Research Article

An Automatic Assessment Method for Spoken English Based on Multimodal Feature Fusion

Qijing Zhang

Department of Basic Science, Yangzhou Polytechnic Institute, Yangzhou 225127, China

Correspondence should be addressed to Qijing Zhang; ypieli@126.com

Received 18 September 2021; Revised 4 October 2021; Accepted 5 October 2021; Published 13 November 2021

Academic Editor: Xin Ning

Copyright © 2021 Qijing Zhang. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

This paper provides a comprehensive explanation of the theoretical foundations of multimodal discourse analysis theory as applied to speaking instructional design. The specific application of multimodal theory in the teaching of elementary English speaking classrooms is explored through the teaching design of elementary English speaking classrooms, the teaching implementation of multimodal teaching design is carried out, and the effect of the teaching practice of elementary English speaking guided by multimodal discourse analysis theory is comprehensively evaluated through classroom observation method, questionnaire survey method, and interview method, combined with the teaching evaluation and teaching implementation effect, which is the multimodal teaching design. The paper also summarizes the findings and shortcomings of the study. Through the teaching design and implementation, the advantages of multimodal teaching are obvious; it can combine with modern advanced teaching techniques to create more realistic communicative situations in the classroom, gather and present various modal resources and information, and ensure rich and diverse language input; students can receive various sensory stimuli in the classroom, deepen their memory and experience of language, increase the interest of classroom teaching, and improve students' participation. It also increases the interest of the classroom and enhances students' participation and motivation. Based on multimodal theory, the author designed a multimodal teaching framework for a semester-long speaking course in the speaking classroom for reference. The fuzzy measures were constructed based on subsets of language segments containing 10 phonemes belonging to the same HDP set. Finally, linguistic scores are given by the Surgeon integral model based on the plausibility of the system and the fuzzy measures. The experimental results based on Sphinx-4 show that the evaluation model yields plausible and stable evaluation results for the 3 test sets at an average correct recognition rate of 84.7% of phonemes.

1. Introduction

In many real-life scenarios, there is a need to assess the speaker's oral expression ability, such as Mandarin exams, oral training, language teaching evaluation, and broadcasting exams. At present, these scenarios are still evaluated by manual scoring in the form of multiple averaging, voting, or one vote, which are too subjective, are often lacking fairness, and cannot give objective feedback to the speaker; in scenarios like Mandarin exams, where the scale of candidates is huge, manual scoring is a huge workload, time-consuming, and costly for on-site scorers. For language learners, many hidden problems in the learning process are often not detected in time, thus affecting the efficiency of language

learning [1]. Give speakers a linguistic score, not a specific numerical evaluation. Through this method, the problem of evaluating specific grammatical phenomena in spoken pronunciation and the pronunciation characteristics of specific groups of people has been successfully solved. The current automatic speech assessment research has many shortcomings, often only referring to the speech level information to assess the speaker; such an assessment method does not involve the semantic, grammatical, and other text-related content and does not reflect the full information of the speaker's verbal expression; Automatic Speech Recognition (ASR) technology is now relatively mature and has been used in many applications such as instant messaging and input methods. Automatic speech assessment studies can combine

text data with speech data to assess the speaker's oral expression ability with the help of ASR technology. In many speech assessment scenarios, scorers often just give back an overall score to the speaker, and such a score has little meaning to the speaker [2]. In addition, most of the existing speech assessment methods still use feature engineering or simply neural networks for modeling, with a tedious feature extraction process and poor model generalization ability. The study of spoken English evaluation algorithms based on key technologies such as speech recognition and natural language processing provides the ability to interact with the user's spoken language, automatic evaluation of the user's spoken language, and the ability to allow the user to automatically discover the crux of his or her lack of fluency [3]. The implementation of this feature will greatly enhance the efficiency of English-speaking learning, promote the maturity of Computer Assisted Language Learning (CALL) technology, solve the current problem of insufficient English-speaking teachers, and promote the development of natural language processing which is important.

The system not only assists in oral training and helps language learners to learn anytime and anywhere but also assists scorers in automatic scoring. This approach not only gives an objective and fair assessment and ensures relative fairness among participants but also saves a lot of costs and improves overall efficiency. In response to the shortcomings of existing speech assessment methods, this paper will propose a series of optimized methods [4]. For the complex speech assessment data in many speech assessment scenarios, this paper designs a set of standardized and effective speech data preprocessing processes to improve the quality of the input data of the subsequent model; for the shortcomings of the existing automatic speech assessment models to go for model innovation, this paper is no longer limited to the traditional feature engineering and machine learning methods and adopts the latest temporal deep learning methods [5–7] for modeling, which combines the textual content of the speaker with speech information which is combined for multimodal learning, and the assessment criterion is no longer an overall score but a multidimensional assessment; for multiparameter speech assessment scenarios, the model can also be optimized by multitask learning at a later stage [8]. The main mode refers to the main mode used by the teacher in the classroom teaching process. It often assumes the mission of conveying key information in the classroom discourse communication, and it is also the focus of the teacher's hope that the students learn. These methodological improvements will further promote the development of the field of automatic speech assessment research and better solve the problems that exist, with certain practical significance and social value.

