Liu, C.-C. et al. (Eds.) (2014). Proceedings of the 22nd International Conference on Computers in Education. Japan: Asia-Pacific Society for Computers in Education

Enhancing English Pronunciation with *Windows Speech Recognition* Training: A Preliminary Study

Hui-Hsien FENG[a] & Ying-Hsueh CHENG[b*]
[a] *Department of English, Iowa State University, USA*
[b] *Teaching and Learning Center, National Pingtung University of Science and Technology, Taiwan*
*sherrycheng85@gmail.com

Abstract: Teacher educators have developed great interest in applying automatic speech recognition software to improve English learners' pronunciation. However, these studies are few compared with those emphasizing on other skills such as reading and writing. This case study thus aims to fill this gap by adopting the *Windows Speech Recognition* (WSR) system to help English learners improve pronunciation skills and promote learner autonomy. Drawing on the Interaction Hypothesis in Second Language Acquisition, this study sought to answer two questions: (1) with the use of WSR in pronunciation training, what is learner performance of pronouncing /n/ and /l/ sounds? and (2) what are learner attitudes toward this training? The training lasted for three weeks, with one hour each time and twice a week. Data sets included a pre-test (screening test) and a post-test based on screen- and audio-recordings of training sessions, a questionnaire, and a semi-structured interview with students. Field notes were also recorded when the entire training process was observed. Findings revealed that learner performance shown in the screening- and posttest is not significant. However, it indicates that with appropriate training, it is possible for learners to understand the feedback provided by WSR and apply the knowledge to judge their own practices. In addition, since learners held positive attitudes towards the use of this software, it is suggested that teachers or tutors can integrate the software with their curriculum for improving English students' pronunciation skills of segmental features, their learner autonomy and learning strategy.

Keywords: Pronunciation training, computer-assisted pronunciation teaching, ESL, windows speech recognition, case study

1. Introduction

With the rapid development of technology, English-as-a-second-language (ESL) teachers have developed vested interest in applying technology in their classrooms. There is no exception in pronunciation classrooms. Computer-assisted pronunciation teaching (CAPT) provides teacher educators opportunities for making pronunciation teaching and learning more approachable. Thus, automatic speech recognition (ASR) software has been widely used because it is not restricted by time and place and thus is able to offer learners practices whenever and wherever necessary. More importantly, it can serve as a private tutor to correct learners' pronunciation. This study thus aims to adopt the *Windows Speech Recognition* (WSR) system to enhance English learners' pronunciation skills and promote learner autonomy.

2. Related Literature

2.1 Use of ASR in Computer-Assisted Pronunciation Teaching

Studies on the use of ASR in pronunciation teaching are relatively few. Two concerns have been raised by previous researchers regarding the use of ASR in computer-assisted pronunciation teaching environments. The first concern is related to the fact that ASR fails to recognize nonnative speakers'

utterance as effectively as native speakers' (Derwing, Munro, & Carbonaro, 2000). Since the database of the ASR software only collects native speakers' speech samples, it is possible that nonnative speakers' utterance cannot be identified by the software. Evaluating the accuracy of *Dragon NaturallySpeaking®* (3rd edition, 1997), Coniam (1999) examined how ten Cantonese speakers read a one-thousand-word article aloud into the computer and compared the accuracy rates produced by these speakers and those by native English speakers. Coniam calculated the accuracy rates of the identified words and clauses printed out by the computer and found that the software had a lower accuracy level in recognizing Cantonese speakers' speech than those produced by native speakers.

