Education and Information Technologies (2019) 24:3731–3743
https://doi.org/10.1007/s10639-019-09955-7

A review of tools and techniques for computer aided pronunciation training (CAPT) in English

Chesta Agarwal[1] · Pinaki Chakraborty[1]

Received: 20 February 2019 / Accepted: 19 June 2019 / Published online: 1 July 2019

Abstract

Widespread use of English in the academia and in business is leading an increasing number of people to learn it as a second or a foreign language. Computer aided pronunciation training (CAPT) systems are used by non-native English speakers for improving their English pronunciation. A typical CAPT tool records the speech of a learner, detects and diagnoses mispronunciations in it, and suggests a way for correcting them. We classified the CAPT systems for English into four categories on the basis of the technology used in them and studied the salient features of each such category. We observed that visual simulation based systems are suitable for young and naive learners, game based systems are advantageous as they can be personalized as per the requirements of the learners, comparative phonetics based systems are suitable for adult learners fluent in another language, and artificial neural network based systems have the highest accuracy in mispronunciation diagnosis and are suitable for experienced and professional learners. We identified the state-of-the-art practices used in CAPT systems, and observed that CAPT systems can detect up to 86% mispronunciations in a speech and help learners to lessen mispronouncing by up to 23%. We recommend collaboration between language teachers and software developers to develop CAPT tools, their wide dissemination and integration with the curriculum at school and university levels, and further investigation on mobile and collaborative CAPT systems.

Keywords Educational software · Computer aided pronunciation training (CAPT) · English as a second language · English as a foreign language · Phonetics

1 Introduction

English is arguably the most successful language in the history of civilization. English has been evolving since the fifth century and is now spoken by more than a billion

✉ Pinaki Chakraborty
pinaki_chakraborty_163@yahoo.com

[1] Division of Computer Engineering, Netaji Subhas University of Technology, New Delhi, India

people spread over all the six inhabited continents. Interestingly, a majority of the English speakers are non-native speakers and live in countries where English is not used as the primary mode of communication. English is used internationally, typically as a lingua franca for business, science and education.

English is used as a second language in countries like India, Malaysia, the Philippines and Singapore where there is either *de jure* or de facto recognition of the language. English is used as a foreign language in most other countries. Non-native speakers learn English because of various reasons including better access to higher education, better employability and prestige. Facilities are now available for learning English as a second language or a foreign language in most countries. However, non-native English speakers often face problems related to grammar, vocabulary and pronunciation. Non-native speakers tend to mispronounce English words because of the differences in the phonetics of their native language and English, and incorrect letter to sound conversion (Li et al. 2017). Moreover, such mispronunciations are often repetitive (Wang and Lee 2015).

Computers have been used for teaching since the 1960s and educational software is being developed ever since. Specialized software tools have been developed to teach different disciplines at different levels following different pedagogical approaches. Software tools have been developed to teach languages as well. Some such tools focus on specific aspects of language training. The first software tool to teach English pronunciation was developed in the early-1970s (Kalikow and Swets 1972). However, actual progress in Computer Aided Pronunciation Training (CAPT) for English happened only since the beginning of the new millennium. Today, sophisticated software tools are available to teach and test English pronunciation. These tools are typically used by non-native speakers. These tools detect and diagnose mispronunciations in the speech of learners, and then help the learners to correct them. Such tools are available for learners belonging to different age groups and may be used for both formal and informal learning.

This paper reviews the CAPT systems developed so far for teaching English pronunciation and classifies them according to the technology used in them. The objective of this paper is to provide English language teachers an overview of different types of CAPT systems and inform them about their utility.

2 Classification strategy

Over the years, CAPT systems have evolved from programs for just detecting mismatch in pronunciation to sophisticated pedagogical systems. There are two parallel design goals for the developers of CAPT systems as follows.

- The first goal is to maximize the accuracy of mispronunciation detection and diagnosis so that the learners can improve their pronunciation as much as possible. Various probabilistic and artificial neural network based models have been used over the years to improve accuracy of CAPT systems.
- The second goal is to make the CAPT systems more interactive and personalized. A CAPT system should ideally prescribe ways to improve pronunciation according to the profile of a learner.

