Scientific Research Publishing

Journal of Applied Mathematics and Physics, 2020, 8, 2671-2682
https://www.scirp.org/journal/jamp
ISSN Online: 2327-4379
ISSN Print: 2327-4352

Adaptive Threshold Estimation of Open Set Voiceprint Recognition Based on OTSU and Deep Learning

Xudong Li[1,2], Xinjia Yang[1,2], Linhua Zhou[1,2*]

[1]School of Science, Changchun University of Science and Technology, Changchun, China
[2]Provincial Demonstration Center for Experimental Mathematics Education (Changchun University of Science and Technology), Changchun, China
Email: *chowlh1718@163.com

How to cite this paper: Li, X.D., Yang, X.J. and Zhou, L.H. (2020) Adaptive Threshold Estimation of Open Set Voiceprint Recognition Based on OTSU and Deep Learning. *Journal of Applied Mathematics and Physics*, **8**, 2671-2682.
https://doi.org/10.4236/jamp.2020.811197

Received: October 22, 2020
Accepted: November 27, 2002
Published: November 30, 2002

Abstract

Aiming at the problem of open set voiceprint recognition, this paper proposes an adaptive threshold algorithm based on OTSU and deep learning. The bottleneck technology of open set voiceprint recognition lies in the calculation of similarity values and thresholds of speakers inside and outside the set. This paper combines deep learning and machine learning methods, and uses a Deep Belief Network stacked with three layers of Restricted Boltzmann Machines to extract deep voice features from basic acoustic features. And by training the Gaussian Mixture Model, this paper calculates the similarity value of the feature, and further determines the threshold of the similarity value of the feature through OTSU. After experimental testing, the algorithm in this paper has a false rejection rate of 3.00% for specific speakers, a false acceptance rate of 0.35% for internal speakers, and a false acceptance rate of 0 for external speakers. This improves the accuracy of traditional methods in open set voiceprint recognition. This proves that the method is feasible and good recognition effect.

Keywords

Voiceprint Recognition, Deep Neural Network (DNN), OTSU, Adaptive Threshold

1. Introduction

Voiceprint recognition is a biometric authentication technology that recognizes the identity based on the human body's own voice characteristics. According to different recognition methods, voiceprint recognition can be divided into two categories: voiceprint verification and voiceprint identification. Voiceprint confirmation is to judge whether a certain speech is spoken by a specific speaker, and voiceprint identification is to judge which person in the set of persons to be identified speaks a certain speech. Voiceprint recognition is generally expounded from two aspects. One is closed-set voiceprint recognition, that is, all the speaker's voices already exist in the model library, and the voice features to be tested are matched with all the speakers in the model library, and the one with highest matching degree is the one to be asked; the other is open-set voiceprint recognition, that is, the voice feature to be tested may not be in the trained model library, which requires a threshold to decide whether to accept or reject. There are two commonly used thresholds: classic threshold and dynamic threshold. The classical threshold [1] is determined by the two error rates of false rejection rate (False Rejection Rate, FRR) and false acceptance rate (False Acceptance Rate, FAR) of falsely rejecting speakers within the set. Generally, the corresponding threshold is used when FRR and FAR have the same value, but sufficient training samples are required to achieve good results. If the training data sample is too small, the point where FRR and FAR are equal may not be obtained, so the threshold is less robust, resulting in reduced system recognition performance. Dynamic threshold [2] is to train a model for each trained speaker in the training stage, calculate the corresponding threshold for each speaker, and compare the test voice with each threshold during recognition. Firstly, the amount of calculation is large and the training time is long; secondly, when the number of training speaker increases, an infinite number of distributions cannot be distinguished in a limited space, and the possibility of overlapping distributions increases. In the testing phase, the voice to be tested must be matched with all the trained models, which also has the problem of time-consuming and high space complexity. This paper intends to determine the threshold by calculating the similarity of the training speech, so as to avoid the problems of poor robustness of the classical threshold and the large amount of calculation of the dynamic threshold matching distance.

