BIOMETRIC VOX SYSTEM FOR THE MCE 2018 CHALLENGE

Roberto Font

Biometric Vox S.L., Spain

ABSTRACT

This paper describes the systems we developed for the MCE 2018 1st Multi-target speaker detection and identification Challenge Evaluation. The first system uses Linear Discriminant Analysis followed by Probabilistic Linear Discriminant Analysis for scoring, while the second system uses an ensemble of Denoising Autoencoders to compute more representative speaker embeddings. Both systems use Top S-Norm for score normalization. Our best system achieves a Top-1 EER of 6.52%, which represents a 46.8% error reduction with respect to the baseline system provided by organizers.

Index Terms— MCE 2018 challenge, speaker recognition, blacklist detection, denoising autoencoder, speaker embeddings

1. INTRODUCTION

The task for the MCE 2018 1st Multi-target speaker detection and identification Challenge Evaluation is to detect if a given speech segment belongs to any of the speakers in a blacklist. The challenge is divided into two related subtasks: Top-S detection, or to detect if the segment belongs to any of the blacklist speakers or to a different speaker; and Top-1 detection, or to detect which specific blacklist speaker (if any) is speaking in the segment. We refer the reader to [1] for a detailed description of the challenge.

This paper presents a complete description of the systems we trained for the evaluation. We used two different backends: a) a Linear Discriminant Analysis (LDA) followed by Probabilistic Linear Discriminant Analysis (PLDA) backend, which is the de-facto standard in speaker recognition, and b) an alternative backend consisting on a Denoising Autoencoder (DAE), trained to map each individual sample to the mean of all samples from that speaker, followed by PLDA scoring. Both systems use Top S-Norm for score normalization.

The rest of the paper is organized as follows. In Section 2 we provide a brief overview of the challenge, in order to make the paper as self-contained as possible. Section 3 provides the description of the systems we developed for the challenge. Finally, Section 4 describes the composition of our final submission.

2. CHALLENGE OVERVIEW

The MCE 2018 data is generated from real-world call center user-agent telephone conversations. Instead of providing the raw audio data, organizers processed the data and generated a fixed-sized 600-dimensional representation of each of the samples. This way, each speech segment is represented by a 600-dimensional vector that is known as i-vector [2]. Although some alternative representations have been recently proposed, i-vectors have been the state of the art representation of speech utterances for the last decade. The details on the training of this *i-vector extractor* can be found in [1].

The challenge data was distributed to the participants divided into three separate subsets: training, development and evaluation. Training and development portions were labeled with speaker identity, while evaluation set was unlabeled. The composition of the different subsets is summarized in Table 1.

	# speakers	# utts. per speaker	Total utts.
Training blacklist	3631	3	10893
Training background	5000	≥ 4	30952
Development blacklist	3631	1	3631
Development background	5000	1	5000
Test	?	?	16017

Table 1. Summary of the different subsets

Two complementary tasks are considered: Top-S detection and Top-1 detection. For Top-S detection, the system must decide if a test sample belongs to any of the blacklist speakers or not. For Top-1 detection, the system must decide if a test sample belongs to a particular blacklist speaker or not [1]. For both tasks, the performance metric is Equal Error Rate (EER).

Finally, the challenge considers two different training conditions: fixed condition, where only the data provided by the organizers can be used for system training and development, and open condition, where the use of any external data is allowed.

3. SYSTEM DESCRIPTION

For our system development, we focused on the Top-1 detection task. We find it a more challenging and complete task, in the sense that improving Top-1 detection will in general improve Top-S detection –at the limit, a perfect Top-1 detector would also have perfect Top-S performance-, but the converse is not true: a perfect binary blacklist/not blacklist detector would provide no information about the identity of the particular blacklist speaker.

We developed two main systems for the challenge. The first one is the usual speaker recognition backend consisting on LDA followed by PLDA. The second system uses a Denoising Autoencoder to increase the discriminative power of the i-vectors prior to PLDA scoring. The next two sections describe in detail each one of these approaches.