On the basis of the technology used in them, we have classified the CAPT systems for English into visual simulation based systems, game based systems, comparative phonetics based systems and artificial neural network based systems (Fig. 1). It may be noted that these categories overlap at times. So, we assigned a CAPT system to a category according to the technology that plays the most important part in it.

3 Visual simulation based systems

A visual simulation based CAPT system records the speech of a learner, analyzes it and provides feedback through explanatory images, animated pictorial characters and comparative videos. The approach is simple, attractive and effective especially for young learners and learners with hearing problems.

The first CAPT system for English developed by Kalikow and Swets (1972) followed this approach. They developed a system that used visual feedback for teaching English pronunciation to Spanish students. The system tried to mimic professional language trainers by providing personalized attention to the learners. The system focused on the pronunciation of vowels. It recorded videos of the learners when they spoke and later played those videos showing them their tongue location and trajectory as they voiced vowels. The system analyzed the speech of the learners isolating the vowels in multisyllabic words, and measured the amount of aspiration and time lapse before voicing of two vowels. Images were used to inform the learners about the corrections they required to make in their pronunciation. Since the system used visual feedback, it could be customized for learners with hearing difficulties. Giuliani et al. (2003) developed a system that used animated pictorial characters for teaching English pronunciation to 8–12 year old Italian children. The children had limited linguistic and technical skills. So, the system presented its feedback as animated pictorial characters, like birds and clowns, showing the steps that the children needed to undertake for pronouncing the words correctly.

4 Game based systems

Suitably designed mobile apps and computer games can be used to teach English pronunciation to learners in both formal and informal settings. A game based system can simulate real-world conversations and teach learners to speak according to the context. A few such systems have been developed recently.

Jing and Yong (2014) developed a mobile app to teach English pronunciation. The app records the speech of a learner and compares it with a pre-recorded speech of a

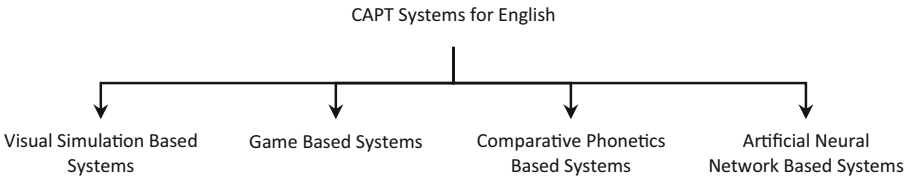

Fig. 1 Classification of CAPT systems for English

human language expert. The app extracts salient features from a speech and uses them in the comparison. The app also uses those features to assess the learners and grades them automatically. The app has been used in the classroom to teach undergraduate students in China. The app was able to improve the pronunciation of the learners at both phoneme level and word level. Recently, Satria et al. (2017) developed a game to augment the conventional mode of teaching English in school. Learners can play the game through voice commands. The learners can progress in the game only if they pronounce correctly. The game internally maintains a database of keywords and matches the pronunciation of the learners with that database. The game provides an interactive and effective way of teaching pronunciation within the setup of a school. The game has been used for improving the vocabulary and pronunciation of 8–13 year old children in Indonesia.

5 Comparative phonetics based systems

The comparative phonetics based approach is sometimes used to teach English pronunciation to adult learners who are fluent in their native languages. In this approach, the phonemes in the native language of a learner are compared to those in English using a stochastic (probabilistic) analysis. The CAPT system then records the speech of the learners, detects mispronunciations and suggests ways for improvement considering their native languages.

