Channel Mismatch Speaker Verification Based on Deep Learning and PLDA

Pengqi Li*, Guanyu Li*, Jiao Han, Tiankai Zhi, Di Wang

Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education, Northwest Minzu University, Lanzhou, Gansu 730000, China

* y182730597@stu.xbmu.edu.cn
lipengqi96@163.com

Abstract. At present, speaker recognition technology is developing continuously, but in real world, existing speaker recognition algorithms are generally disturbed by channel factors. With the increase of current data volume, neural network has also become a key technology for speaker recognition. This paper introduces a speaker recognition method based on supervised training and a channel compensation algorithm. The experimental results show that this method combined with channel compensation algorithm can remove the influence of channel noise and has a good speaker recognition effect. The experimental results show that the combination of x-vector and PLDA channel compensation algorithm can obtain better recognition rate and ensure the robustness of the algorithm.

1. Introduction

In recent years, the Gaussian mixture model (GMMs) has been regarded as the main method for speaker recognition applications [1]. Through continuous research, the GMM-UBM framework based on GMM is used to improve the performance of speaker recognition system. GMM-UBM, also known as the Gaussian mixture model-Universal Background model, was proposed by Reynolds and successfully became a typical model in speaker verification systems. Kenny [2-3] proposed joint factor analysis (JFA) technology, which can limit the difference between speaker and channel to two subspaces in the high-dimensional space of GMM supervector. However, using joint factor analysis, the channel space still contains a small amount of residual speaker speech information. At present, I-vector [4] is the most commonly used method in speaker recognition. Based on the joint factor analysis technique, it is proposed that speaker and session differences can be represented through a separate subspace. I-vector replaces the original variable length sequence of speech information by the fixed length vector, so the i-vector greatly facilitates the modeling and testing process of speaker recognition. Although these algorithms all take into account the influence of channel factors, some back-end compensation algorithms are proposed due to the high complexity of real scenes, such as the feature warping algorithm in the feature domain (Feature warping), mean variance normalization (Cepstrum Variance Normalization, CVN), Linear Discriminant Analysis (LDA) in the model domain, Within-Class Covariance Normalization (WCCN), Probabilistic Linear Discriminant analysis (PLDA), Tnorm, Znorm and other algorithms in the score domain.

With the continuous improvement of hardware equipment performance and the increase of data volume, the deep neural network (DNN) has been gradually promoted internationally. In the field of speech recognition, after the successful application of deep neural network in acoustic modeling, the...
performance of speech recognition has also made great progress. This paper mainly uses the combination of x-vector and PLDA channel compensation technology to improve the speaker recognition effect in the case of channel mismatch.

2. Speaker Embedding and compensation algorithm

2.1. Speaker Embedding x-vector

X-vector system \([7] [15]\) is a speaker recognition system based on DNN, which is trained by DNN to extract the representation of the speaker, and embedding the extracted speaker is called x-vector. As shown in figure 1. The whole system can be divided into two modules.

The first part is the frame-level feature extraction module. Since the speech signal is a time sequence signal and there is also time sequence related information between frames, the network layer here adopts the time delay neural network \([8][9]\) to extract the frame-level features. The time sequence structural information of the speech signal can be learned through multi-layer TDNN, and the final output of this module is the speaker features at the frame level.

The second part is the segment-level feature extraction module. For the frame-level speaker features extracted by TDNN, a statistical layer is used to calculate the mean and standard deviation of these features. This mechanism can transition the frame-level features of the speaker to segment-level features. At the same time, the mechanism of statistical layer can be used to normalize the feature of unequal frame level into sentence level feature of equal length. After the statistics layer, two full connection layers and a Softmax output layer are connected, and feature vectors are extracted from the first full connection layer as speaker representation x-vector.

