Speaker Verification Channel Compensation Based on DAE-RBM-PLDA

Shuangyan Shan, Zhijing Xu

Speaker Verification Channel Compensation based on DAE-RBM-PLDA

Shuangyan Shan[1], Zhijing Xu[1]

College of Information Engineering, Shanghai Maritime University, Shanghai, China
993480720@qq.com

Abstract. In the speaker recognition system, a model combining the Deep Neural Network (DNN), Identity Vector (I-Vector) and Probabilistic Linear Discriminant Analysis (PLDA) proved to be very effective. In order to further improve the performance of PLDA recognition model, the Denoising Autoencoder (DAE) and Restricted Boltzmann Machine (RBM) and the combination of them (DAE-RBM) are applied to the channel compensation on PLDA model, the aim is to minimize the effect of the speaker i-vector space channel information. The results of our experiment indicate that the Equal Error Rate (EER) and the minimum Detection Cost Function (minDCF) of DAE-PLDA and RBM-PLDA are significantly reduced compared with the standard PLDA system. The DAE-RBM-PLDA which combined the advantages of them enables system identification performance to be further improved.

Keywords: Speaker recognition; I-vector; Denoising Autoencoders; Restricted Boltzmann Machine

1 Introduction

Speaker recognition is a kind of biometric technology, which is a technique for extracting effective feature information from speaker voice for speaker recognition. The more popular speaker recognition model is based on the Gaussian Mixture Model-Universal Background Model (GMM-UBM) [1]. Then, Patrick proposed Joint Factor Analysis (JFA) [2], Najim proposed modeling methods such as Identity-Vector (i-vector) [3]. The current i-vector has become the most effective technology which is the text-independent speaker recognition, this framework can be divided into three steps: First, using the GMM-UBM to express the speech acoustics feature sequence as a sufficient statistic, and then converted to low-dimensional features vector i-vector, after the i-vector is extracted, the Probabilistic Linear Discriminant Analysis (PLDA) model is used for channel compensation and the vertexes are obtained by comparing the i-vector generation verification scores of the different speech segments.

In recent years, the deep neural network (DNN) has been successfully applied to the field of speech recognition [4]. In the field of speaker recognition, Lei [5] used DNN to classify phonetic features into different phoneme spaces based on phoneme features, and then extract the acoustic features of different utterances in each space, and propose i-vector based on DNN. This model uses the output of the DNN output layer softmax in the UBM to calculate the various posterior probability, which results in significant performance improvement for the speaker confirmation.

Denoising Autoencoders (DAE) can reconstruct the raw data from the corrupted data by training. The feature of the speaker indicate that i-vector is affected by the influence of the speaker channel information as damaged data. Therefore, channel compensation can be achieved by DAE reconstruction method to obtain a more robust effect, resulting in noise immunity, thereby reducing the channel diversity of the speaker. In [6], RBM-PLDA-based channel compensation technology is proved to be superior to standard PLDA. RBM reconstructs i-vector by separating the speaker information and channel information, and then applies the factor containing the speaker information to the PLDA side for comparison. Based on the advantages of DAE and RBM, this paper proposes a channel compensation method based on DAE-RBM-PLDA to further reduce the influence of speaker channel diversity.

2 I-vector-based Speaker Recognition System

2.1 GMM I-vector Technology

The i-vector is a compact representation of a GMM supervector, containing both the speaker and channel characteristics of a given speech utterance. The model is based on the mean supervector represented by GMM-UBM. The mean super-vector of a speaker’s speech can be decomposed into the following equation:

$$M = m + T\omega \tag{1}$$

Where m is the Universal Background Model (UBM), a GMM mean supervector, T is a low-rank matrix defines the total variability space, ω is a speaker-and channel-dependent latent variable with standard normal distribution, and its posterior mean is i-vector.

In the process of i-vector extraction, we need to use the EM algorithm to estimate the global difference space matrix T, and extract the Baum-Welch statistic. The zeroth-order statistics and the first-order statistics of the speech segment h of the speaker s in the j-th GMM mixed component are as follows:

$$N_{j,h}(s) = \sum P(j \mid x_t) \tag{2}$$

$$F_{j,h}(s) = \sum P(j \mid x_t)(x_t - m_j) \tag{3}$$

Where $P(j \mid x_t)$ represents the posterior probability of generating the x_t in the Gaussian mixture component j in the UBM model:

$$P(j \mid x_t) = \frac{w_j P(j \mid x_t)}{\sum_{j-1}^{M} w_j P(j \mid x_t)} \tag{4}$$

And then the following calculation can be obtained corresponding i-vector.

$$\omega_h = E[W_h] = I^{-1} T^T \sum^{-1} F_h \tag{5}$$

2.2 DNN I-vector Technology

GMM has a strong ability to fit, but its shortcoming is that it cannot effectively model nonlinear or near nonlinear data. Therefore, DNN is applied to acoustic modeling, DNN's multi-layer nonlinear structure makes it a powerful characterization capability, it uses unsupervised generation algorithm for pre-training, and then use the back propagation algorithm for parameter fine tuning.