Tianli et al. (2003) developed a CAPT system based on a comparison of English and Chinese phonemes for teaching English pronunciation to Chinese students. The system first demonstrates to the learners how to voice individual phonemes, multisyllabic words and complete sentences. After a demonstration session, the learners are asked to take a test. The system records the speech of the learners and analyzes the same. The analysis focuses on the utterance of the phonemes and the stress made on the syllables by the learners. The system provides feedback to the learners that they can use to improve their pronunciation. The system tries to mimic human experts while grading the learners. Abe et al. (2003) developed a system for teaching pronunciation in several Asian and European languages including English. The system uses phonological theories and comparative phonetics to improve the pronunciation and dialog of the learners. Wang et al. (2008) observed that mispronunciations are often caused by cross-language phonological comparisons made by non-native English speakers. They compared English and Chinese (Cantonese) phonemes and identified the common mispronunciations that are made by fluent Chinese speakers while speaking English. They compiled a pronunciation dictionary with correct and probable incorrect pronunciations, and used it to detect mispronunciations in the speech of the learners. The system was used to teach English pronunciation to Chinese adult learners. Qian et al. (2010) developed a probabilistic technique for detecting and diagnosing mispronunciations. The system compared Chinese (Cantonese) and English phonemes on the basis of phonological rules defined by expert linguists. The system was used by Chinese undergraduate students learning English. Qian et al. (2016) extended the work by separating the process of mispronunciation detection and mispronunciation diagnosis into two passes. The system was able to detect and diagnosis mispronunciations with a high accuracy.

6 Artificial neural network based systems

An artificial neural network is a model of computation in which a network of artificial neurons, also called nodes, is used to solve computational tasks for which no efficient algorithmic solution exists. The nodes can process information in parallel and are arranged in layers. The first layer is known as the input layer and the nodes in it receive the input signals. The last layer is known as the output layer and the nodes in it emit the output signals. There are one or more layers of nodes between the input layer and the output layer known as the hidden layers. The nodes are connected by connection links which carry signals from one node to another. The connection links have numeric weights associated with them which are multiplied to the amplitude of the signals they carry. Artificial neural networks can be used to detect and diagnose mispronunciations in a speech. An artificial neural network has to be trained first. In the training phase, the system records the speech of human language experts and the weights of the connection links in the artificial neural network are adjusted accordingly following an algorithm known as the backpropagation algorithm. The language experts typically need to voice a few hundred sentences to train the artificial neural network. Once trained, the artificial neural network can be used to detect mispronunciations in the speech of the learners. Artificial neural networks are a powerful model of computation and can be used to diagnose mispronunciations accurately.

In an early study, Akima et al. (1992) developed a system that used an artificial neural network to detect mispronunciation in the speech of Japanese schoolchildren learning English. The system emphasized on the correct voicing of the phonetic vowels. Qian et al. (2012) developed a system that uses stochastic reasoning and a deep neural network, i.e. an artificial neural network with multiple hidden layers, for diagnosing mispronunciation in a speech. The system has been used for teaching undergraduate students in China. Chen and Jang (2015) developed a system which was trained by human language experts on recognizing correct phones. The system used an artificial neural network and a support vector machine, i.e. another model of non-algorithmic computation, for recognizing words based on those phones. The system also automatically grades the learners. Wang and Lee (2015) hypothesized that a learner repeats similar mispronunciations because of the differences in the phonemes present in his/her native language and in English. They developed a system that detects mispronunciations in the speech of a learner with an emphasis on identifying frequently repeated mistakes. The system uses stochastic reasoning and a deep neural network. In a recent study, Li et al. (2017) hypothesized that the three main reasons for speakers to mispronounce are the differences in the phonetics of their native language and English, incorrect letter to sound conversion, and misreading of the text prompts. Li et al. (2017) used a deep neural network to detect mispronunciations arising from these reasons and to suggest the correct pronunciation.

7 Performance analysis

7.1 Technological and human factors

Researchers have tried to assess the CAPT systems in terms of both technological factors and human factors (Table 1).

