Hindawi
Complexity
Volume 2021, Article ID 6623885, 11 pages
https://doi.org/10.1155/2021/6623885

Research Article

Detecting Pronunciation Errors in Spoken English Tests Based on Multifeature Fusion Algorithm

Yinping Wang

School of Foreign Languages, Zhengzhou Sias University, Xinzheng, Henan 451150, China

Correspondence should be addressed to Yinping Wang; wangyinping@sias.edu.cn

Received 19 December 2020; Revised 2 February 2021; Accepted 5 February 2021; Published 15 February 2021

Academic Editor: Wei Wang

In this study, multidimensional feature extraction is performed on the U-language recordings of the test takers, and these features are evaluated separately, with five categories of features: pronunciation, fluency, vocabulary, grammar, and semantics. A deep neural network model is constructed to model the feature values to obtain the final score. Based on the previous research, this study uses a deep neural network training model instead of linear regression to improve the correlation between model score and expert score. The method of using word frequency for semantic scoring is replaced by the LDA topic model for semantic analysis, which eliminates the need for experts to manually label keywords before scoring and truly automates the critique. Also, this paper introduces text cleaning after speech recognition and deep learning-based speech noise reduction technology in the scoring model, which improves the accuracy of speech recognition and the overall accuracy of the scoring model. Also, innovative applications and improvements are made to key technologies, and the latest technical solutions are integrated and improved. A new open oral grading model is proposed and implemented, and innovations are made in the method of speech feature extraction to improve the dimensionality of open oral grading.

1. Introduction

In recent years, computer-assisted teaching systems have become one of the hot research topics in the fields of computer science and education [1]. Particularly, in large-scale language examinations, they have begun to gradually replace teachers in marking and have become a major change in the education sector, which we call Computer-Assisted Language Learning (CALL) systems [2]. Many computer-assisted assessment systems have been used on a large scale in actual teaching and examinations, such as English composition marking systems and computer program marking systems [3]. Such systems are more accurate in grading, and, most importantly, they save human resources and improve efficiency [4]. However, there are still many subjects and question types that are not yet automated, and some subjects, such as oral English, can only have part of their questions marked automatically [5]. There are many oral grading systems for reading and reciting, but there are few grading systems for open-ended oral questions, such as quizzes, repetitions, and individual responses. Ordinate and Speechwriter scoring systems are recognized as typical examples of automated scoring; however, they do not meet the need for open-ended speaking questions [6]. In recent years, with the development of speech recognition technology and the maturity of essay marking systems, it is expected that technically speaking the automatic marking of open-ended speaking questions can be overcome and reaches a practical level [7]. There are many real-life scenarios in which it is necessary to evaluate a speaker's oral expression ability, such as Mandarin exams, oral training, language teaching evaluations, and radio presenter exams [8]. Currently, these scenarios are still evaluated by manual scoring, such as averaging, voting, or one-vote voting, which are too subjective, often lack fairness, and do not give objective feedback to the speaker [9]. The overall scoring is inefficient; for language learners, there are many hidden problems with an oral expression that are often not detected in time, thus affecting the efficiency of language learning [10].

Wason et al. designed an automatic speech evaluation model based on the distribution of spectral density values in the time domain and the perceptual domain, and the correlation between the score of the improved automatic speech evaluation model and the subjective manual score reached 0.824 [11]. Heb-Umbach et al. chose a simple deep learning method to optimize the automatic speech evaluation model for nonnative speakers of [12]. DNN model scores higher correlations than the traditional speech model GMM, and the superiority of the deep learning method is effectively verified in this model [13]. In the 2018 Spoken CALL Shared Task, the participants proposed a better method to build an acoustic model using DNN-HMM and then designed a score mapping module using specific rules and feature engineering methods [14]. In the same year, Anastassia Loukina conducted an automated scoring experiment using seven different regression models, including a random forest, GBDT, and MLP [15]. Through experimental validation with different models, the authors found that, for the automatic speech evaluation task, when the training corpus is large enough, the results of training the models on different training sets and testing them on a unified test set are almost identical [16]. In the same year, Vishwakarma et al. proposed a method for the automatic assessment of the fluency module of spoken English by decomposing the low-rank matrix of correlated penalty terms to remove the subjective interfering factors from the data and improve the performance of the machine scoring of the fluency assessment module [17]. Massa implements an implicit semantic analysis (LSA) system with the core idea of using student-response to texts rather than generic texts from other sources for constructing the LSA model, with some of the texts being manually marked as input to the model [18]. The key issues for the model are the appropriate selection of specific texts to be manually marked and the overall measurement of mark-up effects [19]. The quality of text mark-up is measured by marking up the text until the effect reaches a threshold [20]. In terms of text selection, there are three options: random selection, clustering, or selecting relatively similar text for the tagged data. The advantage of Klein's method is that the text that needs to be manually tagged can be selected automatically with a minimum of effort. The weaknesses of the method are evident in the evaluation of its efficiency, which can only be achieved after the manual completion of the complete annotation of student responses. Another problem is that the parameters in the method are independent; the semantic space dimension and the similarity threshold parameter must be calculated simultaneously.

In this paper, we study audio processing, speech recognition, automatic essay marking, and deep learning techniques to design and implement a multifeatured intelligent automatic marking model based on the data generated from the language training system of Beijing University of Posts and Telecommunications. The model is designed to solve the problem of automatic marking of open-ended oral English questions and to reduce teachers' marking pressure. In this paper, a series of methods will be used to improve the accuracy of the model at various stages of scoring. Before extracting features from speech files, it is necessary to reduce the noise and cut silent fragments of speech to improve the accuracy of speech recognition. In this paper, a malefactor is a malefactor of existing noise reduction techniques, combining traditional noise reduction algorithms with deep learning. It applies to speech noise reduction for automatic grading of speaking language. Also, this paper will use an open, freely available speech recognition engine for speech transcription. The recognition rate of current speech recognition technology does not reach 100%. Furthermore, grammar and vocabulary errors in the recognized text due to pronunciation problems or fluency problems should be excluded when grading the grammar and vocabulary of students' oral answers. Therefore, to score the grammar and vocabulary of the speech-recognized text more accurately, it is necessary to clean the recognized text to a certain extent to maximize the actual expression of the students. The methods used in this paper are spelling correction, removal of onomatopoeia, and removal of consecutive repetitions. In this paper, algorithms that implement these functions are investigated to improve scoring accuracy.