DNN consists of input layer, multiple hidden layer and Softmax output layer. The Softmax layer gives the posterior probability $P(j \mid x_t)$ of the bound three-factor state class on the speech frame, which is used as the corresponding Gaussian occupancy rate, substituting the formulas (2) and (3) to estimate the DNN i-vector zero-order statistics and first-order statistics, and then extract i-vcetor according to formula (5). DNN-based i-vector extraction process and the identification process are as shown in Figure 1:

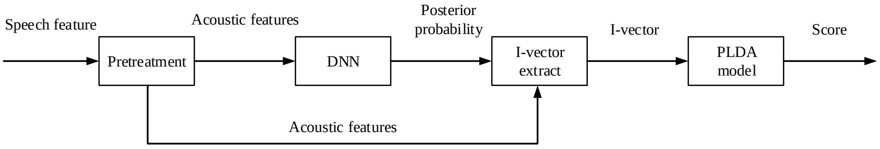

Fig.1. DNN-based Speaker Identification System Flow Chart

3 Analysis of Back - end PLDA Technology

3.1 PLDA Model

PLDA is an i-vector-based channel compensation algorithm, i-vector feature contains speaker information and channel information. We only need to extract the speaker information, so channel compensation is needed to remove channel interference. The simplified PLDA proved to be an effective method of channel compensation [7]. The simplified PLDA model is as follows:

$$\omega_{sh} = \mu + Vy_s + Z_{sh} \tag{6}$$

Where ω_{sh} is the i-vector representing h-th session of s-th speaker, μ is the mean of all training data, matrix V describes the subspace of the speaker, characterizes the differences between human beings, y_s is the hidden speaker factor, z_{sh} is the residual noise. The above parameters satisfy the following distribution:

$$y_s \sim N(0,1) \tag{7}$$

$$z_{sh} \sim N(0, D) \tag{8}$$

The purpose of the PLDA training phase is to estimate the parameter $\theta = \{\mu, V, D\}$ required by the model using the EM algorithm based on the speaker's speech data set for a given sample. Identify the score after the model is trained, the i-vector which is given the same speaker registration and testing are ω_e and ω_s respectively, the formula for calculating the likelihood ratio score is as follows:

$$\text{Score} = \ln \frac{P(\omega_e, \omega_s | H_0)}{P(\omega_e, \omega_s | H_1)} \tag{9}$$

Where H_0 indicates that ω_e and ω_s are from the same speaker, and H_1 is from different speakers. Calculate the likelihood ratio of the two Gaussian functions as the final decision for the score.

3.2 PLDA Based on DAE and RBM

Denoising Autoencoders (DAE) is a self-coding device by special training. Accept the damaged data as input in the input, and training to predict the original undamaged data as an output of the automatic encoder, to produce anti-noise ability, resulting in a more robust data reconstruction effect. DAE training process is shown in Figure 2. Introducing a damage process $C(y \mid x)$, which represents the probability that the given data x will produce a corrupted sample y . The automatic encoder assumes that x is the original input, and the noise reduction automatic encoder uses $C(y \mid x)$ to introduce the damaged sample y . And then taking y as the damage input with noise, taking x as an output, and self-coding for learning and training. The application of DAE to the speaker recognition system back-end model was first proposed in [8], and this paper will continue to explore further improve system performance. The i-vector can be regarded as a damaged data which is influenced by the speaker's channel information in this system. The training can be simplified as follows.

In the experiment, the DAE training starts from generative supervised training of the denoising RBM as shown in Figure 3. This RBM has binary hidden layer and Gaussian visible layer, taking a concatenation of two-valued vectors as an input. The first vector $i(s)$ is the average over all sessions of this speaker, the second vector $i(s,h)$ is an i-vector extracted from the h-th session of s-th speaker. RBM is trained by CD algorithm [9], weight matrix parameters V , W are used to initialize the DAE model.

