PINDROP SUBMISSION TO MCE 2018

Elie Khoury, Khaled Lakhdhar, Andrew Vaughan, Ganesh Sivaraman, Parav Nagarsheth

Pindrop, Atlanta, USA

ABSTRACT

This papers summarizes Pindrop's submission to the MCE 2018 challenge. The main task of this challenge is the open set identification problem that aims to answer the following two questions: Is the speaker on a known list of speakers? If so, whose speaker is she/he? In the context of call centers, the list of bad-actor speakers is also known as Blacklist. While one single system can answers both questions (like the MCE baseline [1]), we looked at them as two separate problems: Blacklist detection and closed-set identification. For the blacklist detection, we propose four different techniques: a PLDA-based system, two DNN flavors, and a system based on Cosine similarity and logistic regression. For the closedset identification, we propose a fusion of PLDA-based and shallow neural network based systems. The proposed systems reduce Top-S EER by up to 60.7% and Top-1 EER by up to 53.3% on the development data over the MCE baseline.

Index Terms— MCE, Blacklist Detection, Open Set Speaker Identification, Mutli-speaker Detection

1. INTRODUCTION

Traditional literature on speaker recognition focuses on evaluating systems for speaker verification, where a speaker makes a claim about their identity. Open-set speaker identification, as a sub-problem of speaker recognition, has received far less attention. The problem is described as open-set because it is possible for a test utterance to not belong to a registered set (blacklist) of speakers. Traditional techniques used in closed-set identification and verification may not work very well because of the a) variability of the size of the blacklist set b) variability of the (unknown) non-blacklisted speaker set. The scope of the MCE challenge constrains the former. In practice, it may be unconstrained.

[2] discuss a system of stacked detectors that output k scores for k blacklist speakers. An identification decision is made by comparing the maximum score to a threshold. In the parallel field of face recognition, the open-set challenge is often addressed using a mix of classification and score normalization techniques [3]. Our approach is two-fold as well. We describe traditional i-Vector similarity based systems (PLDA and cosine) with adaptive symmetric score normalization to improve the correct identification of a blacklisted speaker.

Additionally, we describe two classifiers using DNNs that model the space of blacklisted speakers. We do a score-level fusion of these systems to take advantage of any complementary characteristics. In the evaluation, we present the tradeoffs on selecting these systems for the two metrics chosen for challenge: Top-S errors (blacklist detection) and Top-1 errors (closed-set identification).

2. SYSTEM DESCRIPTION

Our work focuses solely on the fixed condition. We used the full training data (TRAIN) in developing our system. Our testing was done on the development data (DEV), similar to the baseline system. The DEV data was added to TRAIN only in the last moment for the final submission.

2.1. Blacklist Detection

2.1.1. PLDA-Based system

Similar to the MCE baseline, this system is used to score each test utterance against every speaker from the blacklist S_i , and then use the maximum score to decide if it's a blacklist speaker or a genuine speaker. Before applying Probabilistic Linear Discriminant Analysis (PLDA), the raw i-vectors are processed using linear discriminant analysis (LDA) and Length-Normalization. Both LDA and PLDA are trained using the full training data (i.e. both background and blacklist lists. Finally, we normalize the scores using adaptive symmetric normalization. The cohort set used for normalization is an augmented background training data. The augmentation is done by applying at random a weighted sum between background i-vectors and blacklist i-vectors. The rationale behind this augmentation is to generate more challenging negative samples. This system achieves a Top-S EER of 1.32% on DEV.

2.1.2. Simple Feed-Forward DNN

In contrast to the PLDA system, our simple feed-forward DNN system is trained for binary classification. The input to the DNN are length-normalized i-vectors, and the output is either genuine or blacklist classes. The DNN consists of 5 hidden layers, with 4096, 512, 256, 256, and 256 units respectively. Batch-normalization and Dropout are used for

regularization. Similar to the PLDA system, we used data augmentation to double the size of the background samples. At test time, the input values of the softmax activation on the positive class are used for scoring. This system achieves a Top-S EER of 1.26% on DEV.

2.1.3. Wide-Deep DNN

Two systems are presented making use of a "wide-deep" architecture [4]. The first was trained using hamiltonian monte carlo SGD methods described in [5]. The second was trained with simple SGD with an L1 cost function with target values -10 and +10.

Wide-deep architecture seems suited to the problem of this challenge, where the task of the deep network will be to correct the embedding of a simple wide memorization network. This system achieves a Top-S EER of 1.06% on DEV. While training the network, it was noticed that score distributions varied significantly between the training set and the development set. By training with L1 loss, this effect was mitigated, and led to better and more stable development set performance, with a Top-S EER of 0.86%.

