Jianqing Gao*, Genshun Wan, Kui Wu and Zhonghua Fu

Review of the application of intelligent speech technology in education

https://doi.org/10.1515/jccall-2022-0004
Received June 22, 2021; accepted November 22, 2021; published online April 12, 2022

Abstract: Education is the main application field of intelligent speech technology, including automatic speech recognition and speech assessment. In this paper, we introduce some popular application cases of intelligent speech technology in education, such as intelligent micro-lessons, online distance learning and spoken English learning and testing. With the help of intelligent speech technology, people can acquire knowledge faster, learn languages more efficiently, and even communicate by using different languages. Moreover, we propose three kinds of challenging problems in the application of intelligent speech technology, including the recognition of domain-related terms, the recognition of code-switching and the fine-grained mispronunciation diagnosis. Viable state-of-art solutions to these problems are also discussed in detail. Meanwhile, we also discuss what language teachers and researchers can do in solving the above technical problems.

Keywords: automatic speech recognition; code-switching; computer-assisted language learning; domain-related terms; education; mispronunciation diagnosis

1 Basic framework of this paper

With the continuous development of deep learning and the inexorable logarithmic growth of computing power, intelligent speech technology has developed rapidly in recent years (Arel, Rose, & Karnowski, 2010; Hinton, Osindero, & Teh, 2014). Intelligent speech technologies, including automatic speech recognition (ASR) and speech assessment, have been applied in many industries. The widespread

***Corresponding author: Jianqing Gao**, iFlytek Research, iFlytek Co., Hefei, China, E-mail: jqgao@iflytek.com
Genshun Wan and Kui Wu, iFlytek Research, iFlytek Co., Hefei, China, E-mail: gswan@iflytek.com (G. Wan), kuiwu@iflytek.com (K. Wu)
Zhonghua Fu, Xi'an iFlytek Hyper-brain Information Technology Co., Xi'an, China, E-mail: zhfu2@iflytek.com

Open Access. © 2022 Jianqing Gao et al., published by De Gruyter. This work is licensed under the Creative Commons Attribution 4.0 International License.

use of intelligent speech systems improves both work efficiency and lifestyle changes.

This paper first introduces some application cases of intelligent speech technology in education, such as intelligent micro-lessons, online distance learning and spoken English learning and testing. It then gives a detailed analysis of three challenges in the application of intelligent speech technology and the viable solutions to these problems. Finally, we look ahead at the future direction of intelligent speech technology in education.

1.1 Research and application of speech recognition

ASR technology is one of the most mature and advanced techniques that transcribes speech to the corresponding text. Automatic speech recognition has been studied for nearly 70 years. A speech recognition system for isolated words was first developed at Bell Labs in the 1950s (Davis, Biddulph, & Balashek, 1952). The first specified 10 numbers can be recognised by using the template matching algorithm (Vintsyuk, 1968). In order to break through the limitations of speech length, dynamic time wrapping (DTW) was introduced in the 1960s. Compared with the firmly restricted application of the small vocabulary isolated words recognition system, large vocabulary continuous speech recognition (LVCSR) was considered more practical. Since the 1970s, the statistical system based on an acoustic model and a language model was regarded as the leading solution for ASR systems. The acoustic model is modelled on the Gaussian mixture model (GMM) and hidden Markov model (HMM), while the language model is modelled on the N-gram. However, the limitation of the accuracy of a GMM-HMM-based system keeps it from being widely applied (Ferguson, 1980; Rabiner, 1989). Along with the appearance of deep belief networks (DBN) in 2006, both opportunities and challenges have emerged for automatic speech recognition systems. The relevant research based on deep neural network (DNN) not only provides a new idea to use more contextual information, but also reduces the performance loss caused by the inter-frame independence assumption of the GMM-HMM model.