Table 1 Comparison of performance of the CAPT systems

	Technological Factors	Human Factors
Kalikow and Swets (1972)	–	Language experts observed that pronunciation of learners improved in ≈70% scenarios after training
Akima et al. (1992)	–	≈7% decrease in mispronunciation recorded among learners after 2 months of training
Giuliani et al. (2003)	–	Words need to be repeated 1.27 times on an average to obtain correct utterance from learners
Tianli et al. (2003)	≈7% improvement in detection of mispronunciation arising from phoneme substitution over probabilistic scoring techniques	Mispronunciation arising from phoneme insertion is less likely in ≈70% scenarios after training
Wang et al. (2008)	≈86% accuracy in mispronunciation detection	–
Nyugen et al. (2010)	–	Learners gave the tool an average rating of 3.9 out of 5
Qian et al. (2010)	≈15% reduction in false mispronunciation detection and diagnosis compared to machine learning techniques based on hand-coded rules	–
Qian et al. (2012)	≈18% improvement in detection of mispronounced words over probabilistic classification technique	–
Jing and Yong (2014)	–	≈23% improvement in the pronunciation performance of learners
Chen and Jang (2015)	≈15% false mispronunciation detection	Strong correlation with ranking provided to learners by language experts
Wang and Lee (2015)	≈28% mispronunciation detection error rate	–
Qian et al. (2016)	≈15% error in mispronunciation detection and ≈17% error in mispronunciation diagnosis	–
Li et al. (2017)	≈5% false positives and ≈31% false negatives in mispronunciation detection, and ≈14% diagnostic error rate	Phoneme-level mispronunciation reduced from 17% to 11%
Satria et al. (2017)	–	Learners gave the tool an average rating of 4.4 for usability and 4.7 for effectiveness out of 5

Researchers use various technological factors to access how correctly a CAPT system can detect and diagnose mispronunciation in speech. Wang et al. (2008) reported an accuracy of 86% in mispronunciation detection using their CAPT system. For most other systems, the diagnostic error rate varied between 14% (Li et al. 2017) and 28% (Wang and Lee 2015). Some researchers compared the accuracy of their systems with existing techniques for mispronunciation detections (Tianli et al. 2003; Qian et al. 2010, 2012).

Human factors like how the CAPT systems are received by English teachers and learners, and how much learners are benefitted by CAPT systems are matter of interest. Several researchers tried to determine the effectiveness of training using the CAPT systems. Akima et al. (1992) reported 7% decrease in mispronunciation after learners have used their CAPT system for 2 months. Recently, Jing and Yong (2014) reported 23% improvement in pronunciation performance among learners. Language experts agreed with the evaluation performed by CAPT systems (Chen and Jang 2015), while learners appreciated the CAPT system for their effectiveness and usability (Nyugen et al. 2010; Satria et al. 2017).

7.2 State-of-the-art practices

We have identified four practices that have been used by several researchers for improving the effectiveness and usability of CAPT systems. The first two practices improve the technological performance of CAPT systems, while the latter two improve human factors associated with CAPT systems.

- *Pronunciation dictionaries.* A pronunciation dictionary is a database containing the acceptable pronunciations of different words. Several CAPT systems compare the utterances of learners with the entries in a pronunciation dictionary to detect mispronunciation (Wang et al. 2008; Qian et al. 2010; Wang and Lee 2015; Li et al. 2017). Pronunciation dictionaries may also contain frequently occurring mispronunciations to optimize mispronunciation detection.
- *Deep learning.* Artificial neural networks have been used in CAPT systems since the early-1990s (Akima et al. 1992). However, the artificial neural networks used in recent CAPT systems have more neurons and those neurons are typically arranged in multiple hidden layers (Qian et al. 2012; Wang and Lee 2015; Li et al. 2017). CAPT systems often use deep neural networks to achieve higher accuracy.
- *Multimedia support.* Several researchers have made use of the multimedia support provided by computers to improve the usability of CAPT systems (Kalikow and Swets 1972; Abe et al. 2003; Giuliani et al. 2003; Satria et al. 2017). CAPT systems use images, animation and audio and video clips to teach learners how to pronounce correctly.
- *Automated grading.* To relieve language teachers from the task of grading pronunciation assignments of a large number of learners, several CAPT systems have modules to grade learners without any intervention of the teachers (Tianli et al. 2003; Jing and Yong 2014). Chen and Jang (2015) found a strong correlation between the grades given by a CAPT system and a language teacher to learners. Researchers have used both algorithmic and probabilistic techniques to grade pronunciation of learners.

