A Pilot Study of Applying Speech Recognition to Improve Students' English Pronunciation and Oral Fluency in the Freshman English Class

GAO Zhao-Ming
National Taiwan University

Abstract

While many computer-assisted language learning (CALL) systems have emerged, there are very few systems which provide automatic feedback on the pronunciation by learners of a foreign language. Moreover, most existing CALL systems have not incorporated cutting-edge speech technology to help learners detect errors in their pronunciation. To bridge this gap, we have developed a web-based system using speech recognition. The proposed system allows users to learn at any place and any time so long as there is internet connection and a browser. The system enables users to improve their English pronunciation and oral fluency at their own pace with somel guidance of an English teacher. Learners first listen to a sentence or a word. They then do exercises in shadowing. Users' speech was evaluated against the predefined answers based on the phonetic forms identified by the Google speech recognition. If the similarity of the two forms is below a given threshold, the speaker's oral output is considered unsatisfactory and the learner is required to try again. The proposed system was tested in a Freshman English class of thirty-five students. Our preliminary study showed that while 85.7% of the students agreed that the system's performance was acceptable, many of them were uncertain if it could actually improve their pronunciation or oral fluency. Based on the result of this pilot study, we recommend that the system be treated as a supporting tool guided by an experienced English teacher.

運用語音科技改善大一英文學生的英語發音及流利度

高照明
國立臺灣大學

摘要

儘管不少電腦輔助語言教學系統已經問世，提供英語發音及口語訓練的智慧型電腦輔助語言教學系統卻極為少見。即便有，大部份的系統也沒有利用先進的語音科技來幫助學習者找出發音的問題。我們利用語音辨識技術，發展了一個系統讓

學習者可以利用網路與瀏覽器不限時間與地點學習。這套系統的教學功能讓學習者可以在老師指導的情形下依據自己的步調來學習。使用者首先聽一個句子或一個詞的音，接下來對電腦覆誦。系統利用Google引擎語音辨識來評估學習者所說的英語和系統裡面所存答案的語音相似度。如果相似度低於門檻值，表示學習者的發音或句子可能不夠準確，系統會要求使用者重做練習。本研究顯示儘管有85.7%的受試者同意這套系統的表現可以讓人接受，卻有不少人不能確定這套系統是否能有效提升發音的正確性及流利度。基於此項調查的結果，我們建議這套系統做為在英文老師指導下的輔助教學工具。

Introduction

Advance in natural processing language has made intelligent computer-assisted language learning systems not only feasible but also increasingly available. More and more computer-assisted language learning (CALL) systems have employed automatic speech recognition (ASR) including Tell Me More, MyET, EnglishCentral, Versant, among others. Natural processing techniques and speech technologies have even been extended to language assessment (cf. Bernstein et al., 2010; Ginther et al., 2010). In addition, the popularity of mobile devices in recent years has infused CALL with new features and possibilities. The purpose of this paper is to explore the feasibility of applying speech technology in improving the English pronunciation and oral fluency of students in the Freshman English class at National Taiwan University.

Oral fluency is typically characteristic of native speakers and is the goal that every language learner aspires to achieve. Byrne (1986, p. 9) points out that “the main goal in teaching the productive skill of speaking will be oral fluency. This can be defined as the ability “to express oneself intelligibly … reasonably accurately and without too much hesitation.” To attain the goal of oral fluency, Byrne (1986, p. 10) suggests that we have to “bring the students from the stage of where they are mainly imitating a model of some kind, or responding to cues, to the point where they can use the language freely to express their own ideas.” As oral fluency requires a lot of practice and training, Byrne proposes a realistic three-stage training, starting with imitating.

The term “imitating” is reminiscent of drills of audiolingual approaches in the 1940s and 1950s, which emphasize intensive exercises involving listening to sentences, memorizing, and repeating. The audiolingual approaches were criticized for lack of communicative functions and gave way to the communicative approaches emphasizing drills that involve speech functions such as asking directions and making requests and other activities such as role play. Compared with the audiolingual approaches, the communicative approaches place more emphasis on fluency than on grammatical accuracy. Hammerly (1991, p. 9) criticizes the communicative approach by pointing out that it ignores language structures and that "(its) advocates do not seem to care that students mispronounce sounds, use wrong stems or endings, or construct sentences following faulty rules." Summarizing the features of different approaches, Bygate (2010, p. 71) comments that "Audiolingual approaches aimed to develop speaking only in terms of pronunciation and fluent, accurate manipulation of grammar. Situational approaches introduced dialogue patterns into the range of features to be taught, and functional approaches added speech acts into the syllabus."