With the increasing popularity of deep learning, methods based on neural networks, such as feed-forward DNNs (Hinton et al., 2012; Mohamed, Dahl, & Hinton, 2009), recurrent neural networks (RNN) (Hochreiter & Schmidhuber, 1997; Zhang et al., 2016) and convolutional neural networks (CNNs) (Abdel-Hamid, Mohamed, Jiang, & Penn, 2012; Abdel-Hamid et al., 2014) have been proposed in rapid succession. These advanced model structures have significantly improved the performance of ASR systems. However, both acoustic and language models are trained independently, and the connection between the two modules is weak due

to the assumption of conditional independence. Meanwhile, the pronunciation dictionary should be established independently for the different languages or dialects, which affects the popularity of the speech recognition system in different countries.

To solve the existing problems on account of the unreasonable assumption of conditional independence, a hot topic in recent years has been an end-to-end system that can directly convert speech input into the corresponding text without the pronunciation dictionary. End-to-end systems fall into three major categories: connectionist temporal classification (CTC) (Hannun, 2017), recurrent neural network transducer (RNN-T) (Graves, 2012; Sutskever, Vinyals, & Le, 2014) and encoder–decoder based on the attention mechanism (Chan, Jaitly, Le, & Vinyals, 2015). CTC is a widely used and non-autoregressive method, based on strong conditional independent assumptions. It still relies on the additional language model to capture the context information. RNN-T consists of an encoder, a prediction network and a joint network. The joint network combines the other two parts to produce the posterior distribution. The encoder–decoder is the most popular framework among these end-to-end systems. It is based on the attention mechanism and can be divided into three parts: encoder, decoder and the attention module. Compared with the traditional frame-level model, the encoder–decoder system transfers the speech recognition problem into a sequence-to-sequence issue without conditional independence assumption. With the breakthrough of end-to-end systems, the barrier to rapidly building an ASR system was noticeably lowered. Speech input and voice interactive systems based on ASR provide a more convenient and fast intelligent interactive experience to users and have been applied in many fields, including education, finance and medicine. Office products with ASR ability, such as voice recorders and meeting assistant systems, vastly improve office efficiency.

1.2 Research and application of speech assessment

With regard to language learning, the computer-assisted language learning (CALL) system is playing an increasingly important role in providing a more authentic and richer language learning environment for learners. Speech assessment technology is the core module of the CALL system, which refers to using a computer to automatically evaluate the oral proficiency of learners, including pronunciation and oral expression proficiency. Since the 1990s, speech assessment technology has been widely studied and has developed rapidly. According to the chronological order of appearance, speech assessment technology can be divided into two categories: text-dependent and text-independent assessment. The former mainly

refers to pronunciation proficiency evaluation and mispronunciation detection by read aloud tests, such as reading words, sentences and long texts (Neri, Cucchiarini, & Strik, 2002). The latter mainly focus on spontaneous speaking proficiency evaluation by oral tests, such as oral translation, retelling, picture talk and topic expression.

The standard pronunciation acoustic model followed the development of the acoustic model in ASR and forms the basis of pronunciation proficiency evaluation and mispronunciation detection (Li, Chen, Siniscalchi, & Lee, 2017). Recently, some researchers attempted to use end-to-end frameworks for mispronunciation detection (Lo, Weng, Chang, & Chen, 2020). However, the results are not credible due to the small dataset size.

The text-independent speech assessment system is more complicated than the text-dependent assessment. The general approach contains two steps. First, the learners' spontaneous speech is transformed into text by ASR. Then, content-related and pronunciation quality features based on the text are extracted. The content features include content completeness, coherence, lexical resources and grammar range (Chen et al., 2018).

With the ability to automatically score and detect errors for spoken language, speech assessment technology is widely used in large-scale spoken language tests and computer-assisted pronunciation training.

2 Application cases of intelligent speech technology in education

As one of the most mature artificial intelligence technologies, intelligent speech technology has been applied commercially on a large scale. Intelligent speech systems, including the ASR and CALL systems, have been put to wide use in education, healthcare, the judiciary and finance. Especially since the outbreak of COVID-19 in 2020, virtual meetings, call routing systems applied in call centres and online distance learning systems based on intelligent speech technology have been employed in the efforts to prevent the pandemic. This section focuses mainly on three application cases in education.

