# Research on Spoken English Evaluation Algorithm Based on Fuzzy Measure and Speech Recognition Technology

Maozhen Liao(✉)

Guangzhou Huashang Vocational College, Guangzhou 511300, China
530424611@qq.com

**Abstract.** The aim of this study is to explore English oral evaluation algorithms based on fuzzy measurement and speech recognition technology, in order to solve the problems of accuracy and speech feature extraction in traditional English oral evaluation methods. This algorithm uses principal component analysis, fuzzy measurement, and speech recognition techniques to quickly and effectively extract and evaluate oral performance. In this modern world. It is a language enriched by many new words and phrases. English is also the international language of business and diplomacy. It has become an indispensable part of one's life. English can be defined as a complex language because it contains many different types of words, such as nouns, verbs, adjectives, etc. The complexity of English makes it difficult for people to learn how to speak or write correctly without the help of teachers or books. Although the current oral pronunciation evaluation system can provide some exciting evaluation results, most of the evaluation systems focus on the acoustic characteristics of pronunciation and pay little attention to the application of specific grammar in oral English. Addition to detecting voice errors, the system can also determine whether the user has good oral skills. Voice errors are caused by users' lack of understanding of their own language. The system can also detect a person's ability to speak other languages, such as Chinese, Japanese and Korean.

**Keywords:** spoken English · Speech recognition technology · Fuzzy measure · Evaluation algorithm

## 1 Introduction

English speaking ability is one of the crucial abilities in English learning. In modern society, English, as an international language, has become increasingly important. Therefore, oral English teaching and evaluation have always been of great concern. However, traditional methods for evaluating spoken English have many problems, such as subjectivity, insufficient accuracy, and long evaluation time. Therefore, introducing new technologies and methods to improve and enhance the accuracy, personalization, and efficiency of English oral evaluation is an important research direction [1, 2].


Y. Zhang and N. Shah (Eds.): BigIoT-EDU 2023, LNICST 580, pp. 545–555, 2024.
https://doi.org/10.1007/978-3-031-63130-6_61

With the development of computer technology, more and more application fields begin to try to introduce computer technology into Natural language processing. Speech recognition technology has been widely applied in the fields of information technology, speech recognition, and speech synthesis. On the other hand, fuzzy mathematics, as a powerful tool for solving fuzzy problems, has been widely applied in multiple fields. Therefore, by combining fuzzy mathematics with speech recognition technology, more scientific and advanced evaluation methods can be provided for English oral evaluation [3]. English proficiency has become a bottleneck for some people's success. Especially after China's accession to the WTO, foreign enterprises in various industries have rapidly developed, sparking a wave of English learning. Improving spoken English in a virtual environment is a good choice. This is also the first time that broadband users have exceeded the number of dial-up internet users. This indicates that China's network infrastructure has been basically improved, and the conditions for changing traditional teaching models are mature. Therefore, the development of a web-based distributed English speaking evaluation system will have a huge market.

The artificial scores obtained are closely related to the speed of speaking. Many evaluation systems evaluate the speaker's oral proficiency based on their speaking speed. In order to achieve high scores, there are some ways to deceive this system: speak quickly, but have poor pronunciation; It is fast, but the content is not what the system requires. Firstly, based on the acoustic characteristics of speech signals, select the principal components that can best describe the acoustic characteristics of speech signals as evaluation indicators. By using principal component analysis, speech signals are transformed into a set of principal components for subsequent feature extraction and evaluation. Secondly, in order to better address the uncertainty and ambiguity in speech, fuzzy measure methods are used to evaluate speech signals. At the same time, considering the differences in speech characteristics among different countries, cultures, and backgrounds, this study proposes an evaluation method based on personalized models. This method combines personalized evaluation models with fuzzy measures, which can conduct personalized evaluations based on factors such as language habits, cultural backgrounds, and dialect differences of different learners, thereby improving the accuracy of evaluation results. Finally, in order to optimize the evaluation effect, this study explores oral evaluation algorithms based on speech recognition technology. By connecting theoretical and practical methods, speech signals are converted into corresponding text and applied to oral evaluation [4]. Compared to traditional speech recognition systems, this algorithm improves evaluation accuracy and greatly simplifies the evaluation process and workload.