The essence of voiceprint recognition is the process of converting acoustic signals into electrical signals and then performing pattern recognition. The original speech signal is a time-varying signal that contains a lot of redundant information, it must undergo a certain transformation to remove the redundant information in the speech, extract the personality parameters that can characterize the speaker, and then recognize it through pattern recognition and recognition algorithm. The features used for voiceprint recognition are mainly time-domain feature parameters such as energy or amplitude in the time domain, zero-crossing rate, and transform domain feature parameters obtained by performing certain transformations on the original speech signal after framing the original speech signal. Such as linear prediction coefficient, linear prediction cepstrum coefficient [3], and Mel cepstrum coefficient [4]. There are many models for voiceprint recognition, such as dynamic time warping method [5], Hidden Markov Model method [6], vector quantization method [7] and artificial neural network [8] are widely used in voiceprint recognition. In the 1990s, the introduction of Gaussian Mixture-General Background Model and Support Vector Machine [9] enabled voiceprint recognition to enter a new stage of development. Joint factor analysis based on i-vector technology [10], disturbance attribute interference algorithm, eigen channel analysis, these methods improve the robustness of voiceprint recognition. In recent years, DNN [11] [12] has been successfully applied in acoustic modeling. Based on the powerful learning ability of deep learning and the ability to mine deep features of data, voiceprint recognition has gradually transitioned from traditional machine learning to deep learning. Scholars are still trying to combine machine learning and deep learning to achieve better results.

Traditional voiceprint recognition generally uses Gaussian Mixed Model (GMM) [13] [14] and Gaussian Mixture-General Background Model. Because it is very sensitive to noise and belongs to shallow and incomplete learning, the accuracy of model recognition decreases as the number of people increases, the robustness is poor, and the convergence is difficult. DNN has strong expression ability and a high tolerance for noise. Therefore, Deep Belief Network (DBN) is used to extract deep acoustic features and then GMM is used to recognize voiceprints, thereby improving the accuracy of open set voiceprint recognition. Based on the deep acoustic features, this paper proposes an algorithm for determining the adaptive threshold. The experimental results show that the threshold determined by the algorithm has a good performance in the performance of open set voiceprint recognition.

2. OTSU-Based Approach for Threshold Calculation

OTSU [15] [16] was proposed by Japanese scholar Otsu in 1979. It is also called the maximum between-class variance method. This algorithm uses the maximum between-class variance of the average pixel values of the foreground area and the background area as the criterion to calculate the threshold. Based on the idea of OTSU, this paper divides the random number set generated by two different random variables with a certain distance in space into two parts *A* and *B*, and calculates the variance of the two parts. The greater the variance, the better the segmentation effect. The maximum variance under different segmentation conditions is obtained by traversal, and the optimal threshold is determined to achieve a good segmentation of the two sets of random numbers.

For a set of random numbers with a total of *N*, *L* represents the maximum value of random numbers, n_i represents the number of random numbers as *i*, p_i represents the probability of random numbers being *i*, then p_i is

$$p_i = n_i / N . \tag{1}$$

Take the threshold *T*. The proportion of the number belonging to the random number set A to the total random number is recorded as ω_0, and the average value u_0.

$$\omega_0 = \sum_i^T p_i , \quad u_0 = \sum_i^T i p_i / \omega_0 . \tag{2}$$

The proportion of the number belonging to the random number set B to the total random number is recorded as ω_1, and the average value is u_1.

$$\omega_1 = 1 - \omega_0, \quad u_1 = \sum_{i=T}^{L} i p_i / \omega_1 . \tag{3}$$

