Robust speaker recognition algorithm

Wen-Chao Hao, Yi Chen, Lei Wang, Chun-Guang Li, Yue-Qin Feng and Qing-Yun Wang

Communication Engineering College, Nanjing Institute of Technology, Nanjing, Jiangsu,China
E-mail: 836787335@qq.com
www.njit.edu.cn

The accuracy of speaker recognition algorithm would be decreased greatly due to the noise issues. According to noisy environment, a new robust speaker recognition algorithm is proposed in this paper. After Mel-frequency Cepstral Coefficient (MFCC) feature extraction, the features are calibrated with half rised-sine function. Then the features are processed by feature normalization, feature folding and feature mapping. A method of combining BP neural network(NN) with Gaussian mixture model(GMM) is proposed to improve the recognition accuracy and robustness of the model. The neural network works in the probability space of GMM gathers the interactive information between different speakers. The experiment result proves that the proposed algorithm shows more accuracy and robustness.

Keywords: Speaker Recognition; Mel-Frequency Cepstral Coefficients; Gaussian Mixture Model.

1.Introduction

In recent years, with the rapid development of communications and mobile Internet, remote identity authentication based on cell phone or smart phone is experiencing more and more urgent demands in information security field. Many experts have done a lot in the corresponding studies[1-3] of speaker recognition. At present, the most popular features of speaker recognition are MFCC and LPCC[4-6]. In terms of recognition methods, vector quantization[7] , Gaussian Mixture Model(GMM)[7-8] and Hidden Markov Models(HMM) had gradually been applied in the field of speaker recognition. In recent years, support vector machine (SVM)[8] has been widely used due to the use of superscalar features. What's more, the neural network algorithm especially deep learning network algorithm [6] has now become the new direction of speaker recognition algorithm.

Although the current speaker recognition technology has made considerable progress, but there are still many difficulties to apply the speaker recognition technology to real applications. The current achievements are mainly conducted under laboratory conditions which are largely different from actual conditions[1-4]. The noise problem is less considered under the laboratory conditions, while it is unavoidable in the actual conditions. Furthermore, the conditions of speakers such as emotion or health would also influence the recognition accuracy. So there are still many difficult problems to solve to extract more better speaker features and apply it to actual conditions. This paper proposed a method which calibrates the features by feature normalization, feature folding and feature mapping based on MFCC features extraction. Then speaker recognition is conducted with the algorithm based on the combination of GMM and neural networks. The newly proposed algorithm shows higher recognition rate and better noise immunity.

2.Features Extraction and Calibration

The most valuable information for speaker recognition of multi-dimensional MFCC is concluded from the third dimension coefficient to the seventeenth dimension coefficient. Taking the noise and channel mismatch into consideration, this paper calibrates features from the aspects of feature normalization and feature folding to cut down the influence of noise and channel mismatches.

(1)Feature normalization

There are several advantages of feature normalization. First, feature normalization could eliminate features dimension and the physical meaning of features are weakened. Second, individual differences are eliminated to avoid calculation overflow. Last, feature dimension are balanced to keep features in a similar scale. This paper combines the cepstral mean method with feature normalization, conversion formula is as follow:

$$\overline{x_i} = \frac{x_i - x_{avg}}{\sigma}, i = 1, 2, ..., N \tag{1}$$

among them: $x_{avg} = \frac{1}{N}\sum_{i=1}^{N} x_i$, $\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - x_{avg})^2}{N}}$

(2)Feature folding

Feature folding is an effective method in the feature domain processing of speaker recognition. It shows great performance in the cross-channel identification. Steps of feature folding are as follows: (a) The number of vector to fold is calculated. The samples are sequenced from small to large. (b) First,

the transformed values corresponding to the min and max are figured out. Then the distribution function values of these two points are calculated to find out the distribution function values of each sample before transforming. (c) Transform the samples with chosen probability distribution function. The normal distribution is applied in this paper.

3.Improved Speaker Recognition Algorithm

Accuracy of the speaker recognition algorithm is improved by combining the GMM and BP neural network. Interactive information of different speakers in GMM output space is captured by BP neural network. In the case, the BP neural network works in the output space of model rather than original voice signal space.