2.1 Intelligent micro-lesson

Intelligent micro-lessons are currently very popular due to the development of the mobile internet. Intelligent micro-lessons focus on solving one or two problems in

a micro teaching mode with the aid of an app. People use gadgets or a personal computer to watch online videos to acquire knowledge or information. ASR systems can be used to improve the user experience in three aspects when people watch micro-lessons.

First, ASR systems can be used to automatically generate subtitles for these online lectures. The subtitles for videos will help users to better understand the lectures, especially for foreign language courses or lectures that are fast. Second, by transcribing online lectures to text by ASR systems, the makers of micro-lessons can divide the videos into many shorter videos according to the content of the different parts of the videos. These short videos can be distributed and studied more conveniently. Finally, people can search the interesting points of short videos according to the subtitles generated by ASR systems, which greatly improves the search efficiency of short videos.

2.2 Online distance learning

Online distance learning is the integration of multimedia, networks and databases. Compared with traditional offline learning methods, online distance learning has the characteristics of flexibility and availability, regardless of the various needs from different regions at any time. Since online distance learning has the advantages of sharing teaching resources, it is important to give it a much bigger role in education.

Language-related problems are a bottleneck that restricts the development of online distance learning. For many online distance learning resources, it is hard to achieve widespread distribution because of the language issue. Based on existing basic course information and teaching resources in different languages, a cross-language unified system for distance learning needs to be investigated. In order to satisfy more needs from different countries, excellent teaching resources should be provided to the users in an acceptable language. For example, most classes in China are taught in Chinese, and most course materials are provided in Chinese. However, for the students who have not mastered Chinese, it is going to be extremely difficult to extract the key points of the courses. In particular, when the users are expected to discuss certain topics in Chinese, the problem of communication and understanding will be magnified.

To help participants of the course perform better, regardless of the language they speak, a cross-language speech recognition and translation system should be offered to help students connect with the content. Specifically, all the teaching content should first be recognised automatically by ASR as the corresponding text to what the teachers are saying. The recognition results then need to be translated

into the acceptable language for the students by a translation system. Therefore, intelligent speech technology, such as a multilingual speech recognition system and a multilingual translation system, builds this bridge of communication for students from all the world to understand the lesson. At the same time, intelligent speech technology can promote interpersonal communication in class, which will make it possible for a person to speak and write in his or her own language, while the listener will hear and read the message in his or her own language.

2.3 Spoken English test and learning

As mentioned above, speech assessment technology has the ability to automatically score and detect errors in spoken language. It is natural to replace human experts with speech assessment technology in spoken English examinations to improve scoring efficiency. Also, speech assessment technology can play the role of teacher to help learners improve their pronunciation regardless of the time and place.

The traditional spoken English test for entrance examinations has been carried out by way of manual, face-to-face grading or manual grading on the basis of a computer recording, which is not conducive to expanding the examination scale, because scoring by a person is time-consuming and the final scores tend to be affected by the subjective assessment of the scorers. Therefore, before 2011, the college entrance examinations in all provinces of China only had an additional test for spoken English, which was not included in the college entrance examination results. In order to overcome these difficulties, an intelligent speech assessment system was developed to improve scoring efficiency and quality. In 2014, the intelligent speech assessment system developed by IFLYTEK was officially applied to the spoken English test of the Guangdong province college entrance examination (Wang, 2017). More than 600,000 candidates were scored within two days, which greatly reduced the difficulty of spoken English test scoring, making the large-scale spoken English test a reality. At present, with the help of a speech assessment system, an increasing number of spoken English tests are taken in colleges or senior high school entrance examinations.