The other issue concerning the use of ASR in pronunciation teaching is whether the ASR software can provide constructive feedback. This function is considered important because learners need to know which parts of their pronunciation is correct and which is not. Although the ASR software could provide instant feedback (Levis, 2007), it fails to offer constructive and accurate feedback for learners to improve their pronunciation. Researchers have addressed different concerns toward types of feedback display. For example, Neri, Cucchiarini, Strik, and Boves (2002) indicate that it is questionable to provide visual feedback displays such as spectrograms and waveforms. The idea of offering spectrograms and waveforms is to provide two comparable displays, one from the native speaker and the other from the user's utterance. Though showing comparable displays might help nonnative users "imitate" the native speaker, it is a misconception since every native speaker produces spectrograms and waveforms differently. Moreover, visual feedback like these cannot teach users how to pronounce correctly in terms of the location of tongue or shape of lips. This indicates that visual plays are not good enough for providing constructive feedback for learners to improve their pronunciation.

In spite of the aforementioned issues, several pronunciation teachers consider it is beneficial to use ASR (e.g., Franco, Bratt, Rossier, Gadde, Shriberg, Abrash, & Precoda, 2010; Levis, 2007). With the increasing varieties of technology, teachers nowadays can use more accessible and effective systems such as *Windows Speech Recognition*, *Google Voice*, and *Siri* on iPhone 4S. Although these tools were not specifically designed to help pronunciation training, it is possible to develop pronunciation training based on them. In this study, we adopted *Windows Speech Recognition* to train adult English learners because of its accessibility.

2.2 Interaction Hypothesis

Technology has been widely used for language teaching and learning and such application is related to the Interaction Hypothesis theory. Interaction Hypothesis (Long, 1996), a theory of second language acquisition, refers to the idea that development of language proficiency is promoted by face-to-face interaction and communication. It is believed that conditions for acquisition are especially good when interacting in the second language; specifically, conditions are good when a breakdown in communication occurs and learners must negotiate for meaning. For example, when one of the participants in a conversation will say something that the other does not understand; the participants will then use various communicative strategies to help the interaction progress. Strategies used for negotiating meaning may include slowing down speech, speaking more deliberately, requesting for clarification, or paraphrasing (Brown, 2000).

The Interaction Hypothesis theory includes four components: input, interaction, feedback, and output. In the following, we will introduce the role ASR plays with these components.

Regarding Input, ASR software is usually a part of a CAPT package, which provides model inputs for learners to imitate. If not, language teachers and learners could search input from public media or trustworthy websites, e.g. *Dave's ESL Cafe*, *Randall's ESL Cyber Listening Lab*, *Voice of America*, and *Ted Talk*.

Moreover, according to Chapelle (2003), interaction can also happen between the user and the computer in a language learning environment. Thus, ASR software could play the role of an interlocutor and provide interactions with language learners.

One of the crucial parts of the Interaction Hypothesis is the provision of feedback. It is important because it can provide clues for learners to continue to interact with the interlocutor in order for language acquisition to occur. However, if feedback is missing after the interaction happens, acquisition might not take place. The same situation is likely to happen when the learner does not notice the feedback. Thus, feedback is considered crucial for it can create or reduce opportunities for interactions.

Last, output here refers to modified output. It indicates learners' attempt to modify problematic utterance after they receive interactional feedback. ASR software provides this benefit because after learners receive feedback, they can try as many times as possible to practice until the feedback is given.

In sum, modifying speech arising from interactions like communication breakdown helps make input more comprehensible, provide feedback to the learner, and push learners to modify their speech for better output (Long, 1996). ASR software provides input, interaction, feedback, and output for language acquisition to occur.

3. Purpose of the Study

The current study uses *Windows Speech Recognition* (WSR) to train learners' pronunciation production of two sounds: /n/ vs. /l/.

Two research questions guide this study:

1. With the pronunciation training of WSR, what is learners' performance of pronouncing /n/ and /l/ sounds?
2. What are their attitudes toward this training?

4. Methods

4.1 Participants

Participants were recruited from an ESL writing class at a Mid-western U.S. university. To identify potential participants, a pronunciation screening test was carried out to check if they have difficulties pronouncing these particular sounds: /n/ and /l/. According to Swan and Smith (2001), these sounds are commonly mispronounced by Chinese speakers. Therefore, we chose these sounds for the pronunciation training. Seven ESL students were invited to participate in the study, but only five of them completed the entire procedure. These five students (2 males, 3 females) speak Chinese as their first language and they were enrolled in diverse undergraduate programs. They were considered appropriate for the study because they were newly enrolled in the university (in their first or third semester), and they reported some pronunciation problems in their conversations with others.

