Journal of Science and Engineering Research
Vol. 1, No. 1, 2022, pp. 1-14
eISSN: 2957-7187

Deep Learning-based Analysis of Voiceprint Data Mining

Jacky C.K. Tang

The Chinese University of Hong Kong, Hong Kong
tangchikit@emsd.gov.hk

DOI: http://dx.doi.org/10.56828/jser.2022.1.1.1
Article history: *Received (October 16, 2021); Review Result (November 21, 2021); Accepted (January 2, 2022)*

Abstract: In the information age, the intelligent data mining method represented by deep learning is playing an important role in various fields at present. It is necessary to study how to efficiently use the intelligent data mining method to obtain valuable information from massive information. Open-set voiceprint recognition is realized by intelligent data mining technology. Therefore, it is of great practical significance to achieve rapid and accurate identification of the speaker's identity. Because the traditional voiceprint recognition method has insufficient ability to distinguish the speakers inside and outside the set, it often leads to a high false recognition rate. Mining parameters containing more speakers' personality characteristics and how to calculate the threshold become the bottleneck problems of open set voiceprint recognition. Therefore, this paper adopts the deep confidence network stacked by three layers of restricted Boltzmann machines as the deep acoustic feature extractor. The mel-frequency cepstral coefficients of 24-dimensional basic acoustic features are mapped to 256-dimensional feature space, and the parameters of deep acoustic features containing more speaker's personality characteristics are obtained. Then, an open-set adaptive threshold calculation algorithm is obtained. In this paper, the similarity value of deep acoustic features is calculated by the Gaussian mixture model, and the maximum inter-class variance of the similarity value is calculated by the OTSU algorithm. When the inter-class variance is the maximum, the similarity value is the best threshold. The experimental test shows that the algorithm for calculating threshold based on deep learning proposed in this paper has a lower false rejection rate and lower false rejection rate.

Keywords: Intelligent data mining, Voiceprint recognition, Deep neural network, Support vector data description

1. Introduction

With the development of computer technology and the information industry, human society has entered a brand-new information age and people's ability to generate, collect, store and process data has been greatly improved, and data usually contains a lot of useful information. Under this background, people are eager to make a more comprehensive and in-depth analysis to make it useful knowledge. However, due to the lack of methods of mining tacit knowledge in data, rules, and relationships in data cannot be found. This leads to the phenomenon of abundant data but poor knowledge. On the other hand, in recent years, the rapid development of computer and information technology has produced many new concepts and technologies, such as high-performance computers and neural networks. Under the condition that the technical foundation and market demand are satisfied, the concepts and

technologies of data mining are born. Data mining trains various learning algorithms based on a large amount of data and obtains rules and relationships in data, which reflect the potential characteristics of data. It is a deeper expression of useful information contained in data. At present, in the field of artificial intelligence research, intelligent data mining is getting more and more attention. It is a method inspired by intelligent behavior, which can be researched and developed. It can get efficient and convenient solutions by processing data to be better applied to real problems. The successful application of intelligent data mining has greatly promoted the development of artificial intelligence, and intelligent data mining can perform efficient and low-consumption processing of massive data. Therefore, the research and application of intelligent data mining have become a hot spot in various scientific research fields.

2. Related work

2.1. Development of data mining and machine learning

With the continuous maturity of mobile Internet information technology, the computer's ability to collect, store and analyze data has been greatly improved. Analyzing and extracting useful information from these data is a research hotspot at present. Under this trend, the role of data mining technology has become particularly important, which has attracted wide attention.

2.1.1. Development of data mining

In the 1960s and 1970s, the database appeared and matured continuously, and more and more data were collected and stored. When using these data for calculation, the calculation speed and efficiency of the model will be affected by huge data and abnormal data. To make full use of data and improve calculation efficiency, the concept of data mining was put forward at the ACM conference in 1989. Only by mining useful information from data, only in this way can the value of data be better realized. Data mining integrates the theories and technologies of the database, information retrieval, machine learning, neural network, and data visualization, and now it has become a mature interdisciplinary subject. Combining with machine learning, a new discipline of Knowledge Discovery in Databases (KDD) was born [1].