Besides being used in the spoken English test, speech assessment technology has also been used in spoken English teaching. Although schools have been paying more attention to spoken English teaching, for a long time, the development of spoken English teaching faced many difficulties and could not be carried out effectively. First of all, there are many students who are taught in large classes, and it is difficult for every student to have the opportunity to speak in the classroom. Second, without a spoken teaching and test platform, the teacher cannot

effectively arrange and correct the students' oral homework and organise the spoken test to know students' learning status in time. In the end, without the guidance of the teacher, students cannot improve themselves after class. Based on speech assessment technology, a teaching and test platform can be built to help the teacher efficiently carry out spoken English teaching. Students can improve themselves by receiving feedback from the platform. The intelligent scoring and feedback generation process is automatic, objective and accurate and does not require human participation, which greatly decreases teachers' burden of listening, speaking, teaching and examining and provides a good English learning environment for students.

3 Challenges in the application of intelligent speech technology

As well as having an impact on the learning experience of students all over the world, the widespread adoption of intelligent speech technology has brought great changes to education. However, due to the complexity and uncertainty of application scenarios, some challenges in intelligent speech technology application need to be resolved, including the recognition of domain-related terms, the recognition of code-switching and the fine-grained mispronunciation diagnosis.

3.1 Recognition of domain-related terms

In education, every field or subject has its own vocabulary and technical terms. The recognition of domain-related terms plays an important role in building a fully rounded understanding of the subject. The error recognition of keywords, especially domain-related terms, may generate some confusion among students.

Rich resources of textual data are the foundation for improving the recognition performance of domain-related terms. However, data sparseness is a very common problem in practical application scenes. The domain-related terms are often falsely recognised as common words. More seriously, the audio files of domain-related terms are also necessary during the process of training end-to-end systems.

To improve the recognition performance of domain-related terms, one common solution is to use the interpolation method based on an N-gram language model. Specifically, a language model for the related domain is trained and then combined with the common language model. Domain-related data can be collected from users and the internet. However, due to the various domains and

interdisciplinary influences on the research, a course needs to cover textual data from different fields, which will make the interpolation method time-consuming and labour-intensive. Moreover, the performance of the interpolation method, under the encoder–decoder framework, is greatly reduced. The decoder itself acts as a language model that uses attention to summarise the representations of the encoder to predict the output so that the acoustic and language models are coupled together. When an interpolation model acts on the encoder–decoder model, only a small scaling factor can be used to avoid upsetting the balance.

So, how can the audio sequence be simulated based on the text, or how can the decoder be updated without extra audio files becoming a mainstream research direction? Increasing the probability of domain-related terms during the decoding process is obviously also an important issue to improve the performance of keywords. The main effective methods are described as follows.

1) Fusion method

In recent years, many fusion methods continue to emerge in order to improve the accuracy of keywords in the encoder–decoder framework. In simple terms, the fusion method is an effective way to integrate the ASR model with an external language model. The difference between the different methods is the fusion position and the fusion scheme. For shallow fusion (Bahdanau, Chorowski, Serdyuk, Brakel, & Bengio, 2016), the interpolation method is used to combine the decoding score of the encoder–decoder model and the language model to obtain the final score. The results are then ranked uniformly. Cold fusion (Sriram, Jun, Satheesh, & Coates, 2017) and component fusion (Toshniwal et al., 2018) are two more methods of fusion that are based on gating mechanisms. A gating mechanism can be described as a faucet that controls the flow of information. In order to improve the performance of some particular domain, the language models of related fields need to be retrained along with the encoder–decoder model. The parameters of a gating mechanism are used to determine whether to use the language model during the decoding process. For cold fusion, once the language model needs to be updated, all the modules need to be retrained. On the contrary, for component fusion, the related domain language model is simply used to directly replace the common model without retraining during testing. This is mainly because component fusion realises a decoupling between the main model and the built-in language model, which is built with the training data without extra textual data.