4.2 Data Sets and Procedure

This study employed a case study approach (Duff, 2008) to investigate how five students participated in pronunciation training sessions. Data sets included the following:

- Pre-test (screening-test)
- Audio- and screen- recordings of six training sessions
- Posttest
- Questionnaire
- Individual interview
- Field notes based on observation of training sessions

Procedure of the study is summarized in Table 1:

Table 1: Procedure.

Procedure	*Content*
Screening test/Pre-test	Pronounced /n/ and /l/ sounds in minimal pairs and a paragraph.
Training Sessions 1 & 2 (Week 1)	Practiced the build-in tutorials to be familiar with WSR.
Training Sessions 3 & 4 (Week 2)	Practiced /n/ and /l/ sounds with the researcher to provide instruction of specific sounds when WSR could not identify what learners had said for more than five times, or when they raised questions.
Training Sessions 5 & 6 (Week 3)	Practiced /n/ and /l/ sounds; students needed to apply what they learned from the second week.
Posttest (two parts)	Part one included minimal pairs focusing on /n/ and /l/ sounds.
	Part two included two paragraphs; the first was the same as the one used in the screening test, and the second one was new, which also included /n/ and /l/sounds.
Questionnaire	18 five-point Likert-scale items and 3 open-ended questions.
Interview with students	Semi-structured interviews (each lasted 15-20 minutes) for understanding students' perceptions of pronunciation problems, use of WSR, evaluation on the training, and their potential use in the future.

4.3 Data Analysis

Each learner's recording was rated by the researcher according to the focal sounds mentioned above. Paired t-test was conducted to investigate whether any difference exists between the pretest (screening-test) and the posttest. Likert-scale-item responses from the questionnaire were examined through descriptive statistics, whereas responses from the open-ended questions were coded. In addition, interviews with students were transcribed and coded by *in vivo coding* and *themeing the data* (Saldaña, 2009, p. 74 & 139). Questionnaires and interviews were both used to triangulate the data.

5. Results and Discussion

5.1 Learner Performance

The screening and posttest contains two parts: minimal pairs and paragraphs specifically designed to elicit discrimination of /n/ and /l/ sounds.

Table 2: The accuracy rates (%) of /n/ and /l/ in minimal pairs and paragraphs.

Student	Minimal pair: /n/ vs. /l/		Paragraph: /n/ vs. /l/	
	Screening-test	Posttest	Screening-test	Posttest
1	86	93	55	79
2	93	100	91	100
3	93	100	100	100
4	93	100	100	93
5	79	93	64	100
Average	89	97	82	94
SD	6	4	21	9

Table 2 provides the accuracy rates of /n/ vs. /l/ in minimal pairs and paragraph level. The average of /n/ vs. /l/ in minimal pairs and paragraph level in screening-test is 89% (SD = 6%) and 82%

(SD = 21%); and that in posttest is 97% (SD = 4%) and 94% (SD = 9%). Although the difference is not significant, the increasing trend may suggest that by using WSR in pronunciation training with some training to interpret the output, it is possible for students to improve their pronunciation, even by practicing alone.

5.2 Learner Attitudes

Students' attitude is a prominent factor in mastering a second language (Lightbown & Spada, 2006). Multiple sources (questionnaires, field notes, student interviews, and learners' screen- and audio-recordings) revealed that learners held positive attitudes towards the pronunciation training with the use of WSR and believed that the training was effective.

During their practice, learners tended to repeat each word several times. When asking about the reason, they responded that they wanted to see the WSR output, namely what word would show up based on their speech. However, they reported that they were annoyed when WSR showed the wrong words around five times. If this happened, they would ask the researcher why it was like this and how to pronounce the word correctly.