LDA system:

This system uses the backend that can be considered the state of the art in speaker recognition. In particular:

- I-vectors are projected to unit length (lengthnormalized).
- LDA is used to project the i-vectors to a lower dimension and maximize their discriminative power.
- PLDA is used to compute the score and compensate for between-session variability.

In our case, the dimension after the LDA projection is 450 and the PLDA variant employed is the two-covariance model [3]. For model selection and hyperparameter tuning we trained on the training set and evaluated over the development set. Both LDA and PLDA models were trained using background + blacklist training data. The results of this system are shown in Table 2, second row.

For score normalization, we used test-normalization (T-Norm) and symmetric normalization (S-Norm). For both normalization schemes, a set of speakers, in this case the background speakers from the training set, are used to score the test or enrollment segment against each one of the speakers in this cohort. This way, the score after T-Norm is given by

$$S_t = \frac{S - \mu_t}{\sigma_t}$$

and after S-Norm by
$$S_s = \frac{1}{2} \left(\frac{S - \mu_t}{\sigma_t} + \frac{S - \mu_m}{\sigma_m} \right),$$

where μ_t and σ_t are the mean and standard deviation of the scores of the test segment against the cohort and μ_m and σ_m are the mean and standard deviation of the scores of the speaker model against the cohort.

We have used a variant termed Top Norm in which only the top N scores are considered to estimate the mean and standard deviation. For T-Norm we use N=1000 while for S-Norm we use $N_t = 2000, N_m = 3000$. These values were selected by performing a grid search.

Results after score normalization are shown on Table 2, third and fourth columns, where we can see that this extra step brings a considerable improvement.

DAE system:

For this system, we trained a neural network to minimize the cosine distance between each i-vector and the mean of all i-vectors from that speaker. This way, we obtain representations that contain more information about the speaker and are less affected by inter-session variability.

This Denoising Autoencoder (DAE) has a single hidden layer with 2000 units and tanh activation, and an output layer with dimension 600 and linear activation. We train the neural network with Adam optimizer for 5 epochs using only the background training set. As an extra regularization step, we trained an ensemble of 10 of these networks and averaged their predictions to compute the final embeddings. We used Keras with Tensorflow backend.

Once we had the DAE embeddings, we used them to train a PLDA backend as in the previous system. PLDA is trained on background and blacklist training data.

Finally, we use S-Norm as described above. For this system, we use $N_t = 1000$, $N_m = 500$.

Results are shown in the last two rows of Table 2. As we can see, this system obtains slightly better results than the LDA system.

System	Top-S Error	Top-1 Error
Baseline	2.01	12.26
LDA + PLDA	1.82	6.96
LDA + PLDA + S-Norm	1.26	6.72
LDA + PLDA + T-Norm	1.09	6.80
DAE + PLDA	1.73	7.22
DAE + PLDA + S-Norm	1.25	6.52

Table 2. Performance on development set

4. FINAL SUBMISSION

For the final submission, we pooled together the blacklist training and development sets and used this combined set instead of the blacklist training set. This way, speaker models are computed using the 4 utterances available between training and development sets. Apart from this modification, the workflow described above is applied to score the evaluation set.

We experimented with score-level fusion between systems. However, we did not see consistent improvement and decided to submit the scores of individual systems. The content of our submission to the competition is summarized in Table 3.

Submission	System	
Fixed primary	DAE + PLDA + S-Norm	
Fixed contrastive 1	LDA + PLDA + S-Norm	
Fixed contrastive 2	LDA + PLDA + T-Norm	

Table 3. Composition of our final submission

5. 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," *ArXiv e-prints arXiv:1807.06663*. 2018.
- [2] Najim Dehak, Patrick J Kenny, Reda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-End Factor Analysis for Speaker Verification," IEEE Trans. on Audio, Speech, and Lang. Process., vol. 19, no. 4, pp. 788–798, may 2011.
- [3] N. Brümmer and E. De Villiers. "The speaker partitioning problem", In Proc. of the Odyssey Speak. and Lan. Recog. Workshop, Brno, Czech Republic, 2010.