Advances in Social Science, Education and Humanities Research (ASSEHR), volume 182
2018 2nd International Conference on Education, Economics and Management Research (ICEEMR 2018)

Automatic Speech Recognition and Pronunciation Training

Wenqi Xiao*
International College
Xiamen University
Xiamen, Fujian Province, China
wenqixiao@xmu.edu.cn

***Abstract*—Automatic speech recognition (ASR) is providing more speaking opportunities for second language learners. With its progress in more accurate error detecting, the application of ASR in language teaching holds huge potential to practice and improve pronunciation. The purpose of the study is to present the features and application of ASR in second language learning. This paper firstly focuses on presenting the available literature on pronunciation training. Then the effectiveness of the ASR in evaluating the pronunciation data and providing feedback is discussed by referring to some evidence. It is found that state-of-the-art ASR is conducive as a pronunciation measurement, but the feedback provided by ASR is less effective.**

Keywords—pronunciation; ASR; scoring; feedback

I. INTRODUCTION

Computer-assisted pronunciation training (CAPT) has been incorporated in second language learning to promote L2 learners' speaking ability by providing multiple speaking practices and a stress-free environment. As one of the fastest developed pronunciation training technology, automatic speech recognition (ASR) and its assessment arose the research interest dating back to 1990 [1, 2]. Powered by enhanced computer technology, the past decade has witnessed its development with more accurate error detection and more diversified feedback. ASR is being applied in language instruction context by recognizing the learner's speaking output, evaluating the parameters, and providing feedback based on learner's performance.

To provide an overview of ASR in English pronunciation training, pronunciation training will be firstly discussed, and the effectiveness of ASR technology in enhancing English pronunciation will be illustrated by discussing its features in scoring and feedback.

II. PRONUNCIATION TRAINING

A. *Aspects of pronunciation:*

Pronunciation contains different components including segmental aspect, suprasegmental aspect as well as voice-setting aspect. Segmental aspect refers to the vowels and consonants, which are the minimal phonetic unit. The suprasegmental aspect is context-depended, which involves the stress and intonation. The voice-setting aspect means the voice position of individual sound in articulation. In terms of pronunciation error detection, error detection in segmental aspect is more challenge as this aspect has higher level of variability [1].

B. *Pronunciation goal*

When it comes to the pronunciation goal, there is a consensus that the intelligibility is essential for communication rather than orienting toward a native-like accent [3,4]. The research on error rate has revealed that suprasegmental aspect is closely related to intelligibility [5]. In order to develop learner's communicative competence, apart from instruction on segmental aspect, suprasegmental aspect can not be underestimated.

C. *The role of output in pronunciation training*

Exposure to target language is one of the prerequisites for language learning. Krashen noted that language input should be understood and meaningful to the learner[6]. However, Swain has argued that comprehensive input is not sufficient for L2 acquisition [7]. The necessity of output is contributive to language learning for several reasons. L2 learner is able to notice the specific linguistic gaps when they intend to convey meaning. Besides, output enables L2 leaner to testify their hypothesis of how to use the target language.

III. THE EFFECTIVENESS OF THE STATE-OF-THE-ART ASR

In general, ASR first provides scoring based on its assessment and then generates feedback towards learner's pronunciation. Its effectiveness is discussed on the two aspects: scoring and feedback.

A. *Scoring*

Foreign language teachers are considered the human raters for learner's speaking. It is suggested that comprehensive scoring system and rubrics is one of the conditions to ensure the reliability of the human rating [8]. On the other hand, rating training is necessary to guarantee the consistency of the evaluation.

However, it can not be denied that human speech rating is time-consuming and physically-demanding. Assisted by ASR, speech is evaluated in an instantaneous way. Under the score paradigm, the input is firstly elicited to identify the beginning and end of the speech. After the recognizing of the phonetic segmentation, a pronunciation score is generated by comparing the learner's data with the native-speaker's data.

