determine shadowing mistakes. If the distance between sound playback and the user's speech is equal or higher than the threshold, then the situation is judged as "shadowing failure" and the narration will be rewound back. Otherwise, the narration sound will keep playing as the user is following the narration well. The default threshold for determining shadowing mistakes is set to five words, which is concluded from a preliminary experiment that measured the distance when participants make mistakes.

FINDINGS

To evaluate the system, we recruited seven graduate non-native users to practice shadowing with three different methods in randomized order. The three methods are conventional shadowing, WithYou and a manual version of WithYou which provides a key for users to rewind the playback to the last punctuation when they make mistakes.

By measuring their before-practice and after-practice shadowing performance, we found that compared with the conventional method, WithYou helps more people improve shadowing performance. Moreover, users with relatively low shadowing performance before practicing improved to a higher extent than those with high initial shadowing performance. In addition, among all the users who improved their shadowing performance by all the three methods, those who practiced with WithYou improved more.

We also did a subjective evaluation on the system's usability and shadowing difficulty with these three systems. The result shows that compared with the conventional method, shadowing with WithYou is easier and its UI is highly preferred over the conventional one.

RELATED WORK

Automated Speech Recognition (ASR) has been adopted for language learning support in many works. One typical ASR application for language learning support is pronunciation training [7, 1, 11, 9]. Typical examples under this direction include LISTEN [10] – a reading tutor system that listens to a child reading sentences, and then evaluate their pronunciation over specific words. Another example is CALL Mandarin [3], which visualizes speech signal as well as the fundamental frequency (F0) contour in real-time for pronunciation training.

Another kind of application aims to build a system that enables users to have conversations with computer virtually, such as SPELL [4], a system that provides a specific teaching mode for pronunciation training such as intonation or vowel quality.

There are also works that focus on pronunciation evaluation for shadowing [8]. However, this evaluation method works off-line and it does not provide instructions on how to improve pronunciation for each word in real-time.

CONCLUSION AND FUTURE WORK

We proposed WithYou, a shadowing support system that uses speech recognition to track user's speech in real-time and then adapts the narration sound playback to the user's speech during shadowing practice. Our user study shows that WithYou is capable of helping users, especially those with relatively low shadowing performance improve learning efficiency by making shadowing easier to do.

For future improvements, we plan to add real-time error visualization into the system because sometimes it is difficult for users to pinpoint mistakes with only sound feedbacks. Moreover, we plan to provide targeted ability training, such as "pronunciation only" or "rhythm only" with configurable difficulty so that users with different language ability can have targeted support with WithYou.

REFERENCES

1. Eskenazi, M., Kennedy, A., Ketchum, C., Olszewski, R., and Pelton, G. The native accent pronunciation tutor: measuring success in the real world. In *SLaTE* (2007), 124–127.

2. Hamada, Y. The effectiveness of pre-and post-shadowing in improving listening comprehension skills. *LANGUAGE TEACHER 38* (2014), 4.

3. Hansjrg, M., and Hussein, H. CALL mandarin-results and conclusions from a project for developing a computer-assisted pronunciation training program. *Systemtheorie, Signalverarbeitung, Sprachtechnogie. TUD Press* (2013).
4. Hiller, S., Rooney, E., Laver, J., and Jack, M. Spell: An automated system for computer-aided pronunciation teaching. *Speech Communication 13*, 3 (1993), 463–473.
5. Hori, T. *Exploring shadowing as a method of English pronunciation training*. PhD thesis, kwansei gakuin university, 2008.
6. Lee, A., Kawahara, T., and Shikano, K. Julius—an open source real-time large vocabulary recognition engine. In *EUROSPEECH2001: the 7th European Conference on Speech Communication and Technology* (2001).
7. Lee, S., Noh, H., Lee, J., Lee, K., and Lee, G. Postech approaches for dialog-based english conversation tutoring. *Proc. APSIPA ASC* (2010), 794–803.
8. Luo, D., Shimomura, N., Minematsu, N., Yamauchi, Y., and Hirose, K. Automatic pronunciation evaluation of language learners' utterances generated through shadowing. In *INTERSPEECH* (2008), 2807–2810.
9. Meng, H., Lo, W.-K., Harrison, A. M., Lee, P., Wong, K.-H., Leung, W.-K., and Meng, F. Development of automatic speech recognition and synthesis technologies to support chinese learners of english: The cuhk experience. *Proc. APSIPA ASC* (2010), 811–820.
10. Mostow, J., et al. Giving help and praise in a reading tutor with imperfect listening–because automated speech recognition means never being able to say you're certain. *CALICO journal 16*, 3 (2013), 407–424.
11. Tsubota, Y., Kawahara, T., and Dantsuji, M. Practical use of english pronunciation system for japanese students in the call classroom. In *Proc. ICSLP*, vol. 15 (2004), 1689–1692.