The evaluation of the part of spoken English that is hyphenated is motivated not only by the importance of the hyphenation skill in spoken English and its ability to improve the understanding of spoken English but also by the need to address the convergence of natural language arbitrariness and the instability of existing speech recognition systems. The first step is to determine which parts of a sentence can be read together by building a library of rules

for reading together and then submit the grammar annotated by reading together to the Sphinx-4 automatic speech recognition system, which uses its improved grammar builder to build a search grid for the input grammar annotated by reading together. At this point, the recognition results are evaluated using the model constructed in this paper for consecutive reading evaluation. First, the randomness of natural speech pronunciation and the instability of the speech processing system are modeled by fuzzy measures [9, 10] and plausibility, and then, they are fused into the Surgeon integral framework to evaluate the linguistic score instead of specific scores. The next step is to assess the phonemes in spoken English [11] that are easily confused by Chinese learners of English. The first task is to collect all the confusion-prone sounds of Chinese learners of English, classify these sounds into different confusion-prone sets, and build a rule base of confusion-prone sounds from this to dynamically extend the confusion-prone sounds for input utterances. Then, the grammar builder of Sphinx-4 is modified to dynamically accept grammars that have been extended with confusable sounds. The next step is to use Sphinx-4 to recognize the input speech signal and analyze the results using the easy-mix evaluation algorithm. Again, for the easy-mix evaluation algorithm, the Surgeon integral was introduced to obtain a linguistic score.

2. Current Status of Research

Traditional ASR systems based on auditory modality have been successfully applied to some relatively controllable tasks, such as dictation tasks and moderate vocabulary tasks [12]. However, since most current ASR systems are obtained in quiet laboratory environments or trained in some specific environments, while real-life obtained speech is often distorted by various factors such as different channels, different noises, far fields, and narrow bands, these ASR systems often fail to meet the needs of practical applications [13]. In recent decades, many researchers have devoted themselves to exploring a wide variety of methods to combat the interference from various external factors in real life to improve the stability and robustness of ASR systems. Among them, the combination of the Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM), referred to as the GMM-HMM framework, dominates the research field of robust speech recognition techniques. Under the GMM-HMM framework, robust speech recognition techniques can be broadly classified into two major categories: speech recognition techniques based on front-end processing and speech recognition techniques based on back-end processing [14]. The following paragraphs will elaborate on these two types of techniques. Speech recognition techniques based on back-end processing mainly include model compensation techniques and adaptive techniques. The model compensation technique mainly adjusts the parameters of the acoustic model trained in a specific environment to match the test environment to improve the recognition rate [15].

The PMC method first uses data about specific environmental influences (e.g., noise data) to train a specific model and then fuses the model with the pure speech model

TABLE 1

Teaching methods	Teaching method	Phonetics, vocabulary, and grammar rules
Real practice	Listen to	Phonetics, writing, vocabulary, grammar, and meaning communication
	Audiovisual	Grammar, vocabulary, meaning, communicative model, communicative principle, cultural model, etc.
	Situational approach	Vocabulary grammar, meaning model, context suitability vocabulary, grammar, and usage
	Functional ideation	Purpose, communication mode
Clear guidance	Association counseling law	Phonetics, vocabulary, grammar, and usage
	Systemic response	Grammar rules, grammatical usage, and Chinese-foreign comparison
	Suggestion	Understanding of grammar and text, application of grammar, etc.
	Grammar translation method	Provide metalanguage, conceptualize, and theorize linguistic phenomena
Critical framing	Cognitive approach	Discuss issues and draw correct conclusions
	Analytical method	Increase awareness and level through debate, criticism, and evaluation
	Negotiation method	Let students practice and innovate through tasks
Transformation practice	Argument	Improve your level by applying what you have learned in a new field
	Task teaching	Grammar rules, grammatical usage, and Chinese-foreign comparison

parameters in the log spectral domain to compensate for the probability density function of the pure speech model and produce an acoustic model that matches the test environment [16]. The disadvantage of the PMC method is that the parameters of the acoustic model obtained in the complex environment are too many and the computational effort is huge. The VTS technique, on the other hand, approximates the nonlinear relationship between the distorted speech signal and the pure speech signal by using a linear Taylor series expansion of finite length in the logarithmic spectral domain. In contrast to the PMC method, the VTS method has the advantage of integrating the balance between computer computing power and ASR system performance, and its disadvantage of ignoring higher-order terms in the linearized approximation process, which brings a certain error estimate. For this reason, some researchers later used the polynomial approximation (VPS) or statistical linear approximation (SLA) method to approximate this nonlinear function [17]. Model-based compensation techniques require a large amount of speech data with annotations that are relevant to the test environment during model training, which limits the practical applicability of such methods.