2. Multifeature Fusion Speaking Test Detection Algorithm

2.1. Improved Multifeature Fusion Algorithm. The production of human speech is a complex process that the body can execute. It first receives signals from the brain, then extracts gas from the lungs to vibrate the vocal cords, and then allows the laryngeal muscles to express themselves. The characteristics of each person's vocal tract result in a different speech signal, and this information is used to distinguish between speakers. We characterize these differences by the acoustic characteristics of speech. Therefore, it is important to select the acoustic features appropriately in a speaker recognition system. If the selected acoustic features do not adequately characterize the speaker, even a deep learning algorithm cannot achieve a good performance [21]. It has been shown that when the test speech is long enough, the amount of information and discrimination of a single acoustic feature is sufficient to complete the speaker recognition task. The MFCC feature parameters have been applied to more than 90% of current speaker recognition systems. However, in short speech speaker recognition, when only the MFCC feature parameters are used for modeling, the speaker's personality information cannot be fully expressed, which makes it difficult to obtain good recognition results. But increased studies have shown that the brain cannot efficiently digest information from many different sources at the same time. As the human resources manual explains, this method involves assigning a certain time of day to focus on a specific task. You can even put your schedule on your calendar to let other employees know that you are busy. Before you work, this is the key to eliminate all distractions, so you can concentrate on the work in front of you. You can also schedule time blocks to match the natural changes in your energy levels throughout the day. Considering that different features can express the speaker's personality information from different perspectives, multifeature fusion can represent more personality information

about the speaker, which is feasible in testing short speech conditions. However, the simplest method of multifeature fusion is to directly connect multiple acoustic features extracted from each frame of the speech signal into a large vector of high-dimensional features, which are not desirable in practice. Since different features are not orthogonal to each other, the direct connection will affect each other, and the direct connection of different features is a high-dimensional spatial vector; increasing the dimensionality means increasing the complexity of the system. Also, there is a certain amount of repetition between feature components, which generates redundant information. Therefore, the high-dimensional space vector can be mapped to the low-dimensional space by a dimensionality reduction algorithm, and the parts that are more distinguishable between speakers can be selected [22].

According to the human ear's auditory perception mechanism, the human ear perceives speech signals at different frequencies with different perceptual abilities. When the frequency of the speech signal is less than 1 kHz, the frequency and perceptual characteristics of the speech signal are linear; when the frequency is higher than 1 kHz, the frequency and perceptual characteristics of the speech signal are logarithmic. The higher the frequency of the speech signals, the less perceptible it is to the human ear. The relationship between the actual frequency f and Mel's frequency can be expressed by

$$\mathrm{Mel}(f) = 259778g\left(1 + \frac{f}{700}\right)^2. \tag{1}$$

Figure 1 shows the extraction flow of MFCC feature parameters.

The input speech signal $s(n)$ is preprocessed to generate the time domain signal $x(n)$ (length of the signal sequence $N = 256$), and then each frame of the speech signal is processed by Fast Fourier Transform or Discrete Fourier Transform to obtain the speech linear spectrum $X(k)$, which can be expressed as

$$X(k) = \sum_{i=0}^{M} x(n) e^{-j(2\lambda/M)\mathrm{Mel}(f)}, \quad (0 \le n, k \le M). \tag{2}$$

The linear spectrum $X(k)$ is input into the Mel filter bank and filtered to generate the Mel spectrum, and then its low energy is taken to generate the corresponding log spectrum $S(m)$.

The Mel filter set is a set of triangular band-pass filters H_m, where M represents the number of filters, usually 20~28. The transfer function of the bandpass filter can be expressed as

$$H_m(k) = \begin{cases} 0, & (kpf(m-2)), \\ \dfrac{k - f(m-2)}{f(m) + f(m-1)}, & (f(m-2) \le k \le f(m)), \\ \dfrac{k + f(m-2)}{f(m) - f(m-1)}, & (f(m) \le k \le f(m+1)), \\ 1, & (kff(m-2)). \end{cases} \tag{3}$$

The reason for taking the logarithm of the Mel energy spectrum is to promote the performance of the speaker recognition system. The transfer function from the linear spectrum of speech $X(k)$ to the logarithmic spectrum $S(m)$ is as follows:

$$S(m) = \log\left(\sum_{k=0}^{M} |X(k)|^2\right) H_m(k). \tag{4}$$

The nth dimensional edge components $C(n)$ of the MFCC edge parameters are expressed by converting the logarithmic spectrum $S(m)$ to the MFCC edge parameters using the discrete cosine transform (DCT) as follows:

$$C(n) = \left(\sum_{k=0}^{M} |X(k)|^2\right) H_m(k) \cos\frac{n(m + 2/3)}{M}. \tag{5}$$

The time-domain impulse response function of the Gammatone filter is an analogy function that needs to be discredited to facilitate computer processing, and the Laplace transform of (5) is

$$G_i(s) = \frac{3}{4}\left[\frac{(m-1)!}{(s + b - jw)^m} + \frac{(m+1)!}{(s - b + jw)^m}\right]. \tag{6}$$

This is then converted to the Z domain of the Z transform and finally inverted to obtain the discrete impulse response of the Gammatone filter with the following expression:

$$Z = \frac{1}{2\lambda w} \int Z(z) Z^{n-1} \mathrm{d}z. \tag{7}$$

The input speech signal and output speech signal are converted to obtain the output of a Gammatone filter.

As shown in Figure 2, an important advantage of cyclic neural networks is that they have a memory function for historical information, making them well suited for modeling temporal sequencing problems. However, as the length of the time series increases, the number of hidden layers increases as well [23]. In the training process of the network, when calculating the gradient, the weighting parameters recur in the reverse propagation direction layer by layer, leading to a geometric increase or decay of the gradient, which results in a gradient explosion or gradient disappearance. In addition to improving the accuracy of speech recognition, it can also increase the rate of speech recognition. Gradient disappearance can make it difficult to train long-time span with RNNs because the data and eigenvalues that first entered into the RNN model are replaced by the eigenvalues of the data entered later. Traditional models have difficulty in learning the data features and dependencies in a time series if the time between input and associated output is too long. First, we perform noise reduction on the audio file and then input the reduced audio file into the pronunciation scoring module and the speech recognition module to output the pronunciation score and the text after speech recognition, respectively. The text is then cleaned and fed into the fluency scoring module, grammar-vocabulary scoring module, and semantic scoring

