## **System Description for MCE 2018**

Jianbo Ma<sup>1</sup>, Vidhyasaharan Sethu<sup>1</sup>, Eliathamby Ambikairajah<sup>1,2</sup>

<sup>1</sup>School of Electrical Engineering and Telecommunications, UNSW Sydney <sup>2</sup>DATA61, CSIRO, Sydney, Australia

### 1. INTRODUCTION

The aim of multi-target speaker detection and identification is to determine whether a recorded utterance was spoken by one of a large number of "blacklisted" speakers. Additionally, if the answer is positive, then find the speaker among the "blacklisted" speakers [1].

The system proposed by our team is implemented as shown in Fig. 1, in which the test file is tested against all the speakers in the blacklist. A score generator (e.g. cosine distance similarity) is adopted to generate scores for each speaker (in the backlist). One score is chosen or generated and decision is made by comparing a threshold as in [2]. In this implementation, cosine distance similarity (CDS) [3] and Gaussian probabilistic linear discriminant analysis (GPLDA) [4] are used to generate scores. Furthermore, to better distinguish between speakers who are close to each other in the space of provided data format, which is the ivector space [3], a hierarchical system is used.

This hierarchical system is similar to the one previously used in language identification [5]. A high level block diagram of the hierarchical system is shown in Figure 3. The structure of hierarchical system is a tree structure, where the root consists of a single group that contains all speakers. The subsequent lower levels of the hierarchy represent a smaller cluster of speakers and the last level of hierarchy represents individual speaker. Each level of the tree acts as a recognition system, with every node representing a possible hypothesis. The aim of this structure is to divide the detection task into sub-tasks with easier classification sub-tasks closer to the root and harder ones near the leaves [5].

Scores from different systems are fused via logistic regression [6] to form the final submission shown in Fig. 2.

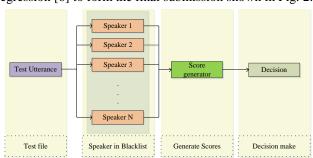
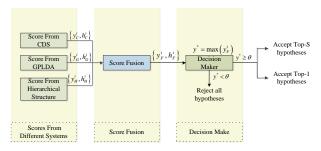
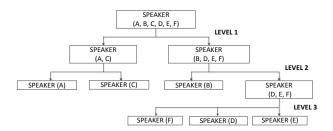


Fig. 1. Diagram of system for MCE 2018



**Fig. 2.** Score fusion and decision make process of submitted system ( $\theta$ : the threshold)



**Fig. 3.** Overview of Hierarchical Framework for Language Identification

### 2. System Description

#### 2.1. Hierarchical system

The hierarchical structure is motivated by the idea of separating detection task into sub-tasks such classification in each group can use discriminative information within that group, which may make the classifier tailored for the group and potentially provides higher classification accuracy. An agglomerative clustering algorithm [7] is used to determine these clusters. GPLDA is used to generate the confusion matrix used in the agglomerative clustering algorithm. Two levels are used in this implementation. Fig. 4 shows the hierarchical structure on the enrolment speakers ('blacklist' speakers). In the first level, 10 groups that generated by the agglomerative clustering algorithm are chosen. Other individual speakers are treated as one group for each speaker. In the second level, each speaker in the first 10 groups is treated as single node. In the test phase, the test file is tested against the groups in the first level. In the second level, the classifier (e.g. GPLDA) is trained to classify speakers within each group and test file is compared with each member of each group. Scores under each node is normalized to [0,1] by logistic regression and scores on each path that starts from root to leaf are multiplied to form the final score for a single speaker in blacklist.

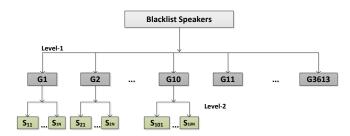


Fig. 4. Hierarchical structure of enrolment speakers

## 2.2. Score generator

In this implementation, cosine distance similarity (CDS) and Gaussian probabilistic linear discriminant analysis (GPLDA) are used to generate scores. When training the GPLDA for the hierarchical system, speakers that are clustered in the same group are treated as they are from the same speaker. Each node in the second level needs to train a separate GPLDA classifier. Members in each node are then treated as different speakers and used to adapt a GPLDA model trained on background speakers of training set by weighted likelihood specified in [8].

#### 2.5. Experimental Setup

Before feeding into the classifier, the raw i-vectors are projected by using linear discriminative analysis (LDA) into 400 dimensions followed by length normalization [9] to make the distribution of the i-vectors more Gaussian. GPLDA models are trained over 10 iterations using the MSR toolbox [10] with 400 dimension speaker factors. When training the GPLDA for the hierarchical system, speakers that are clustered in the same group are treated as the same. For adapting the GPLDA model in hierarchical system, the hyperparameter  $\alpha$  is chosen as 0.25, selected empirically based on performance on the development set. For fusing scores from different systems, a logistic regression model is used and development set is used to train the coefficients of this model.

## 3. Training Data

All the background speakers from the training set are used to train the GPLDA models. Blacklisted speakers from the training set are also added to the GPLDA training in development. For the final evaluation set, all the speakers in blacklist of development set are assigned to each speaker in

the training set by the ground truth provided by organizers. For training the logistic regression model in order to fuse scores, blacklist speakers of training and development set are used to generate training scores.

# 4. Development results

The Development set consist 3631 speakers in blacklist and 5000 speakers in background. Table 1 gives the performance of the system using GPLDA, CDS as score generator and the hierarchical system. Both Top-S and top-1 are reported by using Equal Error Rate (EER).

Table 1: System performances as evaluated on development

System	Top-S (EER %)	Top-1 (EER %)
Hierarchical	1.90	10.46
system		
GPLDA system	2.10	8.66
CDS system	1.70	12.26
Fusion	2.08	8.74

#### 5. REFERENCES

- [1] S. Suwon, R. Douglas, and G. James, "MCE 2018: The 1st Multi-target speaker detection and identification Challenge Evaluation (MCE) Plan, Dataset and Baseline System," *ArXiv e-prints arXiv:1807.06663.*, 2018.
- [2] E. Singer and D. A. Reynolds, "Analysis of multitarget detection for speaker and language recognition," in ODYSSEY04-The Speaker and Language Recognition Workshop, 2004.
- [3] 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, pp. 788-798, 2011.
- [4] P. Kenny, "Bayesian Speaker Verification with Heavy-Tailed Priors," in *Odyssey*, 2010, p. 14.
- [5] S. Irtza, V. Sethu, H. Bavattichalil, E. Ambikairajah, and H. Li, "A hierarchical framework for language identification," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016, pp. 5820-5824.
- [6] C. M. Bishop, "Pattern recognition and machine learning (information science and statistics) springer-verlag new york," *Inc. Secaucus, NJ, USA*, 2006.
- [7] S. Irtza, V. Sethu, S. Fernando, E. Ambikairajah, and H. Li, "Out of Set Language Modelling in Hierarchical Language Identification}," *Interspeech 2016*, pp. 3270-3274, 2016.
- [8] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for i-vector based speaker recognition," in

Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on, 2014, pp. 4047-4051.

- [9] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector Length Normalization in Speaker Recognition Systems," in *Interspeech*, 2011, pp. 249-252.
- [10] S. O. Sadjadi, M. Slaney, and L. Heck, "Msr identity toolbox v1. 0: A matlab toolbox for speaker-recognition research," *Speech and Language Processing Technical Committee Newsletter*, vol. 1, 2013.