Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2024 SCPE. Volume 25, Issues 2, pp. 812–823, DOI 10.12694/scpe.v25i2.2645

A STUDY OF BLOCKCHAIN AND MACHINE LEARNING-ENABLED IOT SECURITY IN TIME-DELAYED NEURAL NETWORK VOCAL PATTERN RECOGNITION TO IMPROVE WEB-BASED VOCAL TEACHING

KAIYI LONG*

Abstract. With the development of information technology, online vocal teaching is becoming more and more popular, but the sound quality of teaching is also becoming more and more demanding. As online vocal instruction becomes more popular, the need for high-quality sound in these digital environments becomes more critical. This research tackles the problem of improving sound quality in real-time vocal teaching by integrating advanced technologies such as Blockchain and Machine Learning within the Internet of Things (IoT) security framework. We created a vocal recognition model using Time-Delay Neural Network (TDNN) and improved it with Generated Feature Vector (GFV). This integration yields a strong GTDNN vocal recognition system that is specifically designed to secure and optimize web-based vocal teaching. Our experiments show that GTDNN outperforms traditional TDNN and i-vector methods in feature vector extraction, adapting well to different speech environments. In various speech settings, GTDNN's Error Rates (EERs) are impressively low at 11.3%, 12.0%, 4.9%, 6.2%, and 6.1%, indicating superior performance over comparison models. GTDNN has an EER of 9.6% for short-duration speech and 2.3% for long-duration speech. Furthermore, the GTDNN system achieves an overall pass rate of 94% for target speech and an impressive rejection rate for non-target speech, ensuring high accuracy in a variety of speech environments.

Key words: TDNN; vocal recognition; network vocal teaching; sound quality fidelity

1. Introduction. At present, the application of network technology in teaching is one of the forms of modernization of education and teaching in China. This form is also reflected in vocal music teaching, which makes it possible to change from traditional offline teaching to online teaching, which greatly increases the flexibility of teaching and the richness of teaching forms. In the process of online vocal teaching, there are inevitably background sounds and other people's voices mixed in the teacher's teaching voice, so online sound quality fidelity technology is needed to improve the sound quality. Voice recognition can identify and distinguish teachers from other voices by extracting their vocal features, thus achieving sound quality fidelity [3, 19]. Voice is a kind of information carrier for communication between people, and voice contains a variety of information, for example, in daily life, people can be distinguished by their voice alone. This is due to the fact that the vocal pattern of each person's voice is different. Therefore, vocal recognition technology can be applied to the task of recognizing speakers. In short, it is the extraction of features in a speaker's voice that characterize the speaker's voice and the analysis and classification of these identity features to distinguish target and non-target speakers [5, 16].

At its core, blockchain is a distributed database that keeps a constantly growing list of records, known as blocks, that are linked and secured using cryptography. Each block contains a cryptographic hash of the previous block, as well as a timestamp and transaction data, resulting in an unchangeable chain of records. This ensures that once a record has been added to the chain, it cannot be changed retroactively without affecting all subsequent blocks, which requires network majority consensus. Because blockchain is decentralized, there is no need for a central authority or intermediary, making it naturally resistant to fraud and corruption. Its applications span a wide range of industries, including finance, healthcare, supply chain management, and others, where it promises increased transparency, security, and efficiency. Vocal recognition has a wide range of application scenarios as scientific research continues to deepen [6]. Especially in recent years, with the continuous development of machine learning theories, machine learning techniques have made new breakthroughs in natural language processing and data mining. In recent years, with the significant breakthrough of deep learning

*The School of Film and Television Art, Hunan Mass Media Vocational and Technical College Changsha, 410100, China (kaiyilong@outlook.com)

theory in speech recognition, more and more researchers have started to use it for voiceprint recognition. There are numerous approaches to use deep learning techniques for voice recognition, but most of the current systems using deep learning still embed neural network models into the identity vector (i-vector) recognition benchmark framework based on Probabilistic Linear Discriminant Analysis (PLDA) models model at [6]. However, the recognition rate of short-time speech is low, and it is easy to be disturbed by noise. Time-Delay Neural Network (TDNN) is a model using multi-layer convolutional neural network, which can carry out convolution operations on both time axis and frequency axis, and has stronger robustness. In order to improve the recognition rate and stability of the voicing system, and strengthen the robustness against noise, this research will build a vocal recognition method based on TDNN to accurately identify the teacher's voice in online lessons and improve the quality of vocal teaching. The combination of these technologies — blockchain for security, machine learning for pattern recognition, and IoT for connectivity — has the potential to significantly improve the dependability and efficiency of web-based vocal teaching platforms. This study aims to improve the overall experience and outcomes of digital vocal education by ensuring secure data exchanges and accurate vocal recognition.

2. Literature Review. In the field of acoustics teaching, researchers have produced results on the reform of teaching mode. Yang h analyzes the quality of multimedia teaching in vocal music class by using elite teaching optimization algorithm to improve the level of acoustic teaching. The main point of view is the teacher's teaching ideas, the level of information-based teaching design, the level of new teaching model, teaching effects and other factors that affect the quality of teaching. The results show that the algorithm is effective and can be used to evaluate the quality of vocal music multimedia teaching [20]. Fu L thinks that the vocal music teaching is changing from teacher-centered to student-centered. The teaching emphasizes the development of students and the change of academic conditions, and pays more attention to the combination of theory and practice. Fu L analyzed the actual situation of vocal music teaching to find out the gap between reality and theory, and thus formulated a more scientific teaching method to improve the teaching model. The study provides a theoretical teaching plan for improving the quality of acoustic teaching and the acoustic level of students [3]. [16], in this paper, a method of system resource allocation based on power iteration is proposed to optimize the resource allocation of vocal music teaching system. The method takes the throughput of the unloading process as the objective function, and achieves the optimal allocation of normal power through iterative optimization. At the same time, a heterogeneous network based on edge server is proposed to improve the low energy efficiency and resource utilization of edge server. The experimental results show that the method is effective and practical.

