Development English Pronunciation Practicing System Based on Speech Recognition

Ngoc Hoang Phan[1(✉)], Thi Thu Trang Bui[1], and V. G. Spitsyn[2]

[1] Ba Ria-Vung Tau University, 80, Truong Cong Dinh, Vung Tau, Ba Ria-Vung Tau, Vietnam
hoangpn285@gmail.com, trangbt.084@gmail.com
[2] National Research Tomsk Polytechnic University, 30, Lenin Avenue, Tomsk, Russia
spvg@tpu.ru

Abstract. The relevance of the research is caused by the need of application of speech recognition technology for language teaching. The speech recognition is one of the most important tasks of the signal processing and pattern recognition fields. The speech recognition technology allows computers to understand human speech and it plays very important role in people's lives. This technology can be used to help people in a variety way such as controlling smart homes and devices; using robots to perform job interviews; converting audio into text, etc. But there are not many applications of speech recognition technology in education, especially in English teaching. The main aim of the research is to propose an algorithm in which speech recognition technology is used English language teaching. Objects of researches are speech recognition technologies and frameworks, English spoken sounds system. Research results: The authors have proposed an algorithm based on speech recognition framework for English pronunciation learning. This proposed algorithm can be applied to another speech recognition framework and different languages. Besides the authors also demonstrated how to use the proposed algorithm for development English pronunciation practicing system based on iOS mobile app platform. The system also allows language learners can practice English pronunciation anywhere and anytime without any purchase.

Keywords: Speech recognition · English pronunciation · Hidden markov models · Neural networks · Mobile application

1 Introduction

1.1 Speech Recognition Technology

Speech recognition technology has been researched and developed over the past several decades. In the 1960's this technology was developed based on filter-bank analyses, simple time normalization methods and the beginning of sophisticated dynamic programming methodologies. In this time technology could recognize small vocabularies (10–100 words) of isolated words using simple acoustic phonetic properties of speech sounds [1].

P. C. Vinh and A. Rakib (Eds.): ICCASA 2019/ICTCC 2019, LNICST 298, pp. 157–166, 2019.
https://doi.org/10.1007/978-3-030-34365-1_12

In the 1970's the key technologies of speech recognition were the pattern recognition models, spectral representation using LPC methods, speaker-independent recognizers using pattern clustering methods and dynamic programming methods for connected word recognition. During this time, we able to recognize medium vocabularies (100–1000 words) using simple template-based and pattern recognition methods [1].

In the 1980's the speech recognition technology started to solve the problems of large vocabulary (1000 – unlimited number of words) using statistical methods and neural networks for handling language structures. The important technologies used in this time were the Hidden Markov Model (HMM) and stochastic language model [1]. Using HMMs allowed to combine different knowledge sources such as acoustics, language, and syntax, in a unified probabilistic model.

In the 1990's the key technologies of speech recognition were stochastic language understanding methods, statistical learning of acoustic and language models, finite state transducer framework and FSM library. In this time speech recognition technology allow us to build large vocabulary systems using unconstrained language models and constrained task syntax models for continuous speech recognition and understanding [1].

In the last few years, the speech recognition technology can handle with very large vocabulary systems based on full semantic models, integrated with text-to-speech (TTS) synthesis systems, and multi-modal inputs. In this time, the key technologies were highly natural concatenative speech synthesis systems, machine learning to improve both speeches understanding and speech dialogs [1].

1.2 Key Speech Recognition Methods

Dynamic Time Warping (DTW)

Dynamic time warping (DTW) is an approach that was historically used for speech recognition. This method is used to recognize about 200-word vocabulary [2]. DTW divide speech into short frames (e.g. 10 ms segments) and then it processes each frame as a single unit. During the time of DTW, achieving speaker independence remained unsolved. DTW was applied for automatic speech recognition to cope with different speaking speeds. It allows to find an optimal match between two given sequences (e.g., time series) with certain restrictions.