The average of the set of random number u is

$$u = \omega_0 u_0 + \omega_1 u_1 . \tag{4}$$

The maximum between-class variance is σ^2:

$$\omega_0 \left(u_0 - u\right)^2 + \omega_1 \left(u_1 - u\right)^2 . \tag{5}$$

The greater the variance between classes, the better the threshold for dividing the set of random numbers is selected. One only need to traverse all i values, calculate the variance of each step, and select the value T of the largest σ^2. The best threshold expression is

$$T^* = \arg\max\left(\sigma_i^2\right). \tag{6}$$

3. Deep Feature Extraction and GMM Similarity Calculation

3.1. DBN-Based Deep Voiceprint Feature Extraction

DBN is obtained by superimposing Restricted Boltzmann Machines (RBM) [17] [18]. Compared with traditional shallow networks, DBN has a better ability to mine potential features of data. Therefore, DBN can be used as a voiceprint depth feature extractor for voiceprint recognition. In the voiceprint recognition task, the MFCC of the input voice is not a binary value, but some real value. The traditional RBM cannot achieve the expected goal. Usually, the voiceprint feature extraction uses the Gauss-Bernoulli RBM model.

1) RBM Structure

RBM is a Markov random field based on an energy function, composed of a visible layer and a hidden layer. When the visible neuron $v = \{v_i\}, i \in d_v$ is in the Gaussian distribution state and the hidden layer neuron $h = \{0,1\}, j \in d_h$ is in the binary state, the connection relationship between the visible layer and the hidden layer is represented by the matrix $W = \{w_{ij}\}$, and the model parameters are: $\theta = \{w_{ij}, a_i, b_j, \sigma_i\}$.

Based on the energy model theory, the energy of Gauss-Bernoulli RBM is defined as:

$$E(v, h \mid \theta) = \sum_{i \in d_v} \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_{j \in d_h} b_j h_j - \sum_{i,j} \frac{v_i}{\sigma_i} h_j w_{ij} , \tag{7}$$

In which, w_{ij} is the weighted coefficient, a_i and b_j are the visible layer unit and hidden layer unit corresponding offset. σ_i refers to the standard deviation of visual layer unit, and d_v and d_h are the number of units of visible layer and hidden layer. From the above formula, the joint probabilities of v and h can be obtained as shown below:

$$p\left(v,h\mid\theta\right)=\frac{\exp\left(-E\left(v,h\mid\theta\right)\right)}{Z\left(\theta\right)},Z\left(\theta\right)=\sum_{v,n}\exp\left(-E\left(v,h\mid\theta\right)\right), \tag{8}$$

In which $Z\left(\theta\right)$ is the partition function, which is used for normalization and calculation of all possible energy allocation between visible and hidden layer neurons. When training RBM, the conditional independence between the visible neurons and the hidden neurons, the conditional probabilities of v and h can be obtained as shown below:

$$p\left(h_j=1\mid v\right)=\text{sigmoid}\left(\sum_{i\in d_v}\frac{v_i}{\sigma_i}w_{ij}+b_j\right), \tag{9}$$

$$p\left(v_i=x\mid h\right)=N\left(\sigma_i\sum_{j\in d_h}w_{ij}h_j+a_i,\sigma_i^2\right), \tag{10}$$

In the formula, $\text{sigmoid}\left(x\right)=1/\left(1+\exp\left(-x\right)\right)$ is the activation function, $N\left(\mu,\sigma\right)$ refers to the Gaussian distribution of the average value μ and the standard variance σ.

In order to solve the problem of training speed of RBM, Hinton proposed the contrast divergence algorithm [19], which approximates the true value by Gibbs sampling, and cannot converge directly.

The update criterion of RBM parameters are as follows:

$$\Delta w_{ij}=\varepsilon\left(\left\langle v_ih_j\right\rangle_{data}-\left\langle v_ih_j\right\rangle_{recon}\right), \tag{11}$$

$$\Delta a_i=\varepsilon\left(\left\langle v_i\right\rangle_{data}-\left\langle v_i\right\rangle_{recon}\right), \tag{12}$$

$$\Delta b_j=\varepsilon\left(\left\langle h_j\right\rangle_{data}-\left\langle h_j\right\rangle_{recon}\right). \tag{13}$$

In the formula, ε is the learning rate, "*data*" represents the expectation of training data, and "*recon*" represents the expectation of model distribution.

