Speech Shadowing Support System in Language Learning

Carson Lee[1], Shinobu Hasegawa[2]
[1]School of Information Science,
[2]Research Centre for Advance Computing Infrastructure
Japan Advanced Institute of Science and Technology
Nomi, Japan
[1]crsnlee@jaist.ac.jp, [2]hasegawa@jaist.ac.jp

***Abstract*—Verbal communication is a major part of a language, but there are not many systems/solutions in the market that caters to self-learning of spoken language. Traditional classroom learning is affected by the cultural background of the learning environment, thus students from different backgrounds might end up speaking a different dialect or accent, and this might result in miscommunication. Speech Shadowing is an experimental technique where a subject repeats speech immediately after hearing it. However, it is a time consuming method as it requires 1-on-1 tutoring. In this paper, we present our approach to utilizing this method for a self-supported learning system and how to utilize technology to improve its efficiency over traditional speech shadowing methods.**

Keywords- Speech Shadowing; language learning; mobile learning

I. INTRODUCTION

The development of speech production throughout an individual's life starts from an infant's first babble and is transformed into fully developed speech by the age of five [1]. It is a type of cognitive skill, and thus, we cannot teach it the same way we would teach sciences or history, as cognitive skill learning is the learning of a skill or knowledge that is hard to symbolize.

Today, the English language is the de-facto lingua franca. Despite the widespread usage of English, there exists many variations of the English dialects, such as British English, Cockney English, American English, Engrish (generally refers to poor Japanese influenced English), Manglish (Malaysian English), and many more. The more formal dialects such as British English and American English are often used as the standard for major English proficiency test such as IELTS and TOEFL. Other dialects have evolved from their original one often due to cultural and environmental influences. For example, Manglish is a result of assimilating the many languages spoken in Malaysia into the English language. Another example would be Japanese English, where students often learn English the aid of furigana. There exist many consonants and vowels that are mutually exclusive in both language, and thus, a student who learns to speak English via furigana often end up having a hard time to be understood by non-Japanese English [2]. For example, a Japanese would often pronounce “eight” as “ei-to (エイト), fight as “figh-to (ファイト)”, or “the” as “za (ザ)”.

Another reason for doing this research is to reduce miscommunication due to different accents/dialects. Looking at the aviation industry, we can observe that many accidents have resulted from communication error. The nuances of a language can be complicated and the same word can carry multiple meanings. Depending on how it is delivered, the message conveyed might vary [3].

Furthermore, in this digital age, information can be disseminated very quickly through the internet and thus many people can spend their downtime (riding on a bus/train, waiting in line, etc.) to absorb more information via their mobile devices. This allows people to learn almost anywhere and anytime. However, some domains are not as easy to be learnt without the presence of an instructor or teacher. There are many applications that cater to language learning. However, the amount of smartphone applications that focuses on improving a learner’s speaking skill is also very limited. Most of these applications focuses on the reading/writing aspect, and the speaking aspect is usually very simple (such as pronunciation of a single word at a time). In teaching a student to speak a foreign language, most attention is devoted to the correct pronunciation of sounds and isolated words. Generally speaking, much less attention is paid to a correct production of intonation [4].

In this research, the aim is to utilize Speech Shadowing to improve verbal communication abilities according to a certain dialect/accent. The scope of this research will cover the development of a system to improve a user’s speaking skill in the English language via Speech Shadowing. In Section II, we will describe what speech shadowing is, and the problems faced by this method. The learning model that will be applied is discussed in Section III. At Section IV, we will describe our approach to solving the problems described earlier, and their algorithms. Section V will be the conclusion to this paper, summarizing it.

II. SPEECH SHADOWING

One way to improve a user’s speaking ability is via Speech Shadowing. Speech shadowing is an experimental technique where a subject repeats speech immediately after hearing it, usually through headphones to reduce noise and/or speech jamming. The reaction time between hearing a word and pronouncing it can be as short as 254ms or even 150ms [5]. While a person is only asked to repeat words, they also automatically process their syntax and semantics. Words repeated during the shadowing practice imitate the

 ISBN: 978-1-61208-541-8

parlance of the overheard words more than the same words read aloud by that subject. We can also observe a similar behaviour in children as they begin to develop their speaking ability. They are often predisposed to imitate/shadow words and speech as a way to guide themselves to enter their cultural community [6]. Since children utilize this method to learn a language, it could be possible to utilize the same method for adults. In fact, learning the patterns of intonation is thought to take place unconsciously by mere imitation. That is, by listening to, and repeating model utterances the foreign-language learner has to acquire a proper intonation.