Hidden Markov Models (HMM)

DTW has been displaced by the more successful Hidden Markov Models-based approach. HMMs are statistical models that output a sequence of symbols or quantities. In HMMs a speech signal can be a piecewise stationary signal or a short-time stationary signal. And speech can be approximated as a stationary process in a short time-scale (e.g., 10 ms).

By the mid-1980s a voice activated typewriter called Tangora was created. It could handle a 20,000-word vocabulary [3]. It processes and understands speech based on using statistical modeling techniques like HMMs. However, HMMs are too simplistic to account for many common features of human languages [4]. But it proved to be a highly efficiency model for speech recognition algorithm in the 1980s [1].

Neural Networks

Neural networks have been used in speech recognition to solve many problems such as phoneme classification, isolated word recognition, audiovisual speech recognition, audiovisual speaker recognition and speaker adaptation [5, 6].

By comparing with HMMs, neural networks make fewer explicit assumptions about feature statistical properties. Neural networks allow discriminative training in a natural and efficient manner, so they are effectiveness in classifying short-time units such as individual phonemes and isolated words [7]. However, because of their limited ability to model temporal dependencies, neural networks are not successfully used for continuous speech recognition.

To solve this problem, neural networks are used to pre-process speech signal (e.g. feature transformation or dimensionality reduction) and then use HMM to recognize speech based on the features received from neural networks [8]. In recently, related Recurrent Neural Networks (RNNs) have showed an improved performance in speech recognition [9–11].

Like shallow neural networks, Deep Neural Networks (DNNs) can used to model complex non-linear relationships. The architectures of these DNNs generate compositional models, so DNNs have a huge learning capacity and they are potential for modeling complex patterns of speech data [12]. In 2010, the DNN with the large output layers based on context dependent HMM states constructed by decision trees have been successfully applied in large vocabulary speech recognition [13–15].

End-to-end Automatic Speech Recognition

Traditional HMM-based approaches required separate components and training for the pronunciation, acoustic and language model. And a typical n-gram language model, required for all HMM-based systems, often takes several gigabytes memory to deploy them on mobile devices [16]. However, since 2014 end-to-end ASR models jointly learn all the components of the speech. It allows to simplify the training and deployment process. Because of that, the modern commercial ASR systems from Google and Apple are deployed on the cloud.

Connectionist Temporal Classification (CTC) based systems was the first end-to-end ASR and introduced by Alex Graves of Google DeepMind and Navdeep Jaitly of the University of Toronto in 2014 [17]. In 2016, University of Oxford presented LipNet using spatiotemporal convolutions coupled with an RNN-CTC architecture. It was the first end-to-end sentence-level lip reading model. And it was better than human-level performance in a restricted grammar dataset [18]. In 2018 Google DeepMind presented a large-scale CNN-RNN-CTC architecture. In the results this system achieved 6 times better performance than human experts [19].

1.3 Speech Recognition Applications

With speech recognition technology computers now can hear and understand what people speak to them and can do what people want they do. The speech recognition technology can be used in a variety way and plays very important role in people's lives. For example, this technology can be used in-car systems or smart home systems to help people do simple thing by voice commands such as: play music or select radio station, initiate phone calls, turn on/off lights, televisions and other electrical devices.

For education, speech recognition technology can be used to help students who are blind or have very low vision. They can use computer by using voice commands instead of having a look at the screen and keyboard [20]. Besides, students who are physically disabled or suffer from injuries having difficulty in writing, typing or working can benefit from using this technology. They can use speech-to-text programs to do their homework or school assignments [21]. Speech recognition technology can allow students to become better writers. They can improve the fluidity of their writing by using speech-to-text programs. When they say to computer, they don't worry about spelling, punctuation, and other mechanics of writing [21]. In addition, speech recognition technology can be useful for language learning. They can teach people proper pronunciation and help them to develop their speaking skills [22].

