On Open-Set Speaker Identification with I-Vectors

Kevin Wilkinghoff

Fraunhofer Institute for Communication, Information Processing and Ergonomics FKIE
Fraunhoferstraße 20, 53343 Wachtberg, Germany
kevin.wilkinghoff@fkie.fraunhofer.de

Abstract

Open-set speaker identification systems first need to decide if an utterance belongs to one of the known so called blacklist speakers and second identify the exact blacklist speaker. In this paper, an open-set speaker identification system based on i-vectors is presented. The system consists of an outlier detector in combination with a classical closed-set speaker identification chain and utilizes an effective preprocessing technique for i-vectors, called linear alignment. Its overall structure is justified both theoretically and experimentally by comparing multiple outlier detectors. In experimental evaluations, our proposed system reaches an improvement of 37.5% for the top-S Equal Error Rate (EER) and a 50% lower top-1 EER over the baseline system of the 1st Multi-target speaker detection and identification Challenge Evaluation and improves upon all other published results obtained on this dataset.

1. Introduction

An open-set speaker identification system first needs to decide whether an utterance belongs to one of multiple target speakers called "blacklist speakers". In a second step, called top-1 detection, the system determines to whom of the blacklist speakers the utterance belongs to. In analogy to that, the first step is referred to as top-S detection where S denotes the total number of blacklist speakers. The difficulty of this task is that there are not only known blacklist speakers and known non-blacklist speakers, whose data is available when training the system, but also unknown non-blacklist speakers [1, 2]. Because of that, the speaker identification system also needs to discriminate the known blacklist speakers against all possible unknown non-blacklist speakers, whose data is not available before testing. In real world applications such as telephone banking or call center conversations, there are almost no restrictions on who is being recorded because many people, known customers and strangers, can call. Hence, it is impossible to gather data of every possible non-blacklist speaker and thus there will always be speakers unknown to the identification system. In conclusion, open-set speaker identification is more challenging but also more realistic than closed-set speaker identification. For an overview of the speaker sets as well as their relations, the reader is referred to Fig. 1.

Much research has been conducted on closed-set speaker identification and speaker verification, which can also be seen as single-target speaker recognition. However, open-set speaker identification (multi-target speaker recognition) has received noticeably less attention although open-set related work does exist [3, 4, 5, 6]. The fact that closed-set classification is inherently easier to solve than open-set classification and that speaker verification is strongly promoted through the NIST Speaker Recognition Evaluation Challenge [7] are probably among the reasons of why this is the case. Previous work on open-set speaker identification often focuses on score normalization techniques [8, 9, 10], which is one possible way to calibrate scores. Score calibration [11, 12] means to map scores to log-likelihood ratios such that a single value can be compared to a fixed decision threshold and is particularly important in speaker verification.

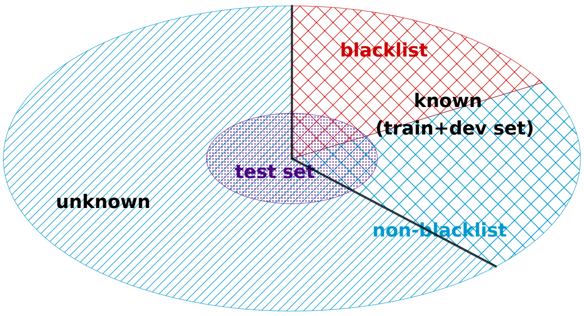

Figure 1: Overview of the different speaker sets.

A challenge specifically targeted at open-set speaker identification is the "1st Multi-target speaker detection and identification Challenge Evaluation" (MCE 2018) [13]. Its dataset consists of customer-agent call-center conversations in the form of i-vectors [14] rather than audio files because of privacy concerns, although this prevents the usage of deep speaker embeddings as x-vectors [15]. The baseline system of the MCE challenge is described in [16]. It uses a nearest neighbor approach via cosine similarity in combination with Multi-Target Score Normalization (M-Norm). Hence, none of the non-blacklist speakers is used for training the system. In fact, only the M-Norm parameters, which are derived from the blacklist speakers exclusively, need to be "trained". The two top-performing systems submitted to the challenge both consist of Probablistic Linear Discriminant Analysis (PLDA) [17] in combination with score normalization and different neural networks. In [18] a two-stage neural network consisting of a speaker embedding and a discriminative block is fused with a PLDA model and background i-vectors are randomly augmented with blacklist i-vectors via a weighted sum. Font [19] applies a denoising autoencoder for preprocessing the i-vectors before inserting them into a PLDA-based speaker identification system.