The data of training set would not be misjudged because specific speaker's data is separated from other speakers' data in the space. While the data of confirming set is possible to be misjudged. The reason for misjudge is the overlapping of different models. The function of BP network uses the misjudged data to extract the interactive information among different speakers. GMM and BP combination model structure is shown in Fig.1:

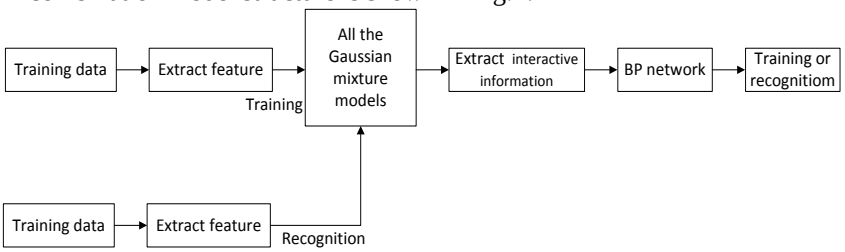

Fig. 1 GMM and BP combination model

The procedure steps are as follows: Assuming the number of training set N and the number of training sample is M. The training sample is $X^m(m=1,2...M)$. N pieces of GMM likelihood probability corresponding to the training sample is $\eta(X^m,b_n)(n=1,2...N)$. The input of BP neural network corresponding to the training sample is $\eta=[\eta_1,\eta_2...\eta_N]$. The input is normalized:

$$\eta_i=\frac{\eta_i}{\eta_{\max}} \tag{2}$$

Among them: $\eta_{\max}=\max\{\eta_i,i=1,2...N\}$

A fuzzy matrix is defined to describe the classification of generating GMM, the expression is:

$$M = \begin{bmatrix} n_{11}\ n_{12}\ \ldots\ n_{1N} \\ n_{21}\ n_{22}\ \ldots\ n_{2N} \\ \ldots\ \ \ldots\ \ \ldots\ \ \ldots \\ n_{N1}\ n_{N2}\ \ldots\ n_{NN} \end{bmatrix} \tag{3}$$

Among them, n_{ij} represents the number of voice which belong to speaker i misjudged to speaker j. The probability of speaker i misjudged to speaker j is figured out according to the fuzzy matrix.

$$p(j \mid X_i) = \frac{\eta_{ij}}{\sum_{n=1}^{N} \eta_{in}} \tag{4}$$

The output of neural network is defined according to conditional probability. The definition is :

$$O(j \mid i) = \begin{cases} 1 & \text{if } \ p(j \mid X_i) > 0 \,\&\&\, j = i \\ -1 & \text{if } \ p(j \mid X_i) > 0 \,\&\&\, j \neq i \\ 0 & \texttt{else} \end{cases} \tag{5}$$

The output of neural network is 1 if the speaker i is recongnized correctly. The output of neural network is -1 to show punishments if the speaker i is misjudged. The tangent function is chosen to be the activation function because the neural networks output ranges from -1 to 1.

The voice is transmitted to each GMM model in the recognition step. Then the probability corresponding to each model is obtained.The group of probability vectors are normalized according to equation (4). They are transmitted to trained BP network. The difference between output vector Y and each speaker's codeword is figured out according to Eq.(6):

$$d(Y, O_n) = \sum_{i=1}^{N} |Y - O(i \mid n)| \tag{6}$$

The criteria of recognition is the smallest distance.

$$K = \arg\min_{1 \le n \le N} d(Y, O_n) \tag{7}$$

4.Experiment Settings and Analysis

4.1. *Experiment Settings*

The ELSDSR (English Language Speech Database for Speaker Recognition)[10] is adapted by the experiments. ELSDSR is a dedicated database for speaker recognition. It is recorded by both teachers and students of mathematics Information College in Technical University of Denmark together. All the recording is conducted in a classroom by microphones. The database contains a

total of 23 people's voice data, including 21 Danes, an Irish and one Canadian. The sampling frequency of speech is 16000Hz. There are10 women and 13 men among the 23 people. Their ages range from 24 to 63.