Academic research on TDNN has produced many results. hu S et al. used TDNN optimized by neural structure search (NAS) based techniques for speech recognition tasks. the NAS was used to automatically learn two hyperparameters of factorized TDNN, namely left-right splicing and contextual offset, and the linear projection dimension of each hidden layer, allowing TDNN to perform in different systems through parameter sharing effective search. Experimental results show that its word error rate is large and model size is greatly reduced, and speech recognition performance is improved [9]. [1] used TDNN to predict the active power demand on a P4 bus in President Prudente. Experimental results demonstrated its validity [11]. [21]used TDNN as a facial expression classifier for an intelligent robot to establish command laws by analyzing and recognizing facial expressions to translate expressions into robot-recognizable language. Experimental results verified its effectiveness and improved the efficiency of recognition. [2] used TDNN and normalization methods for optimizing an automatic speech recognition system. The experimental results showed that the recognition error rate of the optimized system was greatly reduced and its acoustic and language model recording speed was significantly accelerated.

The paper [11] investigated the performance of deep neural networks, convolutional neural networks, temporal convolutional neural networks, and TDNN for English dialect classification. The results showed that TDNN and ECAPA-TDNN classifiers capture a wider temporal context, further improving the performance of the classification models. TDNN improved the performance of SPEC-STFT and SPEC-SFF by 2.8% and 1.4%, respectively. [15] applied focal time delay neural network (FTDNN) to ECG lead set accurate reconstruction. The experimental results showed that the FTDNN method reconstructed leads with correlation values between 0.8609 and 0.9678 and root mean square error values between 123 μV and 245 μV in all cases except for individual subgroups, outperforming other methods and improving the accuracy of ECG leads. [12] applied a modified TDNN to monitor the flow of granular solid material in oil pipes. The method can capture dynamic

acoustic signals. Experimental results showed that its normalized root mean square error for monitoring solid flow rate, solid concentration, pipeline pressure drop and gas velocity were 0.18, 0.17, 0.20 and 0.16, respectively. its performance provides a simple and reliable real-time monitoring system due to the artificial neural network model. [14] used TDNN to build a three-level coding system to decode the signal when rats act. The study first recorded and processed M1 signals from three behavioral types of rats: walking, standing and head shaking, and then divided the signals according to response time points and analyzed them as independent components to extract signal features. Finally, the study finds 16 representative sample signals by dynamic dimension increase algorithm to input a three-level TDNN for training. The results showed that the recognition accuracy of TDNN for these three actions was 51.4%, 80.0% and 54.3%, respectively, which showed that the three-stage TDNN coding model could explain the M1 brain signals of free-motor animals [12].

To sum up, TDNN has rich application results in classification tasks, signal recognition and speech recognition, but few studies have applied it to voice print recognition and vocal music teaching tasks. TDNN can be used to improve the recognition accuracy rate and stability in noisy environment by its strong robustness. This research will apply TDNN in the field of vocal recognition, design a vocal recognition system based on TDNN that can be used for web-based vocal teaching, and provide a better performance method for online lesson sound quality fidelity.

3. Vocal Sound Quality Fidelity Method Based On Voiceprint Recognition With Time Delay Neural Network.

3.1. Design of voice-print recognition model based on time-delay neural network. Vocal recognition refers to the process of extracting features that can represent the identity of a speaker from one or more segments of speech and comparing them with the identity features extracted from the speech of a known speaker to achieve confirmation of the identity of the speaker. Its main processes are specifically as follows: preprocessing of speech signals, feature extraction of speech signals, training and building speaker models, and scoring the unknown speaker against the feature parameters of the known speaker [10]. Vocal recognition feature extraction is a key step in vocal recognition, and it is necessary to obtain the long time dependent nature of speech to obtain a time invariant acoustic model of this nature. TDNN can obtain the long time dependent nature of speech, and can extract the relationship between each feature value related to the time series, and take into account the characteristic factor of the change of feature value due to time change [13]. Therefore, TDNN is chosen for this study to perform vocal feature extraction. TDNN is a feedforward neural network that is computed using interconnected layers composed of many nodes, and its basic structure is shown in Figure 3.1. In the figure, a_i, a_j is the input vector of i dimension and j dimension, w is the connection weight, s_1, s_t is the delay vector, and t is the delay time. When $t = 2$ is used, it means that the current time and the vectors of the two preceding and following frames are combined together and fed into the network. The core idea of TDNN is weight sharing, i.e., the weights of the same parts are the same. Specifically, when $t = 2$, there will be 3 weights in the neural network, which are given to the current time frame and the two frames before and after the current time.