2.1.2. Development of machine learning

In 1943, the hierarchical structure model of neural networks was proposed by Pitts W and Mc Culloch W, which provided the basic basis for the development of machine learning. In the 1950s, the research work related to machine learning was gradually formed. During this period, the concept of "artificial intelligence" was put forward, followed by Turing's "Turing Test" in 1950, which opened a new direction for artificial intelligence. How intelligent a machine can be is the main content of its research. Feigenbaum E A and others put forward that to make a machine intelligent, it is necessary to make the machine master knowledge. In 1957, the general idea of Perceptron was put forward. Rosenblatt F designed the first computer network and Samuel A designed a chess program that can learn independently. This program initially showed people the ability of machine learning. In the 1970s and 1980s, machine learning developed rapidly. In 1962, Hubel and Wiesel put forward the Hubel-Wiese

biological vision model. It laid a foundation for the follow-up study of the neural network model. In 1969, Minsky M and Papert S published Perceptron. In 1980, the first international seminar on Machine Learning was successfully held. In 1986, Machine Learning was officially launched. It means that machine learning has started to rise and become mature all over the world. In the same year, Back Propagation algorithm (BP) was proposed by Rumelhart, Hinton, and Williams. In 1989, the computational model of Convolutional Neural Network (CNN) was first proposed by Professor Le Cun Y, and an efficient training model based on the BP algorithm was deduced and studied. It has been successfully applied to the field of digital handwriting recognition [2]. Even today, BP is still one of the algorithms widely used in practical engineering applications. Because of the influence of the number of samples and the size of the network, it is necessary to set a large number of parameters (such as the learning rate of the neural network, the number of hidden layer nodes and activation functions, etc.). But most of these parameters can only be debugged by experience and trial and error [3][4]. In the mid-1990s, Support Vector Machine (SVM) began to emerge, effectively promoting the development of machine learning. As early as 1963, Vapnik V put forward the related concept of "support vector", and then put forward the principle of VC dimension and structural risk minimization, etc. The kernel technology used in SVM was fully utilized by people. This idea is also integrated into every corner of machine learning. SVM, as one of the machine learning algorithms, has been widely used in various classification and recognition problems [5][6][7][8]. Support Vector Regression (SVR) is developed based on SVM, which is mainly used in data estimation, fitting, and regression prediction. According to the analysis, Computational complexity and sparsity have the greatest influence on the SVR algorithm; so many improved algorithms have emerged continuously in the past few years [9]. With the continuous development of SVM, the algorithm research of artificial neural networks has also made further development. Hinton and others have done a lot of research, put forward a deep learning model, and opened a new era of deep network learning [10][11][12][13][14].

2.2. Research status of voiceprint recognition

The research on voiceprint recognition can be traced back to the 1930s. Voiceprint recognition takes the speaker's voice as the basic data, and extracts the speaker's acoustic characteristics from this data for analysis and modeling, to achieve the purpose of speaker recognition. Voiceprint recognition is one of the key and hot research objects of artificial intelligence technology, which is widely used in financial security, smart home, smart building, and other fields. In the voiceprint recognition system, The technology mainly includes two aspects: acoustic feature extraction and acoustic modeling. In the aspect of acoustic feature extraction, the research work did not make great progress before Davis et al. put forward Mel cepstrum coefficient, which is the cepstrum coefficient that imitates human auditory perception and is the most commonly used low-dimensional voiceprint feature in voiceprint recognition. In the aspect of acoustic modeling, the original artificial neural network and other models have achieved a certain recognition effect. After that, speaker recognition technology based on the Gaussian mixture model-general background model has been further developed. Commonly used models include the Gaussian mixture model, support vector machine [15], joint factor analysis, identity authentication vector-cosine distance scoring, probabilistic linear discriminant analysis, etc. These methods have effectively improved the performance of voiceprint recognition. After that, N.Dehak and other scholars put forward I-vector technology. This technology has been one of the mainstream

technologies in voiceprint recognition since it was put forward. Later, the innovation and practicability of the Gaussian mixture model-general background model were widely recognized and applied in the voiceprint recognition research field. On this basis, some scholars put forward joint factor analysis, and I-vector technology and probabilistic linear analysis can solve the channel mismatch problem. Before deep learning was widely used, the commonly used technologies in voiceprint recognition were based on GMM-UBM and I-vector technology.