Pronunciation error detection methods based on linguistic and distinguishing features perform well in detecting typical errors common to language learners of different language families [18]. As mentioned earlier, this is mainly due to typical errors made by learners when learning a second language due to large pronunciation differences from region to region and language to language. These error types are very limited, and most of the use of this method to detect pronunciation problems requires the creation of a library of error types into which the types of errors that might be expected are included [19]. However, due to the instability of pronunciation error types, the error type library built by the pronunciation error detection method based on discriminative features cannot include all types of the same pronunciation error, and it is very difficult to extend it to the

detection of pronunciation errors for all phonemes. If the types of pronunciation errors are not in the range defined by the established pronunciation error type library, then even serious pronunciation errors cannot be detected by the system. Therefore, pronunciation error detection based on linguistic and distinguishing features is not widely applicable to the detection of all pronunciation errors. Therefore, the current pronunciation error detection methods based on speech recognition are used more than the distinguishing feature method.

3. Analysis of Automatic Assessment of Spoken English with Multimodal Feature Fusion

3.1. Design of a Multimodal Feature Fusion Model for Spoken English. Sentiment recognition algorithms recognize multiple sentiment states by learning modal sentiment features such as video, speech, and text. Firstly, three unimodal emotion recognition algorithms of video, speech, and text are introduced, and then, two multimodal fusion emotion recognition algorithms are analyzed on this basis, and the excellent performance of the tandem-based Dialogue multimodal emotion recognition algorithm is analyzed by comparison to lay the foundation for later improvements. SVM itself is a binary classifier, i.e., the output result is positive or negative. However, in dynamic expression recognition, the output emotion states are defined as calm, happy, and disgusted, so it is necessary to use multiple SVMs together to form a multiclassifier to achieve multiple emotion recognition [20]. It is divided into early feature fusion and later feature fusion. The feature-level fusion method can not only guarantee enough target information but also remove redundant information, thereby improving system performance; the decision-level fusion method is currently the highest level of fusion but compares data analysis, data preprocessing, and feature extraction requirements which are high. The feature vectors corresponding to one of the emotional states are defined as positive sets, and the feature

vectors corresponding to the other emotional states are defined as negative sets during the recognition of calm, happy, and disgusted emotion states, then three training sets can be obtained and used to train three models. When the video images are acquired in real time, their feature vectors are calculated by the above three models, and the maximum value of the output of the three models is taken as the classification result. The selection of the teaching method needs to consider the teaching concept on the one hand and the teaching objectives in the teaching phase on the other hand. And when the multimodal literacy teaching model is used as a reference for instructional design, the choice of teaching methods is also governed by the requirements of its four teaching methods for teaching methods. According to the role played by different pedagogies, the corresponding pedagogies were adapted for the four teaching approaches of real-world practice, explicit instruction, critical framing, and transformative practice, as shown in Table 1.

We can see that each multimodal teaching approach has several pedagogies to choose from; in other words, the teaching methods in multimodal teaching approaches are flexible and varied. The teaching method is only a means to accomplish the classroom goals, and there should not be only one teaching method consistently throughout the teaching process but should be refined to different teaching sessions to choose the appropriate teaching method according to the specific teaching situation [21]. The quality of model input data should be improved, then the basic model used in this article should be introduced, which is a pure speech model based on time series structure. Multimodal vocabulary teaching expands the channels of vocabulary teaching by combining multiple modal symbolic resources, creates a natural and easy learning environment for students, and helps students acquire and perceive information knowledge by creating real language contexts through the design of multimodal symbols, promoting students' participation in classroom interaction, stimulating students' enthusiasm for English vocabulary learning, and strengthening learners' vocabulary input and output ability, thus improving the effectiveness of classroom teaching.

$$T_i^T = \{s_i^T, sz_i^T\}, \quad (1)$$

$$s_i^T = \{V^i s_i^T, g(k, i) sz_i^T\}. \quad (2)$$

Giving speaking practitioners a numerical rating result will not be of much help in improving their speaking skills nor will it help the practitioners to identify their inherent pronunciation deficiencies. In this paper, an automatic recognition engine is chosen as the platform for the experiment, and then, by intervening in the process of building a grammar search tree, words or utterance fragments that have been extended according to a specific grammar are added as a node for the search. The result of the recognition contains characteristic content, which indicates that the pronouncer has followed a specific grammar [22]. Then, based on the inherent instability of the automatic

recognition system, the scoring process is integrated under the fuzzy logic to give a linguistic score to the speaking practitioner rather than a specific numerical evaluation. With this approach, the problem of assessing specific grammatical phenomena in spoken pronunciation and the pronunciation characteristics of a specific population is successfully addressed, as shown in Figure 1.