Recently, all people have their own mobile devices and they can use them anywhere, anytime. Most of mobile apps and devices runs on two main operating systems: iOS and Android OS. These operating systems are equipped with the best speech recognition technology developed by Google or Apple. There are many mobile apps that use these speech recognition technologies for playing games, controlling devices, making phone calls, sending text messages etc.

There are also many software applications to practice English pronunciation on mobile devices. By using these support tools, learners can record all what they say and compare with sample pronunciation of native speakers to correct errors. The applications often display the pronunciation of words, allowing learners to listen to sample pronunciation, then the learners will record their pronunciation and compare themselves with the sample pronunciation. The application has not integrated the voice recognition feature into the software to test the learner's pronunciation.

Because of that, building a mobile app using speech recognition technologies for language pronunciation learning is urgent and perspective. In this paper we present an algorithm that use speech recognition technology to help people determine if they properly pronounce an English sound. The proposed algorithm is used for building mobile app based on speech recognition technology. This algorithm is tested.

2 Proposed Algorithm

In this paper, we propose an algorithm based on speech recognition framework for English pronunciation learning. The framework used to test proposed algorithm in this paper is Apple speech recognition technology [23]. Besides, in this paper we demonstrate how to use the proposed algorithm for development English pronunciation practicing system based on iOS mobile app platform. This proposed algorithm can be applied to another speech recognition framework (e.g. Google speech recognition) and different languages.

The main aim of developing of proposed algorithm to help learners can use speech recognition technologies to test their own English pronunciation and make appropriate adjustments. The application will provide learners with the inherent functions of an English pronunciation training tool and support learners to completely free practice English pronunciation anytime, anywhere.

2.1 Apple Speech Recognition Technology

The Apple speech recognition framework allow to recognize spoken words in recorded or live audio. It can be used to translate audio content to text, handle recognize verbal commands etc. The framework is fast and works in near real time. Besides the framework is accurate and can interpret over 50 languages and dialects [23]. The process of speech recognition task using Apple technology can be presented in Fig. 1.

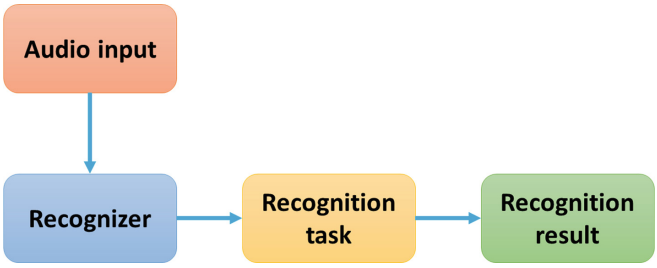

Fig. 1. Process of speech recognition task on speech recognition framework.

Audio Input is an audio source from which transcription should occur. Audio source can be read from recorded audio file or can be captured audio content, such as audio from the device's microphone. The audio input is then sent to Recognizer that is used to check for the availability of the speech recognition service, and to initiate the speech recognition process. At the end, the process gives the partial or final results of speech recognition [23].

2.2 One-Word Pronunciation Assessment

Based on this speech recognition framework, we propose an algorithm to assess the language learner's pronunciation. The process of pronunciation assessment for one word is presented in Fig. 2.

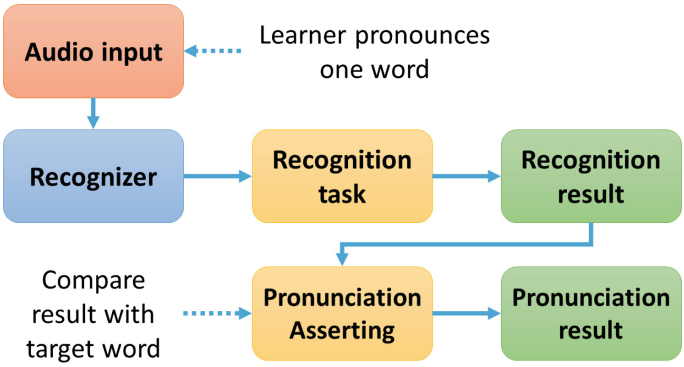

Fig. 2. Process of pronunciation assessment for one word.