While often misunderstood and underestimated, imitation does play an important role in language acquisition. In first language acquisition research, it has long been attested that infants born after a few months start to imitate other people's speech in the environment. Imitation of other people's speech has different forms, function, and applications. For example, imitation is often used in a discourse by speakers with different intonation to express 'approval', 'emphasis', 'question', etc. One frequently used technique of imitation in oral training is shadowing, in which the learners are asked to repeat exactly what they have heard. The learners are often encouraged to imitate as closely as possible the way the speaker speaks and pay attention to the details such as the pronunciations of vowels and consonants, stress, intonation, liaison, and accent. Shadowing has long been used to train interpreters (cf. Kurz, 1992) and to evaluate a learner's proficiency level under the name of "elicited imitation" since 1960s (cf. Murphey, 2001). While there are not many researches on shadowing, the effectiveness of shadowing on improving listening comprehension, oral fluency, and interpreting has been affirmed by several scholars.

For example, according to Murphey (1993), students interviewed in his study report that "conversational shadowing allowed them to assert some control over the process and content of conversations and to build better rapport through reflective listening." Murphey (1993) concludes that that "silent shadowing had a major impact on their learning, increasing attention and retention of material in short-term memory". Celce-Murcia et al. (1996) suggest a method of learning intonations by shadowing, mirroring, imitating native speakers' utterances. Yamada (2010) conducts experiments on the effect of shadowing on listening, and concludes that shadowing is effective to listening comprehension. Hsieh, Dong, and Wang (2013) confirm in their experiment that the technique of shadowing is useful to the learning of intonation, fluency, word pronunciation and overall pronunciation.

The underlying technology required to implement a CALL system with the shadowing function is automatic speech recognition (ASR), which can measure the similarity between utterances by a native speaker and a learner. Several recent papers discuss the applications of ASR in foreign language learning and assessment. For example, Coniam (1998) explores the potential use of ASR as an assessment tool in English as a foreign languages. Dalby and Kewley-Port (1999) investigates the use of ASR technology in foreign language learning, including speakers of American English learning Spanish and speakers of Mandarin Chinese learning English. Their research suggests that ASR-derived feedback can improve pronunciation. Luo et al. (2010) explore the relationship of acoustic features such as phoneme intelligibility and prosodic fluency with the reference scores for learners' shadowing. Ginther et al. (2010) discuss the conceptual and empirical relationships between temporal measures of fluency and oral English proficiency. Franco et al. (2010) describe a speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. Berstein et al. (2010) validate automated speaking tests. Chen (2011) develops a speaking website using Microsoft ASR toolkit. Students and in-service teachers using the system agree that ASR can offer different types of exercises which encourage students to produce more oral output. All these researches suggest that automatic speech recognition technology has become increasingly important in language learning and assessment.

Research Questions

The three research questions this report addresses are the following. (1). What is the attitude of high-intermediate learners towards shadowing? (2). What is the attitude of the subjects towards online shadowing system with automatic feedback? (3). What is the users' evaluation of the proposed system?