Moreover, the participants revealed that they were willing to practice as many times as possible, around eight to ten times on average. They were able to make slight changes and found out in what way the WSR could show the correct word on the screen. This implies that the use of WSR could promote learner autonomy and help learners acquire learning strategies, such as self-correction and self-monitoring.

The effectiveness of the training went beyond our expectations. In addition to the two sounds designed for this study, Participant 3 reported that he learned how to say /th/ sound (as in three) and Participant 2 understood that "four" does not have the same pronunciation as "full." Participants 2 and 5 reported that they learned to distinguish the sounds of these words: "napkin" rather than "lapkin"; "thank you" rather than "sank you." Participant 4 even invited her friends to try WSR when they were in a party. All in all, WSR provided opportunities for the ESL students to practice, modify, and produce correct output.

6. Conclusion and Implication

The use of WSR seemed effective in pronunciation training because it could help learners improve pronunciation skills and promote learner autonomy and develop learners' self-correction and self-monitoring strategies. Although WSR could not accurately capture nonnative speakers' sounds, this motivates learners to practice again and again. With adequate training, learners are able to monitor their utterance and self-correct their speech. The use of WSR helped develop learner autonomy, enhance learning strategy, and improve pronunciation skills of segmental features.

Since this study adopted a case study approach, it only included five participations. For future research adopting quantitative methods, a larger sample size could be considered. This study has some limitations. It was completed within a short time period (three weeks). Therefore, it is suggested that a longer period time for training and a larger sample size may have different impact on the results. Second, since this study only examined the effects of training on learners' pronunciation of two sounds, future studies can incorporate other sounds that may be challenging to English learners, such as /th/ or vowels such as "fat" vs. "fate." Third, this study was conducted in the United States, future research could be undertaken in other countries where students use English as their foreign language (EFL). How these learners use WSR might be different from ESL learners in English-speaking countries. Their perceptions of use and difficulties can provide researchers better understanding. Finally, since this study did not include any native English-speaking instructor, to what extent the combination of a native English-speaking instructor and the WSR can provide opportunities for learners can be examined.

Acknowledgements

We would like to thank the five participants for making this research possible. Without their help, we would not have been able to begin this research.

References

Brown, H. D. (2000). *Principles of language learning and teaching*. White Plains, NY: Longman.

Chapelle, C. (2003). *English language learning and technology: Lectures on applied linguistics in the age of information and communication technology*. Amsterdam: John Benjamins.

Coniam, D. (1999). Second language proficiency and word frequency in English. *Asian Journal of English Language Teaching*, 9, 59-74.

Derwing, T. M., Munro, M. J., & Carbonaro, M. (2000). Does popular speech recognition software work with ESL speech?. *TESOL Quarterly*, *34*, 3, 592-603.

Duff, P. (2008). *Case study research in applied linguistics*. NY: Taylor & Francis.

Franco, H., Bratt, H., Rossier, R., Grdde, V. R., Shriberg, E., Abrash, V., & Precoda, K. (2010). EduSpeak® : A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. *Language Testing*, *27*(3), 401-418.

Levis, J. (2007). Computer technology in teaching and researching pronunciation. *Annual Review of Applied Linguistics*, *27*, 184-202.

Lightbown, P., & Spada, N. M. (2006). *How languages are learned*. Oxford: Oxford University Press.

Long, M. (1996): the role of the linguistic environment in second language acquisition. In W. Ritchie and T. Bhatia (eds), *Handbook of second language acquisition*. San Diego: Academic Press, 413-68.

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The pedagogy-technology interface in computer assisted pronunciation training. *Computer Assisted Language Learning, 15*(5), 441-467.

Saldaña, J. (2009). *The coding manual for qualitative researchers*. London: Sage.

Swan, M., & Smith, B. (2001). *Learner English: A teacher's guide to interference and other problems*. Cambridge: Cambridge University Press.