THUEE SYSTEM DESCRIPTION FOR MCE2018

Liang He, Hongquan Zhao

Department of Electronic Engineering, Tsinghua University (THUEE)
email: heliang@mail.tsinghua.edu.cn, xiaoquan365@gmail.com

ABSTRACT

This paper describes THUEE team's submission to Multi-target speaker detection and identification Challenge Evaluation.

***Index Terms*—** linear discriminant analysis (LDA), local pairwise linear discriminant analysis (LPLDA), probabilistic linear discriminant analysis (PLDA), multiobjective optimization training of probabilistic linear discriminant analysis (MotPLDA)

1. INTRODUCTION

Different from past NIST SREs, the multi-target speaker detection and identification challenge evaluation 2018 (MCE18) have two dominant features [1, 2]. First, the evaluation takes extracted i-vectors as inputs instead of audio segments. This practice is similar to the NIST SRE14 i-vector challenging. Yet, the NIST SRE14 i-vector challenging allows multiple on-line submissions while the MCE18 permits only one submission. Second, the MCE18 contains a multi-target speaker detection task which decides whether the test segment belongs to the blacklist speakers by a top-S stack detector, further more, which blacklist speaker is by a top-1 stack detector. The latter is the feature of this evaluation.

2. DATA USAGES

The MCE18 provides three i-vector setstraining, development and test sets. Each set consists of blacklist and non-blacklist (background) speakers.

For the training set, there are 3,631 blacklist speakers and 5,000 background speakers. Each blacklist speaker has 3 i-vectors, and there are 10,893 i-vectors for blacklist speakers in total in this set. For the development set, there also 3,631 blacklist speakers and 5,000 background speakers. Each speaker has only one i-vector. The blacklist speakers of the training and development sets are the same while he background speakers are not. No information is provided about the distribution of speakers in the test set. All the i-vectors are 600 dimension.

The MCE18 includes the Fixed and Open conditions. In the Fixed condition, we can only use data provided by the MCE18 to train our system. This limitation is removed in the Open condition. We only take part in the Fixed condition test.

The evaluation plan states that "*The participants are free to use the training and development set as they want.*" We would like to use as many data as possible for the test submission. Both the training and development sets are used to train our final detector. However, we have to preserve some data to tell which method or algorithm is more suitable during our system building stage. So, the development set is conserved as the self-evaluation set. In other words, we have two stages. In the first stage (system building stage), we use training set to train system model, e.g. LDA, PLDA and so on, and development set for self-evaluation. In the second stage (test stage), we use both training and development sets to train system model, and test set for evaluation submission. Our following reported figures are all from the first stage.

3. APPROACHES AND ALGORITHMS

There are two reasons that we don't use deep learning method in our system building. One is that according to our previous literature research, the deep learning back-end has no significant advantage compared with the tradition LDA-PLDA backend in the field of speaker recognition so far. The other reason is that we think the data is not sufficient to train a network with good generalization ability.

Our system is a classical back-end, includes length normalization, LDA, PLDA and score normalization in turn. We put focus on the improvement of traditional length normalization-linear discriminant analysis (LDA)-probabilistic linear discriminant analysis method (PLDA) [5, 3, 4]. The improvement involves with our two recent studies: local pairwise linear discriminant analysis (LPLDA) and multiobjective optimization training of probabilistic linear discriminant analysis (MotPLDA) [6, 7] [1] [2]. We try to use LPLDA and MotPLDA to replace LDA and PLDA, respectively.

The object of LDA is to perform dimensionality reduction while minimizing within-class covariance and maximizing between-class covariance. For a target class (or speaker), our task is to make a binary decision about whether a test ut-

[1] LPLDA:https://github.com/sanphiee/LPLDA
[2] MotPLDA:https://github.com/sanphiee/MOT-sGPLDA-SRE14

terance is from a specific target speaker. Generally, the non-target test utterances which are close to the target speaker are easily misjudged. The LPLDA focuses on maximizing the local pairwise covariance, which represents the local structure between the target class samples and neighboring non-target class samples, instead of the between-class covariance which represents the global structure of the data.

The model parameters of PLDA is often estimated by maximizing the log-likelihood function. This training procedure focuses on increasing the log-likelihood, while ignoring the distinction between speakers. The MotPLDA performs training by emphasizing both sides.

4. CONFIGURATIONS, PARAMETERS AND EXPERIMENT RESULTS ON DEVELOPMENT SET

The evaluation provides a baseline code (length normalization and cosine scoring). Our system is built on the provided code.

First, we try to find out the suitable LDA dimension via a length-normalization, LDA and score normalization system. From Table 1, we select 500 as our LDA dimension for the rest experiments.

Table 1. LDA experiment results on development set

LDA	Top-S,EER[%]	Top-1,EER[%]	Conf Error
100	3.20	13.22	477
200	2.90	9.66	345
300	2.92	9.18	321
400	2.85	9.22	330
500	2.76	9.20	325

Second, we compare LDA with LPLDA, see Table 3. From this table, we conclude that the LPLDA outperforms the LDA.

Table 2. LPLDA experiment results on development set

LPLDA	Top-S,EER[%]	Top-1,EER[%]	Conf Error
200	2.48	8.12	291
500	2.41	8.10	291

Third, we examine the effective of PLDA. Our procedure includes length-normalization, LDA, PLDA and score normalization system.

Table 3. PLDA experiment results on development set

PLDA	Top-S,EER[%]	Top-1,EER[%]	Conf Error
500	2.77	9.04	322

Finally, we use MotPLDA instead of PLDA to check the performance. The factor 1.7 and 5 are within-between class ratio parameter introduced in the MotPLDA, Table 4.

Table 4. MotPLDA experiment results on development set

MotPLDA	Top-S,EER[%]	Top-1,EER[%]	Conf Error
500,1.7	3.83	7.50	255
500,5	2.86	8.64	304

5. CONCLUSION AND FUTURE WORK

Our submitted score is a length normalization, LDA and score normalization system. The LPLDA, PLDA and MotPLDA seem more effective. But, they are late by the submission deadline. Our codes are on the Github [3].

In the future, we will study the usage of deep learning method.

6. REFERENCES

[1] Suwon Shon, Najim Dehak, Douglas Reynolds, and James Glass, "Mce 2018: The 1st multi-target speaker detection and identification challenge evaluation (mce) plan, dataset and baseline system," in *ArXiv e-prints arXiv:1807.06663*, 2018.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, May 2011.

[3] A.M. Martinez and A.C. Kak, "Pca versus lda," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 2, pp. 228 –233, feb 2001.

[4] Simon J.D. Prince and James H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proceedings of the IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[5] G.R. Daniel and Carol E.W., "Analysis of i-vector length normalization in speaker recognition systems," in *INTERSPEECH*, 2011, pp. 249–252.

[6] L. He, X. Chen, C. Xu, J. Liu, and M. T. Johnson, "Local pairwise linear discriminant analysis for speaker verification," *IEEE Signal Processing Letters*, pp. 1–1, 2018.

[7] L. He, X. Chen, C. Xu, and J. Liu, "Multiobjective Optimization Training of PLDA for Speaker Verification," *ArXiv e-prints*, Aug. 2018.

[3] https://github.com/sanphiee/MCE18-THUEE