From Figure 1, we can see that the relationships between modalities are complex. If teachers consider the different relationships among all the modalities in their instructional design, it is conceivable that the design task will be extremely large and difficult for teachers to implement, and even teachers may unconsciously shift their main attention from accomplishing the teaching objectives to pursuing the teaching methods in the process of instructional design, and the implementation of classroom teaching will be a situation where various modalities will take over the main one. Therefore, I classify the relationship between modalities used in teaching with two groups of concepts, namely, "primary modality" and "secondary modality" and "automatic modality" and "conscious choice modality." In actual discourse communication, the role of various modalities is different, so there is a primary and secondary distinction. The primary modality is the main modality used by the teacher in the classroom, which often takes on the mission of conveying key information in classroom discourse, and is also the focus of what the teacher wants students to learn. It is not the focus of teachers' and students' attention in teaching, and the communicative information it represents only complements and highlights the information conveyed by the main modality, which teachers hope to use to help students understand the main communicative modality. For example, when a teacher does a model reading aloud, the oral modality is bound to be the primary modality, while other modalities such as the teacher's body movement and facial expressions are auxiliary modalities. Audio activity detection models based on machine learning or deep learning can also be used for segmentation, etc., which can improve the accuracy of audio activity detection and obtain higher-quality audio data.

$$\text{Acc} = \frac{\text{TP} - \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}}, \quad (3)$$

$$\text{Pre} = \frac{\text{TP}}{\text{TP} - \text{FP}}. \quad (4)$$

The plain Bayesian classification method is simple in form and efficient in training and has high model performance. However, the model performance is less stable. Because of its independence assumption, it does not learn as much about the real situation of the data, and there is still some room for improvement. However, it is still frequently used in traditional speech automatic assessment studies with good results.

$$h_{nb}(x) = \arg \min_c p_c \prod_{i=1}^d p(x, c). \quad (5)$$

FIGURE 1: Relationship between modalities in multimodal discourse.

Communication in the real world often involves multiple modal information, for example, human verbal speech contains both audio and textual content information, and the video we see contains both image and audio content, etc. Therefore, this topic is mainly concerned with the research direction of multimodal fusion. Multimodal fusion is the most widely used direction of multimodal fusion, which mainly combines information from multiple modalities to do various tasks, such as classification or regression tasks and target detection. For this research direction of multimodal fusion, there are different levels of fusion according to the type of data involved in data fusion: pixel-level fusion, feature-level fusion, and decision-level fusion, as shown in Figure 2. For feature-level fusion, it can be further divided into prefeature fusion and postfeature fusion according to the fusion at different positions in the network structure. The feature-level fusion method can ensure enough target information and remove redundant information, thus improving the system performance; the decision-level fusion method is the highest level of fusion at present, but it has high requirements for data analysis and data preprocessing and feature extraction. Therefore, in this paper, the feature-level multimodal fusion approach is used for the design of the speech evaluation model. The goal of the feature-level multimodal fusion approach is to fuse the feature information of multiple modalities into higher quality and useful information through some fusion algorithm and use the final fused information for the next decision-making step. The specific process of the feature-level multimodal fusion is illustrated in Figure 2.

The quality of data directly affects the performance of subsequent automatic speech assessment models. Therefore, this paper will design a set of standardized and effective speech assessment data preprocessing processes to provide

an effective guarantee for the quality of the input data of the subsequent speech assessment model. This chapter first analyzes the problems in speech assessment data and improves the quality of model input data by standardizing an effective data preprocessing operation; then introduces the base model used in this paper, which is a pure speech model based on the temporal structure; finally, experiments on the effectiveness of the data preprocessing process are conducted using the previously mentioned base model, and a control variable approach is used to verify the data preprocessing process designed in this chapter. If the continuous reading rules are met, the previous work will be marked with a continuous reading mark to prepare for all possible expansions of continuous reading and the expansion of new words.

The speech data used for automatic speech evaluation modeling are often recorded in real scenarios by recording devices. The data acquisition process is not standardized enough, there are many unexpected situations, and the acquired speech data usually have a lot of noise, such as the current sound of the recording equipment, buzzing background sound, large silent segments inside the audio where no one is talking, and surrounding noise interference. These noisy data are useless for automatic speech evaluation models and even degrade the performance of the models significantly. Therefore, this paper needs to do noise reduction on the speech data to improve the speech quality and provide a reliable guarantee for the performance of the subsequent model. This is a key step in data preprocessing.

3.2. Experimental Design for Automatic Assessment of Spoken English. When analyzing specific audio data, it often found that there is usually a slight noise at the beginning and end of the audio, and there are also many irrelevant low-

FIGURE 2: Types of multimodal fusion.

volume parts in the middle of the sentence, which occupy even longer durations than the active audio [23]. Therefore, to prevent the interference of these noises, the first and last parts of the audio as well as the irrelevant parts in the middle need to be removed, which require the use of audio activity detection techniques. The exact flow of audio activity detection processing is shown in detail in Figure 3.

The audio activity detection algorithm can detect the area range of the active audio from the original audio more accurately, which is convenient for subsequent operations such as slicing and splicing [24]. At present, audio activity detection technology has been an essential part of speech tasks, and audio activity detection technology is constantly evolving. In this paper, I simply use the signal-to-noise ratio value to judge the cut-off point, but I can also use other feature values to cut-off according to the actual situation, and I can also use the audio activity detection model based on machine learning or deep learning to cut-off, etc., all of which can improve the accuracy of audio activity detection and obtain higher quality audio data. By preprocessing the audio activity detection algorithm, the quality of the audio is substantially improved, which paves the way for the subsequent performance improvement of the model.