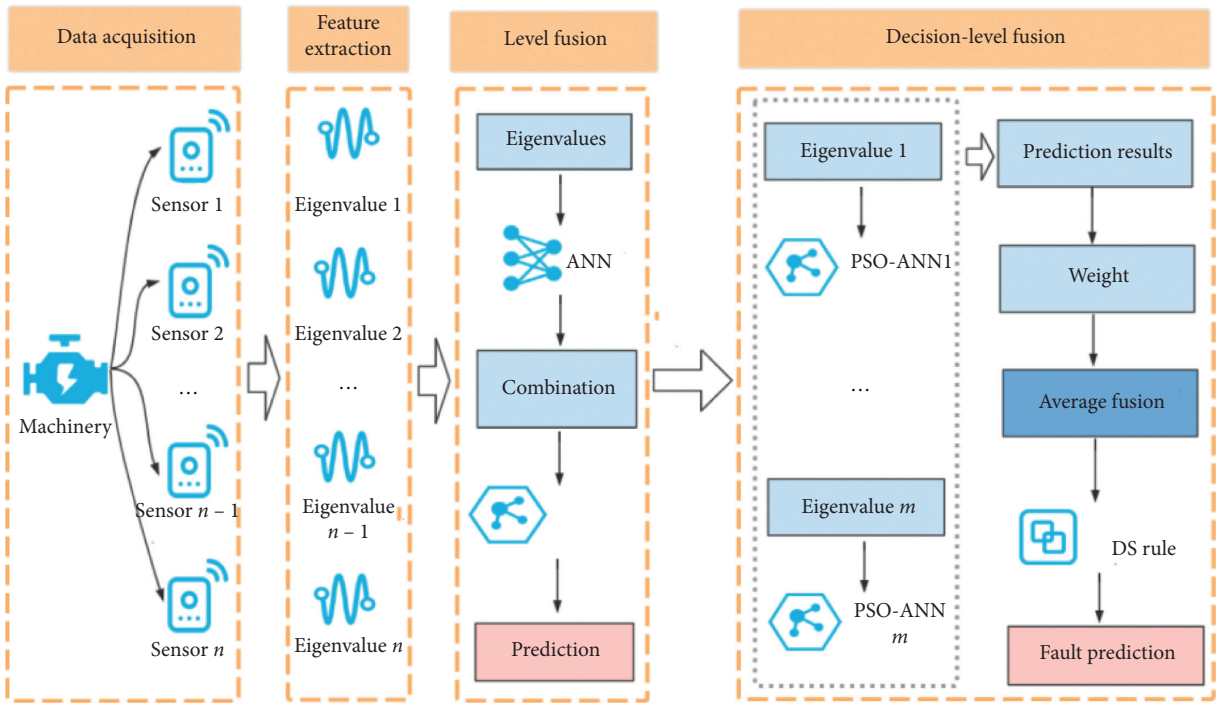

Figure 1: Improved model of the multifeature fusion algorithm.

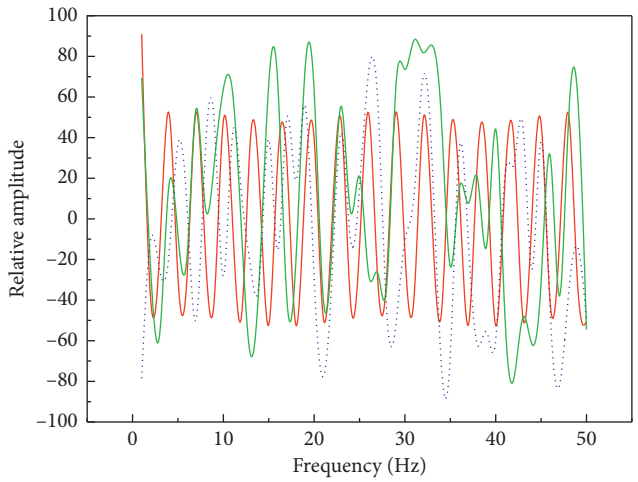

Figure 2: Gammatone filter spectrum for the channel.

module to obtain the fluency score, grammar score, vocabulary score, and the semantic score, respectively. Next, these five values were normalized and used to train our scoring model. Once the scoring model is trained, we can use the model with the feature extraction module to score the students' spoken language. The design of these speech noise reduction, speech recognition, text cleaning, and feature extraction modules will be presented separately. The speaker recognition system can be roughly divided into two parts: feature extraction and pattern recognition: how to mix different types of information, how to extract it, introduce high-level speech signal analysis, and research on enhancing the robustness of voiceprint recognition and reducing the amount of calculation.

Since the text-related feature scoring in this paper relies on the accuracy of the text, the accuracy of speech recognition becomes one of the most important indicators before text scoring. Currently, due to the development of artificial intelligence and speech recognition technology, the accuracy of speech recognition has been greatly improved, and there are many open and free speech recognition platforms available on the market for our use. The recognition effect of using these speech recognition platforms is often better than using open-source tools to train our speech recognition models, as the size of the corpus available for training speech recognition models in the laboratory is much smaller than that of the commercial speech recognition corpus. The accuracy of the speech recognition engine is directly proportional to the size of the corpus used for training. Therefore, this paper investigates the major free speech recognition engines on the market, compares their ease of use, recognition speed, and recognition accuracy, and selects the most suitable one for this study, which is used as the basis for test scoring.

2.2. Spoken English Pronunciation System Design. Fourier analysis of a sound signal can be used to see how the short-time frequency of the sound signal changes over time. Research on the frequency spectrum of sound signals began long before the development of digital signal processing (DSP) technology. Some scholars have used a spectrometer to record and analyze the short-time frequency spectrum of sound signals. A spectrometer is a device that inputs the electrical signal of a voice signal into a set of corresponding filters and, after the output of each filter, records it on paper

in the order of the frequency of the sound signal. The intensity of the sound signal frequency can be determined by observing the grayscale of the track on the recording paper. If the grayscale on the recording paper is deep, the signal frequency is strong, and if it is shallow, the signal frequency is weak. We rotate the recording paper at a constant speed, which is equivalent to recording the frequency value of the sound signal at a different time on the recording paper. When calculating the gradient, the weight parameters recurse layer by layer in the reverse propagation direction, and the gradient gradually becomes smaller. By operating in the above way periodically, we can get a graph of the sound signal recorded by the speech spectrometer, which is the speech spectrum of the sound signal [24]. The sound spectrum is a three-dimensional spectrum, which shows the change of the frequency of the sound signal with time. The horizontal coordinate of the sound spectrum is about the change of time, and the vertical coordinate is the axis of the change of the frequency spectrum with time. The frequency of the sound signal at any one moment can be represented by the shades of color at the same moment's position on the sound spectrum. Because the spectrogram reflects both the time domain waveform and the spectral characteristics of the sound signal on the spectrogram, the spectrogram shows a large amount of information about the characteristics of the sound signal, and we can represent the sound signal by the information contained on the spectrogram [25].

Given these characteristics of the sound spectrogram, we convert the sound signal into a sound spectrogram by processing the collected engineering equipment noise signal, extract the corresponding sound signal features according to the information contained in the sound spectrogram to represent more features of the sound signal, convert the one-dimensional sound signal into an image, and apply the digital image processing technology to the sound signal feature extraction. Get more features for sound signal visualization, as shown in Figure 3.