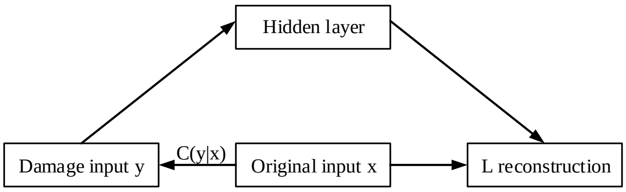

Fig.2. DAE Structure Schematic Diagram

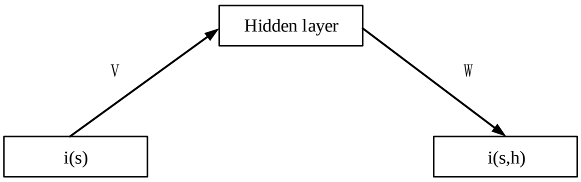

Fig.3. RBM Pre-training

After the pre-training model is expanded as shown in Figure 4, this model can be seen as a standard DAE model to rebuild i-vector. The output uses the speaker's average i-vector to reduce the difference in speaker channel information. Then the back propagation algorithm is used to tune the network parameters. The output of the DAE is whitened and length normalizated and it can be validated directly as a standard PLDA model input (DAE-PLDA), the judgment is based on a pre-set threshold.

RBM is an undirected model consisting of a random layer of visible neurons and a layer of hidden neuron. It can act on the PLDA channel compensation side, the hidden layer is decomposed by the speaker information factor and the channel information factor, as shown in Figure 5. We use the similar algorithm [6] to carry out training, the difference is that it is consistent with the hidden layer values for the previous DAE pre-training, where the hidden layer uses binary values and subjects to the Gaussian Bernoulli distribution. In the recognition phase, the visible layer inputs the i-vector of the speaker and the speaker's speech containing the speaker information at the output as input of the PLDA model (RBM-PLDA) is used to score the comparison.

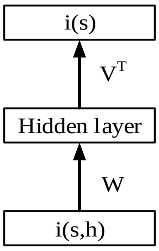

Fig.4. DAE

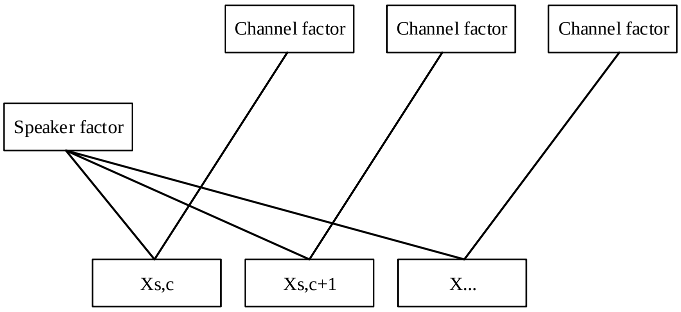

Fig.5. RBM-PLDA

From the above analysis we can see the effective feature extraction principle based on DAE and RBM . Use DAE and RBM mixed method, the first layer is the DAE, after the whitening and length normalization as RBM input, RBM and standard PLDA is combined to form a discriminant model, recorded as DAE-RBM-PLDA. The block diagram for the system is shown in Figure 6.

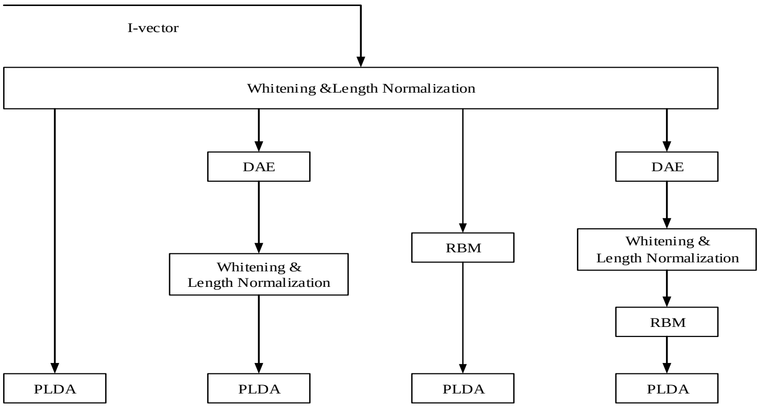

Fig.6. PLDA, DAE-PLDA, RBM-PLDA, DAE-RBM-PLDA Process

4 Experiments and Results

In this paper, DAC13 is used as experimental phonetic database, and the Equal Error Rate (EER) and minimum Detection Cost Function (DCF) are used as performance evaluation indexes.