2.1.4. Cosine Similarity and Logistic Regression

The system presented in this section is a pipeline consisting of a cosine kernel projection step followed by logistic regression. The projection step consists of computing the cosine similarity between the length normalized blacklisted speakers ivectors and the whole training speakers ivectors. In the case of this challenge setup, this results in 3,631 features per training ivector. These features are used to train a logistic regression on the binary task. At test time, the cosine similarity between the test ivector and all the enrolled blacklisted ivectors is computed then fed to the logistic regression to get the posterior probability of belonging to the blacklisted speakers set. This system achieves a Top-S EER of 1.12% on DEV.

2.1.5. System Fusion

We investigate two methods for score fusion. The first one is based on traditional binary *Logistic Regression*, while the second is based on *One-Class Support Vector Machines* with RBF kernel. The rationale behind the choice of one-class classification is that the score distribution of the negative samples may vary, because of the change in background speakers. In contrast to *Logistic Regression*, One-class SVM uses exclusively the positive scores (blacklist scores) for training. The logistic regression fusion achieves a Top-S EER as low as 0.79% on the DEV, while one-class SVM achieves an EER of 1.12%.

2.2. Closed-set Speaker Detection

2.2.1. PLDA-based system

This is the same system described in 2.1.1, with the only exception in the score normalization. Instead of doing normalization using the background training in the cohort, the cohort is constructed using the blacklist speakers. The speaker labels are given by argmax of the scores. This system achieves a Top-1 EER of 6.41%.

2.2.2. NN-based system

This system consists of two shallow neural networks. The first model is used to learn the speaker space. It is trained using the full training data (both background and blacklist speakers). It has 2 hidden layers, with 1024 and 4096 units, respectively. After training, this model is use to extract 4096-D embeddings from the last hidden layer. The second model focuses on classifying between blacklist speakers. Thus, it's trained using only the blacklist training data. The input of this model are the 4096-D embeddings. This system has only one hidden layer with 2048 units. To be able to use it in the fusion, we use the input of the softmax activation as scores. This system achieves a Top-1 EER of 7.20%.

2.2.3. System fusion

The score fusion is trained using logistic regression on the DEV scores. Instead of using all the mismatch cases in the negative data, we focus only on the maximum scores from each of the system. The fusion achieves a Top-1 EER as little as 5.72%.

3. SUBMISSIONS

While we had very optimistic results on DEV with some of the systems, we decided to have a safe choice in our primary submission. The submission is as follows:

• Primary:

- Blacklist: PLDA-based system

- Identification: Fusion of PLDA and NN

• Contrastive 1:

 Blacklist: One Class SVM fusion of the four subsystem

- Identification: Fusion of PLDA and NN

• Contrastive 2:

- Blacklist: Wide-Deep DNN with L1 loss

- Identification: Fusion of PLDA and NN

The results on DEV of the different submitted systems are summarized in Table 1.

| Submission | Top-S | Top-1 |
|---------------|-------|-------|
| Primary | 1.32% | 5.76% |
| Contrastive 1 | 1.12% | 5.72% |
| Contrastive 2 | 0.86% | 5.96% |

Table 1. Results on DEV of the submitted systems

4. 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] Elliot Singer and Douglas A. Reynolds, "Analysis of multitarget detection for speaker and language recognition," in *Odyssey*. 2004, pp. 301–308, ISCA.
- [3] Manuel Günther, Steve Cruz, Ethan M Rudd, and Terrance E Boult, "Toward open-set face recognition," in Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. IEEE, 2017.
- [4] Heng-Tze Cheng, Levent Koc, Jeremiah Harmsen, Tal Shaked, Tushar Chandra, Hrishi Aradhye, Glen Anderson, Greg Corrado, Wei Chai, Mustafa Ispir, Rohan Anil, Zakaria Haque, Lichan Hong, Vihan Jain, Xiaobing Liu, and Hemal Shah, "Wide & deep learning for recommender systems," in *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*. 2016, DLRS 2016, pp. 7–10, ACM.
- [5] Tianqi Chen, Emily B. Fox, and Carlos Guestrin, "Stochastic gradient hamiltonian monte carlo," in *Proceedings of the 31st International Conference on International Conference on Machine Learning Volume 32*. 2014, ICML'14, pp. II–1683–II–1691, JMLR.org.