Multitask learning (MTL) refers to the combination of several single tasks that are relevant to each other, and the information of multiple single-task models is shared for joint learning. The model can learn information from multiple tasks at the same time, and, by influencing the information between these multiple tasks, the model's generalization capability can be improved, thus improving the model's performance. Speaking score is related to fluency and pronunciation. Listening score is related to vocabulary, spelling, grammar, and text. Writing scores are related to grammar, spelling, vocabulary, and text writing. Reading scores are related to grammar, spelling, vocabulary, and text writing. Unlike most simple single-task learning, multitasking can be used to improve model performance in a variety of situations. The specific application of multitask learning in real-life scenarios is usually achieved through the sharing of hidden layers in the network structure, which can be of two types according to the differences in the sharing of hidden layers in the network structure: (1) hard sharing of parameters, which means that all network structures are shared among all tasks, and each task only retains its output layer. To a large extent, it avoids overfitting and improves the generalization ability of the model; (2) soft (soft) sharing of parameters means that each task has its separate model, and the degree of sharing is not as high as that of hard sharing, and the models only interact with each other through regularization terms.

In a real speech evaluation scenario, the scorer often only listens to a segment of the speaker's speech, and the reference for multiple scoring modules is the same speech data, which can also be said to be the basis for scoring each module on a shared basis, which is fully consistent with the hard-share approach. As shown in Figure 3, the overall framework of the parametric hard-share implementation of multitask learning is shown. In front of the network structure is the network sharing layer, and behind the sharing layer is the task-specific layer, from task 1 to task n. The model can learn multiple tasks simultaneously. The more tasks the model learns at the same time, the more correlations the model must consider, and the more difficult it is for the model to learn. Therefore, the risk of overfitting the model is greatly reduced by learning multiple tasks with hard shared parameters.

2.3. Error Detection Indicator Design. Speech data used in automatic speech assessment modeling are often recorded in real-life scenarios using recording devices. The data collection process is not standardized enough, and there are many contingencies. The acquired speech data usually contains a lot of noise, such as the sound of the recording device's electric current, buzzing background noise, largely silent segments in the audio where no one speaks, and the surrounding noise interference. These noises are not useful for the automatic speech evaluation model and may even greatly reduce the performance of the model. Therefore, in this paper, we need noise-reduced speech data to improve the speech quality and provide a reliable guarantee for the performance of subsequent models. This is a critical step in data preprocessing. When analyzing specific audio data, it is often found that there is a slight noise at the beginning and end of the audio, as well as many irrelevant low-volume parts in the middle of sentences, which occupy even longer duration than the active audio. Therefore, to prevent this noise from interfering, it is necessary to eliminate the irrelevant parts at the beginning and end of the audio and in the middle, which requires audio activity detection techniques. The detailed flow of audio activity detection processing is shown in Figure 4.

The speech recognition task involves the conversion between speech modalities and text modalities; therefore, the overall conversion process is complex and difficult to study. The training process is done offline and mainly involves data collection, acoustic model training, and language model training. The specific training process of speech recognition includes four main modules: speech feature extraction, acoustic model building, the language model building, dictionary, and coding [26]. Compared with the most widely used LSTM model in the industry, the DFSMN model has faster training speed and higher recognition accuracy. Using the new DFSMN model of smart audio or smart home

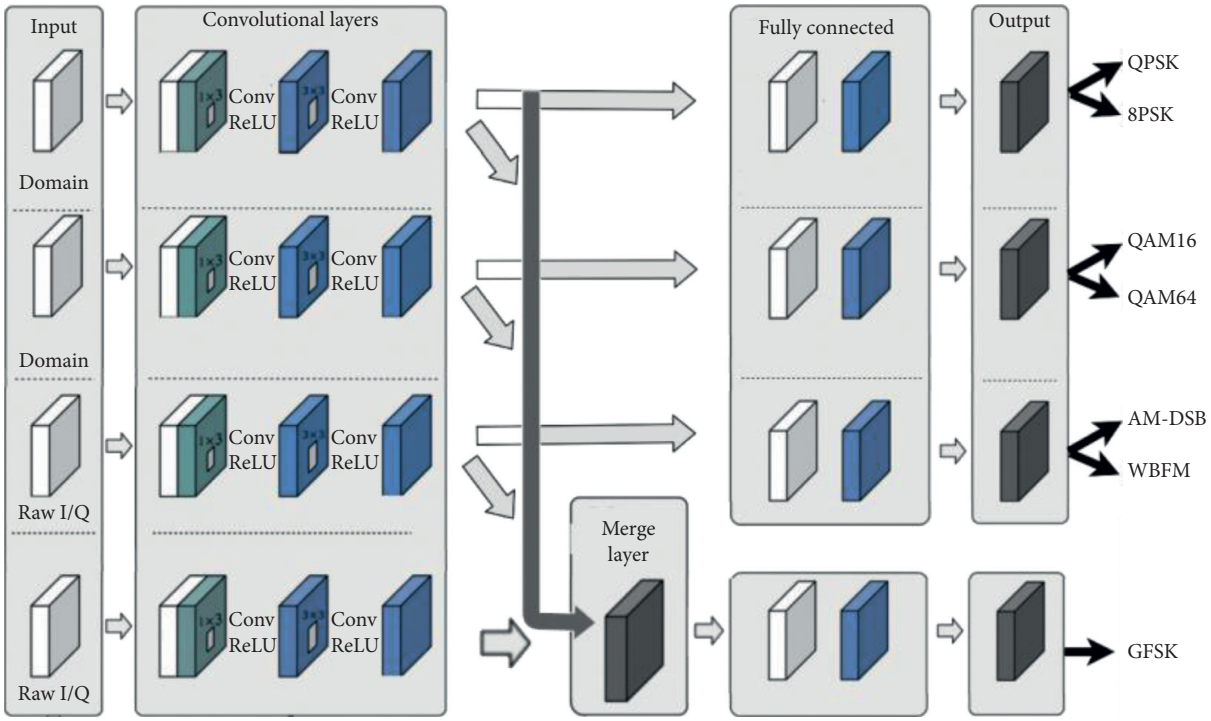

Figure 3: Network structures of hard-shaped forms in multitasking.