Before 2006, the most common cepstrum coefficient was Mel cepstrum coefficient. Using the cepstrum coefficient as an acoustic feature made the early classification performance reach a certain bottleneck. Since Hinton and others put forward the model of deep confidence network in 2006, many scholars introduced deep learning into the modeling framework of voiceprint recognition and built a speaker recognition system based on DNN. Combining the mixed model of the noise-reducing automatic encoder and restricted Boltzmann machine for voiceprint recognition, the deep learning model has significantly improved the recognition accuracy. In recent years, the research of closed set voiceprint recognition technology has achieved good results in recognition accuracy, and commonly used models include the Gaussian mixture model, support vector machine, probabilistic linear discriminant analysis, long-term memory model, cyclic neural network, etc. When deep learning is introduced into voiceprint recognition, the acoustic features are generally extracted twice. Then, through the supervised classifier, some scholars added a layer of softmax layer as a classifier to classify, and the acoustic features extracted by deep learning have higher recognition accuracy in closed set voiceprint recognition. Many researchers have applied a deep neural network to acoustic feature extraction [16][17], such as Novoselov et al., [18], and using the softmax activation function in the last classification layer of the classification neural network can be used as an acoustic feature extractor. Snyder D et al. put forward a new acoustic feature X-vector [19], which maps the variable-length speech signal into a fixed-dimensional space through DNN, and this method can make full use of the training data compared with i-vector technology. Chung J et al. collected the largest voiceprint recognition data set from open-source media at present, and can identify the identity in speech efficiently under various conditions through Convolutional Neural Network (CNN), especially under noisy and unrestricted conditions [20]; In the research of voiceprint feature denoising method, Shen H et al. adopted the endpoint detection algorithm based on spectrum and used Empirical Mode Decomposition (EMD) algorithm to reconstruct the spectrum, which can effectively reduce voiceprint noise [21]. In addition, Wan L et al. proposed a new loss function based on the deep neural network [22], called Generalized End-To-End (GE2E) loss, which makes the training of the speaker verification model more effective than the previous Tuple-Based End-To-End (TE2E) loss function. Later, Bhatt- acharya G et al. proposed research on voiceprint recognition using generated confrontation networks [23]. In the related research of open-set voiceprint recognition, it is defined that unknown speech may come from unknown speakers to the latest progress, current application, and future trend of automatic speaker recognition in 2000, which is called open-set speaker recognition. Patilh et al., introduced a polynomial feature extraction method based on polynomial signal processing and wavelet analysis and proposed a new feature extraction method based on speech spectrum decomposition two years later. In the mode of open set recognition with a polynomial classifier, the speaker classification based on dialect areas in Marathi was further used. The method includes dividing speech signals into approximate Mel cepstrum coefficients by wavelet transform, and modeling each dialect area by using second-order and third-order polynomial expansions of feature vectors; Deng J et al. used the peak-to-side ratio distribution of speech samples

inside and outside the set to determine the rejection threshold in the open-set speaker recognition based on score pattern independent of the text and divided whether the signal to be tested is the signal of the speaker outside the set by the threshold. Bunrits et al., proposed a deep learning model for speaker recognition using a Convolutional Neural Network (CNN) in the research of open-set voiceprint recognition. This paper pointed out that the speech of this method is not limited by the speaker's content, which means that it adopts a text-independent form, which is more difficult to implement than the text-related speaker recognition system. By this method, the speaker's speech is converted into spectrogram images every 2 seconds, and input into the generated CNN model training from scratch. The proposed method is compared with the classification based on mel-frequency cepstral coefficients and Support Vector Machine (SVM).

Although closed-set voiceprint recognition has achieved some good results, the recognition accuracy of open-set voiceprint recognition still needs to be improved, especially for the rejection of speakers outside the set, which is due to the incorrect calculation method and selection of threshold, which leads to the misidentification of speakers outside the set. If the calculation method and selection of threshold are not appropriate, it is easy to mistake the speaker outside the set for the speaker inside the set, and it is also easy to mistake the speaker inside the set for the speaker outside the set. Thus, false rejection and false reception were made. In open-set voiceprint recognition, the two most important factors that affect the final judgment of the model are the acoustic characteristics of modeling and the calculation and selection of the threshold. The extracted features affect the training of the model, and the selection of threshold directly affects the recognition performance of the speaker outside the set.

