# THE DKU-SMIIP SYSTEM FOR MULTI-TARGET SPEAKER DETECTION AND IDENTIFICATION CHALLENGE EVALUATION 2018

Danwei Cai, Ming Li

Data Science Research Center, Duke Kunshan University, Kunshan, China

#### **ABSTRACT**

In this manuscript, we present the system submission with fixed condition for the 1st Multi-target speaker detection and identification Challenge Evaluation. A complete description of the system components, including the algorithms and configurations, is given. Dataset used to train these components are also provided. We also report the performance of the submission system on the development set.

#### 1. SYSTEM COMPOMENTS

In this section, we describe the basic components that we use to build our systems for the 1st Multi-target speaker detection and identification Challenge Evaluation [1]. Our system first apply channel compensation on the provided i-vectors, then Probabilistic Linear Discriminative Analysis (PLDA) is adopted as a back-end modeling approach. The scores of the input i-vector and all the blacklist speakers are calculated as the log-likelihood ratio on the PLDA model. Finally, score normalization is applied to reduce the variability of decision score on different blacklist speakers.

# 1.1. Channel Compensation

We use Locality Sensitive Discriminant Analysis (LSDA) for channel compensation in our submitted systems [2].

As is commonly known, i-vectors model speaker-, languageand channel- dependent information within the same total variability subspace [3]. In order to select the most speaker relevant feature subset for PLDA modeling, LSDA is used to reduce the information irrelevant to the speaker.

LSDA finds k nearest neighbors globally for each sample, constructs within- and between-class graph to model the local geometrical structure. Then it finds a linear transform matrix to map the i-vectors into a subspace in which the margin between i-vectors from different speakers is maximized at each local area. In order to gain good performance, LSDA with k nearest neighbors requires the data samples in each class to be larger than k or close to k, but we can not guarantee that because the number of i-vectors in each speaker is heterogeneously distributed. Considering this inherent characteristic of the training set, we improve LSDA by using adaptive k nearest neighbors for each speaker. Since

the number of i-vectors for each speaker is not the same and sometimes even varies with a wide range, the speakers with fewer i-vectors have little influence in the objective function of LSDA. We further modify LSDA's within-class and between-class weight matrix to handle this issue of unbalanced data.

We set the hyperparameters of the LSDA in [2] as follows:  $k=20, \beta=5, \alpha=0.3$ . The dimension of the i-vectors after LSDA are 500.

## 1.2. Backend Modeling

We use Gaussian PLDA [4] as the back-end modeling method in our submitted system.

The dimensionality reduced i-vectors after LSDA are centered, whitened and unit-length normalized. The Gaussian PLDA model with a full covariance residual noise term is trained on i-vectors of the blacklist and background speakers. The eigenvoice subspace in the PLDA model is assumed to be full-rank.

Suppose we have multiply i-vectors for a blacklist speaker  $\{\mathbf{x}_{i,j}|i=1,2,\cdots,S;j=1,2,\cdots,N_i\}$ , where S is size of blacklist speakers and  $N_i$  is number of i-vectors of blacklist speaker i, the blacklist speaker model  $\mathbf{x}_i$  is the average of all the corresponding i-vectors:

$$\mathbf{x}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{x}_{i,j}$$

After the PLDA is trained, the scores of any given test i-vector  $\mathbf{x}$  and all the blacklist speaker models  $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_S\}$  are calculated as the log-likelihood ratio  $s(\mathbf{x}_i, \mathbf{x})$  on the PLDA model.

# 1.3. Score Normalization

We then apply score normalization to reduce the variability of decision score on different blacklist speakers. Here, we use Adaptive Zero Score Normalization (AZ-norm) [5].

AZ-norm use a cohort  $\mathcal{E}$  to perform score normalization. Here, we use the cohort  $\mathcal{E} = \{\mathbf{x}_{i,j} | i=1,2,\cdots,S; j=1,2,\cdots,N_i\}$  which which is all blacklist i-vectors. For the

**Table 1**. Dataset usage for the development system

Subset	LSDA	PLDA	AZ-norm	Blacklist Speaker Model
Training Blacklist				$\checkmark$
Training Background				

**Table 2.** Dataset usage for the evaluation system

Subset	LSDA	PLDA	AZ-norm	Blacklist Speaker Model				
Training Blacklist				$\sqrt{}$				
Training Background								
Development Blacklist				$\sqrt{}$				

**Table 3**. Performance on MCE 2018 development set

	Top S Detector EER	Top 1 Detector EER	Total Confusion Error					
Baseline	2.01%	12.24%	444					
PLDA	2.40%	7.56%	273					
LSDA + LSDA	5.98%	6.71%	232					
LSDA + PLDA + AZ-norm	5.89%	6.50%	218					

blacklist speaker model  $\mathbf{x}_i$ , the adaptive cohort  $\mathcal{E}_{\mathbf{x}_i}^{top}$  are selected to be the top K closest i-vectors of  $\mathbf{x}_i$  in  $\mathcal{E}$ . The cohort scores based on such selections for the blacklist speaker model  $\mathbf{x}_i$  are then:

$$S(\mathcal{E}_{\mathbf{x}_i}^{top}) = \left\{ s(\mathbf{x}_i, \varepsilon) | \forall \varepsilon \in \mathcal{E}_{\mathbf{x}_i}^{top} \right\}$$

The AZ-norm score of  $s(\mathbf{x}, \mathbf{x}_i)$  is

$$s_{AZ-norm} = \frac{s(\mathbf{x}_i, \mathbf{x}) - \mu[S(\mathcal{E}_{\mathbf{x}_i}^{top})]}{\sigma[S(\mathcal{E}_{\mathbf{x}_i}^{top})]}$$

where  $\mu(\cdot)$  and  $\sigma(\cdot)$  are the mean and stand deviation.

#### 2. DATASET USAGE

MCE 2018 provides 5 subsets of i-vectors, including the blacklist and background subsets in training set, the blacklist and background subsets in development set and the evaluation set. The dataset usage of development system and evaluation system are in the table 1 and 2.

## 3. SYSTEM PERFORMANCE

The system performance on MCE 2018 development set is shown in table 3.

### 4. REFERENCES

[1] S. Shon, N. Dehak, D. Reynolds, and J. Glass, "The 1st Multi-target speaker detection and identification Challenge Evaluation (MCE) Plan," *ArXiv e-prints arXiv:1807.06663*.

- [2] D. Cai, W. Cai, Z. Ni, and M. Li, "Locality Sensitive Discriminant Analysis for Speaker Verification," in 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016.
- [3] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis for Speaker Verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.
- [4] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of I-Vector Length Normalization in Speaker Recognition Systems," in *Proceedings of the Annual Conference* of the International Speech Communication Association, INTERSPEECH, 2011, pp. 249–252.
- [5] D. E. Sturim and D. A. Reynolds, "Speaker Adaptive Cohort Selection for T-norm in Text-Independent Speaker Verification," in *Proceedings of IEEE International Con*ference on Acoustics, Speech, and Signal Processing, 2005., March 2005, vol. 1, pp. I/741–I/744 Vol. 1.