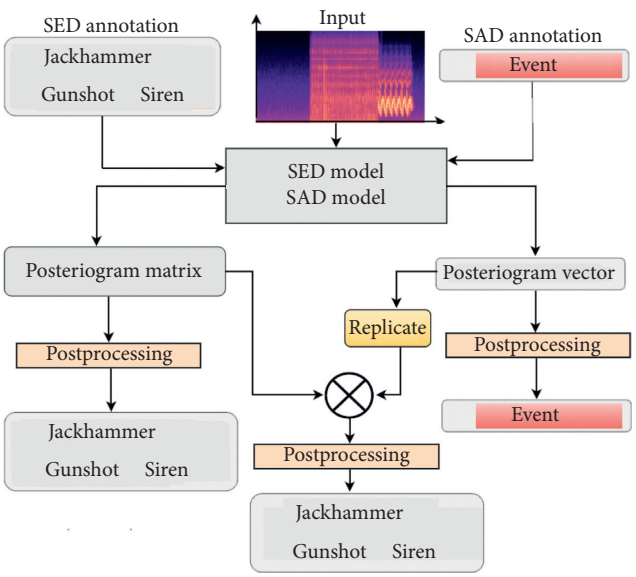

Figure 4: Detailed process of audio activity detection.

equipment, compared with the previous generation technology, the deep learning training speed is 3 times, and the speech recognition speed is 2 times. In Figure 4, the specific process of speech recognition is shown in detail, with the upper process being the training process and the lower process being the recognition process. The specific recognition process of speech recognition is carried out online, insulting speech data, then carrying out data preprocessing and feature extraction operations on the speech data, using the acoustic model and language model trained earlier, decoding the processed speech features and searching the results, and outputting text content. The specific process of speech recognition is (1) to do a series of processing on the audio data, including activity detection, noise reduction, and avoiding irrelevant audio interference. Then, the audio data is subjected to operations such as framing and windowing; finally, common features such as MFCC and LPCC are extracted from the audio, and the audio is converted into a feature matrix; (2) to establish an acoustic model, the input of this module is the extracted speech feature matrix, and the output is phoneme information. In recent years, the commonly used acoustic models are deep learning models, such as DNN and LSTM; (3) also, it establishes a language model, obtains a large number of text content for training, and learns the contextual information at the next level; (4) it establishes a dictionary corresponding to the phonemes and text and decodes the text, using the previously trained language model, the output phonemes of the acoustic model, and the text corresponding to each other. The text is then decoded to output continuous text content.

The texture, color, shape, and other features of an image are the basic properties that characterize the essential features of an image. These features are very important for describing and identifying the properties of an image, and different textures on an image reflect different properties of the image and the visual experience brought to people by this texture is different. Given the structural and textural differences among the visualized images of noise signals from engineering instruments, this paper presents a statistically based texture feature for visualizing sound signals. The visualized sound signal is usually a colored RGB image, which is first converted to a grayscale image before its

image is processed. There are many ways to process the grayscale, including maximum-based, mean-based, and weighted-average-based methods. In this paper, a weighted average algorithm for the RGB components is used, as shown in Table 1.

The horizontal axis is the time axis and the vertical axis is the waveform length. The figure shows that the 10-second original audio contains many useless audio clips and the audio activity detection algorithm detects three active intervals in the original audio. Ultimately, all the detected activity audio fragments are stitched together to generate the new audio, which is only 4 seconds long. This processed 4-second audio is almost noise-free and the audio quality is greatly improved. The audio activity detection algorithm can detect the area of active audio from the original audio more accurately, which is convenient for subsequent operations such as slicing and splicing. At present, audio activity detection technology has been an essential part of the speech task, and audio activity detection technology is also constantly developing; this paper is simply by the signal to noise ratio value to make a cut-point judgment, but also, according to the actual situation, the use of other eigenvalues to cut points can also be based on machine learning or deep learning audio activity detection model to cut points, and so on (improves the accuracy of audio activity detection and obtains higher quality audio data). By preprocessing the audio activity detection algorithm, the audio quality is greatly improved, which paves the way for the subsequent performance improvement of the model.

To a certain extent, Prosody rhythm features can indirectly reflect whether the speaker's speech speed is too fast or too slow, whether the intonation is too high or too low, and whether there are intonation and stops. Therefore, we extracted three speech features: fundamental frequency, loudness, and pitch and merged them into a rhythmic feature to complement the features of speech. The fundamental frequency, also known as the fundamental tone frequency, reflects the frequency between two adjacent openings and closings of the sound gate and can truly describe and characterize the mechanism of sound production. The fundamental frequency is extracted using the autocorrelation method, and the most important parameter in the fundamental frequency extraction process is the frequency range of the band-pass filter (maximum fundamental frequency value and minimum fundamental frequency value, which are set to 52 and 620, resp.), and the window function is a Gaussian function; the loudness feature is often used to reflect the subjective feelings of human beings about the strength or weakness of the speech signal. When the frequency value of the speech signal is fixed, the stronger the sound intensity value, the greater the loudness value; loudness is also related to frequency. In this paper, when extracting the loudness feature, the sound intensity level is used to indirectly represent the loudness feature, and the unit of the sound intensity level is our common decibel dB. The loudness feature is related to the sound intensity level and frequency of speech; therefore, it can be indirectly extracted by the fundamental frequency and the sound intensity level; the pitch is very similar to loudness and is also a reflection of the human subjective feelings towards the speech signal. Pitch characteristics can be very useful for both rhythmic and emotional assessments of a speaker. Therefore, the pitch can also be calculated indirectly from frequency values, roughly logarithmically. The darker the color, the stronger the signal frequency, so that different frequencies can be expressed in different colors.

Table 1: Parameter settings during audio activity detection processing.

Parameter	Parameter value	Parameter description
Top	1457	Signal to noise ratio between audio
Ref	5487	How to select reference audio clips
Frame length	1478	S/N ratio calculation window length
Hop length	2587	Window shift for SNR calculation
Min length	5678	Shortest fragment length

For the manual scoring component, each voice was evaluated on three dimensions: fluency, rhythm, and emotional performance. Fluency is a measure of the fluency of the speaker, which is crucial to the speaker's oral expression; rhythm is a measure of the speaker's pitch and rhythm, which are a higher-level assessment of the speaker's expression. Manual scoring of the dataset was generated by scoring the audio from the audio database by two experts in the field. Considering the subjective influence of manual scoring, we discussed the characteristics of the three evaluation modules with the experts and developed a set of standardized scoring criteria before the experts performed the scoring. For the manual scoring, a 5-point scale was used, with 5 being the highest and 1 the lowest. The average of the scores of the two experts was used as the final manual score.

3. Results and Analysis

3.1. Principal Component Results Analysis. There may be redundant information between the feature components of a feature vector, which affects both the computational cost of the computer and the recognition performance of the speaker recognition system. Principal Component Analysis (PCA) uses the decomposition of feature bases into orthogonal transformation matrices to convert the original feature vector into a low-dimensional noncorrelated and orthogonal linear feature vector. This new low-dimensional feature vector is determined by the variance of the projection and is ordered from largest to smallest. The first principal component corresponds to the direction with the largest variance and so on, and the last components correspond to the direction with the smallest variance. Figure 5 shows the distribution of the two principal components of a set of random data.