2) DBN training

When training a multi-layer DBN, MFCC extracted from speech is used as the input of the first RBM, and the RBM is trained one by one by using unsupervised learning method. The train RBMs are stacked together, which is the pre training of DBN. Then, BP algorithm [20] is used to fine tune the parameters of each layer of DBN, and the error is transmitted back to correct it.

3) Voiceprint depth feature extraction

First, normalize the input original 24-dimensional MFCC features to make the feature distribution of each speaker satisfy $\mu_i=0$ and $\sigma_i=1$. This can avoid the re-estimation of the training sample distribution. DNN is composed of 3 RBMs, the network structure is 24-256-256-256, the output layer is the softmax function, and the voiceprint depth feature output layer takes the last hidden layer. Through this network, the 24-dimensional MFCC features can be converted into 256-dimensional deep acoustic features.

3.2. Speaker Identification Based on DBN-GMM

In the DBN-GMM open set voiceprint recognition model, it is first necessary to denoise and frame the speech signal, extract MFCC from it, then extract deep acoustic features from MFCC through DBN, and finally use GMM to confirm the speaker. In the training phase of the model, a GMM is established for each speaker in the set. The purpose of training is to estimate the parameters of the GMM. In the recognition phase of the model, the speech features to be tested are calculated with the sequence in the model library. The maximum value corresponds to the identified speaker.

GMM is a linear combination of several Gaussian functions used to represent the spatial distribution of acoustic features of each speaker's training speech. Suppose the input voice feature of a speaker is $X=\{x_1,x_2,\cdots,x_N\}$, and x_i is the D-dimensional feature vector. Then the GMM with the M blending degree of the speech feature training can be expressed as:

$$p(x\mid\theta)=\sum_{k=1}^{M}w_k p_k(x\mid\theta_k),\quad \sum_{k=1}^{M}w_k=1, \tag{14}$$

w_k is the weighted factor of $p_k(x_i\mid\theta_k)$, $p_k(x_i\mid\theta_k)$ is the k Gaussian distribution model, which satisfies the following formula:

$$p_k(x_i\mid\theta_k)=\frac{1}{(2\pi)^{D/2}\left|\Sigma_k\right|^{1/2}}\exp\left\{-\frac{1}{2}(x-u_k)'(\Sigma_k)^{-1}(x-u_k)\right\}, \tag{15}$$

In the formula, u_k is the average value, and Σ_k is the covariance matrix. Therefore, GMM can be expressed by the parameter $\theta=\{w_k,u_k,\Sigma_k\}$.

For the module parameter θ, usually maximum likelihood estimation is used to solve the problem, expressed as follows:

$$\theta^*=\arg\max_{\theta}p(X\mid\theta)=\arg\max_{\theta}\prod_{i=1}^{N}p(x_i\mid\theta). \tag{16}$$

Due to the hidden variables in the model, it is difficult to solve the parameters, so EM algorithm is usually used to solve the parameters.

$$w_k=\frac{\sum_{i=1}^{N}p(k\mid x_i,\theta)}{N},\quad u_k=\frac{\sum_{i=1}^{N}x_i p(k\mid x_i,\theta)}{\sum_{i=1}^{N}p(k\mid x_i,\theta)},\quad \Sigma_k=\frac{\sum_{i=1}^{N}p(k\mid x_i,\theta)(x_i-u_k)^2}{\sum_{i=1}^{N}p(k\mid x_i,\theta)}. \tag{17}$$

DBN-GMM uses the deep acoustic features extracted by DBN as the input of GMM. Each speaker's voice features form a specific distribution in a specific space, and these distributions can be used to describe the personality characteristics of the speaker. By training the GMM, we can obtain the GMM similarity value with a high degree of distinction between those who belong to the set and those who do not belong to the set. Figure 1 below is the result obtained by the No. 2 speaker using a GMM with a mixing degree of 4.