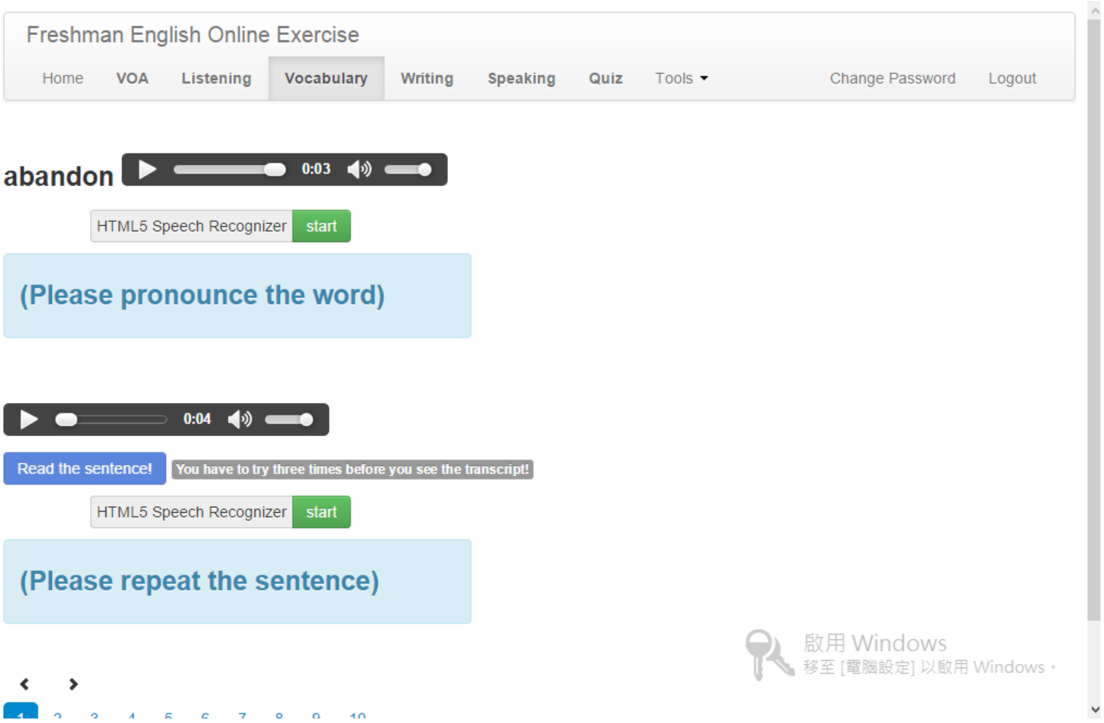
The Design of the System

We have developed a system which provides automatic feedback for a learner's pronunciation based on speech recognition. Our program allows users to learn at any place and any time so long as there is internet connection and a browser. The pedagogical functions of the system have the potential of enabling users to improve their pronunciation and oral fluency at their own pace without much guidance of an English teacher. Several guided exercises are designed to ensure that users can learn the materials efficiently and effectively.

Following Byrne (1986), we combine the features of different approaches in the different stages of oral training. We believe that there are four fundamental stages in training oral fluency. The first stage involves the training of accurate pronunciation of a single word by listening and repeating. The second stage involves imitating a sentence via listening. The third stage involves role play in a given dialogue. The last stage requires interpreting. Based on these procedures, we design a CALL system capable of helping learners improve their English pronunciation and oral fluency. The system described in this report focuses on the first and second stages.

We provide a series of exercises to help learners consolidate their pronunciation. Learners first listen to an audio file spoken by a native speaker of English. To facilitate learning, our system provides the script of the English sentence along with its Chinese translation. Learners first listen to the audio file of a word spoken by a native speaker of English. They imitate the pronunciation of the native speaker to the extent that their pronunciation can be correctly recognized by our system which employs the Google speech recognition engine. They then proceed to the sentence shadowing exercises. If they have difficulties understanding the meanings of the spoken sentence, they can read the Chinese translations of the English example sentences.

Learners are encouraged to listen to an audio clip more than once. If learners still cannot make sense of it after listening to it for several times, they can read the English transcript. Learners do the shadowing exercises at word and sentence level. All these oral exercises employ speech recognition techniques to measure the similarity between the learners' outputs and the predefined answers. For instance, the pronunciation of 'odd' and 'at' are represented as [AA D] and [AE T], respectively, with a space between different phonemes, based on the format of the CMU English pronunciation dictionary. The similarity of the pronunciations of the two words can be simplified as the similarity of two strings of phonetic representations, which can be computed by algorithms such as the Minimal Edit Distance. The Minimal Edit Distance algorithm calculates the minimal steps of editing (i.e. insertion, deletion, and substitution) if the two strings are to be identical. If the similarity of the two strings is below a given threshold, the speakers' pronunciations are considered unsatisfactory and the learners are required to listen to the audio file again. If their pronunciations of the sentence are satisfactory, they proceed to the next exercise. Figure 1 is the interface of the proposed online shadowing system. By using the advanced speech processing techniques, we are able to record learners' performances as well as monitor their progress. The system can facilitate the training of listening and speaking at word and sentence levels.