To achieve the above process, several database including native-speaker corpus, non-native corpus as well as human ratings corpus are required. The native-speaker corpus is adopted as a pronunciation reference, while the nonnative-speaker corpus is used to train the system to detect common pronunciation problems. The human ratings corpus provides human judgments for pronunciation skills.

As for the validity of ASR scoring, there are mixed results. Research indicates that ASR is less reliable when evaluating the suprasegmental aspect such as intonation [9]. However, it is proved that the ASR score is comparable to human rating when the speech data is sufficient [10;11; 12].

Although the accuracy of ASR in scoring still needs improving, it has prominent advantages in detecting the specific error, providing consistent measure for every learner and reducing human factors.

B. *Feedback*

Corrective feedback is a controversial issue when considering its negative effects on generating affective responses (such as embarrassment), but it is also noticeable that corrective feedback is contributive to language learning [13; 14; 15]. In the aspect of pronunciation, it is acknowledged that visual feedback together with the audio feedback are generally contributive to pronunciation improvement [16].However it is necessary to ensure that the feedback from the ASR should be accessible and comprehensible for the learners[17;18].

1) Scoring feedback

Some of the CAPT provides numeric or symbolic score such as a smiling face, or thumb-up sign to indicate the quality of the pronunciation input. Although the scoring feedback is easy to comprehend, learners are unable to revise their pronunciation from this kind of feedback system.

2) Visual feedback

In an ASR system named Better Accent tutor, the system addresses the suprasegmental level of the pronunciation, namely: intonation, stress, and rhythm [18]. The system generates audio-visual feedback after the learner articulates the sound. The speech patterns of both the native speaker and the learner are presented with visualization. The pitch graph containing arrows indicates the intonation of speech. On the basis of the arrows, the learner understands which part of speech is supposed to be raised or lowered, while the stress and rhythm are presented by bars, whose length demonstrate the duration of the speech. By comparing the visual feedback with the native speaker's, the learner is able to modify his/her speech.

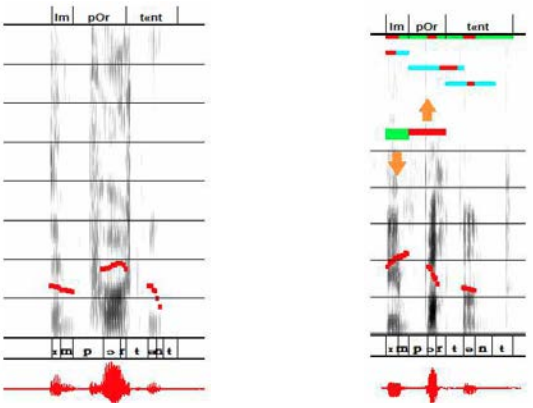

Fig. 1. Sound Spectrum Comparison

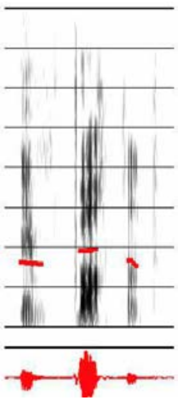

Fig. 2. Sound Spectrum after modification

As shown in Fig. 1[18], the left sound spectrum is the pronunciation of the word: *important* form a native English speaker. The red waves present the realization of stress and intonation in the target word. The right sound spectrum is pronounced by a French speaker. The arrows indicate whether the pith should be raised or lowered. Fig. 2 [18] is the modified voice of the learner.

The results show that the auditory feedback is more effective for advanced learner than the less-proficient learner [18]. However, with the help of combined feedback: the audio and visual, low proficiency learners have made improvement.

Another system emphasizes on segmental level of pronunciation [19]. It detects the mispronunciation in word level such as vowel, consonant and word-stress. The feedback of the system first locates the error by highlighting the word part. In addition, it explains the error by providing some example words (high-frequency words) containing same vowels or consonants. For instance, the correct pronunciation of *flood* is [flʌd], learners mispronounce it as [flu:d]. The feedback of the ASR first marks the mispronounced part *fl**oo**d*. Then it provides high-frequent words containing same vowel such as c**u**p [kʌp] or bl**oo**d [blʌd]. By doing so, the learner can insert the sound of the example word to the target word. It is easier for the learner to notice the mispronounced part and make modification.