In the UBM i-vector system, MFCC and the one-dimensional energy and its first and second order differences, which are total 39-dimensional MFCC features. Voice frame length is 25ms, frame shift is 10ms. In the DNN i-vector system, DNN speaker features are 40-dimensional Filter Bank features and its first and second order differences, which are total 120-dimensional. DNN has 5 hidden layers, each layer includes 2048 nodes. We first compared the performance of the standard PLDA model in the UBM i-vector and DNN i-vector systems. Experiments show that the recognition performance of DNN system is significantly improved by comparing with GMM-UBM system. DNN i-vector PLDA is the baseline system, the performance comparison is shown in Figure 7 and Table 1.

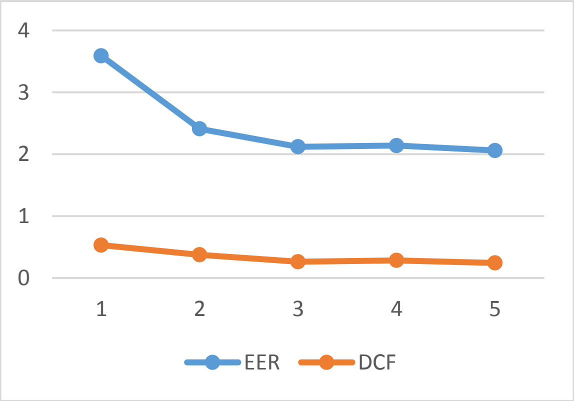

Fig.7. Model Performance Line Chart

Table 1. PLDA, DAE-PLDA, RBM-PLDA, DAE-RBM-PLDA Performance Comparison

System	Back-end Model	EER,%	minDCF
1 UBM	PLDA	4.49	1.432
2 DNN	PLDA	3.32	1.275
3 DNN	DAE-PLDA	2.79	0.863
4 DNN	RBM-PLDA	2.91	0.954
5 DNN	DAE-RBM-PLDA	2.82	0.913

From the experimental results of Table 1, it can be seen that the equal error rate and the minimum detection cost function of DAE-PLDA and RBM-PLDA back-end channel compensation model with deep learning model are significantly lower than those of standard PLDA model system. The performance improvement of the DAE-RBM-PLDA model is more obvious than the baseline system, which is 15.1% higher than the baseline system, which verified the effectiveness of the channel compensation method.

5 Conclusions

In this paper, we proposed the speaker verification channel compensation method based on DAE-RBM-PLDA, the method is combining the advantages of DAE and RBM. This method first performs RBM pre-training and initializes the DAE model with the i-vector processed by whitening and length normalization. The output of DAE is the average i-vector of over all sessions of the speaker, so it reduced the influence of the speaker channel information. And then combined with the RBM, the output i-vector of DAE is used as the input of RBM, hidden layer reconstruction separates speaker information and speaker channel information, select the speaker information needed for the experiment to perform the final likelihood ratio score of the back-end PLDA, further reducing the channel diversity of the speaker. The speaker confirmation experiment on the DAC13 dataset shows that the DAE-RBM-PLDA model, which combines the advantages of both DAE and RBM, can effectively improve the recognition rate.

Acknowledgment

This study was funded by the National Natural Science Foundation of China (Grant No.61404083).

References

1. Reynolds D A, Quatieri T F, Dunn R B. Speaker verification using adapted Gaussian mixture models[J]. Digital signal processing, 2000, 10(1-3): 19-41.
2. Kenny P，Ouellet P，Dehak N，et al． A study of interspeaker variability in speaker verification[J]． Audio，Speech， and Language Processing， IEEE Transaction，2008，16(5) : 980-988．
3. Dehak N，Kenny P，Dehak R，et al． Front-end factor analysis for speaker verification ［J］． Audio，Speech，and Language Processing，IEEE Transactions on，2011，19(4) : 788-798．
4. Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups[J]. IEEE Signal Processing Magazine, 2012, 29(6): 82-97.

5. Variani E, Lei X, McDermott E, et al. Deep neural networks for small footprint text-dependent speaker verification[C]//Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014: 4052-4056.
6. Stafylakis T, Kenny P, Senoussaoui M, et al. PLDA using Gaussian Restricted Boltzmann Machines with application to Speaker Verification[C]//INTERSPEECH. 2012: 1692-1695.
7. Garcia-Romero D, Espy-Wilson C Y. Analysis of i-vector Length Normalization in Speaker Recognition Systems[C]//Interspeech. 2011, 2011: 249-252.
8. Novoselov S, Pekhovsky T, Kudashev O, et al. Non-linear PLDA for i-vector speaker verification[C]//INTERSPEECH. 2015: 214-218.
9. Hinton G E. A practical guide to training restricted boltzmann machines[M]//Neural networks: Tricks of the trade. Springer Berlin Heidelberg, 2012: 599-619.