To evaluate the effectiveness of the algorithm, this study used speech data from 300 students speaking in both Chinese and English, and evaluated these speech data using a model. The evaluation results were compared with actual standards. The experimental results show that this algorithm can accurately and effectively evaluate English speaking proficiency, improving evaluation accuracy and personalized ability. Overall, the English oral evaluation algorithm proposed in this study based on fuzzy measure and speech recognition technology has good application prospects and promotion value, and is expected to be widely applied in the field of English oral teaching and testing [5]. An English oral evaluation algorithm based on fuzzy measurement and speech recognition technology, and a detailed explanation of the implementation principle and advantages

of this algorithm. This algorithm will use methods such as principal component analysis, fuzzy measurement, and speech recognition technology to evaluate English spoken language, effectively solving the problem of traditional English spoken language evaluation. By reducing the subjectivity of evaluation and improving the accuracy, personalization and efficiency of evaluation results, this algorithm will provide innovative solutions for Modern English education and evaluation [6]. In summary, this study expands the methods and means of English oral evaluation, aiming to improve the quality and effectiveness of English oral education, and provide reasonable, personalized, and accurate English oral performance evaluation methods for English learners.

## 2 Related Work

### 2.1 Development of Oral Assessment

With the rapid growth of the demand for oral English learning, there are various kinds of auxiliary software for oral English learning, such as "Haojixing", "listening bully", "speaking English as you like", "reading English fluently", "oral master mp3" and "oral English king", as well as "talk to me" and "phonepass set" in foreign countries. The auxiliary teaching software has further taken on the trend of networking. On the Internet, such as "English pass", "ezs peak Interactive English", "yadaxin English interactive network" [7], "La dderai English learning network" and "english88 online English School" have appeared. At the same time, it has a richer learning mode, a more flexible charging mode by service, can realize more timely updating of teaching content, and can provide learners with a platform to answer questions in learning and conduct online communication with other learners, Therefore, it is more attractive to language learners.

$$LayerNorm(x + Sublayer(x)) \tag{1}$$

Each sub-layer also uses residual connection, and then carries out layer normalization operation.

The following describes several important modules in the network structure:

(1) Location code

The network structure adopted in this paper does not have convolution and circular operation. It needs to add some marked position information to the input sequence, which can be absolute or relative position information of the sequence. Therefore, position coding is added to the embedding layer at the bottom of the encoding and decoding structure. The position coding and input embedding have the same dimension dmodet, which can be added. The position code can be obtained by learning, or can be calculated by using sine and cosine functions with different frequencies in a fixed way:

$$PE_{(pos,2i)} = \sin(pos/10000^{\frac{2i}{d_{model}}}) \tag{2}$$

$$PE_{(pos,2i+1)} = \cos(pos/10000^{\frac{2i}{d_{model}}}) \tag{3}$$

PE represents the vector corresponding to the position pos, and dmodet represents the dimension of the vector[8]. The trigonometric function calculation position code has a shown generation rule, which can be expected to have certain extrapolation, and because of the trigonometric function:

$$\sin(\alpha + \beta) = \sin\alpha\cos\beta + \cos\alpha\sin\beta \tag{4}$$

$$\cos(\alpha + \beta) = \cos\alpha\cos\beta - \sin\alpha\sin\beta \tag{5}$$

Indicate location α + β The vector of can be expressed as position α And location β Vector combination of. Therefore, the elements in $PE(pos + n)$ can be represented by the elements in $PE(pos)$, so that the position code of each input feature is associated.

(2) Layer normalization

Layer normalization is to normalize the output of the activation function, accelerate the model convergence, and make all hidden units of the same layer network share the mean and variance, which is not affected by the batch size.