4. Open-Set Speaker Recognition Experiment Based on OTSU

4.1. Speech Data Set and Acoustic Feature Description

The audio used in the experiment is the Chinese speech data (THCHS-30) published by CSLT of Tsinghua University. In order to find the best model and the parameters corresponding to the model, this paper divides the data into training set, development set and test set (see Table 1 Shown), in which there are 8 people in the training set and 8 audios per person. The development set and the training set are the same 8 people, each with 20 audios; the test set has 10 people (8 within and 2 outside), and each has 60 audios.

The basic acoustic feature is that the frame length is 30 ms, the frame shift is 15 ms, and the first-order difference 24-dimensional MFCC feature is spliced. Assuming that a certain audio segment has n frames, and the MFCC parameter of each frame is $x_i\left(i=1,2,\cdots,n\right)$, the MFCC of this audio segment is recalculated as $\sum_{i=1}^{n}\frac{x_i}{n}$.

DBN uses three-layer RBM (network nodes: 24-256-256-256) stacking, and the output value of the last layer of 256 nodes is used as the deep acoustic feature obtained by the 24-dimensional MFCC after DBN feature extraction.

4.2. OTSU Adaptative Threshold Method Based on DBN-GMM

In the DBN-GMM of a certain speaker, the signal similarity value of the speaker after inspection is approximately subject to a normal distribution, while the signal similarity values of other speakers approximately obey the gamma distribution (as shown in Figure 2). Since there are fewer voices that can participate in training in practice, and the signal similarity value is not enough to accurately represent the distribution of similarity values, two sets of random numbers are

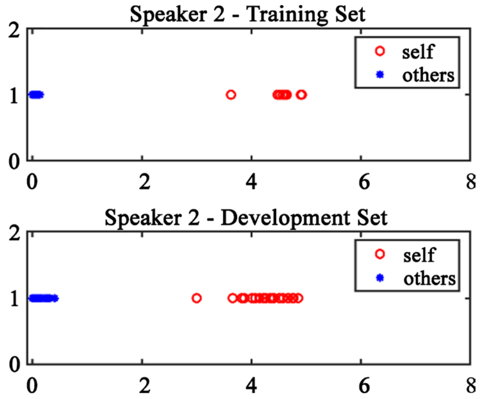

Figure 1. GMM similarity value distribution diagram of speakers inside and outside the training set and the development set.

Table 1. Number of Speakers in and out of the Training Set, Development Set, and Test Set.

	Training Set		Development Set		Test Set	
Number of people	8 within set	0 outside set	8 within set	0 outside set	8 within set	2 outside set
Total audio number	64	0	160	0	480	120

generated according to the distribution of the similarity values of the speaker and other speakers (The histogram is shown in **Figure 3**). Finally, OTSU is used to determine the threshold in the random number set.

The specific implementation steps are as follows:

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

2) The 256-dimensional deep acoustic feature is used as the input of GMM to calculate the similarity value of the feature. The average value of the similarity value of the feature outside the set is L_1, and the average value of the signal similarity value in the set is L_2, and the distribution of similarity values in the set and outside the set is checked according to the similarity values.

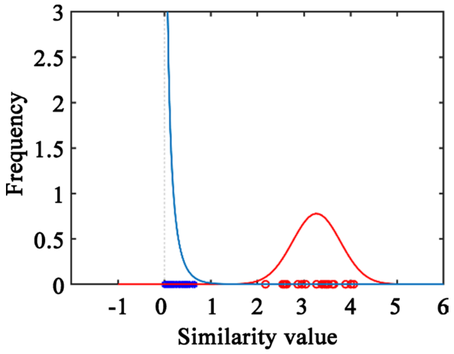

Figure 2. Distribution of similarity values within and outside the set.