At first the language learner pronounces a word which is used to practice pronunciation. Then the learner's pronunciation is handled by speech recognition framework which gives the recognition result. After that, the recognition result is compared with target word to determine if the learner correctly pronounce the target word (Fig. 3).

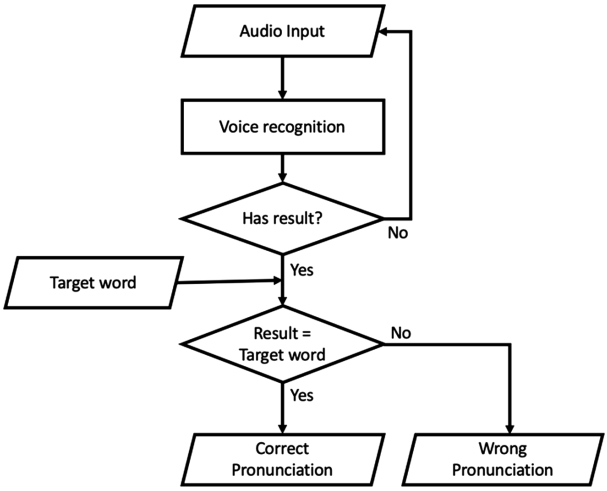

Fig. 3. Learner's pronunciation assessment for one word

2.3 One Sound Pronunciation Assessment

In order to assess one sound pronunciation, we need to assess the pronunciations of list of words which contain the target sound. The process of pronunciation assessment for one sound can be then presented in Fig. 4.

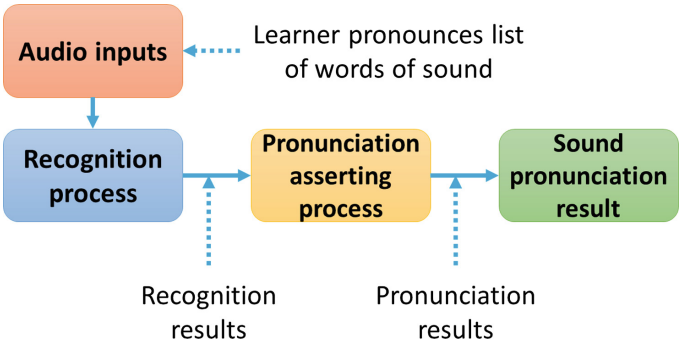

Fig. 4. Process of pronunciation assessment for one sound.

At first the language learner pronounces one word of the list which contains the sound used to practice pronunciation. Then the learner's pronunciation is handled by recognition process. After that the recognition result are processed by pronunciation asserting. The language learner repeats these steps for other words of the list until all words of the list have been pronounced. Based on the pronunciation results of words in the list, we can calculate the sound pronunciation fluency of the language learner by following formula:

Sound pronunciation fluency = Total number of correctly pronounced words /Total number of words in the list.

2.4 English Pronunciation Practicing System

The English language contains 44 sounds divided into three main groups: vowels (12 sounds), diphthongs (8 sounds) and consonants (24 sounds). The vowel sounds consist of two sub-groups: long sounds and short sounds. The consonant sounds consist of three sub-groups: voiced consonants, voiceless consonants and other consonants. The phonemic chart of 44 English spoken sounds is presented in Table 1.

Based on the phonemic chart of spoken English sounds, proposed algorithm for word and sound pronunciation asserting, we developed an iOS app for English pronunciation practicing system. The main aim of this system is to allow language learners can know if they correctly pronounce English sounds. Based on the results, provided by this system, language learners will have proper adjustment to improve their English pronunciation. Besides the app allows language learners can freely practice pronunciation anywhere and anytime.