2.3. Basic principles of neural network

The artificial neural network can simulate the structure and function of a biological neural system and process distributed parallel information, which has a wide application prospect in the fields of pattern recognition, speech recognition, data mining, and machine learning. The artificial neural network is composed of many simple neurons. To realize the purpose of network processing information, the calculation model of neurons and the connection mode of networks should be given separately. Neurons are usually multi-input and single-output nonlinear structures. JIW is the weight of the connection between the J neuron of the previous layer and the I neuron of the current layer; The summation unit is used to calculate the weighted sum among the input signals; Ib is biased; F is a nonlinear activation function, which can limit the output of neurons to a certain range. The above description can be described as formula (1).

$$y_i = f\left(\sum_j x_j w_{ji} + b_i\right) \tag{1}$$

There are dozens of neural network models, which can be divided into three categories according to the network structure: forward neural network, feedback neural network, and self-organizing competitive neural network.

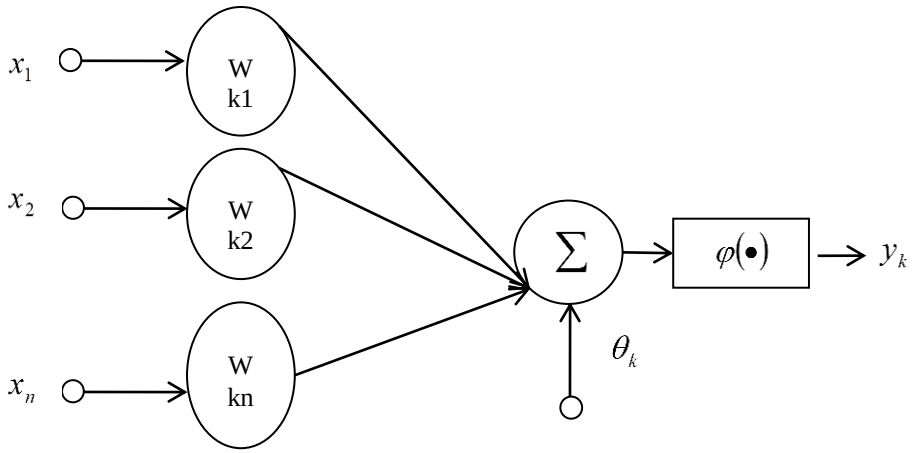

Figure 1: Artificial neuron structure model

In the feed-forward neural network, the output of the previous layer is the input of the next layer, and there is no signal feedback from the latter layer to the previous layer, and the information processing has the direction of layer-by-layer transmission. Typical feed-forward neural networks include perceptron network, BP network, and RBF network.

A feedback neural network is also called a self-associative memory network. Each neuron in the network feeds its output signal into other neurons at the same time. Typical feedback neural networks include Hopfield neural network and the Boltzmann machine.

The self-organizing competitive neural network is a two-layer network composed of an input layer and a competition layer. The neurons in the competition layer are bidirectional connected. The neurons in the competition layer compete for the response to the input, and the neurons that finally get the response represent the classification of the input data.

The learning methods of neural networks mainly include supervised learning, unsupervised learning, and semi-supervised learning. In supervised learning, the error between the expected output and the actual output of the network is obtained according to the existing training samples, and the connection weights are adjusted by minimizing the error function. After many iterations, it converges to a certain weight. Supervised learning can effectively realize the regression and classification functions. The neural network models using supervised learning include the BP network, RBF network, and Hopfield network. In unsupervised learning, it is only necessary to input unlabeled training samples into the network, and the network can learn the sample feature structure by itself. Because there is no expected output, Therefore, it can't be used to approximate functions. Many neural networks use unsupervised learning, such as deep confidence networks, generated confrontation networks, and self-organizing mapping. Semi-supervised learning is a learning method that combines supervised learning and unsupervised learning, and uses a lot of unlabeled data and marked data for pattern recognition.