8 Discussion

8.1 An analysis of CAPT systems for English

We observed that a number of software tools to teach English pronunciation has been developed in the last two decades. However, we believe that there is a need for more such tools given the large number of non-native speakers interested in learning English. Moreover, there is not much awareness about CAPT tools among learners and many English language teachers doubt their efficacy.

Table 2 presents a summary of the CAPT systems developed to date for teaching English pronunciation. We observed that those tools have been developed and used predominantly in Asian and European countries where English is used as a foreign language. No use of such tools in countries where English is used as a second language has been reported in the literature. Over the years, different researchers focused on different aspects of CAPT systems. For example, some researchers emphasized on correct voicing of vowels (Kalikow and Swets 1972; Akima et al. 1992), some researchers tried to identify frequently occurring mispronunciations (Wang and Lee 2015; Li et al. 2017) and some researchers focused on providing learners with personalized feedback (Kalikow and Swets 1972). Speech analysis and mispronunciation detection are computationally intensive tasks. Consequently, the artificial neural networks that have been used by the researchers typically have sophisticated architectures (Qian et al. 2012; Wang and Lee 2015; Li et al. 2017). Training artificial neural networks is a time consuming activity and, in the case of CAPT systems, too much effort is required from the language experts. Qian et al. (2012) experimented with techniques for minimizing the workload of the language experts.

The different techniques used in the CAPT systems have their own limitations. For example, although visual simulation based systems are appealing to naive users, they do not provide detailed analysis and feedback. Game based CAPT systems are preferred by a small section of learners but require a lot of effort to develop. Requirements for a suitable native language corpus and availability of language experts are the limiting factors in developing comparative phonetics based CAPT systems. Alternatively, developing artificial neural network based CAPT systems requires large datasets of phonemes and long training periods.

Some of the CAPT tools target specific sections of learners, while the other tools are generic. Figure 2 arranges the different categories of CAPT systems according to the expected level of maturity of their target users, with the visual simulation based systems being appropriate for the most naive learners and the artificial neural network based systems being suitable for senior students and professionals. As expected, the tools targeting naive learners typically have more attractive user interfaces.

8.2 General computational linguistic tools

There are some general computational linguistic tools that we believe can be extended to teach English pronunciation. For example, Juang and Furui (2000) used a stochastic model to analyze speech signals with an objective to understand speech. In a later study with a similar objective, Shum et al. (2016) used of a

Table 2 A summary of the CAPT systems for English

	Technology Used	Country	Target Learners	Key Contribution
Kalikow and Swets (1972)	Visual simulation	Spain	Students	Visual feedback illustrating pronunciation errors and efficient ways for correcting them
Akima et al. (1992)	Artificial neural network	Japan	Schoolchildren	Detection of mispronunciation using an artificial neural network that has been trained by language experts
Giuliani et al. (2003)	Visual simulation	Italy	Schoolchildren	Feedback provided through multicolored animated pictorial characters
Abe et al. (2003)	Comparative phonetics	Japan	Undergraduate students	Use of phonological theories for teaching pronunciation and dialog in seventeen languages including English
Tianli et al. (2003)	Comparative phonetics	China	Students	Teaching learners to voice individual phonemes, multisyllabic words and complete sentences
Wang et al. (2008)	Comparative phonetics	China	Adult learners	Comparing the speech of learners with a pronunciation dictionary compiled by experts to detect mispronunciations
Qian et al. (2010, 2016)	Comparative phonetics	China	Undergraduate students	Developed a comparative model of Chinese and English phonemes and used it to teach English pronunciation
Qian et al. (2012)	Artificial neural network	China	Undergraduate students	Used stochastic reasoning and a deep neural network for diagnosing mispronunciation
Jing and Yong (2014)	Game based learning	China	Undergraduate students	Mobile app for speech analysis and mispronunciation detection
Chen and Jang (2015)	Artificial neural network	Taiwan	Graduate students	Used an artificial neural network and a support vector machine to recognize words on the basis of phonemes provided by language experts
Wang and Lee (2015)	Artificial neural network	China	–	Used stochastic reasoning and a deep neural network to detect mispronunciation with an emphasis on frequently occurring mistakes
Li et al. (2017)	Artificial neural network	China	Undergraduate students	Tried to identify the reasons for mispronunciation and used a deep neural network to detect mispronunciation
Satria et al. (2017)	Game based learning	Indonesia	Schoolchildren	Game to be played through voice commands