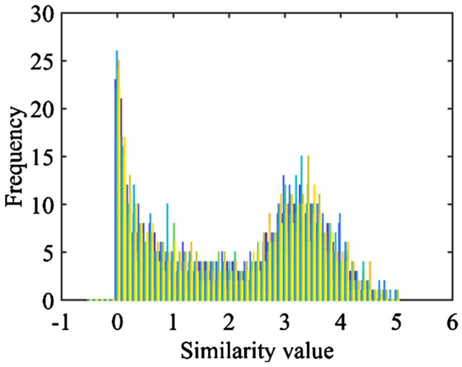

Figure 3. Similarity value histogram.

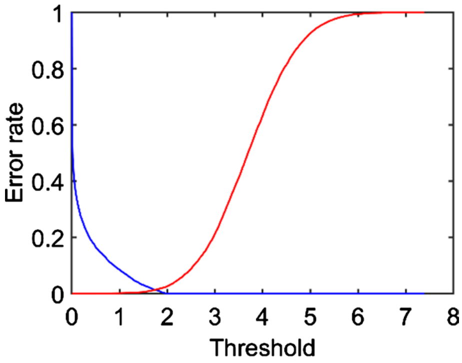

Figure 4. The relationship between false rejection rate and false acceptance rate.

3) According to the distribution of the similarity values of the speaker and other speakers, 1000 random numbers are generated, and the restriction condition is that the maximum value of the random numbers generated by the other speakers is not greater than the minimum value of the similarity value of the speaker. The minimum value of the random number generated by the speaker is not less than the maximum value of the similarity of other speakers.

4) Calculate the probability p_i and average u of each similarity value i in the generated random number set.

5) Calculate the ratio of the similarity value of the speaker and the similarity value of other speakers to the total random number $\omega_0(t)$ and $\omega_1(t)$, and the average value of the similarity value $u_0(t)$ and $u_1(t)$.

6) Calculate the value of σ^2 according to formula (5), where the value range of t is (L_1, L_2), and record the values of σ^2 and t.

7) Compare the value of σ^2, when σ^2 is the largest, calculate the t value at this time. When t takes this value, the variance between classes takes the maximum value, and a good distinction is achieved between inside and outside the set.

4.3. Experimental Results and Analysis

Tables 2-6 show the threshold determined by the method of OTSU, and the result of 5-ford cross-validation. The false acceptance rate of the speakers in the set is 0.35%, and the false rejection rate of the specific speakers is 3.00%. The false acceptance rate of outside speakers is 0.

Currently, the "equivalent error rate" is commonly used to determine the threshold. The calculation formulas for the false rejection rate and false reception rate are as follows.

False rejection rate = the number of corpus of speakers i that are falsely rejected/the number of corpus of speakers i that should be accepted. False acceptance rate = the number of corpus of other speakers who were incorrectly

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

In-set speaker	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8	No. 9	No. 10
Threshold	2.103	1.675	2.379	2.582	2.179	2.179	2.115	2.876
In-set FAR	0	0	0	0	0	0	0	0
FRR	0	0	1.67%	0	0	0	3.33%	5.00%
Outside-set FAR	0	0	0	0	0	0	0	0

Table 3. Model recognition rate of No. 3 and No. 4 speakers outside the set.

In-set speaker	No. 1	No. 2	No. 5	No. 6	No. 7	No. 8	No. 9	No. 10
Threshold	2.229	2.022	2.175	2.429	2.227	2.537	1.869	1.971
In-set FAR	0	0	0	0	0	0	0	0
FRR	8.33%	1.67%	1.67%	1.67%	1.67%	0	3.33%	1.67%
outside-set FAR	0	0	0	0	0	0	0	0

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

In-set speaker	No. 1	No. 2	No. 3	No. 4	No. 7	No. 8	No. 9	No. 10
Threshold	1.975	2.159	1.972	1.416	2.313	2.431	1.743	1.970
In-set FAR	0	0	0	0	12.1%	0	0	0
FRR	1.67%	8.33%	1.67%	0	3.33%	0	1.67%	0
Outside-set FAR	0	0	0	0	0	0	0	0

