INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Pindrop Labs' Submission to the First Multi-target Speaker Detection and Identification Challenge

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

Pindrop, Atlanta, GA, USA
ekhoury@pindrop.com

Abstract

This paper summarizes Pindrop Labs' submission to the multi-target speaker detection and identification challenge evaluation (MCE 2018). The MCE challenge is geared towards detecting blacklisted speakers (fraudsters) in the context of call centers. Particularly, it aims to answer the following two questions: Is the speaker of the test utterance on the blacklist? If so, which speaker is it among the blacklisted speakers? While one single system can answer both questions, this work looks at them as two separate tasks: blacklist detection and closed-set identification. The former is addressed using four different systems including probabilistic linear discriminant analysis (PLDA), two deep neural network (DNN) based systems, and a simple system based on cosine similarity and logistic regression. The latter is addressed by combining PLDA and neural network based systems. The proposed system was the best performing system at the challenge on both tasks, reducing the blacklist detection error (Top-S EER) by 31.9% and the identification error (Top-1 EER) by 46.4% over the MCE baseline on the evaluation data.
Index Terms: MCE, blacklist detection, multi-speaker identification, speaker recognition

1. Introduction

The MCE challenge[1] [1] aims to evaluate how well current technologies are at detecting whether or not a speech utterance was spoken by one of a large number of blacklisted speakers. One good characteristic of the MCE dataset is that the audio recordings are collected from real-world customer-agent conversations in the call center. However, because of privacy concerns, the audio recordings were not shared with participants. Instead, only i-vectors [2] extracted from the customer channel are provided. I-vector is one form of low-rank high-level feature representations that compress the speaker identity. While recent literature in speaker recognition has demonstrated the superiority of DNN-based approaches (such as x-vectors [3]) over Gaussian mixture models [4] based approaches (such as i-vectors), DNN representations are out of the scope of this challenge. More details about the MCE dataset are provided in section 2.

Traditional literature on speaker recognition [5, 6, 7] focuses primarily on developing and evaluating systems for speaker verification, where a speaker makes a claim about her/his identity. This is mainly influenced by the NIST Speaker Recognition Evaluation (SRE) series [6, 7] conducted since 1996, that often evaluates the task of *single speaker detection*.

However, the task of multi-speaker detection, also known as open-set speaker identification, has received far less attention. Among the few related work on speaker identification, score normalization stood out as a crucial component to improve the identification accuracy. Authors in [8, 9, 10] use T-Norm (called *multi-target normalization* in [10]) and Z-Norm. [11] and [12] use Top-Norm (called Adaptive T-Norm in [12]). A more recent work in [13] focuses on extracting DNN-based bottleneck features as a replacement to Mel frequency cepstrum coefficients. In all the aforementioned studies, GMMs were used to compute the likelihood between speaker models and test utterances.

It is worth noting that standard techniques used for identification and verification may not work very well for blacklisted speaker detection, because:

1. The variability of the size of the blacklist set has a great impact on the choice of the machine learning techniques. While some techniques work well with few tens of speakers, they risk to fail with hundreds or thousands of speakers, not only because of the drop in accuracy, but also because of the increase in computational costs. MCE constrains this problem by fixing the number of blacklisted speakers to 3,631.

2. Additionally, the imbalance between blacklisted and non-blacklisted sets is an important factor that is often overlooked. MCE constrains this problem by providing blacklist and non-blacklist sets of comparable sizes. However, in practice, the fraud rate is usually very low (e.g. 1 fraud in 1000 calls).

In this work, we look at the task of open-set speaker identification as two disjoint problems: the first is detecting whether or not the speech utterance is spoken by any blacklisted speaker, and the second is identifying the correct blacklisted speaker. This is motivated by our belief that they do not share the same optimal solution. For instance, we believe that some fraud calls share similar acoustic and phonetic characteristics, and thus can be used to improve the blacklist detection, while they could be harmful for speaker identification.

For both problems, we investigate the use of traditional backends applied to i-vectors including Length-Normalization [14], Linear Discriminant Analysis (LDA) and PLDA [15]. We also study the use of DNN-based backends. Similar to existing work, we do score normalization. However, we use adaptive symmetric normalization (AS-Norm) [16] that was shown to be very effective with i-vectors. We also investigate different score-level fusion strategies to take advantage of the complementarity between systems.

The remainder of the paper is structured as follows: Section 2 describes the MCE dataset, baseline system and evaluation metrics. Section 3 presents our proposed system to the challenge. Section 4 details the submitted systems and their results. Section 5 concludes the paper.