![Fig.1 X-vector speaker recognition system framework](image)
2.2. Probabilistic Linear Discriminant Analysis

The extracted x-vector of speaker features not only contains the information of the speaker, but also contains the information of the pronunciation content and other noise information, such as channel noise and environmental noise. For the speaker recognition task, you only need to pay attention to the information that can verify the identity of the speaker. In order to reduce the impact of irrelevant information on the recognition results, PLDA model is adopted here to reconstruct the extracted speaker characteristics and divide the speaker information into other irrelevant information.

We define the ith speaker's jth sound as $x_{ij}$. Then, according to factor analysis, we define the generation model of $x_{ij}$ as follows:

$$x_{ij} = \mu + Fh_i + Gw_{ij} + \varepsilon_{ij}$$

Where, $w_{ij}$ stands for data mean, $F$ stands for speaker space, $G$ stands for noise space, $\mu$ stands for noise covariance, and $h_i$ stands for the implied variable related to the ith speaker. Represents the implied variable associated with the JTH speech of the first speaker, i.e., the representation of $X$ in the noise space. The vector represents the speaker independent vector (global mean), $F$ and $G$ represent the speaker and channel correlation matrix respectively, and the diagonal matrix $\Sigma$ is the residual variable. The variables $h$ and $x$ represent the speaker and the channel factor respectively.

We refer the readers to [6, 10, 11] for details on the model training procedure.

In speaker recognition tasks, the PLDA model is used for the classifier at the back-end. To calculate the log likelihood ratio of the two sounds to make a decision, the formula is as follows:

$$\text{score} = \log \frac{p(\eta_1, \eta_2 | H_s)}{p(\eta_1 | H_d)p(\eta_2 | H_d)}$$

In the above equation, if there are two test sounds, the hypothesis that the two sounds come from the same space is $H_s$, and the hypothesis that the two sounds come from different spaces is $H_d$. The similarity degree of the two sounds can be measured by calculating the log likelihood ratio. If the score is higher, the probability that the two sounds belong to the same speaker is higher.

3. Experiment Setup

This experiment is mainly based on Kaldi's recipe SRE '16, and the data used in the experiment is AISHELL2[12] dataset. The evaluation method used is equal error rate. In all systems, the speech signal is extracted with 40-dimension MFCC as the input feature, the frame length is 25ms, and the frame shift is 10ms. Besides, the voice activity detection (VAD) based on energy is adopted to filter out the non-voice segment, and then the input feature is processed by CMVN.

3.1. i-vector /UBM baseline system

The i-vector system uses MAP[13-14] and other technologies to perform matrix T estimation. Cosine distance and PLDA were respectively used in the test scoring. Among them, the UBM model contains 1024 Gaussian mixture Numbers. The corpus of training UBM is a subset of training global variance space matrix T, and the dimension of the global difference space matrix is 400. The global variance space matrix is initialized randomly and iterated for 4 times.

3.2. x-vector system

In the x-vector system, TDNN is mainly used to establish the speaker model. The network parameters of the whole system are shown in Table 1. The first 5 layers operate on the speech frame by using the
time-delayed neural network structure, and process the context centered on the current frame $T$. For the whole frame level feature extraction module, 15 frames of context information can be obtained. The frame-level output of all $T$ frames by stacking network layer frames in the statistical layer is calculated with the mean and standard deviation of these features. Through this process, information can be aggregated on the time dimension for subsequent network layers to process on the entire sentence. After the network training is completed, the speaker's representation is extracted from segment6 of the network layer and regarded as an x-vector.

### Table 1. Network parameters of the x-vector system

| The network layer | Time delay parameters | Context-dependent frame number | Number of nodes |
|-------------------|-----------------------|--------------------------------|-----------------|
| frame1            | \{t-2,t-1,t+1,t+2\}  | 5                              | 512             |
| frame2            | \{t-2,t,t+2\}        | 9                              | 512             |
| frame3            | \{t-3,t,t+3\}        | 15                             | 512             |
| frame4            | \{t\}                | 15                             | 512             |
| frame5            | \{t\}                | 15                             | 1500            |
| Stats pooling     | \[0,T\]              | T                              | 3000            |
| segment6          | \{0\}                | T                              | 512             |
| segment7          | \{0\}                | T                              | 512             |
| softmax           | \{0\}                | T                              | -N              |

Tables 2 show the performance of the baseline ivector system trained on the AISHELL2 train set (Ioschannel), the x-vectors/PLDA system trained on the same dataset. The results on AISHELL2 test set(microphone (44.1kHz, 16-bit); Android-system mobile phone (16kHz, 16-bit), iOS-system mobile phone) show consistent improvement of x-vector/ PLDA system compared to i-vector system baseline. The relative improvement amounts to 35% reduction in EER.