The main contributions of this paper are the following: First and foremost, we present an open-set speaker identification system[1] for i-vectors, which performs significantly better than the baseline system of the MCE challenge and all other published

[1] An open source Python implementation of our proposed system is available here: https://github.com/wilkinghoff/mce2018

systems evaluated on the same dataset. Second, we theoretically motivate the system's overall structure and justify its choice by presenting and evaluating multiple outlier detection models. As a third contribution, a simple preprocessing technique for i-vectors, called linear alignment, is presented whose effectiveness is shown in additional experiments.

2. The speaker identification system

2.1. Motivating the system's structure

Let $x \in \mathbb{R}^D$ be an i-vector and y_i be the label of blacklist speaker $i \in \{1, ..., M\}$. Then, predicting the most likely blacklist speaker can be formalized as

$$\underset{i \in \{1,...,M\}}{\operatorname{argmax}} \; P(Y = y_i, B = \text{true} | X = x) \tag{1}$$

where Y, B, X denote random variables. Here, Y corresponds to blacklist speaker labels, X to the observed data and B is a binary random variable indicating whether the data belongs to one of the blacklist speakers or not. We will now rewrite this equation to gain more insights on how to tackle the problem.

The chain rule for probabilities implies that

$$\begin{aligned} &P(Y = y_i, B = \text{true}, X = x) \\ =&P(Y = y_i | B = \text{true}, X = x) P(B = \text{true} | X = x) P(X = x). \end{aligned} \tag{2}$$

Hence, we obtain

$$\begin{aligned} &P(Y = y_i, B = \text{true} | X = x) \\ =&P(Y = y_i | B = \text{true}, X = x) P(B = \text{true} | X = x). \end{aligned} \tag{3}$$

The first term of the right hand side is the posterior probability that needs to be estimated in closed-set classification problems. Note that the silent assumption that the data belongs to a known class is explicitly stated in contrast to the usual notation. Determining the second term of the right hand side is called outlier detection [20]. In conclusion, open-set classification can be decomposed into two subtasks: 1) closed-set classification and 2) outlier detection. Intuitively, this also makes sense since one needs to decide whether a sample belongs to any known class (outlier detection) and, if this is the case, output the correct class (closed-set classification).

2.2. Closed-set classification

For most classification problems, much research has been conducted in the closed-set setting, which led to many well-working techniques. Speaker identification is not an exception. Therefore, setting up a subsystem for closed-set speaker identification is straight forward. Since the speech data comes in the form of i-vectors, the options to recognize speakers with them are fairly limited. Traditionally, one can use cosine similarity for comparing i-vectors as done in the baseline system. Better performing techniques such as Linear Discriminant Analysis (LDA) and Probablistic Linear Discriminant Analysis (PLDA) [17] are the state-of-the-art for closed-set speaker identification with i-vectors. In our system, both techniques are used as follows. First, an LDA model of dimension 600 is trained with Scikit-learn [21] to discriminate among the 3631 blacklist speaker classes. Then, all LDA projected i-vectors are length normalized before applying two-covariance PLDA [22, 23] as implemented in [24]. The PLDA model is trained for 20 iterations. For identification, the blacklist speaker belonging to the maximum score is chosen (maximum likelihood).

Accepting or rejecting utterances based on a fixed decision threshold requires all scores to be calibrated appropriately. One possibility to do this is to normalize the scores. An overview of score normalization techniques can be found in [25]. In experiments conducted in this paper, Adaptive Symmetric Normalization (AS-Norm) [26, 27] is found to be the best performing one. Moreover, it has been successfully applied in an open-set speaker identification setting [18, 19]. AS-Norm makes use of a set of utterances, called cohort, to normalize the scores. More concretely, AS-Norm is defined as