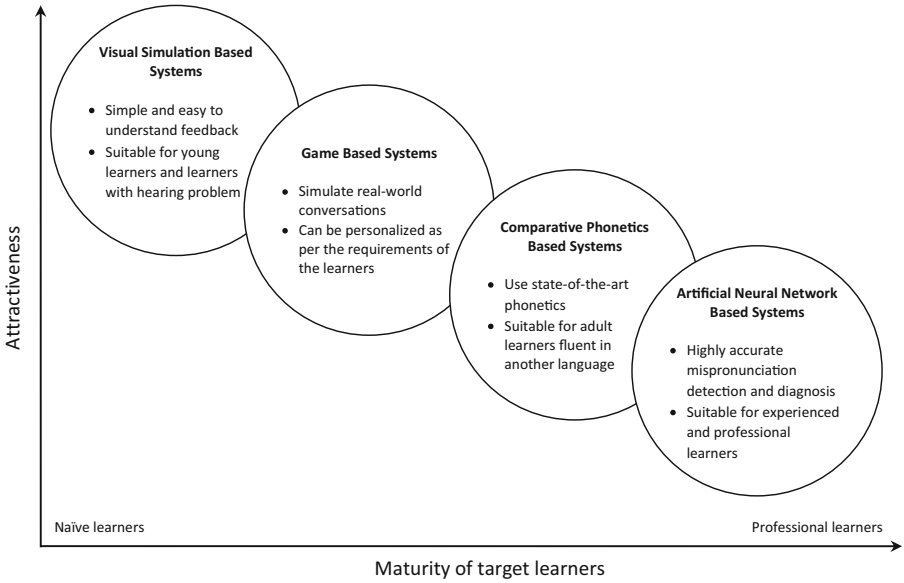

Fig. 2 Salient features of different categories of CAPT systems for English. The categories are arranged according to their attractiveness and the expected level of maturity of the target learners

combination of spectral analysis, stochastic reasoning and a deep neural network to analyze speech of learners and build a language recognition system. These systems can be adapted to accurately diagnose mispronunciations. Nyugen et al. (2010) developed a mobile app for teaching English as a foreign language to Vietnamese learners. The app personalizes the content presented to the learners based on their location and level of expertise, and the time of the day. Liaw (2014) developed a speech recognition based online reading system to supplement teaching English as a foreign language to elementary school students. In a recent experiment, Lee et al. (2017) used a recurrent neural network, i.e. an artificial neural network in which the nodes that use signals produced by one another form a cycle, to analyze the messages posted by the learners on social networking websites. Such techniques that present material to learners based on the context can be used in CAPT as well. Jain et al. (2018) used an augmented reality mobile app to teach novel words to kindergarteners. Such augmented reality mobile apps may be used to teach pronunciation as well. Nakai et al. (2018) discussed tools to display videos of the tongue and other internal organs as people voice sounds found in different languages. The tools can help in a systematic study of pronunciation.

8.3 CAPT systems for other languages

In the last two decades, some sophisticated CAPT systems have been developed for languages like Arabic, Chinese, French and Japanese. We believe that the technologies used in these systems can also be used to develop CAPT systems for English. For example, Abdou et al. (2006) used stochastic reasoning to develop a system for teaching Arabic pronunciation with an emphasis on uttering verses. A few sophisticated game based CAPT systems have been developed for Chinese

and French. Chiu et al. (2007) developed a computer game to teach Chinese pronunciation. The system shows short video clips and the learners have to speak and voice words as per the scene. Later, Su et al. (2015) developed a more advanced game based system to teach Chinese pronunciation. The system presents a scenario to the learners and starts a conversation. The system continuously analyzes the speech of the learners and makes them speak specific sentences that will improve their pronunciation. Athanasopoulos et al. (2017) developed a three-dimensional game to teach French pronunciation. The learners take part in a conversation as a part of the game, and the system analyzes their speech and provides feedback for improving their pronunciation. In a recent study, Samsudin and Mano (2017) developed a visual simulation based system to teach Japanese pronunciation using animated texts. The system can be used efficiently by both naive and expert learners.