2) Hot words optimisation method: CLAS

Another popular method to improve the accuracy of keywords is the Contextual Listen Attend Spell (CLAS) model. CLAS jointly optimises the encoder–decoder model along with the embedding of the context (Pundak, Sainath, Prabhavalkar,

Kannan, & Zhao, 2018) for which domain-related textual data is not necessary. In the process of decoding, the attention mechanism is used to focus on not only the acoustic feature, but also the additional vocabulary. The fusion of this information can offer a more accurate and more targeted decoding strategy. The extra vocabulary can be determined dynamically by the users without restraint. However, false triggering will become more serious with the expansion of the scale of the extra vocabulary, which means some common words are mistakenly recognised as hot words. Common words and hot words that sound alike often lead to this problem. In order to alleviate the problem of false triggering, the pronunciation of the words can be introduced to enhance the discrimination between the words that sound alike. Similarly, decreasing the vocabulary size is also an effective method. That is to say, most of the candidate hot words can be removed by the attention coefficient, and the remaining candidates are then normalised to sharpen the distribution.

Accurate and reliable data play an important role in the recognition of domain-related terms, so automatic data cleaning is also an important step in data mining. However, confusion words could only be collected from language teachers and researchers. How to convert language skills, experience and deep understanding of different cultures into a cleaning strategy is a subject worthy of study.

3.2 Recognition of code-switching

In education, domain-related terms are often expressed in English because we habitually have trouble in finding the appropriate words and phrases to give expression to our thoughts in our native language. Therefore, code-switching increasingly appears more frequently in our lives. As an effective communicative strategy, code-switching usually refers to using two or more language varieties in one conversation. For a traditional ASR system, the different language recognition systems are mainly independently modelled. The most important question in code-switching is: How to achieve an effective merge with different languages in a single ASR system, and how to gain reasonable text data with a different language in one sentence?

The key to coding-switching for traditional ASR is to combine the two different language acoustic modelling units. For example, English modelling units can be rebuilt according to Chinese pronunciation. Another strategy is partial sharing, which means the same pronunciation part in different languages can be shared while the different pronunciation parts are independently modelled. To solve the problem of sparse resources in Chinese–English code-switching, both the translation and speech synthesis systems can be used in combination to generate more code-switching text. However, the size of the modelling unit in the traditional speech

recognition system often focuses on the phoneme level. The modelling form of coding-switching based on the phoneme level may limit language discrimination. The emergence of encoder–decoder provides new opportunities to improve performance. The selection of the modelling units is freer and more stable. Though the encoder–decoder makes it possible for us to promote and extend the code-switching system, there are still many problems to be solved, such as data sparseness and pronunciation confusion. The main effective methods are described as follows.

1) Generation of Chinese–English code-switching data
The particularity of code-switching text mainly focuses on grammatical forms and structural characteristics, which makes it more difficult to generate the reasonable expression text by using translation. In simple terms, the effectiveness of code-switching text must first cater to language habits. Generative adversarial networks (GAN) can be taken into account to simulate Chinese–English data (Guo et al., 2018). It is mainly used to find the reasonable position of words that can be replaced with another language, according to a real scenario. Related variants of GAN, such as Cycle GAN (Zhu, Park, Isola, & Efros, 2017), guarantee the effectiveness of the method, even without an abundant parallel corpus. Further, speech synthesis based on code-switching text gains the richness of textual data and improves the performance of the encoder–decoder system.

2) Improvement of pronunciation confusion
For code-switching scenes, poor pronunciation is one of the most common problems we are faced with because of phonetic conversion. Therefore, in order to provide the decoding path with more choices, a confusion matrix with common pronunciation confusion pairs can be provided to improve error identification. Specifically, the distance of the similarity between Chinese characters and English sub-words is calculated and selected to express the degree of confusion. Chinese and English path candidates are all supported to avoid complete independence between languages. For example, when the next token is evaluated as Chinese, English candidates are also provided with a confusion matrix to preserve the possible path.

3.3 Fine-grained mispronunciation diagnosis

In pronunciation evaluation, fine-grained mispronunciation diagnosis, that is, pointing out the mispronunciation location and the reason for mispronunciation, can help learners to quickly improve their pronunciation. Currently, although the accuracy of automatic scoring is relatively high, fine-grained mispronunciation diagnosis is not near the level of practical application, with the detected error

number and locations not reaching the results when marked by an expert. Furthermore, the exact cause of the detected error is not always accurately given. These challenges are mainly due to two reasons (Korzekwa et al., 2021). First, it is very difficult to achieve high accuracy for phoneme recognition due to limited contextual information and training data with a noisy label. Thus, phoneme recognition-based mispronunciation detection methods cannot achieve the desired performance. Second, the lack of sufficient real pronunciation error training data further reduces the recognition accuracy of mispronounced phonemes. So far, two potentially effective methods have been described as follows.