To make Sphinx-4 be able to recognize concatenation, it first needs to be able to accept concatenated grammar and then build a search grid based on the input grammar nodes, so that when the input speech contains concatenated components, it can be detected and recognized by the system. This is achieved by adding methods to the original class that can dynamically generate new grammar nodes so that the system can accept grammar nodes processed by concatenation rules [25]. The more important parameter in the process of fundamental frequency feature extraction is the setting of the frequency range of the band-pass filter, which

is set to 52 and 620, respectively, and the window function adopts Gaussian function. Then, the Sphinx-4 system itself recognizes the input speech and produces the corresponding results. Since the evaluation is to be performed for concatenation, the recognition results need to be compared with the utterances processed by the concatenation rules to give a result of the evaluation. For the continuous reading system, the core modules of this paper are continuous reading annotation and recognition result analysis. The main functions implemented in this model are, for a given grammatical utterance, to annotate all the concatenation possibilities in that grammatical text according to the existing concatenation rules, to generate all possible concatenation extensions, and to add new synthetic words to the system's lexicon. For an input utterance, the system's dictionary is first used to find out the phonetic symbols of each word in the utterance, and then, every two adjacent words (the last phonetic symbol of the former word and the first phonetic symbol of the latter word) are processed in turn according to the concatenation rules, and if they match the concatenation rules, the former word is marked as concatenable for all concatenation possibility expansions as well as new word expansions. The utterance that has been marked for alliteration is then extended for alliteration, i.e., the utterance is extended to the set of utterances that can contain all alliteration possibilities.

The commonly used resonance peak extraction methods are the band-pass filter bank method, inverse spectrum method, and linear prediction method. They have their advantages and disadvantages, the earliest used resonance peak extraction method is the band-pass filter bank method, the band-pass filter bank method extracts resonance peak features with good flexibility, but the performance is worse than the linear prediction method. The inverse spectrum

FIGURE 4: Linear frequency vs. Mel frequency.

base-frequency feature extraction process are the settings of the frequency range of the band-pass filter (maximum base-frequency value and minimum base-frequency value), which are set to 52 and 620, respectively, and the window function, which is a Gaussian function; the loudness feature is often used to reflect the subjective human perception of the strength of the speech signal [27]. When the frequency value of the speech signal is certain, the stronger the sound intensity value is, the greater the loudness value is; loudness is also frequency-dependent. In this paper, the loudness feature is extracted indirectly by using the sound intensity level, the unit of which is the decibel dB. The loudness feature is related to the sound intensity level and frequency of the speech, so the loudness feature can be extracted indirectly by using the fundamental frequency and the sound intensity level; the pitch is like the loudness, and is also a value that reflects the subjective perception of the human to the speech signal. Pitch features can be very helpful for both rhythmical assessment and emotional assessment of the speaker. Therefore, the calculation of pitch can also be obtained indirectly through frequency values, roughly in a logarithmic relationship.

4. Analysis of Results

4.1. Performance Results of Multimodal Feature Fusion Models for Spoken English. The multimodal emotion recognition algorithm is optimized by describing the implementation process of the interactive emotion recognition parameter optimization method. The algorithm is then subjected to statistical and comparative experiments to demonstrate the usability and better recognition performance of the algorithm. In the interactive sentiment recognition parameter optimization method, scalars and model graphs are visualized by the tensor board tool, and then, a judgment mechanism is used to empirically judge these visual graphs to achieve timely adjustment of hyperparameters and optimization of the model framework. As a result, the overall scoring efficiency is low; for language learners, there are many hidden oral expression problems in the learning pro-

cess that cannot be discovered in time, which affects the efficiency of language learning. The process of processing and visualization of data by the tool is an important part of the whole parameter optimization method. The visualization of the stored logs using tools uses two main components, scalar and graph. The function contains the directory in which the generated files are stored and default to run. The plotted scalar change graph and the plotted weight histogram and multiquantile line graph are more intuitive and convenient in finding problems with the model during training and monitoring the distribution of weights and gradients and the direction of the network updates. Figure 5 depicts the trend of accuracy and loss function changes in the training set; it shows the distribution of weight values over the range of each value frequency of occurrence. As for the visual modality in classroom teaching, 42% of the students had no preference for visual color modality; they did not have a clear need for the colors presented on the PPT screen and did not have a clear preference for using colors to convey linguistic meaning. One of the students who chose the “no preference” option also explained his choice, stating that the colors of the PPTs used in the classroom now met his expectations.