The activation function can make the neural network have stronger classification ability. ReLU function and Softmax function are both commonly used activation functions. Among them, the ReLU function can effectively overcome the problem of gradient disappearance and can transfer the error better than the Sigmoid function. Therefore, the ReLU function is chosen as the activation function of TDNN in this study, and its formula is shown in equation 3.1.

$$\mathrm{Max}(0, x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x \geq 0 \end{cases} \tag{3.1}$$

For the output classification function of TDNN, the Softmax classification function was chosen for the study, and its formula is given in equation 3.2.

$$y_i = \frac{\exp(x_i)}{\sum \exp(x_i)} \tag{3.2}$$

To design a vocal pattern extraction model based on TDNN, the model can be divided into two modules: speaker identity feature vector extraction and back-end scoring module. The former can utilize Mel-Frequency

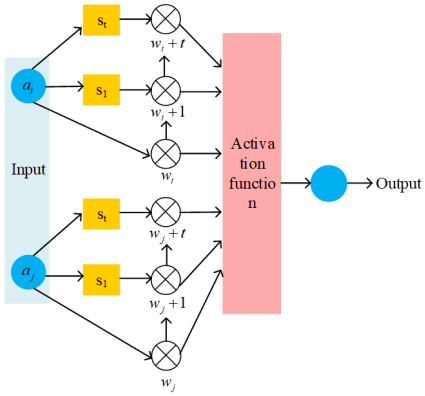


Fig. 3.1: Basic structure diagram of TDNN network

Cepstral Coefficients (MFCCs) features of fixed speech signals as input features for TDNN. each delay layer in TDNN reads the information features in the speaker's speech while expanding the context of the frame, which facilitates capturing the speaker's identity information. The delay layer is followed by a pooling layer, where TDNN collapses the features along the time axis by computing the mean and standard deviation statistics of the features to obtain a feature vector containing global speaker information. In order to get the speaker embedding correct, the pooling layer is followed by a fully connected layer. The back-end scoring module can be chosen from the PLDA model. The scoring principle of the PLDA model is to assume that the training data speech consists of speaker's speech, where each speaker has different segments of his or her own speech. Then, the b speech of the a speaker is defined as $x_a b$. Then, the generative model of can be defined as equation 3.3 based on the factor analysis.

$$x_{ab} = \mu + Fh_a + Gw_{ab} + \epsilon_{ab} \tag{3.3}$$

In equation 3.3, μ is the numerical mean, F and G are two spatial feature matrices, represents the feature representation of in speaker space, and is the noise covariance. In the recognition scoring stage, if the likelihood of two voices having the same h_a features is greater, then the two voices belong more definitely to the same speaker. The degree of likelihood is used to calculate the score by the log-likelihood ratio, which is given in equation 3.4.

$$S = \log\left(\frac{p(a_1, a_2 \mid T_1)}{p(a_1 \mid T_2)p(a_2 \mid T_2)}\right) \tag{3.4}$$

In equation 3.4, S is the log-likelihood ratio. T_1 and T_2 are two hypothetical events, i.e., the vector a_1, a_2 belongs or does not belong with a speaker. p represents the probability of the event. the PLDA model is based on two basic assumptions: first, that the channel influence is independent of the speaker; and second, that the total variability factor of the speaker cannot fluctuate excessively. However, these two assumptions are not fully valid. Therefore, the study improves it by eliminating the process of solving the spatial feature matrix G in equation 3.3, without reference to the differences between different speech sounds of the same speaker, to obtain a simplified PLDA model. The equation of the improved model is given in Equation 3.5.

$$x_{ab} = \mu + Fh_a + \epsilon_{ab} \tag{3.5}$$

The training phase of the neural network requires the selection of a suitable loss function, and this time the softmax cross-entropy function is chosen as the loss function, the formula of which is given in equation 3.6.

$$E = \sum_{n=1}^{N}\sum_{i=1}^{n} y_i \log\left(\frac{e^{z_i}}{\sum_k e^{z_k}}\right) \tag{3.6}$$

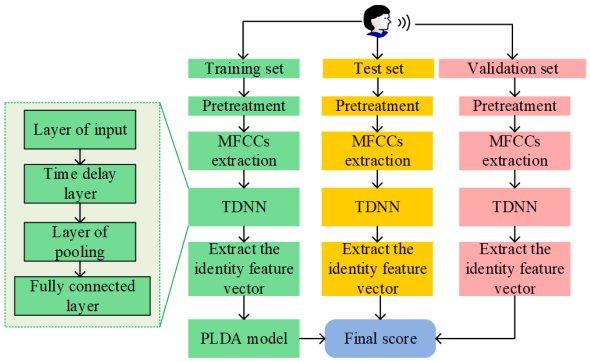


Fig. 3.2: Process of voice print recognition model based on TDNN

In equation 3.6, n is the number of data samples, and each data x_n contains m frames of sound feature data. If the speaker's label is now i, then y_i is 1; otherwise, it is 0. z_i denotes the input value of the first i node, and z_k denotes the input values of other nodes. Combined with the above design, the obtained time-delay neural network-based voice recognition model flow is shown in Figure 3.2. In the training phase, the speech data of the training set is inputted into the time-delay neural network with the extracted MFCCs after speech preprocessing and Mel cepstral coefficient extraction, and the output data of the softmax layer is used for the training of the PLDA back-end. In the testing phase, it is also necessary to first perform the speech preprocessing operation on the test set data to extract the MFCCs in the voice and input the MFCCs into the time-delay neural network, unlike the test phase, the feature vectors extracted by the neural network are presented after the hidden layer7 of the time-delay neural network and used directly for scoring the PLDA model.