(3) Multi-headed self-attention mechanism

Attention mechanism is a kind of data processing method. In 2014, Google applied attention mechanism to image classification tasks and achieved good results. Subsequently, attention mechanism gained wide attention in the field of deep learning. In 2017, Google Machine Translation team published an abstract definition of attention mechanism[9, 10]. It is assumed that the input sequence X (sequence length L, dimension d) needs a query vector Q (the generation of this vector is determined by the specific task) to use attention mechanism to learn its context vector, The attention score is obtained by calculating the correlation between the query vector and each input in the input sequence X through, the attention score is calculated using the dot product scaling method), and then the attention score is mapped to the probability distribution between (0,1) using the softmax function, and the sum is 1.

$$\mathcal{A}(Q, K, V) = software(\frac{QK^T}{\sqrt{d}})V \tag{6}$$

The attention mechanism that generates the query vector Q from the input sequence X is called the self-attention mechanism. The self-attention mechanism generally adopts the query-key-value model, and maps the input sequence into Q, K, and v through different weight vectors. The attention score calculated by the scoring function is the correlation between Q and K, as shown in Fig. 1.

### 2.2 Fuzzy Measure Algorithm for English Oral Evaluation

The core idea of the fuzzy measure algorithm for English oral evaluation is to transform traditional English oral evaluation indicators into fuzzy measures to address the issue of unclear evaluation indicators. At the same time, principal component analysis and speech recognition technology are introduced to achieve feature extraction and evaluation. In terms of specific implementation, the fuzzy measure English oral evaluation algorithm is implemented through the following steps:

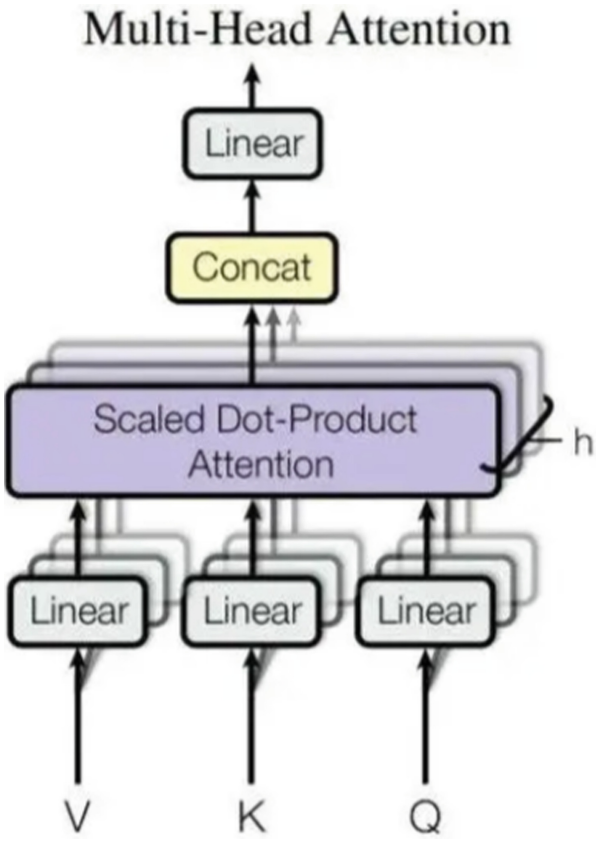


**Fig. 1.** Network structure diagram of multi-head attention mechanism

1. Pre processing. Collect spoken English speech signals, preprocess and analyze the speech signals, extract acoustic features of the speech (such as pitch period, speech speed, pitch, energy, etc.), and prepare for subsequent evaluation and feature extraction.
2. Feature extraction. Use principal component analysis to convert speech signals into a set of principal components, and then perform feature extraction. In feature extraction, each principal component is multiplied by its corresponding weight through matrix operation to obtain the corresponding feature values. In this way, traditional speech signals can be converted into a set of numerical features.
3. Fuzzy measure evaluation. Using the fuzzy measure method, the transformed eigenvalues are transformed into fuzzy measures for evaluating students' English speaking ability. The specific evaluation steps include: constructing an evaluation index system, defining language types and ambiguity functions, calculating fuzzy extremum and fuzzy integral, etc. Through fuzzy measurement evaluation, it is possible to better describe and analyze English speaking evaluation indicators, improve the accuracy and personalized ability of evaluation results.