9 Conclusion

Several CAPT systems have been developed and used successfully for teaching English pronunciation to learners belonging to different age groups in different countries. However, we believe that more effort is needed on the development of CAPT systems and they should be used more proactively. Ongoing developments in digital signal processing, artificial intelligence, augmented and virtual reality and other branches of computer science will be helpful in developing more sophisticated CAPT systems in the near future. We have some recommendations regarding the development, dissemination and use of CAPT systems for English as follows.

- Language teachers and computer engineers must collaborate to develop high quality CAPT tools. CAPT software should be disseminated through online platforms like GitHub and Google Play so that they can reach a wider audience.
- CAPT tools should be integrated with the curriculum in schools and universities. Language teachers should be made aware of the benefits of using CAPT tools. CAPT tools may be used by the learners under the supervision of a teacher during the class hours as well as for self study later.
- CAPT tools should be developed and used in countries where English is used as a second language. Many of these countries have a large number of students interested in learning English but inadequate English language teachers. CAPT tools for English will be immensely helpful for the students in these countries.
- Smartphones have computation power comparable to that of a personal computer and built in support for multimedia. One mobile app for English CAPT has been already developed and used successfully. Mobile based CAPT systems can be used for both formal and informal learning. Implementing CAPT software as mobile apps will also allow more people to use them.
- More research and innovation related to CAPT is necessary. English language teachers should offer massive open online courses (MOOCs) on English pronunciation and use CAPT tools in them. Scope of social media based and collaborative CAPT should also be investigated.

References

Abdou, S. M., Hamid, S. E., Rashwan, M., Samir, A., Abdel-Hamid, O., Shahin, M., & Nazih, W. (2006). Computer aided pronunciation learning system using speech recognition techniques. In: Proceedings of the ninth international conference on spoken language processing, pp. 849–852.

Abe, S., Nakata, S., Kigoshi, T., & Mochizuki, H. (2003). Designing and developing multilingual e-learning materials: TUFS language education pronunciation module - introduction of a system for learning Japanese language pronunciation. In: Proceedings of the Third IEEE International Conference on Advanced Learning Technologies, pp. 462–462.

Akima, Y., Watanabe, S., Tsubota, A., & Sone, M. (1992). Application of neural networks to the teaching of English pronunciation. In: Proceedings of the Singapore ICCS/ISITA Conference, vol. 2, pp. 553–557.

Athanasopoulos, G., Hagihara, K., Cierro, A., Guerit, R., Chatelain, J., Lucas, C., & Macq, B. (2017). 3D immersive karaoke for the learning of foreign language pronunciation. In: Proceedings of the international conference on 3D immersion, pp. 1–8.

Chen, L. -Y., & Jang, J. -S. R. (2015). Automatic pronunciation scoring with score combination by learning to rank and class-normalized DP-based quantization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23*(11), 1737–1749.

Chiu, C. -F., Lee, G. C., & Yang, J. -H. (2007). Design and implementation of video-enabled web-based pronunciation debugging system. In: Proceedings of the Seventh IEEE International Conference on Advanced Learning Technologies, pp. 374–378.

Giuliani, D., Mich, O., & Nardon, M. (2003). A study on the use of a voice interactive system for teaching English to Italian children. In: Proceedings of the Third IEEE International Conference on Advanced Learning Technologies, pp. 376–377.

Jain, D., Patil, A. P., Nawal, D. J., & Chakraborty, P. (2018). ARWAK: An augmented reality wordbook smartphone app for kindergarteners. *Journal of Multi Disciplinary Engineering Technologies, 12*(2), 59–66.