(Please pronounce the word)

You have to try three times before you see the transcript!

(Please repeat the sentence)

Figure 1. The proposed online shadowing system

Participants

Thirty-five freshman students at National Taiwan University participated in this study. Most of them majored in engineering and about one sixth of them studied law. Male students constituted about 80%. The proficiency level of these students was high-intermediate, as evidenced by the fact that most of them either passed or nearly passed the reading and listening comprehension sections of the GEPT high-intermediate test.

Discussion and Analysis

As stated earlier, we want to explore the following three research questions in this pilot study. (1). What is the attitude of high-intermediate learners towards the shadowing method? (2). What is the attitude of students towards the proposed online shadowing system with automatic feedback? (3). What is the users' evaluation of the proposed system?

Regarding the first research question, most students have very different attitudes towards shadowing based on words and shadowing based on sentences. 74.3% of them agree that shadowing based on words can improve their pronunciation. In contrast, only 37.2% of them agree that shadowing based on sentences can improve their pronunciation. When asked if they like the online shadowing exercises based on words, 52.9% of them were positive. Only 32.5% like online shadowing exercises based on sentences without script. However, 62.1% like online shadowing exercises based on sentences if they are provided with a script. When asked if the system can help improve their oral fluency, only 42.9% were positive and 37.1% of them were unsure. Only 48.6% found the automatic feedback function of the system helpful to learning and 42.9% of them remained uncertain. The statistics showed that many students were skeptical that the shadowing method and the proposed shadowing system could improve their pronunciation and oral fluency. They did not easily subscribe to the idea that speech recognition technology would improve their pronunciation.

Table 1. Results of the Survey

1. Repeating the pronunciation of a vocabulary word can help me improve my pronunciation of consonants, vowels, and stress.

Strongly disagree	1	2.9%
disagree	4	11.4%
not sure	4	11.4%
agree	17	48.6%
Strongly agree	9	25.7%

2. Repeating an English spoken sentence without the script can help me improve my pronunciation.

Strongly agree 3

Strongly disagree	1	2.9%
disagree	6	17.1%
not sure	15	42.9%
agree	10	28.6%
Strongly agree	3	8.6%

3. I like the online exercise of repeating the pronunciation of a vocabulary word.

Strongly disagree	0	0%
disagree	6	17.1%
not sure	7	20%
agree	15	42.9%

Strongly agree	7	20%

4. I like the online exercise of repeating a spoken sentence with script.

Strongly disagree	0	0%
disagree	7	20%
not sure	8	22.9%
agree	18	51.4%
Strongly agree	2	5.7%

5. I like the online exercise of repeating a spoken sentence without script.

Strongly disagree	1	2.9%
disagree	8	22.9%
not sure	15	42.9%
agree	10	28.6%
Strongly agree	1	2.9%

6. The system can help me improve my oral fluency in English.

Strongly disagree	1	2.9%
disagree	6	17.1%
not sure	13	37.1%
agree	14	40%
Strongly agree	1	2.9%

7. The automatic feedback provided by the system is helpful to learning.

Strongly disagree	1	2.9%
disagree	2	5.7%
not sure	15	42.9%
agree	14	40%
Strongly agree	3	8.6%

8. The pronunciations of the words in the system are good.

Strongly disagree	0	0%
disagree	1	2.9%

not sure	5	14.3%
agree	18	51.4%
Strongly agree	11	31.4%

9. Overall, the performance of the system is acceptable.

Strongly disagree	0	0%
disagree	0	0%
not sure	5	14.3%
agree	27	77.1%
Strongly agree	3	8.6%

Conclusion and Future Research

Our initial investigation comes to the same conclusion as Coniam (1998) and Chapelle and Chung (2010) that current speech recognition systems are not reliable enough to be used in high-stake assessment. However, as suggested by Coniam's (1998) research, the accuracy rates of ASR systems correlate with learners' proficiency levels. This is because most ASR systems are trained using the speech produced by native speakers. Oral outputs which are more similar to native speakers in pronunciation and fluency tend to get higher accuracy rates.