In the GMM-UBM-based speaker recognition system, the choice of the degree of blending has a great impact on recognition performance. This is because the accuracy of the speaker feature distribution model is directly related to the

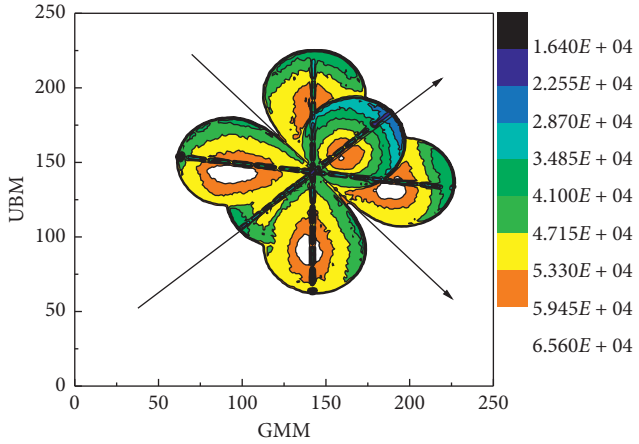

Figure 5: Principal component results.

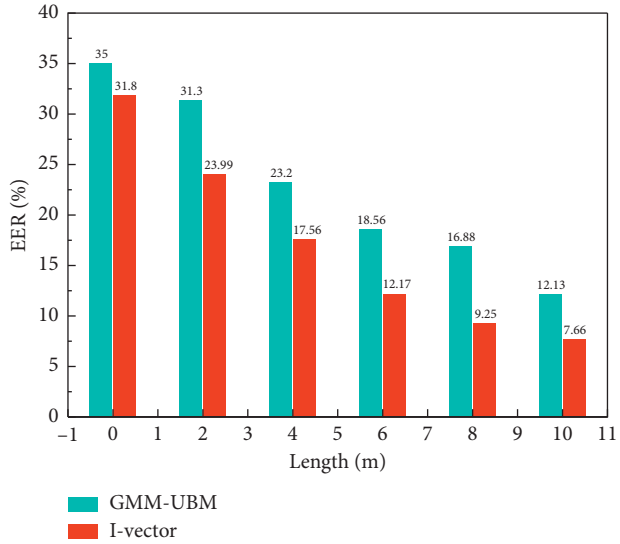

Figure 6: Comparison of the performance of two baseline systems (EER%).

blending degree, so this experiment mainly tests the system performance by different blending degrees. The 13-dimensional MFCC feature parameters and their first-order and second-order differences are used to assemble the 39-dimensional feature parameters. Figure 5 shows the performance of the GMM-UBM interpreter recognition system with different degrees of blending, based on the EER system evaluation. From the comparative observation, the higher degree of blending has better recognition performance than the lower degree of blending. As the blending degree increases, the EER of the system decreases, but the rate of change of EER decreases gradually, which indicates that the Gaussian mixture model has reached the best fit. If the blending is too low, the Gaussian blending model parameters are too simple, and the model distribution does not adequately characterize the speaker parameters of the trained speech, resulting in underfitting. The voice data needs to be cleaned before data preprocessing and feature extraction operations. On the contrary, if the mixture is too high, the Gaussian mixture model parameters are too complex, which leads to overfitting and weakening of the generalization ability of the model. Therefore, for the training of the GMM-UBM model, the choice of blending degree is crucial. The experimental results show that when the mixing degree is 1024, the performance of the speaker recognition system has reached saturation and increasing the mixing degree will only increase the complexity of the computational system, which will drastically reduce the recognition performance of the system.

Based on the above theories, the speaker recognition model can have an impact on the performance of the recognition system. In this experiment, the performance of the two baseline models has been experimentally evaluated under the condition of testing short speech, as shown in Figure 6.

As shown in Figure 6, the performance of the speaker recognition system is greatly affected by the test speech length. When the speech length is less than 2 s, the performance of two speech recognition systems drops dramatically. When comparing two different models with the same speech length, since the GMM-UBM model cannot suppress the interference of channel variation on speaker recognition, and the extracted feature parameters cannot fully fit the speaker feature distribution when the test speech length is less than 10 s, the I-Vector-based speaker recognition model does not strictly distinguish between speaker variation information and channel variation information, and the I-Vector-based speaker recognition model does not distinguish the speaker variation information from the channel variation information. The I-Vector model translates the high-dimensional feature space into a low-dimensional vector for study, which reduces both time and space complexity; thus, the I-Vector-based speaker recognition system has better robustness in testing short speech conditions.

The contextual harmony weight value α and the error detection thresholds $\beta 1$ and $\beta 2$ are experimentally determined. The above context values and below context values of the wrong words are simulated, and the context distribution of the wrong words is shown in Figure 7.

The smaller the context value of the wrong word, the smaller the contextual harmony value, which plays a greater role in determining the wrong point of identification. The greater weight of the following contexts is used to calculate the contextual harmony for error checking, which determines the value of α in the range [0, 0.5].

3.2. Oral English Pronunciation System Performance Results Analysis. First, we segment the speech by the speaker's pause and sound intensity. We use by dub a bath processing library to detect the sound intensity of the target sound at a certain moment, to segment the bath more accurately. We repeatedly adjusted the duration of silence and the sound intensity that was judged to be silent and finally set the parameters for judging the cut point at a sound intensity of less than −60 dB and a duration of more than 100 ms and added a cut point as long as the voice met this condition at

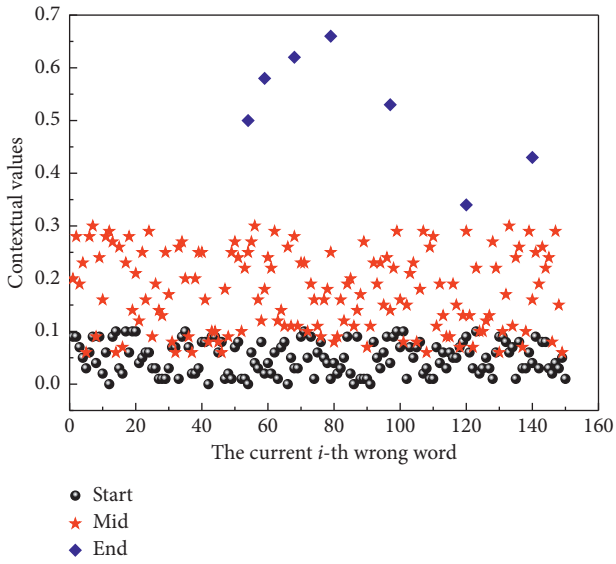

Figure 7: Distribution of speech recognition errors.