3.2. Model optimization and identification system construction. TDNN considers multi-frame features, which enhances the complexity of the computation. Therefore, the study introduces the Generated feature vector (GFV) to improve it. The basic principle of the Generated feature vector is to perform correlation analysis between the Identity authentication vector (i-vector) and the extracted vectors of the TDNN model using Canonical correlation analysis (CCA) to generate a new feature vector. CCA allows the two feature vectors to learn from each other. To achieve voice recognition, it is necessary to find an invisible vector space in which each speaker is a point, i.e., a set of basis vectors of this vector space can be used to represent this speaker [18]. The identity feature vector i-vector is the vector belonging to this vector space that can characterize the speaker. In the registration and testing phases, the i-vector will be excluded from the system and only the feature vector extracted by the TDNN network will be considered and linearly transformed using the transformation matrix obtained by typical association analysis methods. This paper utilizes this transformed output as a generative feature vector, which captures some properties of the identity feature vector extracted by the i-vector model during the transformation process. The steps are shown in Figure 3.3. After inputting the speech, the matrix based on the identity vector W_a and the matrix extracted by TDNN W_b are transformed by using CCA learning, and then this transformation matrix is used for generative feature vector extraction.

The matrix W_a and the matrix W_b need to satisfy the constraints of equation 3.7.

$$\max_{W_a, W_b} \operatorname{corr}(W_a \phi_a, W_b \phi_b) \tag{3.7}$$

In Eq. 3.7, ϕ_a is the feature vector extracted from the i-vector and ϕ_b is the feature vector extracted from the TDNN of the same speech. The CCA transform can transfer the information from the i-vector model to the TDNN model and vice versa. The expression of the generative feature vector is given in equation 3.8.

$$\phi_g = W_b \phi_b \tag{3.8}$$

In equation 3.8,ϕ_g is the generative feature vector and also the feature vector extracted by TDNN with i-vector features. i-vector and TDNN networks extract feature vectors that are both zero-centered, but the

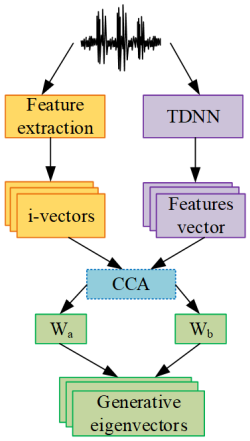


Fig. 3.3: Schematic diagram of GFV principle

dimensionality of the two is different. Applying CCA can maximize the correlation between the two. the principle of CCA is that if there are two random vectors $X = \{x_1, x_2, ...\}^T$ and $Y = \{y_1, y_2, ...\}^T$, CCA can redefine the new variables M and N by a linear combination of X and Y. M and N see equation 3.9.

$$\begin{cases} M = a^T X \\ N = b^T Y \end{cases} \tag{3.9}$$

In Eq. 3.9, a^T and b^T are the two parameters that CCA needs to find to maximize the correlation coefficient of the two vectors. The expression of the correlation coefficient r is given in Eq. 3.10

$$r = \operatorname{corr}\left\langle a^T X, b^T Y \right\rangle = \frac{E(a^T X b Y^T)}{\sqrt{E(a^T X a X^T)}\sqrt{E(b^T Y b Y^T)}} \tag{3.10}$$

In equation 3.10, E denotes the unit matrix. The constraints of this equation are shown in equation 3.11.

$$\begin{cases} a^T \sum_X a = 1 \\ b^T \sum_Y b = 1 \end{cases} \tag{3.11}$$

In equation 3.11, $\sum_x$ and $\sum_T$ are the covariances of the matrices. $\sum_x$ The expressions are given in Eq. 3.12.

$$\sum_X = E(X^T X) \tag{3.12}$$

For $\sum_Y$ the expression of is shown in equation 3.13.

$$\sum_Y = E(Y^T Y) \tag{3.13}$$

The relevant parameters for the maximization are given in equation 3.14.

$$r = a^T \sum_{XY} b = a^T E(XY^T) b \tag{3.14}$$

In CCA, the aim is to find mutually orthogonal pairs of the maximum correlated linear combinations of variables in X and Y. In the study, X and Y refer to the corresponding i-vector and TDNN network extracted

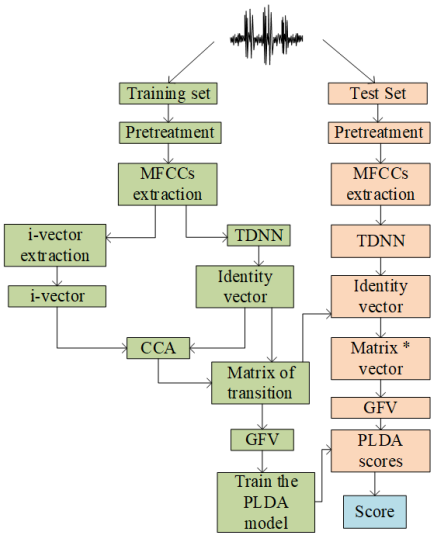


Fig. 3.4: Architecture of GTNN-based voice recognition system

feature vectors of the same speech segment, respectively. After introducing GFV into the voicing recognition model, a voicing recognition system based on time-delay neural network is designed. The system is developed in Python language under Linux environment and uses Advanced RISC Machine (ARM) processor for data processing. The system first obtains the audio from peripherals and stores it in the memory unit, then drives the AXI-DMA unit to transmit the data to the on-chip processor after processing by the processor, then uses the ARM processor to control the CNN hardware accelerator to read the features and weights, and finally sends the structure back to the on-chip memory and then sends it back to the memory via AXI_DMA. Process the output further. The processing flow of the system for voiceprint recognition is shown in Figure 3.4