A. *Traditional Speech Shadowing*

In the traditional speech shadowing method, an instructor is needed to sit there to evaluate the student performing speech shadowing. Fig. 1 illustrates the usual steps for a speech shadowing session and they are as follows:

1) Playback of a speech/conversation recording

2) Student performs speech shadowing (repeats the heard speech with minimal delay as clearly and loudly as possible)

3) Instructor listens to the shadowed speech and provides evaluation/feedback to the student

4) The student attempts to improve based on the given feedback and retries the process on a later date.

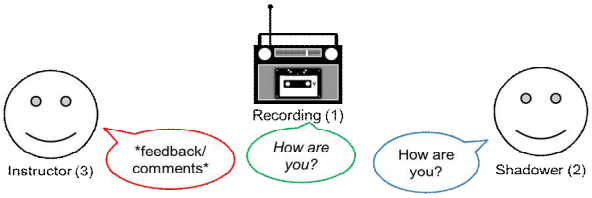

Figure 1. Traditional Speech Shadowing Session Setup

Due to the fact that one instructor can only focus on one student at a time during a speech shadowing session, the process becomes inherently expensive. The instructor should also be highly trained and/or be very experienced with the language and dialect that he is instructing on. This only adds to the cost of speech shadowing. Furthermore, because speech shadowing is still largely an experimental technique, there exists no formal feedback/evaluation method. Verbal and/or written feedback comments may be subjective and thus prove to be ambiguous at times. This makes it hard to keep track of past performance that could be used to help the student improve.

III. LEARNING MODEL

The learning model used in this research would be the Cognitive Apprenticeship Theory. It is the process where a master of a skill teaches it to an apprentice via 5 steps/stages, which is modelling, coaching, reflection, articulation and exploration [7].

- Modelling – Demonstrating the thinking process
- Coaching – Assisting and supporting student cognitive activities as needed (includes scaffolding)
- Reflection – Self-analysis and assessment
- Articulation – Verbalizing the results of reflection
- Exploration – Formation and testing of one's own hypothesis

The focus of this research will be modelling, coaching, and reflection, whereby the original speech would be the model, the scaffolding being the coach, and self-evaluation being the reflection.

Coaching would be done via scaffolding with the 4 elements being used to control the difficulty. The 4 elements would be discussed in Section IV.A. Initially the user would be given a questionnaire to judge their own level and a speech of appropriate difficulty will be given to the user to shadow without any scaffolding. After the initial rating, the user will then be given scaffolding suited to his level.

At this phase of the research, reflection would be self-evaluation. The user would be given some visual aids such as the audio waveform in order to evaluate his own performance and then he would answer a questionnaire. Feedback such as graphs will then be provided to show the user his current performance in various aspect of speech such as intonation, tempo, and pronunciation. The user can also track his past performances. These metrics would be fed back to the system in order to determine the coaching needed for the next shadowing session.

IV. OUR APPROACH

Due to the impracticality of the traditional speech shadowing for language learning on a larger scale, we propose a system that is able to replace the role of the instructor of the traditional method. At the same time, we want the system to provide a more tailored learning method for the student using it, so that he/she may learn and improve faster. The lack of an instructor also allows the student to learn independently, and due to the simplicity of our proposed system, the system can also be implemented on a mobile system, allowing students to learn anywhere and anytime. This will be approached by 2 methods

A. *Speech Shadowing System*

The system would contain recordings of speeches to be listened to, and the speeches will be sorted by difficulty levels according to their length, speed, and difficulty of the words or sentences. The system would also pickup and record the speech shadowed by the student so that it can be analysed to provide feedback and evaluation.

The difficulty level of the speeches will be determined by the following elements of speech:

- Length of speech
- Speed/tempo of speech
- Difficulty of words used
- Number of stresses/intonation in sentence

The reason the elements are chosen are explained as follows. The length of speech can directly affect the difficulty of the speech as it increases the cognitive load as it becomes longer. The speed and tempo of a speech also affects the difficulty of a speech as speech rate (the number

 ISBN: 978-1-61208-541-8

of words spoken per minute) has been used extensively in the previous research of oral fluency [8] [9] [10]. Previous research also found that speech rate positively correlated with other measures of fluency, such as length of speech without pauses, hesitations, or repeats [11] [12]. Difficulty of words that appear is also taken into consideration as it can affect the understanding of a shadowed speech.

The number of stresses and intonation in a sentence can affect the difficulty of a speech because linguistic, syntactic and semantic information is more easily conveyed when a speaker produces the correct variations in pitch in a speech utterance [13]. Of all the elements of a target language, the intonation appears to be the most difficult to acquire [14]. First, because the intonation in infants is learned at a very early stage in the language-acquisition process [15], it is most resistant to change. Second, as a result of the fact that suprasegmental patterns are particularly deep-rooted, foreign language learners often superimpose the prosodic features of their mother language on the sounds of the foreign language. For this reason, foreign-language learners are often not aware of any differences in intonation between the mother language and the foreign language [4]. This makes the number of stresses in a sentence directly related to the difficulty of shadowing a sentence.