In contrast, 58% of the students wanted to use visual modalities such as images, graphics, and animations in the classroom, and most of them showed a preference for images and animations. 8% of the students who chose the “Other” option said in the interview that the current use of images, graphics, and animations in the classroom was good enough and did not need to be further improvements are not needed. The use of the auditory modality of video and music in the classroom is preferred by 67% and 58% of the students, respectively. Videos and music are pleasurable and most students prefer to use these media to help with oral learning. In the personal interview, 8% of the students who chose the “other” option in question 7 said that they would like the teacher to use videos in class when the videos used for teaching are interesting. This shows that students prefer videos with a coordinated combination of visual and auditory modalities, as well as music modalities that are pleasurable, as shown in Figure 6. The feature extraction process is cumbersome and the generalization ability of the model is poor. Research on the spoken English evaluation algorithm based on key technologies such as speech recognition and natural language processing, so that it can interact with the user’s spoken language, automatically evaluate the user’s spoken language, and allow the user to automatically discover Ability to lie in the crux of not being fluent in spoken English.

Unlike existing studies, the clustering process is performed in the recognition model. Compared to inductive learning, the proposed adaptive model theory-based transductive learning is more suitable for application processes where individual differences are evident, as it does not focus on the overall accuracy of the model but on the individual situation. To improve the quality of subsequent model input data and to innovate the model based on the shortcomings of the existing automatic speech assessment model, this article is no longer limited to traditional feature engineering and machine learning methods and uses the latest time series

FIGURE 5: Visualization result graph.

FIGURE 6: Distribution of database validity labels.

deep learning methods for modeling. This satisfies the need for adaptive sentiment recognition: the ability to use multiple user samples to measure the sentiment state of an individual improves the small-sample limitation of subject-dependent training. As shown in Figure 6, in this paper on transduction learning, the features of the test samples and the training set samples work together to determine the best subset, which can be constructed as a specific local classifier through the provided high-dimensional mapping, kernel matrix, local relevance computation, and optimization of the objective function process. The purpose of transduction learning is not only to perform adaptive sentiment recognition but also to include outlier exclusion, thus reducing errors caused by individual differences and improving the performance of multimodal sentiment recognition.

4.2. Experimental Results of Automatic Assessment of Spoken English. Finally, there is also the processing of the recognition results. In line with the result processing of consecutive reading, the system's recognition results also need to be regularized, and the algorithm of regularization is the same as

the algorithm of great matching of consecutive reading; if the number of matching words is less than 60%, the system rejects the recognition. The difference is that the system only compares the recognition result with the standard phoneme string, looks for the position function inside the standard phoneme string in the recognition result, and judges whether two words are identical or not, not whether they are spelled the same, but also considers the confusable phoneme set, and the phonemes in the same HDP set can be regarded as the same phoneme. Another difference is that confusion-prone sounds are evaluated in HDP clusters, and the size of the cluster used in this experiment is 10. This subsection uses the base model of speech evaluation based on temporal structure introduced earlier for data preprocessing experiments, which is a pure speech model based on GRU (MA_SE+audio+GRU). Comparative experiments are designed to verify the effectiveness of each operation in the data preprocessing process through a control variable approach, mainly verifying the effectiveness of these two preprocessing operations of audio activity detection and data sampling, and the multimodal experiments in later chapters will verify the effectiveness of the text operation introduced by speech recognition. The experimental results for the three scoring modules are presented in Table 2.

The results of the fluency scoring module in the preprocessing process validation are shown in Table 2. Both the removal of data sampling and the removal of audio activity detection have a significant impact on the performance of the fluency module, and the removal of these two modules results in a decrease in the agreement rate, adjacent agreement rate, and root mean square error of the model, with a greater decrease for the removal of the audio activity detection process. The experimental results show that both audio activity detection and data sampling have a significant improvement on the model performance, with the audio activity detection process showing a greater improvement. Then the system collects the pronunciation of the spoken language practitioner, recognizes the speech signal, and finds the best matching recognition result.

From the overall experimental results of the preprocessing process validation, the removal of the audio activity detection processing module resulted in a significant decrease in model performance for all three metrics, with an average decrease of 20.53% in agreement rate, an average decrease of 3.09% in adjacent agreement rate, and an average increase of 0.1918 in RMSE error values, indicating that the audio activity detection processing had a significant effect on the performance of the speech assessment model. The data resampling module also had a relatively large impact on the three assessment modules, with the performance of two indicators, emotional performance and rhythm, indicating that the data distribution of these two indicators was very uneven, and the data sampling operation increased the agreement rate of the three indicators by an average of 9.02%. The average increase in the adjacent agreement rate was 1.13%. In summary, audio activity detection preprocessing has the greatest impact on the performance of the speech evaluation model, with significant improvements in all three-evaluation metrics, with the most

TABLE 2: Results of the rhythm sense module in the preprocessing process validation.

Exclude modules	Agreement rate	Adjacent agreement rate	Root mean square error
Data sampling	4.64	7.27	3.96
Audio activity detection	1.98	9.94	3.89
MA_SE+audio+GRU	14.51	1.98	7.79

FIGURE 7: Results of the emotional performance module in the experiment.

significant improvement in agreement rate; the performance improvement caused by data resampling is not as significant as audio activity detection; however, it also plays a role. Therefore, the data preprocessing operations designed in this paper all have some improvement on the performance of the automatic speech evaluation model, as shown in Figure 7.

