A Study on the Translation of Spoken English from Speech to Text

Ying Zhang

School of Automotive Engineering, Henan Mechanical & Electrical Vocational College, Zhengzhou, Henan 451191, China
E-mail: zying_zy@hotmail.com

Received 25 September 2024; Accepted 03 December 2024

Abstract

Rapid translation of spoken English is conducive to international communication. This paper briefly introduces a convolutional neural network (CNN) algorithm for converting English speech to text and a long short-term memory (LSTM) algorithm for machine translation of English text. The two algorithms were combined for spoken English translation. Then, simulation experiments were performed by comparing the speech recognition performance among the CNN algorithm, the hidden Markov model, and the back-propagation neural network algorithm and comparing the machine translation performance with the LSTM algorithm and the recurrent neural network algorithm. Moreover, the performance of the spoken English translation algorithms combining different recognition algorithms was compared. The results showed that the CNN speech recognition algorithm, the LSTM machine translation algorithm and the combined spoken English translation algorithm had the best performance and sufficient anti-noise ability. In conclusion, utilizing a CNN for converting English speech to texts and LSTM for machine translation of the converted English text can effectively enhance the performance of translating spoken English.

Keywords: Speech recognition, spoken English, machine translation, speech-to-text.

Journal of ICT Standardization, Vol. 12_4, 429–442.
doi: 10.13052/jicts2245-800X.1244
© 2025 River Publishers

1 Introduction

With the fast pace of globalization and the growing frequency of international interactions, one's proficiency in spoken English has emerged as a crucial metric for assessing competence in international communication [1]. However, due to language differences and the immediacy of spoken English, there are obstacles in communication and understanding of spoken English. In the past few years, the development of artificial intelligence and machine learning technology has led to the emergence of a novel approach for translating spoken English – speech-to-text technology [2]. First, speech-to-text technology is used to transform spoken English into written text, and then the text is translated by machine. This kind of translation method can convert spoken English into written text in real time, which has the characteristics of immediacy and high accuracy [3] and can significantly enhance the efficiency of English communication. Vu et al. [4] proposed a technique that applies named entity recognition (NER) and part-of-speech (POS) tagging for enhancing translation accuracy in Vietnamese sentences. Wang [5] proposed the use of a recurrent neural network (RNN) to recognize speech and employed a connection-time classification algorithm to conduct force alignment between the input speech sequence and the output text sequence. Shimizu et al. [6] proposed cross-language transfer learning for end-to-end speech translation, in which model parameters move from the pre-training phase of speech translation in one language pair to the fine-tuning phase of speech translation in another language pair. Slim and Melouah [7] introduced an incremental transfer learning method for low-resource language translation, which utilizes various relevant corpora and employs an incremental fine-tuning strategy to transfer language features from a grandparent model to a child model. Xiao et al. [8] systematically compared and discussed various non-autoregressive translation models from different perspectives.

The above-mentioned literature has conducted research on machine translation algorithms. Some studies combine POS tagging with machine translation algorithms, some focus on machine translation of speech signals, and others concentrate on fast training methods. In this paper, the research direction is focused on machine translation of English speech. The adopted method first converts English speech into text characters and then translates the English text using a machine translation approach, aiming to reduce the difficulty of translating English speech directly. This article briefly introduces the convolutional neural network (CNN) algorithm for English speech-to-text and the long short-term memory (LSTM) algorithm for machine translation

of English texts. The two algorithms were combined for spoken English translation, followed by simulation experiments. The novelty of this article lies in dividing spoken English translation into two parts. First, a CNN is used in the speech recognition module to convert English speech into text. Then, an LSTM is utilized as an encoder and decoder to translate the converted English text, aiming to reduce the difficulty of English translation. The purpose of this article is to enhance the efficiency and accuracy of translating spoken English.

2 Spoken English Speech Recognition and Translation

Translation of spoken English can take the form of end-to-end, that is, directly convert the audio signal of spoken English into the corresponding translation [9]. Although this method is relatively direct, audio signals need to be divided in order to correspond with the translated characters, and its continuity makes the separation less easy than that of text characters. Moreover, the relevant corpus whose speech corresponds with the translated text is relatively limited, making it difficult to train the algorithm [10]. Another form of spoken translation is to first convert spoken English into English text and then translate the English text to obtain the target text. In this form, the translation of spoken English is divided into two relatively independent steps. A speech recognition algorithm is used to textualize the spoken English, and then a machine translation algorithm is used to convert the spoken English text into the target translation.