### Table 2. The EER (%) of i-vector and x-vector systems on the test set.

| Back-end | EER% |
|----------|------|
| i-vector | 1.51  |
| Cosine   | 1.193 |
| PLDA     | 3.614 |
|           | 2.186 |
| x-vector | 13.5  |
| Cosine(LDA) | 0.8  |

### 4. Conclusions

This paper mainly introduces the effects of existing speech recognition techniques in cross-channel scenarios. The speaker recognition system using x-vector model of supervised training effectively extract the speaker-discriminative speaker representation, improved the recognition performance. Compared with traditional Cosine technology as the back end, PLDA compensation model was used to effectively reduce channel factors and background noise interference, and reduce the influence of signal distortion on speaker model building. This method makes the speaker recognition performance have better robustness.

### Acknowledgments

This work was supported by The National Natural Science Fund (NO.61633013), this work was also supported by Northwest MinZu University the central college basic scientific research operating expenses special fund self-help graduate student project (No. Yxm2019114).

### References

[1] Campbell W M, Sturim D E, Reynolds D A. Support vector machines using GMM supervectors for speaker verification[J]. IEEE Signal Processing Letters, 2006, 13(5):308-
Kenny P. Joint factor analysis of speaker and session variability: Theory and algorithms[J]. CRIM, Montreal, (Report) CRIM-06/08-13, 2005.

[3] Kenny P, Boulianne G, Ouellet P, et al. Speaker and session variability in GMM-based speaker verification[J]. IEEE Transactions on Audio Speech & Language Processing, 2007, 15(4):1448-1460.

[4] Dehak N, Kenny P, Dehak R, et al. Front-end factor analysis for verification[J]. Audio, Speech, and Language Processing, IEEE Transactions on, 2011, 19(4):788-798.

[5] Furui S. Cepstral analysis technique for automatic speaker verification[J]. Acoustics Speech & Signal Processing IEEE Transactions on, 1981, 29(2):254-272.

[6] Prince S J D, Elder J H. Probabilistic Linear Discriminant Analysis for Inferences About Identity[C]// IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007. IEEE, 2007.

[7] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, X-vectors: Robust embeddings for speaker recognition," in Proc. ICASSP, 2018.

[8] Waibel A, Hanazawa T, Hinton G, et al. Phoneme recognition using time-delay neural networks[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 2002, 37(3):328-339.

[9] Snyder D, Garcia-Romero D, Povey D. Time delay deep neural network-based universal background models for speaker recognition[C]// Automatic Speech Recognition & Understanding. IEEE, 2016.

[10] S Ioffe, “Probabilistic linear discriminant analysis,” in Proc. ECCV, Part IV, LNCS 3954, 2006, pp. 531-542.

[11] S. J. D. Prince, Computer vision: models, learning, and inference, Cambridge University Press, 2012.

[12] Du J, Na X, Liu X, et al. AISHELL-2: Transforming Mandarin ASR Research Into Industrial Scale[J]. 2018.

[13] Gauvain J L, Lee C H. Bayesian Learning for Hidden Markov Model with Gaussian Mixture State Observation Densities[J]. Speech Communication, 2007, 11(2-3):205-213.

[14] Gauvain J L, Lee C H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains[J]. IEEE Transactions on Speech & Audio Processing, 1994, 2(2):291-298.

[15] Snyder D, Garcia-Romero D, Povey D, et al. Deep Neural Network Embeddings for Text-Independent Speaker Verification[C]// Interspeech 2017. 2017.