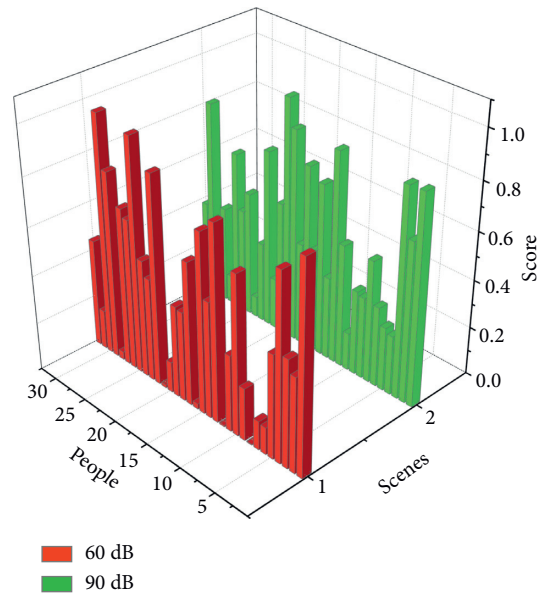

Figure 8: Pronunciation scoring results chart.

some point in time and after cutting to ensure the consistency of the speech fragment, which is convenient for subsequent comparison with the template voice. We add a silent region of 1 s before and after each speech segment, as shown in Figure 8.

After noise reduction, speech recognition is performed. The word error rate of the text has been significantly reduced. This shows that audio quality has a great influence on the speech recognition system, particularly, for some very low quality, high quality, and low-cost products. The RNN noise reduction can effectively compensate for the noise in the examination room and background noise of the recording device, which can greatly improve the robustness of the whole multifeature intelligent grading model. As mentioned above, this paper conducted a text cleaning experiment to compare the vocabulary and grammar scores of the automatic marking system before and after text cleaning, as shown in Figure 9.

As can be seen in Figure 9, both the vocabulary score and the grammar score improved after the text cleaning, but the improvement in the vocabulary score was more pronounced because after we corrected the text, there were no more misspelled words in the text, so the vocabulary score went up considerably. For grammar scores, our text cleaning did not change the grammatical structure of the original text, so grammar scores increased only slightly. Figure 10 shows the word error rate of our text before and after text cleaning of 100 manually transcribed utterances, as well as the average word and grammar score before and after cleaning and the comparison between the mean lexical and grammatical score before and after the cleaning. Laplacian Gaussian operator is a kind of second-order derivative operator, which will produce a steep zero crossing at the edge. Laplacian operator is isotropic and can sharpen boundaries and lines in any direction without direction sex. This is the biggest advantage that distinguishes Laplacian operator from other algorithms.

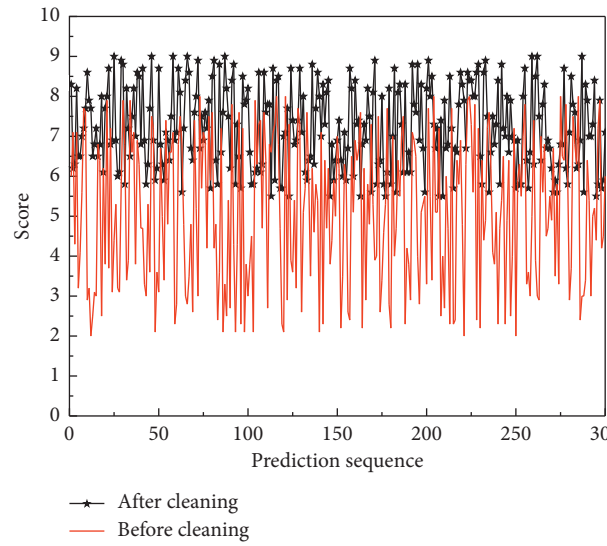

Figure 9: Word scores before and after washing.

The boost classification model has better accuracy, precision, recall, and F-value score than the support vector machine and decision tree models. Finally, the experiments are conducted using the boost classification model in which the model has better accuracy, precision, recall, and F-value score than the support vector machine and decision tree models. The accuracy rate reflects the percentage of cases that are predicted to be positive by the model and the percentage of cases that are predicted to be positive by the original sample. In this paper, the sample size of the test set in which the model identifies the similarity between the reference answer score text and the student-response text is 82.68%, indicating that the model has good accuracy in predicting the similarity between the reference answer and the student response. The recall rate reflects the fact that the

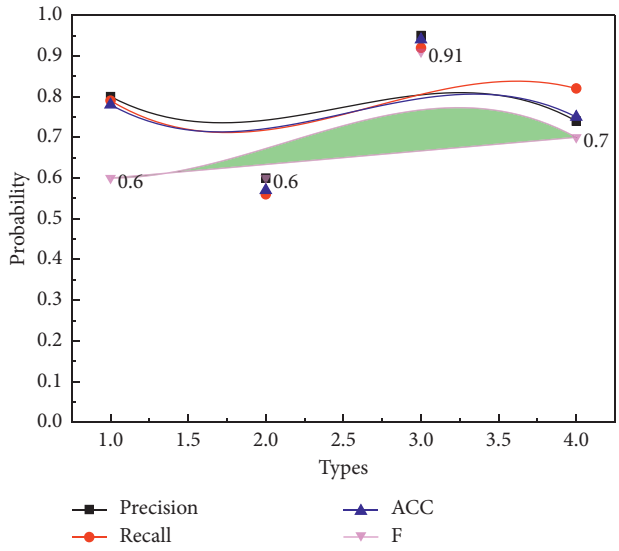

Figure 10: Comparison of classification model effects.

model correctly identified only 60% of similar references and student-response text pairs, indicating that the model is not very good at identifying manually marked text. This may be due to a possible bias in the score of the manually marked reference answers. The F-value is a weighted summed average of recall and accuracy, which reflects a combination of accuracy and recall.

4. Conclusion

When the test and training speech lengths are sufficient, speaker recognition can achieve a good recognition result. In everyday life, the length of the speaker's test speech tends to be short. Currently, to make speech recognition technology more user-friendly, researchers have started to focus on short speech speaker recognition. This paper focuses on the detailed analysis and research on feature extraction and model selection for short speech speaker recognition, mainly using a variety of features as input features of the acoustic model and experimental simulation through the built short speech speaker recognition system, and finally verifying the effectiveness of the improved short speech speaker recognition algorithm. The selection of the matching model has a direct impact on the performance of the speaker recognition system. The feature parameters extracted from the speaker's speech signal can be used for speaker identification only when the corresponding speaker model is constructed. Since the GMM-UBM model cannot suppress the interference of channel variation on speaker recognition and the length of the test speech is less than 10 s, the extracted feature parameters do not fully fit the speaker feature distribution, and the I-Vector-based speaker recognition model does not strictly distinguish between speaker variation information and channel variation information. The I-Vector model also reduces the computational complexity to some extent by transforming the high-dimensional feature space into a low-dimensional vector for study. Experiments demonstrate that the I-Vector-based speaker recognition system has better robustness in testing short speech conditions.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest reported in this paper.