$$\begin{aligned} s(e,t)_{\text{as-norm}} := \frac{1}{2}\Bigg(&\frac{s(e,t) - \mu(s(e, \mathcal{C}_{\text{top}(e,n_1)}))}{\sigma(s(e, \mathcal{C}_{\text{top}(e,n_1)}))} \\ &+ \frac{s(e,t) - \mu(s(t, \mathcal{C}_{\text{top}(t,n_2)}))}{\sigma(s(t, \mathcal{C}_{\text{top}(t,n_2)}))}\Bigg) \end{aligned} \tag{4}$$

where $s(e,t)$ denotes the score between enrolment utterance e and test utterance t, and μ and σ denote mean and standard deviation of the scores, respectively. Furthermore, $\mathcal{C}_{\text{top}(e,n_1)}$ denotes the $n_1 \in \mathbb{N}$ utterances $\{c_k \in \mathcal{C} : k = 1, ..., n_1\}$ of the cohort $\mathcal{C}$ with highest scores $s(e, c_k)$ ($\mathcal{C}_{\text{top}(t,n_2)}$ is defined analogously). For closed-set speaker identification, it is sufficient to learn to discriminate among the known blacklist speakers because it is a priori known that one of them must be present. Since data from non-blacklist speakers does not contain information about the blacklist speakers, this data does not contain any helpful information when training a discriminative model to decide to whom of the blacklist speakers a given utterance corresponds to. Thus, the whole closed-set classification chain is trained without the usage of any non-blacklist speaker and consequently the cohort consists of all files belonging to blacklist speakers only, as also proposed in [18]. In our experiments, $n_1 = 700$ and $n_2 = 9000$ have been used as cohort sizes. These values have been determined by maximizing the performance on the development set.

2.3. Outlier detection

In the following, we will present a few possible models for detecting outliers whose performance will be evaluated later. We will see that some models (obviously) fail to improve upon the baseline system. Nevertheless, we included them here because negative results are still valuable results.

2.3.1. Cosine-similarity based model

A good starting point for outlier detection is the baseline model [16]. It measures the cosine similarity of a test i-vector to each speaker's mean i-vector. In addition to that, M-Norm is applied to the scores. This means that the cosine similarities between the speakers' mean i-vectors and all training i-vectors that belong to some blacklist speaker are computed first. Then, all scores are normalized by subtracting the mean and dividing by the standard deviation of these blacklist scores. In all of our experiments, we only centered the scores with the mean as this led to slightly better results. Note, that the organizers of the challenge also noticed this behavior on the development set but still used regular M-Norm for the baseline system. After that, the highest score, which corresponds to the most similar i-vector, is taken as a result. Finally, this nearest neighbor score is compared to a threshold to either accept the i-vector as belonging to one of the blacklist speakers or not.

2.3.2. *PLDA-based model*

PLDA is the state-of-the-art back-end technique in speaker verification and thus also suitable for detecting outliers. As done for closed-set classification, we trained a two-covariance PLDA model for 20 iterations. Again, AS-Norm (see Eq. 4) with $n_1 = 2800$ and $n_2 = 600$, which have been determined by minimizing the top-S Equal Error Rate (EER) on the development set, has been applied. But in contrast to the closed-set PLDA model, we used all available files (both blacklist and background) for training the model. We will show later that this is indeed beneficial. Furthermore, the cohort consists of all background files instead of the files belonging to blacklist speakers.

2.3.3. *Autoencoder*

Another way to detect outliers is to utilize an autoencoder [28]. This autoencoder is trained in an unsupervised manner to encode the i-vectors belonging to the blacklist speakers to a smaller dimension and reconstruct them again. When actually trying to detect outliers, the assumption is that i-vectors belonging to known blacklist speakers have a relatively low reconstruction error whereas non-blacklist speakers cannot be reconstructed sufficiently well. Hence, the reconstruction error serves as a score for outlier detection.

Another possibility to train the autoencoder is to decode to the corresponding blacklist speakers' mean i-vectors instead of reconstructing the i-vectors themselves. This can also be viewed as a denoising autoencoder where means of i-vectors serve as estimates of the true i-vectors representing the blacklist speakers. Note that transforming i-vectors with a denoising autoencoder has been shown to be an effective preprocessing technique for speaker identification [19, 29, 30].

Table 1: Architecture of the autoencoder.