2.1 Speech Recognition of Spoken English

In this article, the CNN algorithm is used to recognize spoken English. Compared with other deep learning algorithms, it uses a convolutional structure to obtain local features of the input signal and can combine local features into global features to retain the feature information of the input signal to the greatest extent [11]. In the overall structure of the CNN algorithm, in addition to the conventional input and output layers in deep learning algorithms, there are also a convolution layer and a pooling layer as the core structure. The formula of convolutional operation is:

$$x_l = f(x_{l-1} \otimes k_l + b_l), \tag{1}$$

where x_l and x_{l-1} are the convolution features extracted by the current convolution layer and the previous convolution layer respectively, k_l is the

weight of the current convolution layer during convolution, b_l is the bias of the current convolution layer during convolution, $\otimes$ represents the convolution operation [12], and $f(\ )$ is the activation function. The specific speech-to-text process of spoken English using the CNN algorithm is shown below.

(1) A spoken English audio is input and preprocessed by subframing using the Hamming window [13]. The Hamming window function is:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \tag{2}$$

where $w(n)$ refers to the Hamming window function, n is the audio sampling point, which needs to be within the window length range, and N is the window length [12].

(2) After the audio signal sub-frame processing, the Mel-frequency cepstral coefficient (MFCC) features are extracted from the signal frame.

(3) The extracted MFCC features are input into the CNN algorithm for forward computation. First, the convolutional features are extracted in the convolutional layer using the convolution kernel and Equation (2), and then they are compressed in the pooling layer through the pooling box [14]. The fully connected layer computes the distribution probability of characters through softmax and outputs the characters with the highest probability.

(4) If the CNN algorithm is in the use stage, the speech recognition result can be obtained by the previous step; if it is in the training stage, the output result is compared with the real result, and the weight parameters in the CNN algorithm are reversely adjusted according to the error.

2.2 Machine Translation of Spoken English Recognition Text

After the recognition text of spoken English is obtained by the speech recognition algorithm, the text is machine translated. In this paper, a Sep2Sep model is adopted to carry out machine translation. The Sep2Sep model consists of an encoder and a decoder. The basic principle of text translation is to convert the source text into a string of fixed-length vector encoding through the encoder and then convert the vector encoding into the translation through the decoder [15]. The advantage of this model is that it uses the intermediate vector encoding of fixed length to avoid the problem that the length between the source and target texts is difficult to correspond one by one. Deep learning algorithms are usually used in the encoder and decoder.

For machine translation, a sequence of characters is converted to another sequence of characters, so it is more suitable to use the LSTM algorithm that works based on context. The relevant equation of the LSTM algorithm for forward calculation of the character sequence is:

$$\begin{cases} i_t = f(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \\ C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \\ f_t = f(W_f \cdot [h_{t-1}, x_t] + b_f) \\ o_t = f(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t = o_t \cdot \tanh(C_t) \end{cases}, \tag{3}$$

where i_t is the output of the input gate, h_{t-1} is the status of the previous hidden layer, x_t denotes the current input, b_i, b_C, and b_f are the corresponding biases, $\tilde{C}_t$ is the temporary state of the neural node after inputting x_t, C_t is the updated state of the neural node after inputting x_t, C_{t-1} is the previous neural node state, W_i, W_t, and W_f are the corresponding weights, f_t is the output of the forget gate, o_t is the output of the output door, and h_t is the final output or the next hidden state.

The machine translation process of a spoken English recognition text by the Sep2Sep model is as follows.

(1) A spoken English recognition text is input, followed by vectorization by Word2vec [16].
(2) The text vector is input into the encoder for forward computation. In this paper, in order to make the intermediate vector given by the encoder contain semantic information as much as possible, a bidirectional LSTM (BiLSTM) algorithm is used for forward computation of the text vector. Compared with the traditional LSTM algorithm, the BiLSTM algorithm performs forward calculations on the forwardly input sequence and the reversely input sequence, respectively, using Equation (3). The sum of the obtained forward h_t and reverse h_t is averaged, and the result is the intermediate vector of the output of the encoder.
(3) The intermediate vector is input into the decoder, i.e. the traditional LSTM algorithm, and the forward calculation adopts Equation (3). The calculated hidden state uses the softmax function to calculate the distribution probability of characters in the translation at the fully connected layer, and finally the string with the highest probability is output.