We propose that the system runs on a smartphone so that it can make the learning process more accessible as year-by-year digital media audiences are increasingly coming from mobile devices [16]. Setting up a headset is also easier and less costly compared to a desktop-based system as most smartphone owner would already have access to a headset. This also ensures students can learn on the go, although they should use the system in an isolated environment to avoid disturbing others.

An account would be created for progress tracking purposes. First time user of the system would take a standardised test and answer a short questionnaire to determine his/her initial level and proficiency (system initialization). The test would be a speech shadowing session without any support from the system. The difficulty of the speech would also be a predetermined medium level speech.

Under a normal use-case condition (post-initialization), students would login to the system and be presented with a list of recommended speeches to shadow, which are determined by the student's proficiency and level. The amount and type of scaffolding provided during a shadowing session is affected by the student's proficiency and level along with the difficulty of the speech attempted. Take for example Student A is rated by the system as a level 6 user (out of 10 possible levels, with 1 being lowest and 10 being highest) attempts a speech of difficulty level 2 (out of 5 difficulties with 1 being easiest and 5 being the hardest). Student A would get no scaffolding as his proficiency should be sufficient to attempt the speech with ease. However, if Student A attempts a level 5 difficulty speech, all scaffolding would be activated to help Student A with his shadowing attempt. In the optimal scenario, Student A should be attempting speeches with difficulty level that matches his own proficiency level, as the effect of learning via speech shadowing can be affected by having too much scaffolding.

TABLE I. USER LEVEL AND SPEECH DIFFICULTY LEVEL MATCHING

(User Level) / 2	Scaffolding	Notes
> speech level	No scaffolding	Scaffolding provided depends on user's proficiency on speech elements as well
= speech level	Partial Scaffolding	
< speech level	More / All Scaffolding	

Fig. 2 shows the types of scaffolding that is provided by the system:

1) Speech transcript
2) Pronunciation help
3) Highlighting sentence stress points
4) Speed control for recordings

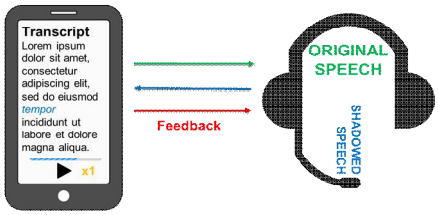

Figure 2. Scaffolding 1,3, and 4 being used in a shadowing session

Figure 2 provides an example of 3 types of scaffolding being used to coach the user. The Transcript is there to help the user know what exactly he/she is saying while the highlighted word is the part of speech where a stress is needed. A playback speed is also displayed and can be used to change the speed of the playback to help the user cope with higher difficulty speeches.

B. *Performance Evaluation*

In order to provide the student with a valuable feedback and evaluation without an instructor, a way to grade the speech shadowing session needs to be devised. Using 3 metrics, the user's performance can be measured more accurately and the training time needed can be shortened as the student knows what he has to focus on to improve. The 3 metrics that is used in this system are:

- Intonation
- Pronunciation
- Tempo

The user would evaluate the 3 metrics on his own by comparing his shadowed speech to the original recording. Using a simple questionnaire, the student would rate his own performance compared to the sample recording. The system will provide some visualisation of the data in order to make the process easier.

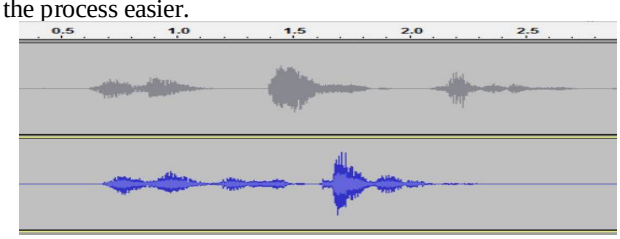

Figure 3. **Visualization** of intonation differece

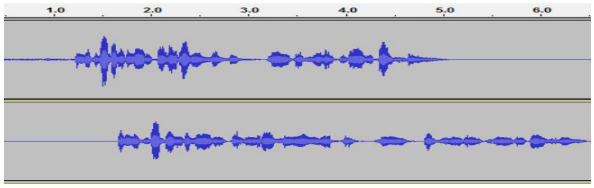

Figure 4. Visualisation of tempo difference

After the evaluation is done, the system would use the data to determine if a user has levelled up and thus have some of the scaffolding removed. The data would also be archived so that users can keep track of their past performance and pinpoint where their weakness is.