The system performance is measured by probability of correct identification:

$$\text{Error rate} = \frac{\text{Misjudged number}}{\text{Total number of samples}} \tag{8}$$

4.2. *Speaker recognition experiments*

The GMM model with mixture degree of 64 can describe the data distribution better. So the mixture degree of 64 is adapted by the combined model. Neural network works in probability model output space. The dimensions of input data is 10.Over-fitting would be caused by too many neurons. So, the node structure is 20-20-20.

Algorithm performances under different noise conditions are contrasted. The GMM(M = 64), BP(100-50-50) and GMM+BP(20-20-20) models are chosen to contrast. Different level Gaussian noise(SNR 5dB, 10dB, 20dB, 40dB) are added to the test voice. The error rate and training time of the three models are recorded in table 1:

Tab. 1 Contrast of three different models

Model\Noise	none	40dB	20dB	10dB	5dB	Training time
GMM	0.4533	0.4745	0.5167	0.71	0.8333	16.5s
BP	0.1207	0.3194	0.4236	0.5623	0.6820	133s
GMM+BP	0.1123	0.1981	0.2698	0.3455	0.4684	91s

The performance of speaker recognition system is decreased seriously due to the introduce of noise. The error rate raises obviously when SNR is smaller than 20 dB. The BP network shows better performance than GMM in high noise level. While the performance is still worse than no noise condition. The possible reasons are analyzed from two aspects. First, the added noise would result in large differences between distribution of test data and training data. Second, extra neurons would result in over-fitting of the data. Recognition accuracy is decreased due to over-fitting of the data. The GMM+BP model is not sensitive to voice changing. It shows higher robustness than BP network. Because the GMM+BP model uses less neurons would not result in over-fitting. Over-fitting would not happen because the GMM+BP model uses less neurons. So the GMM+BP model shows better performance than BP model. The GMM+BP spends only three fourths training of BP model. So the GMM+BP model is more

excellent than normal BP model no matter from the training time or the robustness. The GMM+BP model shows better performance than BP model.

5.Summary

This paper proposes a improved speaker recognition algorithm to increase the robustness of speaker recognition algorithm in noisy environment. A combined model is proposed by ultilizing the disadvantages of GMM and BP neural network. The improved speaker recognition algorithm shows better performance and higher robustness than traditional algorithms.

Acknowledgments

This research project was founded in part by National Natural Science Foundation (No. 61375028) and the Scientific Research Funds of Nanjing Institute of Technology (No.TZ20160012 and No.CKJC201505).

References

1. D'arca E, Robertson N M, Hopgood J R. Robust indoor speaker recognition in a network of audio and video sensors. Signal Processing, 2016; 129: 137-149.
2. Mandasari M I, Saeidi R, Van Leeuwen D A. Quality measures based calibration with duration and noise dependency for speaker recognition. Speech Communication, 2015; 72: 126-137.
3. Dufour R, Estève Y, Deléglise P. Characterizing and detecting spontaneous speech: Application to speaker role recognition. Speech Communication, 2014; 56: 1-18.
4. Vijayan K, Raghavendra Reddy P, Sri Rama Murty K. Significance of analytic phase of speech signals in speaker verification. Speech Communication, 2016; 81: 54-71.
5. Madikeri S R. A fast and scalable hybrid FA/PPCA-based framework for speaker recognition. Digital Signal Processing, 2014; 32: 137-145.
6. Liu Y, Qian Y, Chen N, Fu T, Zhang Y, Yu K. Deep feature for text-dependent speaker verification. Speech Communication, 2015; 73: 1-13.
7. S.M A, K A, Rajendran R, Mohan A, P.S A, K M S, Aziz F. Efficient online and offline template update mechanisms for speaker recognition. Computers & Electrical Engineering, 2016; 50: 10-25.
8. You C H, Li H, Lee K A. Relevance factor of maximum a posteriori adaptation for GMM–NAP–SVM in speaker and language recognition. Computer Speech & Language, 2015; 30(1): 116-134.

9. Hossen A, Al-Rawahi S. A Text–Independent Speaker Identification System Based on the Zak Transform. Signal Processing: An International Journal, 2010; 4: 68-74.