3 Simulation Experiment

3.1 Experimental Data

The experiment was conducted on a laboratory server, and the data used in the experiment included a speech dataset, a English–Chinese parallel corpus dataset, and a self-built dataset. TIMIT (https://catalog.ldc.upenn.edu/LDC93S1) was used as the speech dataset. The sampling parameters of this dataset were 16 kHz and 16 bits. In total, 630 people were involved, and 6300 sentences were included. Each sentence underwent manual classification and marking at the phoneme level. The English–Chinese Parallel Corpus from Tsinghua University (http://thumt.thunlp.org/) was adopted. For the self-built database, 8000 sentences were randomly selected from the parallel corpus, and the speech data of the selected sentences were collected in a recording studio at 16 kHz and 16 bits.

3.2 Experimental Setup

The relevant parameter settings for the speech recognition algorithm and machine translation algorithm are displayed in Table 1. In the speech recognition algorithm, the rectified linear unit (Relu) function was used as the activation function when performing convolution in the convolutional layer.

When the proposed algorithm was tested, the speech recognition algorithm and machine translation algorithm were tested separately, and then they were combined for a comprehensive test. The speech recognition algorithm was compared with the hidden Markov model (HMM) model and the back-propagation neural network (BPNN). The machine translation algorithm was compared with the algorithm adopting RNN as the encoder and decoder. In the comprehensive test of the proposed algorithm, the algorithms used as comparison were only different in the speech recognition part, using BPPNN and HMM model respectively. In addition to testing the pure speech dataset used in this paper, this paper also added white noise to this dataset to test the noise resistance of the speech recognition and translation algorithm. The signal-to-noise ratio (SNR) of the speech signals added with white noise was set to 0, 5, 10, 15, and 20 dB, respectively, to test the translation performance of the algorithm under different SNRs.

3.3 Evaluation Criteria

A word error rate (WER) was employed to assess the performance of the speech recognition algorithm, and a bilingual evaluation understudy (BLUE)

Table 1 Parameter settings of the speech recognition algorithm and machine translation algorithm

	Parameter	Setting	Parameter	Setting
Speech recognition algorithm	MFCC feature dimension	39	Convolution layer 1	64 convolution kernels (2 × 3), a moving step length of 2
	Convolution layer 2	64 convolution kernels (2 × 3), a moving step length of 2	Pooling layer 1	A pooling box (2 × 2), a moving step length of 2, and maximum pooling
	Convolution layer 3	32 convolution kernels (2 × 3), a moving step length of 2	Convolution layer 4	32 convolution kernels (2 × 3), a moving step length of 2
	Pooling layer 2	A pooling box (2 × 2), a moving step length of 2, and maximum pooling	Fully connected layer	Using the softmax function
Machine translation algorithm	The vector dimension of Word2vec	300	The input layer of the encoder	300 nodes
	The forward hidden layer of the encoder	Two layers, 512 nodes per layer, and sigmoid	The reverse hidden layer of the encoder	Two layers, 512 nodes per layer, and sigmoid
	The hidden layer of the decoder	Three layers, 512 nodes per layer, and sigmoid	The output layer of the decoder	Using the softmax function [17]

was used to measure the performance of machine translation. The maximum order of the n-gram grammar is expressed as N, the weight of the n-gram grammar is expressed as ω_n, the phrase proportion of the n-gram grammar is expressed as p_n, and the penalty factor is expressed as B. The equation of BLUE is:

$$BLEU = B \cdot \exp\left(\sum_{n=1}^{N} \omega_n \log p_n\right). \tag{4}$$

3.4 Test Results

First, the speech recognition algorithm was tested and compared with the BPNN and HMM model. The datasets used in the test were the TIMIT dataset

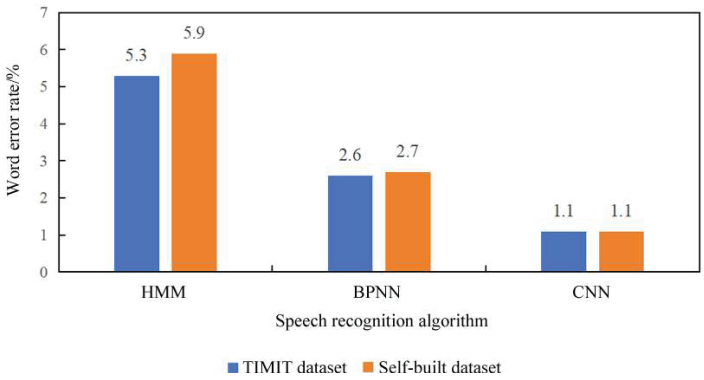

Figure 1 Recognition performance of three speech recognition algorithms.