Table 1. Phonemic chart English sounds

English sounds	Vowels	Short sounds	ɪ	e	æ	ʌ	ʊ	ə	ɒ	
		Long sounds	i:	ɜ:	u:	ɔ:	ɑ:			
	Diphthongs		eɪ	ɔɪ	aɪ	eə	ɪə	ʊə	əʊ	aʊ
	Consonants	Voiceless consonants	p	f	θ	t	s	ʃ	ʧ	k
		Voiced consonants	b	v	ð	d	z	ʒ	ʤ	g
		Other	m	n	ŋ	h	w	l	r	j

The English pronunciation practicing system consists of 44 lessons according to 44 spoken English sounds (Fig. 5a). Each lesson has its own practicing exercises and depending on the sound these exercises normally divided into the following types: the sound is at the beginning of words; the sound is in middle of words; the sound is at the end of words; the sound is followed by a vowel/consonant; the sound is after a vowel/consonant (Fig. 5b).

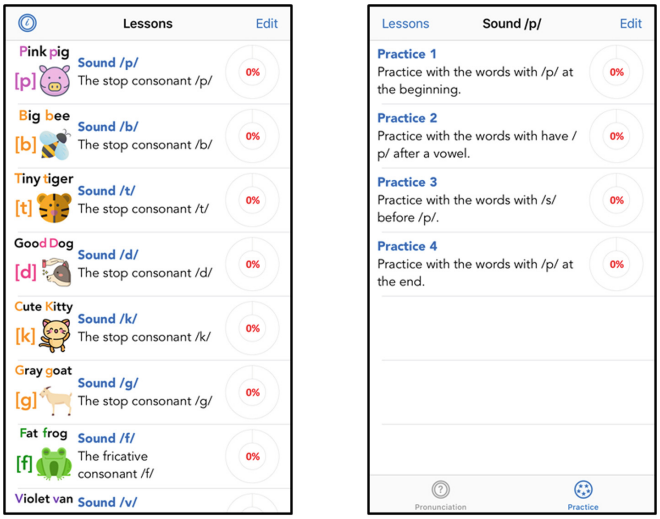

Fig. 5. English pronunciation practicing system: (a) list of lessons, (b) examples of exercise types of sound p.

The language learners must practice with all words in the list of exercise, and then the system will automatic give recognition and pronunciation results according each word (Fig. 6). After that the system calculates the pronunciation fluency for each sound and shows the results to the language learners (Fig. 7).

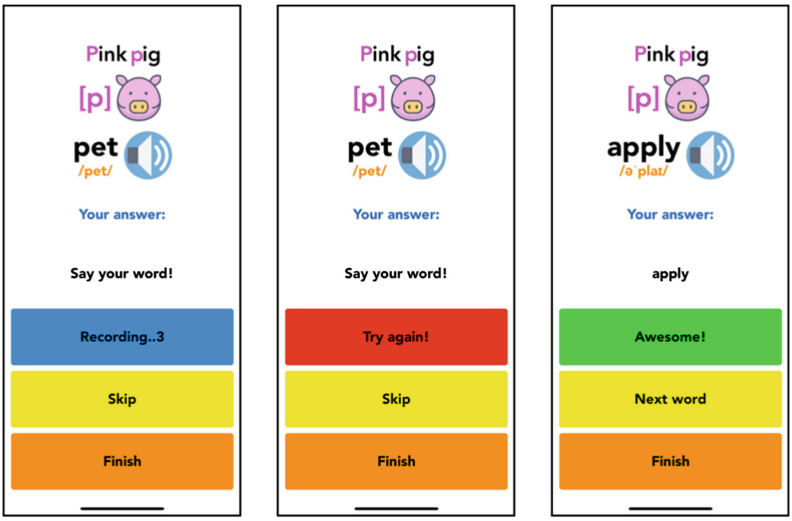

Fig. 6. Example of one practice: (a) practice overview and mode, (b) practice answer mode, (c) pronunciation result of one word.

a **b** **c**

Fig. 7. Example of pronunciation assessment: (a) pronunciation result for one practice, (b) pronunciation result for practices, (c) pronunciation result for sound.