3. Voiceprint recognition based on deep learning and OTSU

In this paper, the open-set voiceprint recognition based on deep learning is studied, and an adaptive threshold calculation based on the OTSU algorithm is proposed. First, the deep confidence network composed of three layers of restricted Boltzmann machines is used to extract the deep features of speech as the acoustic features used in the experiment. Then, the

similarity value of speech features is calculated by training the Gaussian mixture model, and the threshold value is further calculated by the OTSU algorithm. The comparative experiment shows that this method is feasible and has a good recognition effect.

3.1. Preprocessing of voiceprint signal

(1) Pre-emphasis

Pre-emphasis is a signal processing method, which can compensate for the excessive attenuation of high-frequency components of speech signals during transmission. Before analyzing speech signals, the high-frequency components of speech signals should be compensated to make up for the losses caused by passing through the mouth and vocal cords, and pre-emphasis does not affect noise, so the output signal-to-noise ratio is effectively improved. Usually, a first-order high-pass filter is used for pre-emphasis, and its transmission function is as follows:

$$H(z) = 1 - az^{-1} \tag{2}$$

Where A is the pre-emphasis coefficient, and the value range is 0.9 < A < 1. The calculation of pretreatment is as follows:

$$\hat{S}(n) = S(n) - aS(n-1) \tag{3}$$

(2) Framing and windowing

After preprocessing, the speech signal is framed and windowed. The speech signal is not stationary, but it is nearly stationary within 10ms to 30ms, that is, the speech signal has short-term stationarity. Framing refers to dividing the speech signal into several segments, each segment is called a "frame", and the characteristic parameters of a segment are the time series of characteristic parameters composed of the characteristic parameters of each frame. Framing is often carried out by sliding windows. The overlapping part of two frames before and after is called frameshift. The ratio of frameshift to frame length is usually between 1: 2 and 1: 3. After framing, windowing is needed. The purpose of windowing is to make the amplitude of a frame signal smoother at both ends, to make each peak on the spectrum after a fast Fourier transform finer. To prevent spectrum leakage, it is necessary to choose the appropriate window function. The commonly used window functions are as follows:

1. Rectangular window

$$w(n) = \begin{cases} 1, 0 \le n \le N-1 \\ 0, \quad other \end{cases} \tag{4}$$

2. Hanchuang

$$w(n) = \begin{cases} 0.5 - 0.5\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \\ 0, \quad other \end{cases} \tag{5}$$

3. Hamming window

$$w(n)=\begin{cases}0.54-0.64\cos\left(\frac{2\pi n}{N-1}\right), & 0\le n\le N-1\\ 0, & other\end{cases} \tag{6}$$

4. Blakeman window

$$w(n)=\begin{cases}0.42-0.5\cos\left(\frac{2\pi n}{N-1}\right)+0.08\left(\frac{2\pi n}{N-1}\right), & 0\le n\le N-1\\ 0, & other\end{cases} \tag{7}$$

5. Triangular window

When n is odd:

$$w(n)=\begin{cases}\frac{2n}{N+1}, & 0\le n\le \frac{N+1}{2}\\ \frac{2(N-n+1)}{N+1}, & \frac{N+1}{2}\le n\le N\end{cases} \tag{8}$$

When n is even:

$$w(n)=\begin{cases}\frac{2n}{N+1}, & 0\le n\le \frac{N}{2}\\ \frac{2(N-n+1)}{N+1}, & \frac{N}{2}\le n\le N\end{cases} \tag{9}$$

(3) Endpoint detection

The purpose of endpoint detection is to determine the starting point and ending point of speech and distinguish between speech segments and blank segments. Effective endpoint detection can not only reduce the amount of data in the voiceprint recognition system and save processing time but also eliminate the interference of blank segments or noise segments and improve the performance of the voiceprint recognition system. Commonly used endpoint detection methods include endpoint detection based on short-term average zero-crossing rate, endpoint detection based on entropy, endpoint detection based on complexity, and endpoint detection based on time-frequency variance, etc.