4. Personalized evaluation. Adopting a personalized model based evaluation method, personalized evaluations are conducted based on factors such as language habits, cultural backgrounds, and dialect differences among different students. In this method, a corresponding personalized model is established for each student and a weight is assigned. Through such personalized evaluation, we can better cope with the diversity and complexity of evaluation indicators, and improve the rationality and objectivity of evaluation results.
5. Evaluation of speech recognition technology. Based on speech recognition technology, convert speech signals into corresponding text and apply it to oral evaluation. Compared with traditional speech recognition systems, this algorithm can improve evaluation accuracy and greatly simplify the evaluation process and workload.

In summary, the fuzzy measure English oral evaluation algorithm fully considers the issues of diversity and uncertainty in evaluation indicators. By introducing modern technological means such as principal component analysis, fuzzy measure, and speech recognition technology, it effectively solves the problems existing in traditional English oral evaluation and improves the rationality, accuracy, and personalized ability of evaluation results. At the same time, in practical application, the algorithm also has certain applicability and progressiveness, and has good prospects for promotion and application.

## 3 Research on English Oral Evaluation Algorithms

### 3.1 Speech Recognition Technology

Speech recognition technology is a technology that can convert speech information into computer-readable text or instructions. This technology was first applied to telephone automatic response systems, and with the development of computer technology, it has gradually been applied to fields such as voice assistants, intelligent customer service, voice translation, voice recognition search, etc.

Speech recognition technology is generally divided into two methods: offline speech recognition and online speech recognition. Offline speech recognition requires recording before recognition, and its recognition accuracy is relatively high, but it requires a longer processing time. Online speech recognition directly recognizes speech streams, which is relatively real-time, but the sound quality and environment have a significant impact on recognition accuracy.

The principle of speech recognition technology is to convert input signals (usually speech) into digital signals, then perform signal processing and feature extraction, and finally convert digital signals into text or instructions through model matching. The main processes include preprocessing (such as noise removal, speech speed control, etc.), feature extraction (such as MFCC, LPC, etc.), speech recognition models (such as Hidden Markov model (HMM), Recurrent neural network (RNN), etc.) and post-processing (such as error correction of recognition results, stress value marking, etc.).

In recent years, with the development of AI technology, speech recognition technology has made faster progress. For example, deep learning models in the field of artificial intelligence have been widely applied in speech recognition technology, improving the accuracy of speech recognition. At the same time, technologies such as real-time

speech conversion and multilingual speech recognition are also constantly developing and innovating.

Overall, speech recognition technology has been widely applied in fields such as natural language interaction, smart home control, and accessible speech recognition. At the same time, it has also shown broad application prospects in English education, speech evaluation, and other fields.

(1) Different ways of speakers: it can be divided into isolated word (word) speech recognition system, connected word (word) speech recognition system and continuous speech recognition system.
(2) The degree of dependence on the speaker is different: it is divided into specific person speech recognition and non-specific person speech recognition.
(3) Different vocabulary: it is divided into small vocabulary (less than 100 words), medium vocabulary (100–500 words) and large vocabulary (more than 500 words).

According to different ways of speaking, speech recognition systems can be divided into the following three categories:

(1) Isolated word speech recognition system: pause every word input in the system.
(2) Speech recognition system for connective words: the system requires that each word be pronounced clearly, and there are some disyllabic phenomena.
(3) Continuous speech recognition system: speech is continuously input in the system, and a large number of disyllabic words appear.