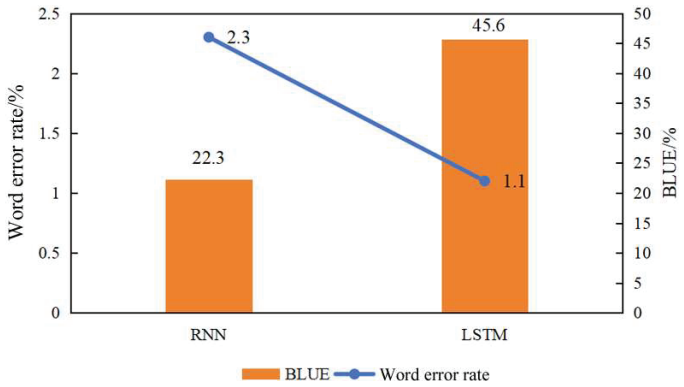

Figure 2 Performance of two machine translation algorithms.

and self-built dataset, respectively. The WER of the speech recognition algorithms are shown in Figure 1. No matter whether it was the TIMIT dataset or the self-built dataset, the HMM model always had the highest WER, followed by the BPNN algorithm, and the CNN algorithm had the lowest rate. In addition, faced with the two datasets, the HMM model had a relatively high WER for speech recognition of the self-built database; the BPNN algorithm had a slightly higher rate for the same database, but the difference was not large. The CNN algorithm almost had no difference in the WER for speech recognition of the two datasets.

Moreover, the performance of the machine translation algorithm was tested and compared with the algorithm that used an RNN as both an encoder and a decoder (Figure 2). The machine translation algorithm based on LSTM had a lower WER and a higher BLUE.

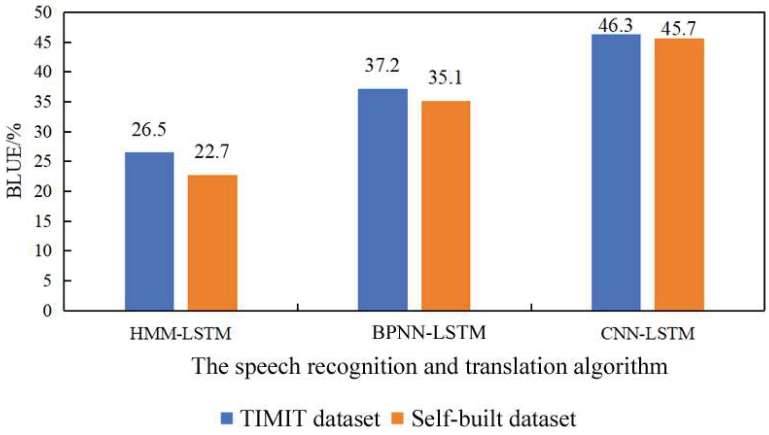

Figure 3 Performance of the spoken English translation algorithm under three different speech recognition algorithms.

The spoken English translation algorithm combining the speech recognition algorithm with the machine translation algorithm was tested, and the performance of this algorithm under different speech recognition algorithms was compared (Figure 3). For both TIMIT dataset and self-built dataset, the performance of the CNN algorithm integrated with the LSTM algorithm was always the best, followed by the BPNN algorithm integrated with the LSTM algorithm, and the HMM speech recognition algorithm combined with the LSTM algorithm was the worst. In addition, under the same English translation algorithm, the translation performance for the self-built dataset was slightly worse, and the CNN-LSTM algorithm had the least difference in performance.

The translation performance of the speech recognition and translation algorithms on English speech with noise was also tested (Figure 4). As the SNR of speech increased, the translation performance of the CNN-LSTM algorithm almost remained stable, while that of the HMM-LSTM and BPNN-LSTM algorithms gradually improved. Under the same SNR, the translation performance of the CNN-LSTM algorithm was the best, followed by the BPNN-LSTM algorithm, and the HMM-LSTM algorithm was the worst.