3.2. Experimental process and result analysis

The voice used in the experiment is Chinese voice data (THCHS-30) published by Tsinghua University CSLT. To find out the model with the best effect and the corresponding parameters of the model, the data are divided into a training set, development set, and test set (as shown in Table 3-1), among which there are 8 people in the training set, each with 8 voices; The development set and training set are the same for 8 people, each with 20 voices; Test set of 10 people (8 people in the set and 2 people out of the set), each with 60 voices.

Table 1: Number of speakers' voices inside and outside the training set, development set, and test set

	Training set		Development set		Test set	
Total number of people	There are 8 people in the collection	Collect 0 people outside	There are 8 people in the collection	Collect 0 people outside	There are 8 people in the collection	2 people outside the collection
Total speech number	64	0	160	0	480	120

In the DBN-GMM of a target speaker, it is verified that the similarity value of speech samples belonging to the target speaker approximately obeys the normal distribution, while the similarity value of speech samples belonging to the non-target speaker approximately obeys the gamma distribution. Because the number of speech samples that can be trained in practice is small, it is not enough to accurately represent the distribution of the similarity value of samples. Therefore, two random number sets are generated according to the distribution of the similarity values of the speech samples of target speakers and non-target speakers, and then the OTSU algorithm is used to calculate the threshold in the random number sets. The concrete steps are as follows:

(1) 24-dimensional basic acoustic features MFCC are trained by DBN to obtain 256-dimensional depth acoustic features.

(2) 256-dimensional depth acoustic features are used as the input of GMM, and the similarity values of speech samples are calculated. The average similarity values of non-target speakers' speech samples are recorded to be 1L, the average similarity values of target speakers' speech samples are 2L, and the distribution of similarity values between target speakers and non-target speakers' speech samples is tested according to the similarity values.

(3) According to the distribution of similarity values of the target speaker and non-target speaker's voice samples, 1000 random numbers are generated respectively, with the restriction that the maximum value of the non-target speaker's random number is not greater than the minimum value of target speaker's voice sample similarity, and the minimum value of target speaker's random number is not less than the maximum value of non-target speaker's voice sample similarity.

(4) The probability P_i and average value u of each similarity value I in the generated random number set.

(5) Calculate the proportion sum of similarity values of speech samples belonging to target speakers $w_0(t)$ and $w_1(t)$ non-target speakers in total random numbers $u_0(t)$ and $u_1(t)$ the average sum of their similarity values.

(6) The value is calculated according to formula (3-4), where the value range of t is (L_1, L_2), record, δ^2 and value of t.

(7) The value of comparison, when the value δ^2 is the maximum, calculates the value of t at this time. When t is the value, the variance δ^2 between classes is the maximum, and a good distinction can be achieved between inside and outside the set.

The following tables are five groups of experimental results of eight speakers in the set obtained by the OTSU algorithm.

Table 2: Model recognition rate of speakers No.1 and No.2 outside the set

Internal speaker	No.3.	No.4.	No.5.	No. 6	No. 7	No. 8	No. 9	No. 10
threshold value	2.103	1.675	2.379	2.582	2.179	2.959	2.115	2.876
Intra-set error acceptance	0	0	0	0	0	0	0	0
Error rejection rate	0	0	1.67%	0	0	0	3.33%	5.00%
Out-set error acceptance	0	0	0	0	0	0	0	0

Table 3: Model recognition rate of speakers outside the sets 3 and 4

Internal speaker	No.1	No.2.	No.5.	No. 6	No. 7	No. 8	No. 9	No. 10
threshold value	2.229	2.022	2.175	2.429	2.227	2.537	1.869	1.971
Intra-set error acceptance	0	0	0	0	0	0	0	0
Error rejection rate	8.33%	1.67%	1.67%	1.67%	1.67%	0	3.33%	1.67%
Out-set error acceptance	0	0	0	0	0	0	0	0

Table 4: Model recognition rate of speakers outside the set No.5 and No.6

Internal speaker	No.1	No.2.	No.3.	No.4.	No. 7	No. 8	No. 9	No. 10
threshold value	1.975	2.159	1.972	1.416	2.313	2.431	1.743	1.970
Intra-set error acceptance	0	0	0	0	12.1%	0	0	0
Error rejection rate	1.67%	8.33%	1.67%	0	3.33%	0	1.67%	0
Out-set error acceptance	0	0	0	0	0	0	0	0