From the experimental results, all three scoring modules jointly have a great improvement on the emotional performance module, with significant improvement in all three indicators of agreement rate, adjacent agreement rate, and root mean square error. Therefore, it is necessary to use multiple SVMs to form a multiclassifier to realize the recognition of multiple emotions. In the process of recognizing the three types of emotional states, calm, happy, and disgust, the feature vector corresponding to one of the emotional states is defined as a positive set, and the feature vector corresponding to the other emotional states is defined as a negative set. It shows that emotional performance is a result of a combination of many factors, so information from multiple scoring modules is needed for joint learning. Combining only the information of emotional performance and rhythm

or fluency two by two, the agreement rate of both emotional performance scoring modules decreased significantly and the adjacent agreement rate increased significantly. It shows that the union of two modules is not sufficient for the emotional performance module, which tends to score adjacent scores and cannot make accurate predictions and that there is a significant decrease in the root mean square error after the union of rhythmicality and emotional performance.

5. Conclusion

This paper provides a comprehensive explanation of the theoretical foundations of multimodal discourse analysis theory as applied to the design of speaking instruction and explores the specific application of multimodal theory in the teaching of elementary English-speaking classrooms through the design of elementary English-speaking classrooms. Firstly, separate convolutional layers are used to process visual and auditory data, which not only allow independent modeling on the auditory and visual channels but also allow the transfer of audiovisual synchronicity information to occur in the convolutional layers and finally capture the mutual long-time dependencies therein using a higher-level shared fully connected layer. Multimodal vocabulary teaching expands the channels of vocabulary teaching by combining multiple modal symbol resources, creating a natural and relaxing learning environment for students, and creating a real language context by using the design of multimodal symbols to help students acquire and perceive information and knowledge and promote students to participate in classroom interaction. The advantages of multimodal teaching are obvious: it can combine modern advanced teaching technologies to create more realistic communicative situations in the classroom, gather and present resource information in various modalities, and ensure rich and diverse language input; students can receive various sensory stimuli in the classroom to deepen their memory and experience of language; it increases the interest of classroom teaching and improves students' classroom participation and motivation. The framework of multimodal teaching in the speaking classroom is a reference for how to use multimodal teaching in the speaking classroom. To reflect the practicality of the experimental model, I designed an automatic pronunciation correction web system, which is of great significance to improve the quality of English learners' oral pronunciation and learning efficiency and to accelerate the transformation and upgrading of English teaching methods, considering the electronic teaching method of the Internet.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The author does not have any possible conflicts of interest.

Acknowledgments

This study was supported by the Jiangsu Social Science Application Research Project “Construction of the ‘Three Coordination’ Education Mode of Higher Vocational English Course” (21SWC-52), Jiangsu Modern Educational Technology Research Project “Research on Hybrid Teaching of Humanistic Quality Courses in Higher Vocational Colleges Based on Multi-platform Collaboration” (2019-R-72621), and Research Project of Teaching Reform of Yangzhou Polytechnic Institute “Research on Cultivation Mode of Humanistic Quality of Higher Vocational Students Based on Multi-platform Cooperation” (2020XJJG57).