As shown in Figure 3.4, the system only has the extraction process for i-vectors in the training phase, but not in the test set, which reduces the amount of operations in the system. In the training phase, the i-vector feature extraction module is introduced mainly to calculate the transformation matrix W_a, which is used for generative feature vector generation. In the training phase of the model, the speech in the training set is first preprocessed to extract the acoustic parameters MFCCs in the speech, and then the i-vector of the speech is extracted and the TDNN network is used to extract the feature vectors of the speech, and two transformation matrices W_a and W_b are found using CCA to make the maximum correlation between the two vectors. One of the matrices W_b will be used to generate the test phase generative vectors. In the test phase, the speech is also preprocessed first to extract the acoustic parameters MFCCs in the speech, and the MFCCs extracted from the speech are input to the TDNN network as the input layer of the neural network. The TDNN network extracts the feature vectors that represent the identity of the speaker, and multiplies the feature vectors with the transformation matrix W_b calculated in the training phase to calculate the generative feature vectors in the speech. Finally, the generative feature vectors are input to the back-end scoring PLDA model for scoring, and the scoring results are used to determine whether the speech belongs to the target speaker to achieve the purpose of voice recognition.

The use of TDNN for vocal feature extraction is a significant step forward in vocal recognition technology. TDNN can extract time-invariant acoustic models, which are critical for identifying and verifying speaker identities with high accuracy, by capturing the long time-dependent nature of speech.

4. Analysis of Experimental Results of GTNN-based Voice Recognition System.

4.1. Experimental environment and parameter determination. The environment for this experiment is shown below: the operating system is Linux Ubuntu18.0464, the GPU is NVIDIA TeslaP100, the CPU

Table 4.1: Basic parameters of sound library

Voice Library	Total number of speakers	Number of men	Number of women	Total number of voices	Number of voices per capita	Recording Environment	Total time/h
TIMIT	630	440	192	6300	10	Pure	3.96
Aishell	402	188	210	141600	355	Pure	177
VoxCeleb	1250	690	566	153520	120	Contains noise	353

Table 4.2: Experimental results of frame level layer delay configuration

Program Number	Delay layer number						loss	EER
	1	2	3	4	5	6		
1	{T, T+2}	{T-1, +3}	{T-2, T+4}	{T-3, T+6}	{T-4, T+6}	{T-5, T+7}	0.33	0.123
2	{T, T+2}	{T, T+2}	{T-1, T+3}	{T-1, T+3}	{T-2, T+4}	{T-2, T+4}	0.30	0.164
3	{T, T+2}	{T-1, T+3}	{T-1, T+3}	{T-2, T+4}	{T-2, T+4}	{T-2, T+4}	0.22	0.160
4	{T, T+2}	{T-1, T+3}	{T-1, T+3}	{T-2, T+4}	{T-2, T+4}	{T+1}	0.19	0.131
5	{T-1, T+3}	{T-1, T+3}	{T-2, T+4}	{T-2, T+4}	{T-3, T+5}	{T-3, T+5}	0.15	0.092
6	{T-1, T+3}	{T-1, T+3}	{T-2, T+4}	{T-3, T+5}	{T+1}	{T, T+1}	0.13	0.072

is Intel(R) Xeon(R) CPU is E5-2697V424, and the running memory is 96G. the software is developed using Python. The speech recognition data are obtained from existing publicly available voice databases, including the English dataset TIMIT voice library, VoxCeleb voice library and Chinese voice library Aishell, and the noise database is selected from MUSAN. the MUSAN dataset includes 929 various noise files with a sampling frequency of 16 KHz. all audio files are in WAV format. The basic parameters of other speech libraries are shown in Table 4.1. the TIMIT speech database has a sampling frequency of 16 kHz and includes a total of 630 target speakers, each recording 10 voices. voxCeleb is an audiovisual dataset covering video and audio data, and only the speech data are taken in the storage experiment. the Aishell speech library is a Chinese speech library covering various types of factual information.

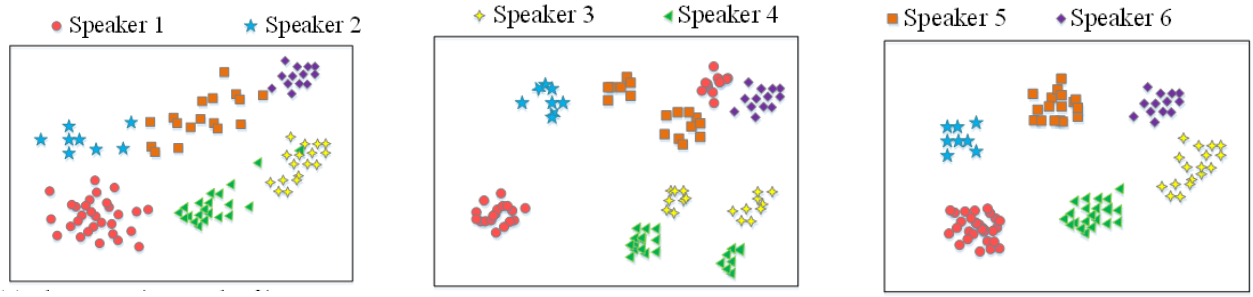
The data of the speech library is divided into training set, registration set and test set, the training set is used to train TDNN and PLDA models, the registration set is used as the reference data set for recognizing voice patterns, and the test set is used to test the system performance. The Equal Error Rate (EER) is used as the main evaluation index of recognition effect. When the EER is larger, it means that the recognition system is less effective, and when the EER is smaller, it means that the recognition system is more effective.