References

[1] M. S. Mahmud, H. Fang, and H. Wang, "An integrated wearable sensor for unobtrusive continuous measurement of autonomic nervous system," *IEEE Internet of Things Journal*, vol. 6, no. 1, pp. 1104–1113, 2018.

[2] J. Pagan, R. Fallahzadeh, M. Pedram et al., "Toward ultra-low-power remote health monitoring: an optimal and adaptive compressed sensing framework for activity recognition," *IEEE Transactions on Mobile Computing*, vol. 18, no. 3, pp. 658–673, 2019.

[3] S. Chatterjee, S. Sarkar, S. Hore et al., "Structural failure classification for reinforced concrete buildings using a trained neural network based multi-objective genetic algorithm," *Structural Engineering and Mechanics*, vol. 63, no. 4, pp. 429–438, 2017.

[4] D. J. Miller, Z. Xiang, and G. Kesidis, "Adversarial learning targeting deep neural network classification: a comprehensive review of defenses against attacks," *Proceedings of the IEEE*, vol. 108, no. 3, pp. 402–433, 2020.

[5] Y.-H. Lai, Y. Tsao, X. Lu et al., "Deep learning-based noise reduction approach to improve speech intelligibility for cochlear implant recipients," *Ear & Hearing*, vol. 39, no. 4, pp. 795–809, 2018.

[6] Z. Gobi, C. Zhou, and A. Wieser, "F2S3: robustified determination of 3D displacement vector fields using deep learning," *Journal of Applied Geodesy*, vol. 14, no. 2, pp. 177–189, 2020.

[7] K. A. Anderson, "Skill networks and measures of complex human capital," *Proceedings of the National Academy of Sciences*, vol. 114, no. 48, pp. 12720–12724, 2017.

[8] A. Maier, C. Syben, T. Lasser, and C. Riess, "A gentle introduction to deep learning in medical image processing," *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 86–101, 2019.

[9] T. T. Q. Bui, N. T. Thang, and T. H. Le, "A robust PCA-SURE thresholding deep neural network approach for mental task brain computer interface," *Journal of Informatics and Mathematical Sciences*, vol. 11, no. 3-4, pp. 383–406, 2019.

[10] X. Wang, Y. Han, V. C. M. Leung, D. Niyato, X. Yan, and X. Chen, "Convergence of edge computing and deep learning: a comprehensive survey," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 2, pp. 869–904, 2020.

[11] R. Wason, V. Jain, G. S. Narula, and A. Balyan, "Deep understanding of 3-D multimedia information retrieval on social media: implications and challenges," *Iran Journal of Computer Science*, vol. 2, no. 2, pp. 101–111, 2019.

[12] R. Haeb-Umbach, S. Watanabe, T. Nakatani et al., "Speech processing for digital home assistants: combining signal processing with deep-learning techniques," *IEEE Signal Processing Magazine*, vol. 36, no. 6, pp. 111–124, 2019.

[13] E. Burbage, N. B. Anuar, A. W. Abdul Wahab et al., "An overview of audio event detection methods from feature extraction to classification," *Applied Artificial Intelligence*, vol. 31, no. 9-10, pp. 661–714, 2017.

[14] A. Metros, T. Heittola, E. Benetos et al., "Detection and classification of acoustic scenes and events: outcome of the DCASE 2016 challenge," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 2, pp. 379–393, 2017.

[15] D. M. Denmark, E. A. Holm, and S. R. Niezgoda, "Perspectives on the impact of machine learning, deep learning, and artificial intelligence on materials, processes, and structures engineering," *Integrating Materials and Manufacturing Innovation*, vol. 7, no. 3, pp. 157–172, 2018.

[16] R. Vishwakarma and A. K. Jain, "A survey of DDoS attacking techniques and defence mechanisms in the IoT Network," *Telecommunication Systems*, vol. 73, no. 1, pp. 3–25, 2020.

[17] A. Massa, G. Oliveri, M. Salucci, N. Anselmi, and P. Rocca, "Learning-by-examples techniques as applied to electromagnetics," *Journal of Electromagnetic Waves and Applications*, vol. 32, no. 4, pp. 516–541, 2018.

[18] L. Möckl, A. R. Roy, and W. E. Moerner, "Deep learning in single-molecule microscopy: fundamentals, caveats, and recent developments," *Biomedical Optics Express*, vol. 11, no. 3, pp. 1633–1661, 2020.

[19] L. Deng and D. Li, "Multimedia data stream information mining algorithm based on jointed neural network and soft clustering," *Multimedia Tools and Applications*, vol. 78, no. 4, pp. 4021–4044, 2019.

[20] L. Shi, J. K. Nielsen, J. R. Jensen, M. A. Little, and M. G. Christensen, "Robust Bayesian pitch tracking based on the harmonic model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 11, pp. 1737–1751, 2019.

[21] J. Calvo-Zaragoza, J. H. Hajič Jr, and A. Pacha, "Understanding optical music recognition," *ACM Computing Surveys*, vol. 53, no. 4, pp. 1–35, 2020.

[22] N. F. Faust, A. S. M. Jaya, M. I. Jarrah, and H. S. Akbar, "Thin film roughness optimization in the TiN coatings using genetic algorithms," *Journal of Theoretical and Applied Information Technology*, vol. 95, no. 24, pp. 6690–6698, 2017.

[23] X. Guo and N. Ansari, "Localization by fusing a group of fingerprints via multiple antennas in indoor environment," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 11, pp. 9904–9915, 2017.

[24] M. Norman, E. Namjoo, and S. Mohammadi, "Trust classification in social networks using combined machine learning algorithms and fuzzy logic," *Iranian Journal of Electrical and Electronic Engineering*, vol. 15, no. 3, pp. 294–309, 2019.

[25] L. Liu, O. de Vel, Q.-L. Han, J. Zhang, and Y. Xiang, "Detecting and preventing cyber insider threats: a survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 2, pp. 1397–1417, 2018.

[26] M. Azad-Manjiri, A. Amiri, and A. Saleh Sedghpour, "ML-SLSTSVM: a new structural least square twin support vector machine for multi-label learning," *Pattern Analysis and Applications*, vol. 23, no. 1, pp. 295–308, 2020.