Table 5: Model recognition rate of speakers outside the set No.7 and No.8

Internal speaker	No.1	No.2.	No.3.	No.4.	No.5.	No. 6	No. 9	No. 10
threshold value	1.937	1.976	1.866	0.962	1.998	2.251	1.554	1.880
Intra-set error acceptance	0	0	0	0	0	0	0	0
Error rejection rate	8.33%	8.33%	0	13.3%	0	0	8.33%	10.0%
Out-set error acceptance	0	0	0	0	0	0	0	0

Table 6: Model recognition rate of speakers outside the set No.9 and No.10

Internal speaker	No.1	No.2.	No.3.	No.4.	No.5.	No. 6	No. 7	No. 8
threshold value	2.229	2.644	2.363	2.308	2.393	2.632	3.082	2.614
Intra-set error acceptance	0	0.24%	0	0	0	0	0	0
Error rejection rate	8.33%	5.00%	13.3%	1.67%	0	0	1.67%	0
Out-set error acceptance	0	0	0	0	0	0	0	0

Under the same experimental environment as OTSU's algorithm for calculating the threshold, the experimental test shows that the algorithm for calculating the threshold with an equal error rate has a rejection rate of 3. 96% for intra-set errors, 0. 38% for intra-set errors and 0. 73% for out-of-set errors. The experimental results show that the method based on the DBN-GMM model combined with the OTSU algorithm proposed in this paper has a recognition rate of 99. 9% for intra-set speakers. The equal error rate method has a recognition rate of 99. 18% for in-set speakers and a rejection rate of 98. 54% for out-of-set speakers. Both in-set speaker recognition and out-of-set speaker rejection are superior to the traditional equal error rate method.

4. Conclusion

This paper mainly studies the mining of speaker's speech features under the framework of deep learning and based on this feature, open set speaker recognition is conducted to explore ways to improve the system's recognition performance for speakers inside and outside the set. The deep neural network is adopted as the deep confidence network, which is used as the feature extractor to mine deep acoustic features containing more speaker's personality features from the original input MFCC parameters. Considering the shortcomings of traditional threshold calculation methods, In this paper, under the experimental condition of text-independent open-set speaker recognition, the deep learning method and machine learning method are combined to train the Gaussian mixture model to calculate the similarity value of deep acoustic features. The distribution map of the similarity value is bimodal, so the maximum inter-class variance of the similarity value can be calculated by the OTSU algorithm. When the inter-class variance is the maximum, the similarity value at this time is the best threshold, and the similarity value of voiceprint features is distinguished most perfectly. The open-set voiceprint recognition method proposed in this paper is different from the traditional open-set voiceprint recognition method.

(1) The threshold calculation method proposed in this paper has the characteristics of adaptive adjustment, which varies according to different speakers, and has better robustness than the fixed threshold.

(2) Experiments show that the method proposed in this paper can improve the recognition accuracy of speakers inside and outside the set to a certain extent compared with the traditional threshold calculation method.

Since the development of speaker recognition research, the data environment has changed from a simple noiseless environment to a complex noisy environment, from long-term small amount to short-term large amount, from text-related to text-independent, from closed set voiceprint recognition to open set voiceprint recognition, which is becoming more and more difficult and demanding. Many practical problems need to be solved urgently, such as: extracting speaker feature parameters from the complex noisy environment; Enhancing the adaptability of the speaker model base, and when new speakers are added, the trained codebook can be automatically added. Improve the robustness of the algorithm while reducing its complexity of the algorithm. Because of the time relationship, the research on the calculation method of open set speaker threshold is not deep enough in this paper. I hope to find an algorithm with a high recognition rate, strong robustness, and low complexity in future work and study.

Reference

[1] H. Witten, E. Frank, & A. Hall. (2016). Data Mining. *Cambridge: Morgan Kaufmann*.

[2] Y. Le Cun, Y. Bengio, & G. E.Hinton. (2015). Deep learning. Nature. 521(7553), 436-444.

[3] C. Moro, H. El Fil, & V. Francioso. (2021). Influence of water-to-binder ratio on the optimum percentage of Nano-Ti O2 addition in terms of compressive strength of mortars: A laboratory and virtual experimental study based on ANN model. *Construction and Building Materials*, 267, 120960.