[1] http://www.mce2018.org

http://dx.doi.org/10.21437/Interspeech.2019-3179

2. MCE Challenge

2.1. Dataset

The partition of the MCE dataset is detailed in Table 1. The training data (TRAIN) consists of 41,845 utterances: 10,893 for blacklist speakers and 30,952 for genuine (background) speakers. It worth noting that the number of utterances per speaker is the same across speakers and equal to three. This balance in terms of number of known fraud calls per fraudster is typically not the case in real-world applications. The development data (DEV) consists of 8,631 utterances: 3,631 for blacklist speakers and 5,000 for genuine speakers. It is interesting to note that there is no overlap between the genuine speakers of both TRAIN and DEV sets. The evaluation data (EVAL) consists of 16,017 utterances: 3,631 for blacklist speakers and 12,386 for genuine speakers. The partial details on EVAL were made available after the challenge.

The provided i-vectors are 600-dim extracted using a standard recipe from Kaldi[2].

2.2. Baseline System

The baseline system provided as part of the challenge is based on [10]. This work proposes a system of stacked single-speaker detectors $D_i, \ldots, D_S$, where S is the number of blacklisted speakers. Each detector D_i is providing a similarity score l_i of how likely the test utterance was spoken by the target blacklisted speaker T_i. A first decision is made by comparing the maximum score to a threshold. The MCE organizers call it the "Top-S" decision. If it is higher than the threshold, the test utterance is considered to be spoken by a blacklist speaker. In case the former condition is met, a second decision is made by identifying the speaker that has the highest score. The MCE organizers call it the "Top-1" decision. The similarity score is computed using cosine measure followed by multi-target score normalization (M-Norm). More detail about M-Norm can be found in [1].

2.3. Evaluation Metrics

The evaluation metrics of MCE are inspired by [10], and they assess the Top-S and Top-1 decisions. The distinction between both decisions is very important, because in some applications (e.g. financial institutions), the user cares more to detect a fraud event (i.e. Top-S is correct), than to detect the identity of the fraudster (i.e. Top-1 is correct).

Similar to the traditional evaluation of speaker verification systems, the official metrics used in this challenge are EERs. The Top-S EER is straight-forward and is computed using only the false acceptance (i.e. a genuine call being mis-classified as fraud) and false rejection (i.e. a fraud call being mis-classified as genuine). However, the Top-1 EER uses an additional confusion error, that accounts for the cases where an actual caller is correctly detected as a blacklist speaker, but the speaker is not correctly identified.

3. Proposed System

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]https://github.com/kaldi-asr/kaldi/tree/master/egs/sre10/v1

3.1. Blacklist Detection

3.1.1. 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.

3.1.2. 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.

3.1.3. Wide-Deep DNN

Two systems are presented making use of a "wide-deep" architecture [17]. The first was trained using hamiltonian monte carlo SGD methods described in [18]. 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%.

3.1.4. 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.

Table 1: *Partitionning of the MCE dataset.*

Set	Subset	Nb. of Speakers	Nb. of Utterances/Speaker	Total Nb. of Utterances
TRAIN	Blacklist	3,631	3	10,893
	Genuine	5,000	≥ 4	30,952
DEV	Blacklist	3,631	1	3,631
	Genuine	5,000	1	5,000
EVAL	Blacklist	3,631	1	3,631
	Genuine	N/A	N/A	12,386

3.1.5. 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%.

3.2. Closed-set Speaker Detection

3.2.1. PLDA-based system

This is the same system described in 3.1.1, 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%.

3.2.2. 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%.

3.2.3. 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%.

4. Experimental Results

4.1. Submitted Systems

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

4.2. Results

The results on DEV of the different submitted systems are summarized in Table 2.

Table 2: *Results on DEV of the submitted systems.*

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

We observed that for the blacklist speaker detection, the contrastive 2 system performed the best on the development set with a top-S EER of 0.86%. For the blacklisted speaker identification problem, the Contrastive 1 system performed the best on the development set with a top-1 EER of 5.72% Among the three systems submitted, the primary system did not achieve the best results on the development set, but the results were in the same range as the contrastive systems. With the limited amount of training data, over-fitting to the training data was a serious threat with the neural network systems. Since the i-vectors are already a speaker embedding which have been trained to discriminate speakers, further non-linear transformations on the i-vectors using neural networks could possibly lead to overfitting because of the limited available training data. Further, the i-vectors provided in the training and development set might probably be only a fraction of the number of speakers that the i-vector system was trained on. The provided i-vectors probably cover only as sparse subspace of the i-vector space. Thus the NN and the wide-deep DNN systems learn a discriminative nonlinear transformation of only a subspace of the i-vector space which works really well with the training and development data but might not work well for unseen i-vectors. Based on this rationale we chose the PLDA based system for blacklisting and the fusion of PLDA and NN for blacklisted speaker identification as the primary system submissions.