4 Discussion

With the rapid development of artificial intelligence technology, speech recognition and translation technology has made significant progress and is

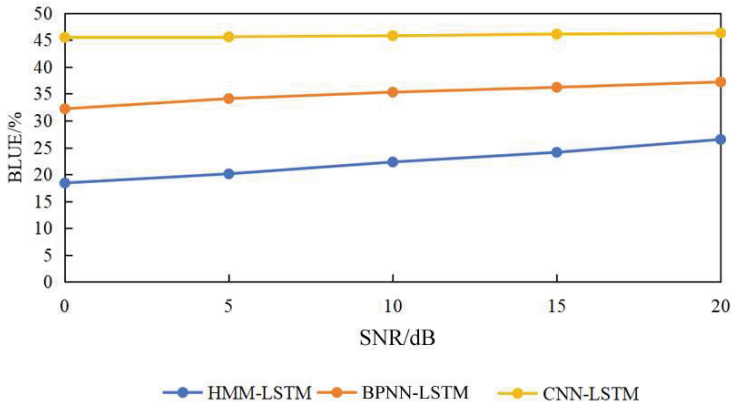

Figure 4 Translation performance of different algorithms for English speech with different SNRs.

gradually being applied in various fields. In today's increasingly globalized world, there is a growing demand for spoken English translation as an important international communication tool. Traditional translation methods rely on manual operation, which not only consumes time and effort but is also inefficient when facing large-scale translation tasks. The emergence of speech-to-text technology provides a new solution for spoken English translation. Through this technology, spoken content can be converted into text in real-time and then translated, thus achieving efficient and accurate cross-language communication. As a natural language processing technology, speech recognition technology can convert human speech signals into text and achieve interchange between speech and text. In spoken English translation, the speech-to-text technology first converts spoken language into text through the speech recognition module and then translates the text using a machine translation system. In the English translation algorithm proposed in this paper, the CNN algorithm was used in the English speech recognition module, while the LSTM algorithm was used in the machine translation module. Then, simulation experiments were conducted on the algorithm for English translation. Firstly, a comparison was made with HMM and BPNN to test the performance of the speech recognition module in converting speech to text. Then, a comparison was made with an RNN to test the translation performance of the text machine translation module. Finally, a comparison was made among three different algorithms in terms of their translation performance. Compared with the HMM and BPNN, the CNN performed better. Compared with the RNN, the LSTM exhibited superior translation

performance. Among the three algorithms for spoken English translation, the CNN-LSTM algorithm had the best overall performance and was least affected by interference.

The reasons causing the above results are analyzed. For the speech recognition module, the HMM models the speech signal through statistical methods. When recognizing speech, it derives the character with the highest probability based on the probabilities given by modeling. The BPNN explores hidden rules in training samples through the activation function in hidden layers, i.e. the mapping rules between speech signals and English characters in this article. As a deep learning algorithm, the CNN is capable of uncovering hidden speech-to-character mapping rules within training samples. In comparison to the BPNN, the CNN could automatically extract local features from speech signals through convolving operations using kernels and then combine them into global features, thus obtaining more comprehensive mapping rules. Regarding text machine translation modules, the LSTM offers advantages over the RNN as an encoder and decoder algorithm because it can handle sequential data as the RNN and address issues like gradient vanishing or exploding during long sequence data processing through gate mechanisms.

5 Conclusions

To further enhance the efficiency and accuracy of spoken English translation, this article combined the CNN algorithm with the LSTM algorithm. Simulation experiments were carried out. In the experiment, the speech recognition performance of the CNN algorithm was compared with that of the BPNN algorithm and the HMM, and the LSTM algorithm was compared with the RNN algorithm in terms of machine translation performance. Moreover, the performance of the spoken English translation algorithm combining different speech recognition algorithms was compared. No matter whether TIMIT or the self-built database was used, the HMM had the highest WER, followed by the BPNN algorithm, the CNN algorithm was the lowest, and the CNN algorithm had almost no difference in the WER of speech recognition for the two datasets. The machine translation algorithm using LSTM had a lower WER and a higher BLUE. No matter whether TIMIT dataset or the self-built dataset was used, the CNN-LSTM algorithm had the best performance, and the performance difference between the two datasets was also the smallest. The CNN-LSTM algorithm could maintain stable performance when translating speech with different SNRs, and it performed better than the other two algorithms.