Before starting the experiments, preliminary tests are needed to determine the optimal parameters of the TDNN [23, 24, 25]. The parameters to be determined include the delay time of each layer of the TDNN and the position of the extracted feature vectors. The first 6 layers of the TDNN model are the delay layer. In order to verify the effect of the different delay time of each layer on the experimental results and to find the optimal selection scheme, 6 schemes are tested in this experiment and the EER value as well as the loss value of each scheme are calculated. The experimental results of the delay configuration of the frame-level layers obtained are shown in Table 4.2.From the above table it is easy to see that the recognition effect of the delayed layer delay scheme selection of scheme 6 is optimal, so the delay parameters of each layer of the hidden layer of the delayed neural network are the same as in experiment 6.

In order to extract feature vectors with good distinguishability and represent the identity of the speaker, the effect of extracting feature vectors after fully connected layer 1 and fully connected layer 2 is tested for voice recognition, respectively. The dimensionality of the feature vector corresponds to the number of nodes in the fully connected layer. In order to extract the appropriate dimensionality of the feature vector, the effect of the number of nodes in the fully connected layer on the recognition results was tested with 256, 350 and 512 nodes, respectively. The VoxCeleb speech library was used for both testing and training, and the scoring backend was a PLDA model. The final model recognition results EER are shown in Table 4.3 below: from Table 4.3, it can

Table 4.3: Experimental results of feature vector extraction location

/	Feature vector extraction location	
Number of nodes	Fully connected layer 1	Fully connected layer 2
256	0.819	0.820
352	0.800	0.770
512	0.700	0. 728



(a) The extraction result of i-vector (b) The extraction result of TDNN (c) GTDNN

Fig. 4.2: Classification effect of extracted feature vectors

be seen that the best recognition results are obtained when the number of nodes is 512 and the feature vectors are extracted after the fully connected layer 1.

Therefore, the TDNN's delay parameter is chosen as option 6, and the feature vector extraction position is fully connected layer 1 when the number of nodes is set to 512.

4.2. Effect analysis of feature vector classification. To be able to visualize the performance progress of GTDNN in extracting feature vectors effect, the speaker model is visualized using t-snet (Distributed Stochastic Neighbor Embedding) high-dimensional data visualization technique to examine the speaker's identity vector feature extraction effect. Six speakers are randomly selected from the database, each with at least thirty speech items, and the feature vectors in the speech are extracted using three methods, i-vector, TDNN, and GTDNN, respectively. Then, to verify the effectiveness of the system, GTDNN is compared with TDNN, the continuous Markov model with real-time embedding from the literature (He and Dong 2020), and the hybrid acoustic model based on PDP coding from the literature [22]. In order to detect the robustness of the recognition system to noise and to compare the recognition effect of different systems, five different voice libraries were set up in this experiment, namely TIMIT voice library, TIMIT voice library with noise, Aishell voice library, Aishell voice library with noise and VoxCeleb voice library. t-sne visualization results are shown in Figure 4.2. The feature vectors extracted by GTDNN have strong differentiability, and the classification effect of its extracted feature vectors is significantly better than the classification effect of the feature vectors extracted by TDNN and i-vector based methods.

4.3. The recognition effect analysis of the system. The EER results for different environments are shown in Figure 6, where nTIMIT and nAishell denote the TIMIT speech bank with noise and the Aishell speech bank with noise. All systems perform better in the pure speech environment than in the noisy environment. In the overall performance, GTNN is the best recognition among several models. the continuous Markov model with GTNN and real-time embedding and the hybrid acoustic model based on PDP coding outperform the traditional TDNN model. Taking the VoxCeleb speech library, which is closest to the real environment, as an example, the EER of GTDNN is 6.1% in the test environment of the VoxCeleb speech library, which is 2.5% lower than the traditional TDNN model. In other speech libraries, their EERs are 11.3%, 12.0%, 4.9% and

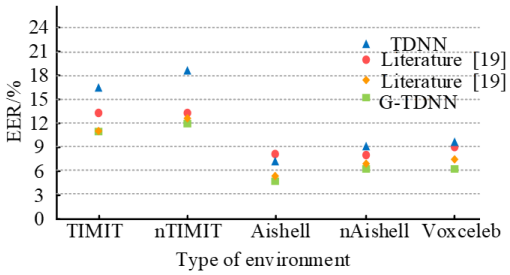


Fig. 4.3: EER results in different environments

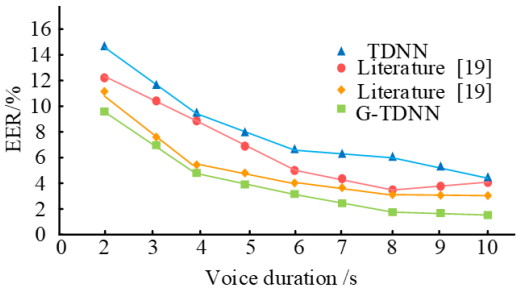


Fig. 4.4: Recognition results with different speech durations

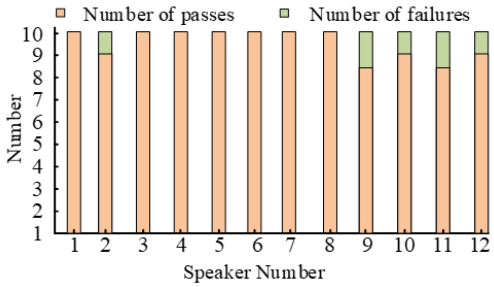
6.2%, which are all at a lower level.