1) Generation pronunciation error data through more advanced speech synthesis technology

At present, speech synthesis technology could generate speech with a high quality, and speaker diversity can be customised quickly. After generating texts with a variety of common and unfrequented errors, a large amount of mispronunciation data can be synthesised to make up for the dilemma of insufficient real pronunciation error data. Based on synthetic speech and real speech, the training of the mispronunciation detection model can be improved.

2) Usage of speech recognition pre-trained model

The speech recognition model is trained by a vast volume of training data, and its different hidden layers abstract the pronunciation content at different levels, which can be combined to provide an accurate representation of fine-grained pronunciation. With the pronunciation representation from the speech recognition pre-trained model, the amount of training data for the pronunciation diagnosis model can be reduced. During the training stage of the speech recognition pre-training model, in order to further meet the downstream mispronunciation diagnosis task, the long-context and short-context models can be combined to improve the diversity of the model's representation. The long-context model will accurately recognise correctly pronounced phonemes by taking advantage of long-context information. In contrast, in order to increase sensitivity to mispronunciation, the short-context model only utilises short-context information to recognise mispronounced phonemes.

There is no denying that we can obtain most pronunciation error pairs and effective teaching schemes to correct mispronunciation from those on the front lines in language teaching and study. This prior information can be taken as specified memory units to optimise the speech assessment system and build the self-learning framework. To capture more regular details, we should seek the balancing point of a variety of teaching experiences.

4 Conclusion

This paper mainly focuses primarily on two typical speech technologies: automatic speech recognition and speech assessment. First, three application cases of intelligent speech technology in education are introduced. Second, three challenging problems in the education application are proposed, including the recognition of domain-related terms, the recognition of code-switching and the fine-grained mispronunciation diagnosis. State-of-the-art methods and directions for future development are also discussed.

In addition, a personalised service based on speech recognition, such as the optimisation of accent, is still one of the most important research subjects. More intelligent solutions, such as improving the recording quality of micro-lessons by a microphone array and generating customised learning methods for different students by natural language processing, need to be further researched.

References

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., & Penn, G. (2012). Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4277–4280). Piscataway: IEEE.

Abdel-Hamid, O., Mohamed, A. R., Jiang, H., Deng, L., Penn, G., & Yu, D. (2014). Convolutional neural networks for speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *22*(10), 1533–1545.

Arel, I., Rose, D. C., & Karnowski, T. P. (2010). Deep machine learning – A new Frontier in artificial intelligence research [research Frontier]. *IEEE Computational Intelligence Magazine*, *5*(4), 13–18.

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., & Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4945–4949). Piscataway: IEEE.

Chan, W., Jaitly, N., Le, Q. V., & Vinyals, O. (2015). Listen, attend and spell. *arXiv preprint arXiv: 1508.01211*.

Chen, L., Zechner, K., Yoon, S. Y., Evanini, K., Wang, X. H., Loukina, A., & Gyawali, B. (2018). Automated scoring of nonnative speech using the SpeechRater SM v. 5.0 engine. *ETS Research Report Series 2018*(1), 1–31.

Davis, K., Biddulph, R., & Balashek, S. (1952). Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, *24*(6), 637–642.

Ferguson, J. D. (1980). *Application of hidden Markov models to text and speech*. Princeton, NJ: Princeton University Press.

Graves, A. (2012). Sequence transduction with recurrent neural networks. *Computer Science*, *58*(3), 235–242.

Guo, J., Lu, S., Cai, H., Zhang, W., Yu, Y., & Wang, J. (2018). Long text generation via adversarial training with leaked information. In *Proceedings of the AAAI conference on artificial intelligence*. Palo Alto: AAAI.