References

[1] J. Sangeetha, S. Jothilakshmi, 'Speech translation system for English to Dravidian languages', Appl. Intell., vol. 46, no. 3, pp. 1–17, 2016.
[2] A. S. Dhanjal, W. Singh, 'An automatic machine translation system for multi-lingual speech to Indian sign language', Multimed. Tools Appl., vol. 81, no. 3, pp. 4283–4321, 2022.
[3] Y. Wu, Y. Qin, 'Machine translation of English speech: Comparison of multiple algorithms', J. Intell. Syst., vol. 31, no. 1, pp. 159–167, 2022.
[4] V. H. Vu, Q. P. Nguyen, K. H. Nguyen, J. C. Shin, C. Y. Ock, 'Korean-Vietnamese Neural Machine Translation with Named Entity Recognition and Part-of-Speech Tags', IEICE T. Inf. Syst., vol. E103.D, no. 4, pp. 866–873, 2020.
[5] S. Wang, 'Recognition of English speech – using a deep learning algorithm', J. Intell. Syst., vol. 32, no. 1, pp. 225–237, 2023.
[6] S. Shimizu, C. Chu, S. Li, S. Kurohashi, 'Cross-Lingual Transfer Learning for End-to-End Speech Translation', J. Nat. Lang. Process., vol. 29, no. 2, pp. 611–637, 2022.
[7] A. Slim, A. Melouah, "Low Resource Arabic Dialects Transformer Neural Machine Translation Improvement through Incremental Transfer of Shared Linguistic Features', Arab. J. Sci. Eng., vol. 49, no. 9, pp. 12393–12409, 2024.
[8] Y. Xiao, L. Wu, J. Guo, J. Li, M. Zhang, T. Qin, T. Y. Liu, "A Survey on Non-Autoregressive Generation for Neural Machine Translation and Beyond', IEEE T. Pattern Anal., vol. 45, no. 10, pp. 11407–11427, 2023.
[9] S. Matsuda, T. Hayashi, Y. Ashikari, Y. Shiga, H. Kashioka, K. Yasuda, H. Okuma, M. Uchiyama, E. Sumita, H. Kawai, S. Nakamura, 'Development of the "VoiceTra" Multi-Lingual Speech Translation System', IEICE T. Inf. Syst., vol. E100.D, no. 4, pp. 621–632, 2017.
[10] K. Soky, M. Mimura, T. Kawahara, C. Chu, S. Li, C. Ding, S. Sam, 'TriECCC: Trilingual Corpus of the Extraordinary Chambers in the Courts of Cambodia for Speech Recognition and Translation Studies', Int. J. Asian Lang. Process., vol. 31, pp. 2250007:1–2250007:21, 2022.
[11] Y. Zhou, Y. Yuan, X. Shi, 'A multitask co-training framework for improving speech translation by leveraging speech recognition and machine translation tasks', Neural Comput. Appl., vol. 36, no. 15, pp. 8641–8656, 2024.

[12] T. Kano, S. Takamichi, S. Sakti, G. Neubig, T. Toda, S. Nakamura, 'An end-to-end model for cross-lingual transformation of paralinguistic information', Mach. Transl., vol. 2018, no. 2, pp. 1–16, 2018.
[13] H. H. O. Nasereddin, A. A. R. Omari, 'Classification Techniques for Automatic Speech Recognition (ASR) Algorithms used with Real Time Speech Translation', 2017 Computing Conference, vol. 2017, pp. 200–207, 2017.
[14] C. Long, S. Wang, 'Music classroom assistant teaching system based on intelligent speech recognition', J. Intell. Fuzzy Syst., vol. 2021, no. 14, pp. 1–10, 2021.
[15] L. M. Lee, H. H. Le, F. R. Jean, 'Improved hidden Markov model adaptation method for reduced frame rate speech recognition', Electron. Lett., vol. 53, no. 14, pp. 962–964, 2017.
[16] Y. Wang, Y. Lu, 'UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation', Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 2014.
[17] N. Hammami, M. Bedda, F. Nadir, 'The second-order derivatives of MFCC for improving spoken Arabic digits recognition using Tree distributions approximation model and HMMs', International Conference on Communications and Information Technology, vol. 2012, pp. 1–5, 2012.

Biography

Ying Zhang, born in June 1984, graduated from Zhengzhou University with a Master's degree in June 2013. She is working at Henan Mechanical & Electrical Vocational College as a lecturer. She is interested in English education, and British and American literature.