In our pilot study, the accuracy rates for individual words are in general acceptable. However, words that begin or end with consonants, especially fricatives such as s,z,f,v, ch,sh, have much poorer accuracy rates. ASR systems often give incorrect feedback in these situations. In addition, ASR systems do not seem capable of detecting mispronunciation due to incorrect stress. We believe that these are the reasons why many students were skeptical about the learning effects of our system. It should be noted that a side effect might occur when learners' oral outputs fail to be recognized by ASR systems after many trials. Learners might lose their patience if they are too much frustrated. In order to avoid this potential problem, we propose to allow users to read the script of the spoken sentence.

To conclude, we believe that the proposed system has the potential for improving students' pronunciation and oral fluency if it is treated as a supporting tool guided by an experienced English teacher who knows how to make the best use of the tool while circumventing its limitations.

References

Bernstein, Jared, Van Moere, Alistair, and Cheng, Jian. (2010). Validating automated speaking tests. Language Testing, Vol. 27, No. 3, pp. 355-377.

Bygate, Martin. (2010). Speaking. In Kaplan, Robert. (ed). The Oxford Handbook of Applied Linguistics, p. 63. -73.

Celce-Murcia, Marianne, Donna Brinton and Janet M. Goodwin. (1996). Teaching pronunciation: A reference for teachers of English to speakers of other languages. Cambridge: Cambridge University Press.

Chapelle Carol, Chung, Yoo-Ree. (2010). The promise of NLP and speech processing technologies in language assessment, Language Testing, Vol. 27, No. 3, pp. 301-315.

Chen, Hao-Jan. (2011). Developing and Evaluating an Oral Skills Training Website Supported by Automatic Speech Recognition Technology. RECALL, Vol. 23 , No 1, pp. 59-78.

Chung, Da-Un. (2010). The Effect of Shadowing on English Listening and Speaking Abilities of Korean Middle School Students. English Teaching. CALICO, Vol. 65, No. 3, p. 97.

Coniam, D. (1998). The Use of Speech Recognition Software as an English Language Oral Assessment Instrument: An Exploratory Study. CALICO, Vol. 15, No. 4, pp. 7.-23.

Dalby, Johnathan and Kewley-Port, Diane. (1999). Explicit Pronunciation Training Using Automatic Speech Recognition Technology. CALICO, Vol. 16, No. 3, pp. 425.-445.

English Learning Programs at the VOA Chinese Website. http://www.voachinese.com/section/english-teaching/2506.html

Franco, Horacio, Bratt, Harry, Rossier, Romain, Gadde, Venkata Rao, Shriberg, Elizabeth, Abrash, Victor, and Precoda, Kristin. (2010) EduSpeak® : A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications Language Testing, pp. 401-418.

Ginther, April, Dimova, Slobodanka, and Yang, Rui. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring, Vol. 65, No. 3, pp. 379-399.

Hamada, Yo. (2012). "An Effective Way to Improve Listening through Shadowing." The Language Teacher, Vol. 36, No. 1. pp. 3-10.

Hughes, Rebecca. (2002). Teaching and Researching Speaking. Routledge.

Hsieh, Kung-Ting, Dong, Da-Hui, and Wang, Li-Yi. (2013). A Preliminary Study of Applying Shadowing Technique to English Intonation Instruction. Taiwan Journal of Linguistics, Vol. 11 , No. 2, pp. 43-66.

Luo, Dean, Yamauchi, Yutaka and Minematsu, Nobuaki. (2010). Speech Analysis for Automatic Evaluation of Shadowing. In Proceedings of SLaTE.

Lynch, Tony. Listening: Sources, Skills, and Strategies. In Kaplan, Robert. (ed). The Oxford Handbook of Applied Linguistics, p. 74. -87.

Murphey, T. (2001). Exploring conversational shadowing. *Language Teaching Research*, *5*(2), 128-155.

Shiki, O., Mori., Y., Kadota, S., & Yoshida, S. (2010). Exploring differences between shadowing and repeating practices. *Annual Review of English Language Education in Japan, 21,* 81-90.