IV. Conclusion

Pronunciation training requires both segmental and suprasegmental aspects to achieve intelligence in communication. Apart from input, output performs an essential role in noticing and hypothesis testing and reflective functions.

Introducing ASR in pronunciation training has distinctive advantages in converting traditional language class into a learner-centered environment with more speaking chances and less anxiety. Advanced by more accurate error detecting technique, ASR can be adopted to measure learner's pronunciation. However, the feedback provided by ASR is not comparable to the teacher's in terms of comprehensibility and effectiveness. What kind of feedback generated by ASR is more comprehensive and contributing to guide learners to improve is still an issue that needs further investigation.

References

[1] S. Witt-Ehsani, "Automatic Error Detection in Pronunciation Training: Where we are and where we need to go," International Symposium on automatic detection on errors in pronunciation training, 2012.

[2] L. Neumeyer, et al. "Automatic scoring of pronunciation quality," Speech *Communication,* vol. 30.pp. 83-93, 2000.

[3] M. Tracey, Derwing, and M. J. Munro, "Second language accent and pronunciation teaching: A Research-Based Approach," *Tesol Quarterly,* vol.39, pp: 379-397, 2005.

[4] J. Murray Munro, and T. M. Derwing, "Foreign accent, comprehensibility, and intelligibility in the speech of second language learners," *Language Learning,* vol. 49, pp:285-310, 2010.

[5] Raux, and T. Kawahara, "Automatic intelligibility assessment and diagnosis of critical pronunciation errors for computer-assisted pronunciation learning." Interspeech, 2003.

[6] E. Tragant, and C. Muñoz, Second Language Acquisition and Language Teaching. Oxford University Press, 1981.

[7] M. Swain, "Communicative competence: Some roles of comprehensible input and comprehensible output in its development," In M.A. Gass & C.G. Madden (Eds.), Input in second language acquisition, Rowley, MA: Newbury House, pp. 235–253, 1985.

[8] P. Valette, Comparing the Phonetic Features of English, French, German and Spanish. London: G. Harrap, 1965.

[9] S. I. Kim, "Automatic speech recognition: reliability and pedagogical implications for teaching pronunciation," Journal of Educational Technology & Society, vol. 9, pp:322-334, 2006.

[10] C. Cucchianni, "Quantitative assessment of second language leaners' fluency: an automatic approach." Journal of the Acoustical Society of America, vol.107, p. 989, 1998.

[11] D. Cordier, Speech Recognition Software for Language Learning: toward an evaluation of validity and student perceptions. University of South Florida, 2009.

[12] L. Neumeyer, H. Franco, V. Digalakis, & M. Weintraub, "Automatic scoring of pronunciation quality," Speech Communication, vol. 30, pp.: 83-93, 2000.

[13] R. Lyster, & K. Saito, "Oral feedback in classroom SLA: a meta-analysis," Studies in Second Language Acquisition, vol. 32, pp: 265-302, 2010.

[14] M. J. Norris, & L. Ortega, "Effectiveness of l2 instruction: a research synthesis and quantitative meta-analysis." Language Learning, vol. 50, pp.: 417-528, 2000.

[15] J. Turscott, "What's wrong with oral grammar correction," Canadian Modern Language Review, vol.*55*, p.: 437, 2006.

[16] N. Ambra, et al. "The pedagogy-technology interface in computer assisted pronunciation training," Computer Assisted Language Learning, vol.15, pp.: 441-467, 2002.

[17] J. Kommissarchik, & E. Komissarchik, "Better accent tutor – Analysis and visualization of speech prosody," Proceedings of InSTILL, Dundee, Scotland, pp. 86–89, 2000.

[18] W. Menzel, D. Herron, P. Bonaventura, & R. Morton, "Automatic detection and correction of non-native English pronunciations," Proceedings of Instill, 2000.

[19] Bonneau, & V. Colotte, "Automatic feedback for l2 prosody learning," Speech & Language Technologies, 2011.