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

In-set speaker	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 9	No. 10
Threshold	1.937	1.976	1.866	0.962	1.998	2.251	1.554	1.880
In-set FAR	0	0	0	0	0	0	0	0
FRR	8.33%	8.33%	0	13.3%	0	0	8.33%	10.0%
Outside-set FAR	0	0	0	0	0	0	0	0

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

In-set speaker	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8
Threshold	2.229	2.644	2.363	2.308	2.393	2.632	3.082	2.614
In-set FAR	0	0.24%	0	0	0	0	0	0
FRR	8.33%	5.00%	13.3%	1.67%	0	0	1.67%	0
Outside-set FAR	0	0	0	0	0	0	0	0

accepted/the number of corpus of other speakers that should be rejected. Under the same experimental conditions, after experimental testing, the algorithm of equal error rate calculation threshold has a false rejection rate of 3.96% for specific speakers, a false acceptance rate of 0.38% for within speakers, and a false acceptance rate of 0.73% for outside speakers.

The experimental results show that the proposed method based on the DBN-GMM model combined with OTSU to determine the threshold has a recognition rate of 99.32% for the speakers in the set, and a rejection rate of 100% for the speakers outside the set. The method of equal error rate has a recognition rate of 99.18% for within speakers and a rejection rate of 98.54% for outside speakers. Although the time complexity of the algorithm in this paper is $O(n)$, if N random numbers are generated, the optimal threshold requires $N(N+1)$ addition and subtraction operations, $4N$ times multiplication and division operations, and N times squaring operations, which requires a large amount of calculation. However, this algorithm is better than the traditional equal error rate method in the case of a small increase in complexity, whether it is the identification of within speakers or the rejection of outside speakers.

5. Conclusion

This paper studies the determination of the open set Voiceprint recognition threshold and proposes a dynamic threshold calculation model for the open set voiceprint recognition based on OTSU. This model has strong characterization ability for data, which further improves the recognition effect. In view of the insufficient mining of interlocutor's personality characteristics and poor modeling ability in traditional GMM-based recognition, a voiceprint recognition model combining DBN and GMM is proposed, and a nonlinear RBM-based DBN deep learning model is constructed. Its powerful modeling capabilities can dig out deep-level information of features, which is more suitable for voiceprint recognition. The GMM is trained to calculate the similarity value of the signal, and OTSU is used to calculate the maximum inter-class variance of the signal similarity value to determine the threshold, and it is tested and verified in the CSLT public voice database. Experimental results show that the algorithm for determining the threshold described in this article has higher recognition accuracy than the algorithm for determining the threshold with an equal error rate, and this method is feasible.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Jagadiswary, D. and Saraswady, D. (2016) Biometric Authentication Using Fused Multimodal Biometric. *Procedia Computer Science*, **85**, 109-116. https://doi.org/10.1016/j.procs.2016.05.187

[2] Lin, L., Wang, S.X. and Wang, X.L. (2006) Real-Time Implementation of Open Set Speaker Recognition System Based on DSP. *Journal of Jilin University* (*Information Science Edition*), No. 24, 252-258. (In Chinese)

[3] Atal, B.S. (1974) Effectiveness of Linear Prediction Characteristics of the Speech Wave for Automatic Speaker Identification and Verification. *Journal of the Acoustical Society of America*, **55**, 1304-1312. https://doi.org/10.1121/1.1914702

[4] Lokesh, S. and Ramya Devi, M. (2019) Speech Recognition System Using Enhanced Mel Frequency Cepstral Coefficient with Windowing and Framing Method. *Cluster Computing*, **22**, 11669-11679. https://doi.org/10.1007/s10586-017-1447-6

[5] Geppener, V.V., Simonchik, K.K. and Haidar, A.S. (2007) Design of Speaker Verification Systems with the Use of an Algorithm of Dynamic Time Warping (DTW). *Pattern Recognition and Image Analysis*, **17**, 470-479. https://doi.org/10.1134/S1054661807040050