References

- [1] H. Zhang, “Research on spoken English analysis model based on transfer learning and machine learning algorithms,” *Journal of Intelligent & Fuzzy Systems*, vol. 38, no. 6, pp. 7377–7387, 2020.
- [2] S. Rasipuram and D. B. Jayagopi, “Automatic multimodal assessment of soft skills in social interactions: a review,” *Multimedia Tools and Applications*, vol. 79, no. 19, pp. 13037–13060, 2020.
- [3] A. R. Butt, A. Arsalan, and M. Majid, “Multimodal personality trait recognition using wearable sensors in response to public speaking,” *IEEE Sensors Journal*, vol. 20, no. 12, pp. 6532–6541, 2020.
- [4] Z. Li, L. Tian, Q. Jiang, and X. Yan, “Fault diagnostic method based on deep learning and multimodal feature fusion for complex industrial processes,” *Industrial & Engineering Chemistry Research*, vol. 59, no. 40, pp. 18061–18069, 2020.
- [5] M. Zhao, A. Jha, Q. Liu et al., “Faster mean-shift: GPU-accelerated clustering for cosine embedding-based cell segmentation and tracking,” *Medical Image Analysis*, vol. 71, p. 102048, 2021.
- [6] W. Chu, P. S. Ho, and W. Li, “An adaptive machine learning method based on finite element analysis for ultra low-k chip package design,” *IEEE Transactions on Components, Packaging and Manufacturing Technology*, vol. 11, no. 9, pp. 1435–1441, 2021.
- [7] M. Zhao, Q. Liu, A. Jha et al., “VoxelEmbed: 3D instance segmentation and tracking with voxel embedding based deep learning,” 2021, <https://arxiv.org/abs/2106.11480>.
- [8] X. Liu, A. Yu, X. Wei, Z. Pan, and J. Tang, “Multimodal MR image synthesis using gradient prior and adversarial learning,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 6, pp. 1176–1188, 2020.
- [9] Y. Jiang, X. Gu, D. Wu et al., “A novel negative-transfer-resistant fuzzy clustering model with a shared cross-domain transfer latent space and its application to brain CT image segmentation,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 1, pp. 40–52, 2021.
- [10] Y. Jiang, Y. Zhang, C. Lin, D. Wu, and C.-T. Lin, “EEG-based driver drowsiness estimation using an online multi-view and transfer TSK fuzzy system,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1752–1764, 2021.
- [11] Y. Wang, “Detecting pronunciation errors in spoken English tests based on multifeature fusion algorithm,” *Complexity*, vol. 2021, Article ID 6623885, 2021.
- [12] Q. Liang, N. Xu, W. Wang, and X. Long, “Multimodal information fusion based on LSTM for 3D model retrieval,” *Multimedia Tools and Applications*, vol. 79, no. 45–46, pp. 33943–33956, 2020.
- [13] C. Zhang, Z. Yang, X. He, and L. Deng, “Multimodal intelligence: representation learning, information fusion, and applications,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 478–493, 2020.
- [14] G. Yuan, X. Liu, Q. Yan, S. Qiao, Z. Wang, and L. Yuan, “Hand gesture recognition using deep feature fusion network based on wearable sensors,” *IEEE Sensors Journal*, vol. 21, no. 1, pp. 1–547, 2020.
- [15] S. Y. Han, N. S. Kwak, T. Oh, and S. W. Lee, “Classification of pilots’ mental states using a multimodal deep learning network,” *Biocybernetics and Biomedical Engineering*, vol. 40, no. 1, pp. 324–336, 2020.
- [16] Q. Zhang, J. Xiong, Y. Cai, J. Shi, S. Xu, and B. Zhang, “Multimodal feature learning and fusion on B-mode ultrasonography and sonoelastography using point-wise gated deep networks for prostate cancer diagnosis,” *Biomedical Engineering/Biomedizinische Technik*, vol. 65, no. 1, pp. 87–98, 2020.
- [17] F. Tang, S. Liang, T. Zhong et al., “Postoperative glioma segmentation in CT image using deep feature fusion model guided by multi-sequence MRIs,” *European Radiology*, vol. 30, no. 2, pp. 823–832, 2020.
- [18] G. B. da Fonseca, G. Sargent, R. Sicre, Z. K. G. Patrocínio Jr., G. Gravier, and S. J. F. Guimarães, “Hierarchical multi-label propagation using speaking face graphs for multimodal person discovery,” *Multimedia Tools and Applications*, vol. 80, no. 2, pp. 2797–2820, 2021.
- [19] I. Bakkouri and K. Afdel, “Computer-aided diagnosis (CAD) system based on multi-layer feature fusion network for skin lesion recognition in dermoscopy images,” *Multimedia Tools and Applications*, vol. 79, no. 29, pp. 20483–20518, 2020.
- [20] G. A. Kumar and P. V. Sridevi, “E-fuzzy feature fusion and thresholding for morphology segmentation of brain MRI modalities,” *Multimedia Tools and Applications*, vol. 80, no. 13, pp. 19715–19735, 2021.
- [21] T. H. Trojahn and R. Goularte, “Temporal video scene segmentation using deep-learning,” *Multimedia Tools and Applications*, vol. 80, no. 12, pp. 17487–17513, 2021.
- [22] R. Bibi, Z. Mehmood, R. M. Yousaf, T. Saba, M. Sardaraz, and A. Rehman, “Query-by-visual-search: multimodal framework for content-based image retrieval,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 11, pp. 5629–5648, 2020.
- [23] J. Li, H. Peng, H. Hu, Z. Luo, and C. Tang, “Multimodal information fusion for automatic aesthetics evaluation of robotic dance poses,” *International Journal of Social Robotics*, vol. 12, no. 1, pp. 5–20, 2020.
- [24] S. Halfon, M. Doyran, B. Türkmen, E. A. Oktay, and A. A. Salah, “Multimodal affect analysis of psychodynamic play therapy,” *Psychotherapy Research*, vol. 31, no. 3, pp. 313–328, 2021.
- [25] U. Kumaran, S. Radha Rammohan, S. M. Nagarajan, and A. Prathik, “Fusion of mel and gammatone frequency cepstral coefficients for speech emotion recognition using deep C-RNN,” *International Journal of Speech Technology*, vol. 24, no. 2, pp. 303–314, 2021.

- [26] R. H. Huan, J. Shu, S. L. Bao, R. H. Liang, P. Chen, and K. K. Chi, "Video multimodal emotion recognition based on bi-GRU and attention fusion," *Multimedia Tools and Applications*, vol. 80, no. 6, pp. 8213–8240, 2021.
- [27] Z. Wang, Z. Yin, and Y. A. Argyris, "Detecting medical misinformation on social media using multimodal deep learning," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 6, pp. 2193–2203, 2021.