[4] N. Parashar, N. Aslfattahi, & S. M. Yahya. (2021). ANN modeling of thermal conductivity and viscosity of mxene-based aqueous IO Nanofluid. International Journal of Thermophysics, (2021), 42(2), 1-24.

[5] O. Rostami & M.Kaveh. (2021). Optimal feature selection for SAR image classification using biogeography-based optimization (BBO), artificial bee colony (ABC), and support vector machine (SVM): A combined approach of optimization and machine learning. *Computational Geosciences*, 1-20.

[6] A. Srinivasa Reddy, & P. Chenna Reddy. (2021). MRI brain tumor segmentation and prediction using modified region growing and adaptive SVM. *Soft Computing*, 1-14.

[7] H. Rizwan, C. Li, & Y. Liu. (2021). Online dynamic security assessment of wind integrated power system using SDAE with SVM ensemble boosting learner. *International Journal of Electrical Power & Energy Systems*, 125, 106429.

[8] X. Song, Y. Zheng, & W. Xue. (2021). Identification of risk genes related to myocardial infarction and the construction of an early SVM diagnostic model. *International Journal of Cardiology*, 328, 182-190.

[9] H. Han, X. Cui, & Y. Fan. (2019). Least squares support vector machine (LS-SVM)-based chiller fault diagnosis using fault indicative features. *Applied Thermal Engineering*, 154, 540-547.

[10] Z. I. Attia, A. Sugrue, & S. J. Asirvatham. (2018). Noninvasive assessment of dofetilide plasma concentration using a deep learning (neural network) analysis of the surface electrocardiogram: A proof of concept study. *PLOS One*, 13(8), e0201059.

[11] Y. Chen & Z. Wang. (2018). Quantitative analysis modeling of infrared spectroscopy based on ensemble convolutional neural networks. *Chemometrics and Intelligent Laboratory Systems*, 181, 1-10.

[12] J. Acquarelli, T. van Laarhoven, & J. Gerretzen. Convolutional neural networks for vibrational spectroscopic data analysis. *Analytica Chimica Acta*, 954, 22-31.

[13] H. C. Shin, H. R. Roth, & M. Gao. (2016). Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics, and transfer learning. *IEEE Transactions on Medical Imaging*, 35(5), 1285-1298.

[14] Y. Fujisawa, Y. Otomo, & Y. Ogata. (2019). Deep-learning-based, computer-aided classifier developed with a small dataset of clinical images surpasses board-certified dermatologists in skin tumor diagnosis. *The British Journal of Dermatology*, 180(2), e52-e52.

[15] P. Tome, J. Fierrez, & R. Ver-Rodriguez. (2014). Soft biometrics and their application in person recognition at a distance. *IEEE Transactions on Information Forensics and Security*, 9(3), 464-472.

[16] Y, Zhu, T, Ko, & D. Snyder. (2018). Self-attentive speaker embedding for text-independent speaker verification. *Interspeech*, 3573-3577.

[17] D. Snyder, D. Garcia-Romero, & D. Povey. Deep neural network embedding's for text-independent speaker verification. *Interspeech*, 999-1003.

[18] S. Novoselov, A. Shulipa, & I. Kremnev. (2018). On deep speaker embedding for text-independent speaker recognition. *Odyssey 2018 the Speaker and Language Recognition Workshop*.

[19] D. Snyder, D. Garcia-Romero, & G. Sell. X-vectors: Robust DNN embedding for speaker recognition. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5329-5333.

[20] J. S. Chung, A. Nagrani, & A. Zisserman. (2018). Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622.

[21] H. Shen, B. Wang, & J. Wang. (2018). Research on the robustness of voiceprint recognition technology. *Proceedings of the 2018 International Conference on Algorithms, Computing and Artificial Intelligence*, 1-5.

[22] L. Wan, Q. Wang, & A. Papir. (2018). Generalized end-to-end loss for speaker verification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4879-4883.

[23] G. Bhattacharya, J. Monteiro, & J. Alam, (2019). Generative adversarial speaker embedding networks for domain robust end-to-end speaker verification. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE*, 6226-6230.

This page is empty by intention.