Jing, X., & Yong, L. (2014). The speech evaluation method of English phoneme mobile learning system. In: Proceedings of the IEEE Workshop on Advanced Research and Technology in Industry Applications, pp. 546–550.

Juang, B. -H., & Furui, S. (2000). Automatic recognition and understanding of spoken language – A first step toward natural human-machine communication. *Proceedings of the IEEE, 88*(8), 1142–1165.

Kalikow, D. N., & Swets, J. A. (1972). Experiments with computer-controlled displays in second language learning. *IEEE Transactions on Audio and Electro Acoustics, 20*(1), 23–28.

Lee, H. -Y., Tseng, B. -H., Wen, T. -H., & Tsao, Y. (2017). Personalizing recurrent-neural-network based language model by social network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(3), 519–530.

Li, K., Qian, X., & Meng, H. (2017). Mispronunciation detection and diagnosis in L2 English speech using multi-distribution deep neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 25*(1), 193–207.

Liaw, M. -L. (2014). The affordance of speech recognition technology for EFL learning in an elementary school setting. *Innovation in Language Learning and Teaching, 8*(1), 79–93.

Nakai, S., Beavan, D., Lawson, E., Leplâtre, G., Scobbie, J. M., & Smith, J. S. (2018). Viewing speech in action: Speech articulation videos in the public domain that demonstrate the sounds of the international phonetic alphabet (IPA). *Innovation in Language Learning and Teaching, 12*(3), 212–220.

Nyugen, V. A., Pham, V. C., & Ho, S. D. (2010). A context aware mobile learning adaptive system for supporting foreigner learning English. In: Proceedings of the IEEE International Conference on Computing and Communication Technologies, Research, Innovation, and Vision for the Future, pp. 1–6.

Qian, X., Soong, F., & Meng, H. (2010). Discriminative acoustic model for improving mispronunciation detection and diagnosis in computer-aided pronunciation training (CAPT). In: Proceedings of the eleventh annual conference of the international speech communication association, 757–760.

Qian, X., Meng, H., & Soong, F. (2012). The use of DBN-HMMs for mispronunciation detection and diagnosis in L2 English to support computer-aided pronunciation training. In: Proceedings of the thirteenth annual conference of the international speech communication association, pp. 775–778.

Qian, X., Meng, H., & Soong, F. (2016). A two-pass framework of mispronunciation detection and diagnosis for computer-aided pronunciation training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24*(6), 1020–1028.

Samsudin, N. S. B., & Mano, K. (2017). Animated texts application in visualizing speech features for foreign language learning. In: Proceedings of the IEEE region 10 conference, pp. 1778–1783.

Satria, F., Aditra, H., Wibowo, M. D. A., Luthfiansyah, H., Suryani, M., Paulus, E., & Suryana, I. (2017). EFL learning media for early childhood through speech recognition application. In: Proceedings of Third International Conference on Science in Information Technology, pp. 568–572.

Shum, S. H., Harwath, D. F., Dehak, N., & Glass, J. R. (2016). On the use of acoustic unit discovery for language recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24*(9), 1665–1676.

Su, P. -H., Wu, C. -H., & Lee, L. -S. (2015). A recursive dialogue game for personalized computer-aided pronunciation training. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23*(1), 127–141.

Tianli, Z., Jia, L., Yanfeng, L., Shunping, H., & Chaolei, L. (2003). An automatic pronunciation teaching system for Chinese to learn English. In: Proceedings of the IEEE international conference on robotics intelligent systems and signal processing, vol. 2, pp. 1157–1161.

Wang, Y. B., & Lee, L. S. (2015). Supervised detection and unsupervised discovery of pronunciation error patterns for computer-assisted language learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 23*(3), 564–579.

Wang, L., Feng, X., & Helen, M. (2008). Automatic generation and pruning of phonetic mispronunciations to support computer-aided pronunciation training. In: Proceedings of the ninth annual conference of the international speech communication association, pp. 1729–1732.