The recognition EER results for different speech durations are shown in Figure 4.4. Overall, it seems that the EERs of all four systems are decreasing as the duration increases, and the recognition effect of each model is getting better. the TDNN model approach is worse for short-time speech of 2 s, and the recognition EER is the largest. the EER of GTDNN is the smallest among the four models for both short-time and long-time speech, with EER value is the largest at 9.6% and the smallest at 2.3%. The EERs of the continuous Markov model with real-time embedding and the hybrid acoustic model based on PDP coding are always in between those of TDNN and GTDNN. It can be seen that GTDNN has better recognition performance regardless of the speech environment and speech duration.

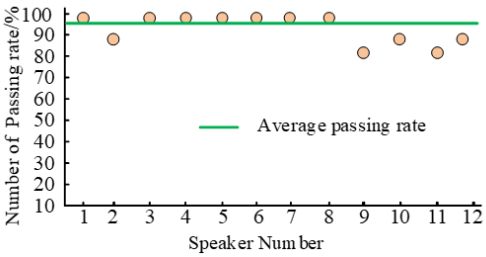
The results of the correct speech recognition pass rate of GTDNN are shown in Figure 4.6, from which it can be seen that the recognition pass rate of most people is above 90%, and the overall recognition pass rate can reach 94%, which can meet the needs of daily human-computer interaction. It can be seen that GTDNN has a high recognition rate and recognition stability.

The rejection rate of detecting non-target speech is shown in Table 4.4, and it can be seen from Table 4 that the rejection rate of GTDNN is relatively high, with an overall rejection rate of 94% and a rejection rate of 90% for most non-target speech, and the stability is also relatively strong, but there is still a risk of misidentification, and in practical applications one can choose to integrate other recognition methods to further improve the recognition accuracy. In summary, the application test using the system proposed in the study has a good recognition performance and the recognition effect is also stable, but there is still a partial false pass, and the evaluation threshold after scoring can be adjusted for different application scenarios in practical applications. In the system with the increase of the judging threshold, the false rejection rate of the system will increase to, while the smaller the judging threshold, the passing rate of correct speech will increase and the false acceptance rate of the system will decrease.

5. Conclusion. In the information age, online lessons have become a new mode of teaching in all majors, and this is also true for vocal majors. However, vocal music teaching requires a higher quality of sound for



(a) Number of passes and failures



(b) Failure Rate

Fig. 4.6: Recognition pass rate results

Table 4.4: Rejection rate of non-target speech

Speaker Number	Number of passes	Number of failures	Rejection rate	Speaker Number	Number of passes	Number of failures	Rejection rate
1	0	10	1.0	8	1	9	0.9
2	0	10	1.0	9	0	10	1.0
3	1	9	0.9	10	2	8	0.8
4	1	9	0.9	11	0	10	1.0
5	2	8	0.8	12	1	9	0.9
6	0	10	1.0	13	0	10	1.0
7	0	10	1.0	Total	8	122	0.94

teaching. Excluding other noise by recognizing the target person's vocal pattern is a feasible way to improve the sound quality of vocal teaching in online lessons. In this study, TDNN is improved using generative feature vectors to obtain a vocal pattern recognition system based on GTNN. The experimental results show that the feature vector extraction of GTDNN is better than TDNN and i-vector methods. The EERs of GTDNN are 11.3%, 12.0%, 4.9%, 6.2% and 6.1% in TIMIT speech bank, TIMIT speech bank with noise, Aishell speech bank, Aishell speech bank with noise and VoxCeleb speech bank, respectively, which are all in the lower level. the EERs of GTDNN for both short-time speech and long-time speech EER is the smallest among all four models, with the largest EER value of 9.6% and the smallest of 2.3%. The recognition pass rate for most of them is above 90%, and the overall recognition pass rate can reach 94%. the rejection rate of GTDNN is relatively high, with an overall rejection rate of 94% and a rejection rate of 90% for most impersonators. It can be seen that GTDNN has more stable recognition performance in various environments and can recognize the target speaker's voice and exclude the interference of other non-target speaker's voice, providing good vocal sound quality for online vocal teaching. Although GTDNN can extract more recognizable speaker voice features, its limitation is that it requires relatively high data volume, long training time for the model, and high equipment requirements. In the follow-up study, further optimization is needed for the learning ability of GTDNN. Future research could concentrate on improving the TDNN algorithms' efficiency and accuracy. Experimenting with different neural network architectures, tuning hyperparameters, or using more advanced machine learning techniques to improve the system's ability to handle diverse vocal patterns and accents could be part of this.

Fundings. The research is supported by: Hunan Philosophy and Social Science Foundation Project "Research on Inheritance and innovation of Suining traditional music culture under the background of integrated development of culture and tourism", Project No.: 18WTC28.