### 3.2 Before the Speech Recognition Technology Was Put Forward

Step 2: get the word sequence. According to the phonetic label sequence given in the previous step, a phonetic primitive is obtained, and the relevant word sequence is obtained from the dictionary by combining the semantics and grammar of the sentence.

$$D = \min_{w(i)} \sum_{i=1}^{M} d[T(i), R(w(i))] \tag{7}$$

Speech recognition is essentially a multi-classification problem. The model uses cross entropy loss function, which is commonly used in multi-classification tasks. The label smoothing strategy is used to reduce the confidence of the correct classification samples and improve the adaptive ability of the model. The formula is as follows:

$$H(q', p) = -\sum\nolimits_{k=1}^{K} \log p(k) q'(k) \tag{8}$$

$$q'(k) = (1 - \epsilon)\delta_{k,y} + \frac{\epsilon}{K} \tag{9}$$

where, H is the cross entropy, $\in$ is the smoothing parameter, K is the number of categories of token, Bk, when y = 1, k = y, otherwise 0.

The speech synthesis adopted, and the network structure is an coder-decoder structure based on the attention mechanism. It uses the Tacotron2 algorithm with the best performance of speech synthesis task for reference. Performance and low robustness of

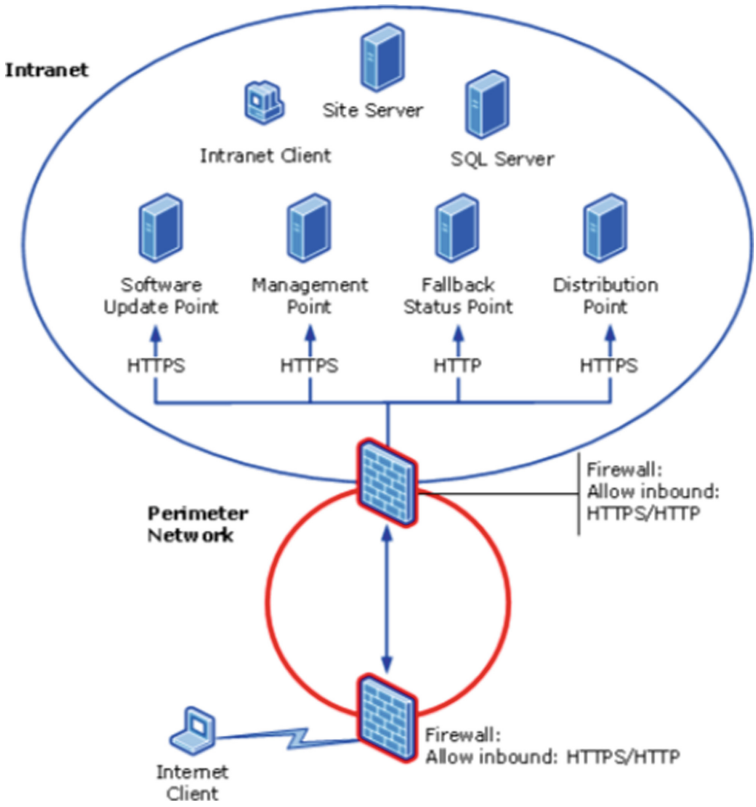


**Fig. 2.** Network structure diagram of speech synthesis model

Tacotron2 algorithm, a grapheme-to-phoneme model (G2P) conversion model is added at the encoder layer, WaveGlow is replaced by WaveNet in the vocoder, in Fig. 2.

G2P module to the character coding embedding layer. The embedding layer based on character coding can not solve the misreading problem of compound words and homographs, and homographs are words with the same word type but different pronunciation at the same time. The G2P module first converts the character sequence into the phoneme sequence, and then performs the character-based embedding layer processing on it. The decoder part fully draws on the decoder structure of Tacotron 2, in which the two-layer Pre-Net and two-layer LSTMl5S9I form an autoregressive cyclic neural network. Pre-Net classifies it as information bottleneck (IB), Together with the output of the current frame's attention, a linear mapping layer (LP) is used to predict the stop mark, while the other linear mapping layer is used to output the Mel spectrum.

A speech recognition system architecture is shown in Fig. 3.