[6] Zeinali, H., Sameti, H. and Burget, L. (2017) HMM-Based Phrase-Independent i-Vector Extractor for Text-Dependent Speaker Verification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **25**, 1421-1435. https://doi.org/10.1109/TASLP.2017.2694708

[7] Soong, F. and Rosenberg, A. (1988) On the Use of Instantaneous and Transitional Spectral Information in Speaker Recognition. *IEEE Transactions on Acoustics, Speech Signal Processing*, **36**, 871-879. https://doi.org/10.1109/29.1598

[8] Dey, N.S., Mohanty, R. and Chugh, K.L. (2012) Speech and Speaker Recognition System Using Artificial Neural Networks and Hidden Markov Model. 2012 *Interna-*

tional Conference on Communication Systems and Network Technologies, Rajkot, 11-13 May 2012, 311-315. https://doi.org/10.1109/CSNT.2012.221

[9] Wan, V. and Campbell, W.M. (2000) Support Vector Machines for Speaker Verification and Identification*. Neural Networks for Signal Processing X. Proceedings of the* 2000 *IEEE Signal Processing Society Workshop*, Vol. 2, 775-784.

[10] Ghahabi, O. and Hernando, J. (2017) Deep Learning Backend for Single and Multisession I-Vector Speaker Recognition. *IEEE ACM Transactions on Audio Speech and Language Processing*, **25**, 807-817. https://doi.org/10.1109/TASLP.2017.2661705

[11] Zhu, H., Akrout, M., Zheng, B., *et al.* (2018) Benchmarking and Analyzing Deep Neural Network Training. *IEEE International Symposium on Workload Characterization*, Raleigh, 30 September-2 October 2018, 88-100. https://doi.org/10.1109/IISWC.2018.8573476

[12] Le Cun, Y. and Bengio, Y. (2015) Hinton G. Deep Learning. *Nature*, **521**, 436-444. https://doi.org/10.1038/nature14539

[13] Sadıç, S. and Gülmezoğlu, M.B. (2011) Common Vector Approach and Its Combination with GMM for Text-Independent Speaker Recognition. *Expert Systems with Applications*, **38**, 11394-11400. https://doi.org/10.1016/j.eswa.2011.03.009

[14] Reynolds, D.A. (1995) Speaker Identification and Verification Using Gaussian Mixture Speaker Models. *Speech Communication*, **17**, 91-108. https://doi.org/10.1016/0167-6393(95)00009-D

[15] Otsu, N. (1979) A Threshold Selection Method from Gray-Level Histograms. *IEEE Transactions on Systems*, *Man*, *and Cybernetics*, **9**, 62-66. https://doi.org/10.1109/TSMC.1979.4310076

[16] Xu, X.Y., Xu, S.Z., Jin, L.H., *et al.* (2011) Characteristic Analysis of Otsu Threshold and Its Applications. *Pattern Recognition Letters*, **32**, 956-961. https://doi.org/10.1016/j.patrec.2011.01.021

[17] Hinton, G.E. and Salakhutdinov, R.R. (2006) Reducing the Dimensionality of Data with Neural Networks. *Science* (*New York*, *N.Y.*), **313**, 504-507. https://doi.org/10.1126/science.1127647

[18] Hinton, G.E. (2002) Training Products of Experts by Minimizing Contrastive Divergence. *Neural Computation*, **14**, 1771-1800. https://doi.org/10.1162/089976602760128018

[19] Hinton, G.E., Osindero, S. and The, Y.W. (2006) A Fast Learning Algorithm for Deep Belief Nets. *Neural Computation*, **18**, 1527-1554. https://doi.org/10.1162/neco.2006.18.7.1527

[20] Achkar, R., El-Halabi, M., Bassil, E., *et al.* (2016) Voice Identity Finder Using the Back Propagation Algorithm of an Artificial Neural Network. *Procedia Computer Science*, **95**, 245-252. https://doi.org/10.1016/j.procs.2016.09.322