C. Evaluation Algorithms – Determining user level

1) Post system initialization

$$S_{cs} = (S_i \times W_i) + (S_t \times W_t) + (S_p \times W_p)$$

$$UL = \frac{S_{cs}}{10}$$

$$S_{pp} = S_{cs}$$

2) Next Iterations

$$S_{cs} = (S_i \times W_i) + (S_t \times W_t) + (S_p \times W_p)$$

$$UL = \frac{(S_{cs} + S_{pp}) \times W_{pp}}{10}$$

$$S_{pp} = \frac{S_{cs} + S_{pp}}{2}$$

TABLE II. USER LEVEL AND SPEECH DIFFICULTY LEVEL MATCHING

Variables	Definition
S_i	Score – intonation
S_t	Score – tempo
S_p	Score – pronunciation
S_{cs}	Score – current session
S_{pp}	Score – past performance
N_s	Total Number of Sessions
W_i	Weightage – intonation
W_t	Weightage – tempo
W_p	Weightage – pronunciation
UL	User Level

The variables are calculated after the user takes the standardised test during the system initialization phase.

V. CONCLUSION

In conclusion, speech shadowing could be a good method for learning and improving one's speaking proficiency. However, the traditional method of it is not suitable to implement on a larger scale. Therefore, we propose the idea of a speech shadowing support system so that we can overcome the constraint. By breaking down the elements in a speech, the system will be able to provide a more tailored coaching method to individual students. By further splitting up the user into different levels, the learning curve would not be as steep, making the task of learning much less daunting.

In future research, automated evaluation by the system will replace self-evaluation system and it would provide a more standardised method for evaluation and thus give better feedback for reflection. Just like in self-evaluation, the system would use intonation, tempo, and pronunciation to evaluate the user, and the result of the automated evaluation would be fed back into the system for the next session.

ACKNOWLEDGMENT

The work is supported in part by Grant-in-Aid for Scientific Research (C) (No. 26330395) from the Ministry of Education, Science, and Culture of Japan.

REFERENCES

[1] Kevin Shockley, L. S. Imitation in shadowing words. Perception & Psychophysics , pp. 422-429, 2004.

[2] Marslen-Wilson, W., Linguistic Structure and Speech Shadowing at Very Short Latencies. Nature Vol, pp. 244, 522-523, 1973.

[3] McMillan, David. Miscommunications in Air Traffic Control. Diss. Queensland University of Technology, 1998.

[4] Gerard W. G. Spaai and Dik J. Hermes, A Visual Display For the Teaching of Intonation. CALICO Journal, Volume 10 Number 3, 1993.

[5] Shudong, Wang, Michael Higgins, and Yukiko Shima. "Teaching English pronunciation for Japanese learners of English online." JALT CALL Journal 1.1 (2005): pp. 39-47.

[6] Trevor A. Harley and L. J., Decline and fall: A biological, developmental, and psycholinguistic account of deliberative language processes and ageing. Aphasiology, 2011.

[7] Dennen, Vanessa P., and Kerry J. Burner. "The cognitive apprenticeship model in educational practice." Handbook of research on educational communications and technology 3 pp. 425-439, 2008.

[8] R. Ejzenberg and H. Riggenbach, "The juggling act of oral fluency: A psycho-sociolinguistic metaphor," Perspectives on Fluency, Ann Arbor University of Michigan, pp. 287-314, 2000.

[9] Freed B. Riggenbach H. 'Is fluency, like beauty, in the eyes (and ears) of the beholder?', Perspectives on Fluency, Ann Arbor University of Michigan pp. 287-314, 2000.

[10] Lennon P. 'Investigating fluency in EFL: A quantitative approach,' , Language Learning, 1990, vol. 40, pp. 387-417, 1990.

[11] Freed, B.F., Segalowitz, N. and Dewey, D.P., Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. Studies in second language acquisition, 26(02), pp.275-301, 2004.

[12] Segalowitz N , Freed B . 'Context, contact, and cognition in oral fluency acquisition,' , Studies in Second Language , 2004 , vol. 27, pp. 175 -201, 2004.

[13] Crystal, D. Prosodic Systems and Intonation in English. Cambridge: Cambridge University Press, 1969.

[14] Leon, P. R., and P. Martin, "Applied Linguistics and the Teaching of Intonation." The Modern Language Journal, 56, 3, pp. 139-144, 1972.

[15] Lieberman, P. (1967). Intonation, Perception and Language. Cambridge, MIT Press. _____, and S. B. Michaels, "Some Aspects of Fundamental Frequency and Envelope Amplitude as Related to the Emotional Content of Speech." Journal of the Acoustical Society of America, 34, pp. 922-927, 1962.

[16] comScore, Cross-Platform - Future in focus 2016.