3 Conclusion

In this paper, we propose an algorithm based on speech recognition framework for English pronunciation learning. This proposed algorithm can be applied to another speech recognition framework (e.g. Google speech recognition) and different languages. Besides we also demonstrate how to use the proposed algorithm for development English pronunciation practicing system based on iOS mobile app platform.

This system allows language learners can determine if they correctly pronounce English sounds. Based on these results, the language learners will have proper adjustment to improve their English pronunciation. The system also allows language learners can practice English pronunciation anywhere and anytime without any purchase, which they can not do in the classroom.

References

1. Juang, B.H., Rabiner, L.R.: Automatic speech recognition–a brief history of the technology development (2015). https://web.ece.ucsb.edu/Faculty/Rabiner/ece259/Reprints/354_LALI-ASRHistory-final-10-8.pdf
2. Benesty, J., Sondhi, M.M., Huang, Y.: Springer Handbook of Speech Processing. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-49127-9
3. Jelinek, F.: Pioneering Speech Recognition (2015). https://www.ibm.com/ibm/history/ibm100/us/en/icons/speechreco/
4. Huang, X., Baker, J., Reddy, R.: A Historical perspective of speech recognition. Commun. ACM **57**(1), 94–103 (2014)

5. Hanazawa, T., Hinton, G., Shikano, K., Lang, K.J.: Phoneme recognition using time-delay neural networks. IEEE Trans. Acoust. Speech Sig. Process. **37**(3), 328–339 (1989)
6. Wu, J., Chan, C.: Isolated word recognition by neural network models with cross-correlation coefficients for speech dynamics. IEEE Trans. Pattern Anal. Mach. Intell. **15**(11), 1174–1185 (1993)
7. Zahorian, S.A., Zimmer, A.M., Meng, F.: Vowel Classification for Computer based Visual Feedback for Speech Training for the Hearing Impaired, ICSLP, 2002
8. Hu, H., Zahorian, S.A.: Dimensionality reduction methods for HMM phonetic recognition. In: ICASSP (2010)
9. Sak, H., Senior, A., Rao, K., Beaufays, F., Schalkwyk, J.: Google voice search: faster and more accurate. Wayback Machine (2016)
10. Fernandez, S., Graves, A., Hinton, G.: Sequence labelling in structured domains with hierarchical recurrent neural networks. In: Proceedings of IJCAI (2007)
11. Graves, A., Mohamed, A., Schmidhuber, J.: Speech recognition with deep recurrent neural networks. In: ICASSP (2013)
12. Deng, L., Yu, D.: Deep Learning: Methods and Applications. Found. Trends Sig. Process. **7** (3), 197–387 (2014)
13. Yu, D., Deng, L., Dahl, G.: Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning (2010)
14. Dahl, G.E., Yu, D., Deng, L., Acero, A.: Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Sig. Process. **20**(1), 30–42 (2012)
15. Deng, L., Li, J., Huang, J., Yao, K., Yu, D., Seide, F.: Recent advances in deep learning for speech research at microsoft. In: ICASSP (2013)
16. Jurafsky, D., James, H.M.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Stanford University (2018)
17. Graves, A.: Towards end-to-end speech recognition with recurrent neural networks. In: ICML (2014)
18. Yannis, M.A., Brendan, S., Shimon, W.N., de Freitas, N.: LipNet: End-to-End Sentence-level Lipreading. Cornell University (2016)
19. Brendan, S., et al.: Large-Scale Visual Speech Recognition. Cornell University (2018)
20. National Center for Technology Innovation Speech Recognition for Learning (2010). http://www.ldonline.org/article/38655/
21. Follensbee, B., McCloskey-Dale, S.: Speech recognition in schools: an update from the field. In: Technology and Persons with Disabilities Conference (2018)
22. Forgrave, K.E.: Assistive technology: empowering students with disabilities. The Clearing House **7**(3), 122–126 (2002)
23. Apple Inc: Speech framework (2010). https://developer.apple.com/documentation/speech