The official results on EVAL of the different submitted systems[3] are summarized in Table 3. Interestingly, our primary

[3]Our experimental systems do not reflect the performance of our products.

system was the best performing system in the challenge, with lowest Top-S and Top-1 EERs.

Table 3: *Official results on EVAL of the submitted systems.*

Submission	Top-S	Top-1
Primary	4.25%	6.05%
Contrastive 1	10.27%	12.23%
Contrastive 2	14.69%	15.12%

5. Conclusions

In this paper we have addressed a challenging problem of speaker blacklisting and blacklisted speaker identification. We were provided only the i-vectors of the speakers and their labels with no knowledge about or access to the i-vector speaker embedding system. We addressed the challenge by training separate systems for speaker blacklisting and speaker identification. We trained a PLDA based system, a shallow NN system, and a wide-deep DNN system. We applied length normalization, and adaptive symmetric normalization for the PLDA based system. We also performed data augmentation by applying at random a weighted sum between background ivectors and blacklist i-vectors to augment the training data for the PLDA system. The final submission consisted of a fusion of the PLDA, NN and the wide-deep DNN systems. The submitted primary system was the best performing system of the challenge with a top-S EER of 4.25% and a top-1 EER of 6.05%.

6. References

[1] S. Shon, N. Dehak, D. Reynolds, and J. Glass, "Mce 2018: The 1st multi-target speaker detection and identification challenge evaluation," in *to appear in Interspeech*, 2019.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 19, pp. 788–798, 2011.

[3] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," *ICASSP*, 2018.

[4] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, pp. 19–41, 2000.

[5] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Communication*, vol. 52, no. 1, pp. 12–40, 2010.

[6] S. O. Sadjadi, T. Kheyrkhah, A. Tong, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, "The 2016 nist speaker recognition evaluation," in *INTERSPEECH*, 2017.

[7] S. O. Sadjadi, C. S. Greenberg, D. A. Reynolds, E. Singer, L. P. Mason, and J. Hernandez-Cordero, "The 2018 nist speaker recognition evaluation," in *submitted to INTERSPEECH*, 2019.

[8] J. Fortuna, P. Sivakumaran, A. M. Ariyaeeinia, and A. Malegaonkar, "Relative effectiveness of score normalisation methods in open-set speaker identification," in *ODYSSEY04-The Speaker and Language Recognition Workshop*, 2004.

[9] J. Fortuna, P. Sivakumaran, A. Ariyaeeinia, and A. Malegaonkar, "Open-set speaker identification using adapted gaussian mixture models," in *Ninth European Conference on Speech Communication and Technology*, 2005.

[10] E. Singer and D. A. Reynolds, "Analysis of multitarget detection for speaker and language recognition," in *Odyssey*. ISCA, 2004, pp. 301–308.

[11] Y. Zigel and M. Wasserblat, "How to deal with multiple-targets in speaker identification systems?" in *2006 IEEE Odyssey-The Speaker and Language Recognition Workshop*. IEEE, 2006, pp. 1–7.

[12] H. Do, I. Tashev, and A. Acero, "A new speaker identification algorithm for gaming scenarios," in *2011 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2011, pp. 5436–5439.

[13] T. Yamada, L. Wang, and A. Kai, "Improvement of distant-talking speaker identification using bottleneck features of dnn." 2013.

[14] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems," in *INTERSPEECH*, 2011, pp. 249–252.

[15] S. Prince and J. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *IEEE ICCV*, vol. 0, 2007, pp. 1–8.

[16] Z. Karam, W. Campbell, and N. Dehak, "Towards reduced false-alarms using cohorts," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 4512–4515.

[17] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, R. Anil, Z. Haque, L. Hong, V. Jain, X. Liu, and H. Shah, "Wide & deep learning for recommender systems," in *Proceedings of the 1st Workshop on Deep Learning for Recommender Systems*, ser. DLRS 2016. ACM, 2016, pp. 7–10.

[18] T. Chen, E. B. Fox, and C. Guestrin, "Stochastic gradient hamiltonian monte carlo," in *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ser. ICML'14. JMLR.org, 2014, pp. II–1683–II–1691. [Online]. Available: http://dl.acm.org/citation.cfm?id=3044805.3045080