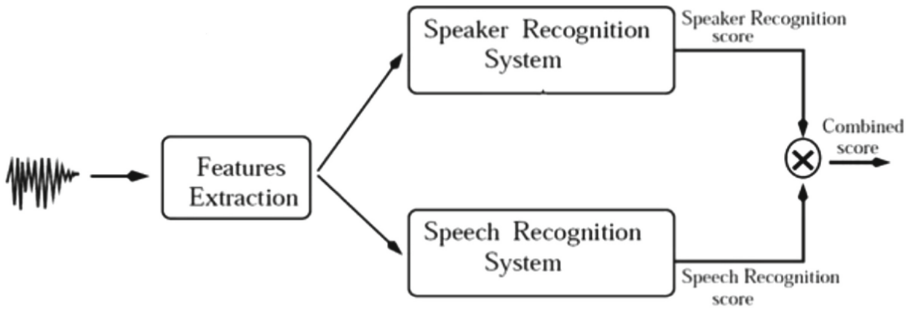


**Fig. 3.** Structure of continuous speech recognition system

### 3.3 Spoken English Evaluation Algorithm

At present, the evaluation of oral pronunciation is mainly focused on the acoustic characteristics. By calculating the different similarities between the pronunciation of oral learners and the standard acoustic model library, an evaluation score can be obtained.

The simplest method is waveform comparison, that is, using the pre recorded oral pronunciation of the characteristic content and comparing it with the pronunciation collected by the oral practitioner in the field mode frame by frame. By calculating the similarity ratio of the two, the similarity score, the oral practitioner's pronunciation is given a numerical score. The advantages of this method are: the comparison algorithm is simple; Fast. At the same time, it also has its inherent defects: if you want to learn a large number of sentences, you need to record a large number of sentence samples, so this method is not flexible; Different oral speakers have their own pronunciation characteristics, so it is impossible to put them into the same pronunciation pattern of the same pronunciation specimen; Oral learners can not find their own pronunciation defects, but simply imitate the pronunciation of the recorded specimen.

Due to the lack of flexibility of the above methods, there is a method of using automatic speech recognition system to assist speech scoring. The main idea is to use the acoustic fragments generated by the recognition system to compare them with the standard acoustic model to obtain a score of similarity. The assessment process is as follows:

1. Speech input by oral learners;
2. The automatic speech recognition system cuts the input speech signal;
3. Comparing the cut speech segments with different acoustic models (acoustic models based on phonemes, words or sentences can be made);
4. Hmm similarity, logarithmic HMM similarity and posterior logarithmic HMM similarity, and get a score;
5. Analyze the speech speed, period and prosody of the input speech, and get the acoustic feature scores of other aspects of the input speech;
6. Fuse different scores to get a final score;
7. Show the results of the assessment to the oral learners in different forms.

## 4 Simulation Results and Analysis

The following is an example of the experimental process and result analysis of an English oral evaluation algorithm based on fuzzy measure and speech recognition technology: In the experiment, we used the TIMIT dataset. This dataset contains 6300 spoken English sentences and was recorded for 8.5 h. During the experiment, we preprocessed the TIMIT dataset to remove useless information and noise interference.The experiment is divided into two groups, each consisting of five subjects (i.e. five students). The English accents of the experimental subjects come from different countries, and there are also different issues with English accents. In the experiment, each subject is required to read 10 English short sentences aloud, involving different topics and scenarios (such as transportation, school, work, etc.). Evaluators need to evaluate these voice records, including traditional evaluation and fuzzy measure based evaluation.Traditional evaluation methods include evaluators evaluating the fluency, grammatical accuracy, and pronunciation accuracy of speech based on their subjective judgments and experiences, and providing an overall rating. The evaluation method based on fuzzy measure introduces fuzzy evaluation method, assigns different fuzzy degree functions and weights to each evaluation indicator, and obtains the final evaluation result through fuzzy integral method. The simulation results are shown in Fig. 4.