REFERENCES

[1] Bonfim, B., Bratifich, R. & Silva, M. HG Silva. *PREVISO DE CONSUMO DE ENERGIA UTILIZANDO REDE NEURAL COM RETARDO DE TEMPO (TDNN). Colloquium Exactarum.* **12**, 63-70 (2021)

[2] Cheng, G., Zhang, P. & Ji, X. Automatic Speech Recognition System with Output-Gate Projected Gated Recurrent Unit. *IEICE Transactions On Information And Systems E.* **102** pp. 355-363 (2019)

[3] Fu, L. Discussion on the Differences between Theory and Practice in Vocal Music Teaching. region - Educational Research and Reviews. (2021)

[4] Fu, L. Discussion on the Differences between Theory and Practice in Vocal Music Teaching. *Region - Educational Research And Reviews.* **3**, 6-9 (2021)

[5] Fu, L. Research on the Reform and Innovation of Vocal Music Teaching in Colleges. region - Educational Research and Reviews. (2020)

[6] Hanifa, R., Isa, K. & Mohamad, S. review on speaker recognition: technology and challenges. *Computers Electrical Engineering.* **90**, 1-10700 (2021)

[7] Hanifa, R., Isa, K. & Mohamad, S. review on speaker recognition: technology and challenges. *Computers Electrical Engineering.* **90**, 1-10700 (2021)

[8] He, Y. & Dong, X. Real time speech recognition algorithm on Embedded System Based on continuous Markov model. microprocessors and Microsystems. (0)

[9] Hu, S., Xie, X., Cui M. J Deng, S. Liu S. Liu & S. Liu, J. Yu M Geng M Geng X Liu X Liu X Liu H Meng . *Neural Architecture Search For LF-MMI Trained Time Delay Neural Networks.* **30** pp. 1093-1107 (2022)

[10] Kathania, H., Kadiri, S., Alku, P. & Others A formant modification method for improved ASR of children's speech. *Speech Communication.* **136**, 98-106 (2022)

[11] Kethireddy, R., Kadiri, S. & Gangashetty, S. Deep neural architectures for dialect classification with single frequency filtering and zero-time windowing feature representations. *The Journal Of The Acoustical Society Of America.* **151**, 1077-1092 (2022)

[12] Kuda, T. Aminu. *Don , Kuda McGlinchey , Andrew Andrew Cowell Cowell . Acoustic Signal Processing With Robust Machine Learning Algorithm For Improved Monitoring Of Particulate Solid Materials In A Gas Flowline. Flow Measurement Instrumentation.* **65** pp. 33-44 (2019)

[13] Kumar, C. Hybrid models for intraday stock price forecasting based on artificial neural networks and metaheuristic algorithms. pattern Recognition Letters. (2021)

[14] Kuo, T., Lo, R. & Chen, L. Shang-Hsien Cai Shang-Hsien Cai . *Activity Command Encoding Of Cerebral Cortex M.* **32**, 1-20500 (2020)

[15] Smith, G., Heever, D. & Swart, W. The Reconstruction of a 12-Lead Electrocardiogram from a Reduced Lead Set Using a Focus Time-Delay Neural Network. *Acta Cardiologica Sinica.* **37**, 47-57 (2021)

[16] Sun, J. Research on resource allocation of vocal music teaching system based on mobile edge computing. computer Communications. (2020)

[17] Sun, J. Research on resource allocation of vocal music teaching system based on mobile edge computing. *Computer Communications.* **160**, 342-350 (2020)

[18] Tatli, A., Kahvecioglu, S. & Karakoc, H. Time-Series Prediction for Amount of Airworthiness Based on Time-Delay Neural Networks. *Elektronika Ir Elektrotechnika.* **2020**, 28-32 (0)

[19] Tomlinson, J. Music therapist collaboration with teaching assistants for facilitating verbal and vocal development in young children with special British Journal of Music Therapy. (2020)

[20] Yang, H. Quality Analysis of Multimedia Teaching in Vocal Music Class Combining Elitist Teaching-Learning-Based Optimization Algorithm. *Hindawi.* **4781**, 1-65947 (2021)

[21] Zekhnine, C. & Berrached, N. Human-Robots Interaction by Facial Expression Recognition. international Journal of Engineering Research in Africa. 2020. (0)

[22] Zhu, W., Jin, H., Chen J. L Luo, J. Wang J. Wang, J. Wang Q. Lu & Lu, Q. Lu A Li . *A Hybrid Acoustic Model Based On PDP Coding For Resolving Articulation Differences In Low-resource Speech Recognition.* **192** pp. 4 (2022)

[23] Lee, S., Abdullah, A. & Jhanjhi, N. A review on honeypot-based botnet detection models for smart factory. *International Journal Of Advanced Computer Science And Applications.* **11** (2020)

[24] Azeem, M., Ullah, A., Ashraf, H., Jhanjhi, N., Humayun, M., Aljahdali, S. & Tabbakh, T. Fog-oriented secure and lightweight data aggregation in iomt. *IEEE Access.* **9** pp. 111072-111082 (2021)

[25] Gaur, L., Solanki, A., Wamba, S. & Jhanjhi, N. Advanced AI techniques and applications in bioinformatics. (CRC Press,2021)

Edited by: Kumar Abhishek
Special issue on: Machine Learning and Blockchain based Solution for Privacy and Access Control in IoT
Received: Sep 28, 2023
Accepted: Jan 14, 2024