Hannun, A. (2017). Sequence modeling with CTC. *Distill*, *2*(11), e8.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., & Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, *29*(6), 82–97.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2014). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Korzekwa, D., Lorenzo-Trueba, J., Zaporowski, S., Calamaro, S., Drugman, T., & Kostek, B. (2021). Mispronunciation detection in non-native (L2) English with uncertainty modeling. In *ICASSP 2021-2021 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 7738–7742). Piscataway: IEEE.

Li, W., Chen, N. F., Siniscalchi, S. M., & Lee, C. H. (2017). Improving mispronunciation detection for non-native learners with multisource information and LSTM-based deep models. In *Interspeech* (pp. 2759–2763). Stockholm, Sweden: Interspeech.

Lo, T. H., Weng, S. Y., Chang, H. J., & Chen, B. (2020). An effective end-to-end modeling approach for mispronunciation detection. *arXiv preprint arXiv:2005.08440*.

Mohamed, A. R., Dahl, G., & Hinton, G. (2009). Deep belief networks for phone recognition. In *Nips workshop on deep learning for speech recognition and related applications*, *1*(9), 39.

Neri, A., Cucchiarini, C., & Strik, H. (2002). Feedback in computer assisted pronunciation training: Technology push or demand pull? In *ICSLP* (pp. 1209–1212), Denver, Colorado: ICSLP.

Pundak, G., Sainath, T. N., Prabhavalkar, R., Kannan, A., & Zhao, D. (2018). Deep context: End-to-end contextual speech recognition. In *2018 IEEE spoken language technology workshop (SLT)* (pp. 418–425). Piscataway: IEEE.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286.

Sriram, A., Jun, H., Satheesh, S., & Coates, A. (2017). Cold fusion: Training seq2seq models together with language models. *arXiv preprint arXiv:1708.06426*.

Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27 (pp. 3104–3112).

Toshniwal, S., Kannan, A., Chiu, C. C., Wu, Y., Sainath, T. N., & Livescu, K. (2018). A comparison of techniques for language model integration in encoder–decoder speech recognition. In *2018 IEEE spoken language technology workshop (SLT)* (pp. 369–375). Piscataway: IEEE.

Vintsyuk, T. K. (1968). Speech discrimination by dynamic programming. *Cybernetics*, *4*(1), 52–57.

Wang, Z. (2017). 人工智能技术在考试中的应用 [Application of artificial intelligence technology in examinations]. *China Examinations*, *307*(11), 30–36.

Zhang, Y., Chen, G., Yu, D., Yao, K., Khudanpur, S., & Glass, J. (2016). Highway long short-term memory rnns for distant speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5755–5759). Piscataway: IEEE.

Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223–2232). Piscataway: IEEE.

Bionotes

Jianqing Gao
iFlytek Research, iFlytek Co., Hefei, China
jqgao@iflytek.com

Jianqing Gao received a D.Eng. degree in electronics and information from the University of Science and Technology of China (USTC). He is the vice dean of IFLYTEK AI Research. His research interests include automatic speech recognition, speech and language information processing and spoken dialogue systems.

Genshun Wan
iFlytek Research, iFlytek Co., Hefei, China
gswan@iflytek.com

Genshun Wan received a B.Eng. degree in communication and information systems from Jiangsu University. He is the director of research of IFLYTEK AI Research. His research interests include automatic speech recognition and speech and language information processing.

Kui Wu
iFlytek Research, iFlytek Co., Hefei, China
kuiwu@iflytek.com

Kui Wu received a master's degree in electronics and information from the University of Science and Technology of China (USTC). He is the senior researcher at IFLYTEK AI Research. His research interests include speech assessment, automatic speech recognition, speech and language information processing.

Zhonghua Fu
Xi'an iFlytek Hyper-brain Information Technology Co., Xi'an, China
zhfu2@iflytek.com

Zhonghua Fu received a PhD degree in computer science from Northwestern Polytechnical University (NPU). He is the vice dean of IFLYTEK AI Research Speech Group. His research interests include audio and speech signal processing.