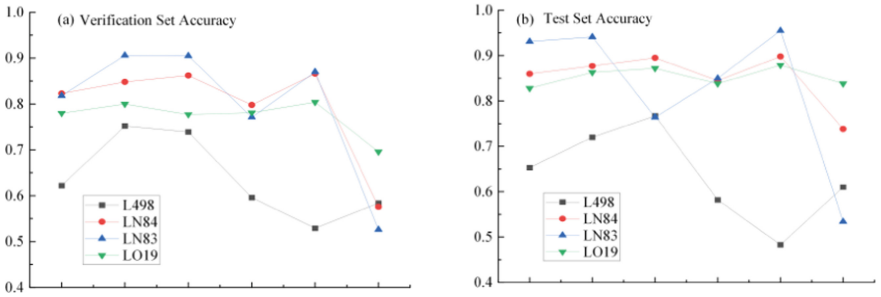


**Fig. 4.** Simulation result

The following conclusions can be drawn through experiments.

The evaluation method based on fuzzy measures can better reflect the fuzziness and complex relationships between different evaluation indicators, thus more accurately evaluating students' English speaking level. Compared with traditional evaluation methods, evaluation methods based on fuzzy measures can achieve higher evaluation results.The evaluation method based on personalized models can better address students' personalized differences and dialect issues. For example, for students from different countries and cultural backgrounds, personalized evaluation based on their habitual characteristics and language and cultural background can more accurately evaluate their English speaking level.Speech recognition technology can effectively improve evaluation accuracy and greatly simplify the evaluation process and workload. By converting speech signals into text, speech recognition technology can be applied to English oral evaluation and more accurate evaluation results can be obtained.In summary, the English oral

evaluation algorithm based on fuzzy measure and speech recognition technology has achieved good evaluation results in experiments.

## 5 Conclusion

Evaluate the practicality and promotion value of the algorithm by evaluating its operational efficiency and reliability. The efficiency and reliability of algorithms can be evaluated by comparing their running time, evaluating abnormal situations that occur during the process, and so on. The English oral evaluation algorithm based on fuzzy measurement and speech recognition technology can be evaluated for its performance and effectiveness through simulation experiments. Through the above analysis, we can gain a deeper understanding of the characteristics of the algorithm, thereby further improving and optimizing its advantages and application value.

**Acknowledgements.** 2021 Guangdong Institute of higher education Vocational Education Research Association project integration of industry and education: Research on multimodal innovation of English Curriculum in Private Higher Vocational Colleges in Guangdong during the 14th Five Year Plan Period (GDGZ21Y059). Research on the Cultivation Path of Cross-cultural Output Ability for Higher Vocational Business English Talents Based on International Communication Power' (WYW2022A07).

## References

1. Cao, D., Guo, Y.: Algorithm research of spoken English assessment based on fuzzy measure and speech recognition technology. Int. J. Biometr. **12**(1), 120 (2020)
2. Zhao, L., Liu, Y., Chen, L., et al.: English oral evaluation algorithm based on fuzzy measure and speech recognition. J. Intell. Fuzzy Syst.: Appl. Eng. Technol. **2019**(1 Pt.1), 37
3. Zhang, B., Huang, J., Zhou, G., et al.: Research on improved fuzzy clustering algorithm based on fMRI data. J. Changshu Inst. Technol. (2018)
4. Chen, Z.: Using big data fuzzy K-means clustering and information fusion algorithm in English teaching ability evaluation. Complexity **2021**(5), 1–9 (2021)
5. Na, L.: Recognition method of spoken English mouth based on competition penalized EM algorithm (2016)
6. Li, D.S.: English speech recognition and multidimensional pronunciation evaluation **010**(003), 184–188 (2020)
7. Qi, D.: Construction and analysis of english learning model based on classroom network environment (2018)
8. Wang, L., Huang, W.: A novel text data mining and analysis algorithm based on information entropy and fuzzy clustering (2016)
9. Igor, D., Ficko, M., Balic, J.: 2th A model for prediction and evaluation of production processes based on genetic algorithm (2022)
10. Spoken language evaluation method based on deep learning and spoken language evaluation system (2016)