Layer	Output Shape	#Parameters
Input	600	0
Dropout (0.5)	600	0
Length normalization	600	0
Dense (Leaky ReLU: 0.01)	500	300,500
Dense (Leaky ReLU: 0.01)	400	200,400
Dense (Leaky ReLU: 0.01)	300	120,300
Dense (Leaky ReLU: 0.01)	200	60,200
Dense (Leaky ReLU: 0.01)	300	60,300
Dense (Leaky ReLU: 0.01)	400	120,400
Dense (Leaky ReLU: 0.01)	500	200,500
Dense (Leaky ReLU: 0.01)	600	300,600
		∑ 1,363,200

The structure of the autoencoder we used is shown in Tab. 1 and has been designed by optimizing the performance on the development set. Since there are not many samples of i-vectors per blacklist speaker, training a separate autoencoder for each blacklist speaker leads to serious overfitting and results in a poor performance. Thus, all blacklist speakers were treated as belonging to one "blacklist" class. To reduce the overfitting effects even more, we used dropout [31] and a length normalization layer afterwards. The autoencoder is trained by minimizing the cosine similarity between the input i-vectors and their reconstructions for 1000 epochs. It is implemented with Tensorflow [32] and Keras [33]. We also experimented with a Variational Autoencoder [34] but this did not result in a better performance than a regular autoencoder.

2.3.4. *A naive discriminative model*

In contrast to the previously described models, the following model is trained to directly discriminate between the two speaker classes "blacklist" and "non-blacklist" by using examples of non-blacklist speakers. Here, the underlying assumption is that the variability of all non-blacklist speakers is sufficiently captured by the samples of non-blacklist speakers provided for training. Thus, the model learns to discriminate between blacklist and non-blacklist speakers. Note that it is well known that this is not a valid assumption because there is nothing inherent to the blacklist speakers that separates them from the non-blacklist speakers.

For detecting the blacklist speakers, first an LDA model is trained on the blacklist speaker classes plus one additional class corresponding to the non-blacklist speakers. Using this LDA model, the dimension of the i-vectors is reduced to a relatively small value of 40. This particular dimension has been chosen because it yielded the lowest top-S EER when evaluating on the development set. Then, a two-class SVM with RBF-kernels as implemented in Scikit-learn [21] is trained using the parameter settings $C = 1$ and $\gamma = 0.01$. This SVM also outputs probabilities for each i-vector stating how likely it is that this i-vector belongs to any of the blacklist speakers. As it is done with the baseline model, these probabilities are then utilized as scores to either accept or reject an i-vector by using a threshold.

2.4. Putting it together

The structure of our proposed speaker identification system can be found in Fig. 2. Its fundamental idea is to use an effective preprocessing along with an ensemble of scores derived from the PLDA-based outlier detection model and the closed-set speaker classification chain has been described in Sec. 2.2. The preprocessing consist of linear alignment, which will be described in the following paragraph, and length normalization. After that, our ensembling strategy will be presented.

2.4.1. *Linear alignment*

The goal of linear alignment is to improve the outlier detection capabilities by reducing intra-class variability of the blacklist speakers. Inter-class variability does not need to be reduced because it is simply not important when detecting outliers. Thus, linear alignment is less restricted when reducing intra-class variability than LDA where inter-class variability is also minimized. To reduce intra-class variability, an affine transformation is estimated that minimizes the Euclidean distance between all blacklist i-vectors and their corresponding speaker's mean i-vector. In mathematical terms, this corresponds to solving

$$\underset{A\in\mathbb{R}^{D\times D},b\in\mathbb{R}^{D}}{\operatorname{argmin}} \sum_{i=1}^{M}\sum_{j=1}^{N_i}\left\| Ax_{ij} + b - \frac{1}{N_i}\sum_{k=1}^{N_i} x_{ik}\right\|_2 \qquad (5)$$

where $x_{ij} \in \mathbb{R}^D$ denotes the jth of the N_i i-vectors belonging to blacklist speaker $i \in \{1, ..., M\}$. For this purpose, a single-layered neural network with no nonlinearity has been trained by minimizing the mean squared error for 400 epochs to estimate the linear transformation. Although one can also use a deeper neural network structure with nonlinearities and multiple hidden layers, this did not improve but degrade the performance. A probable reason for this behavior is overfitting caused by the

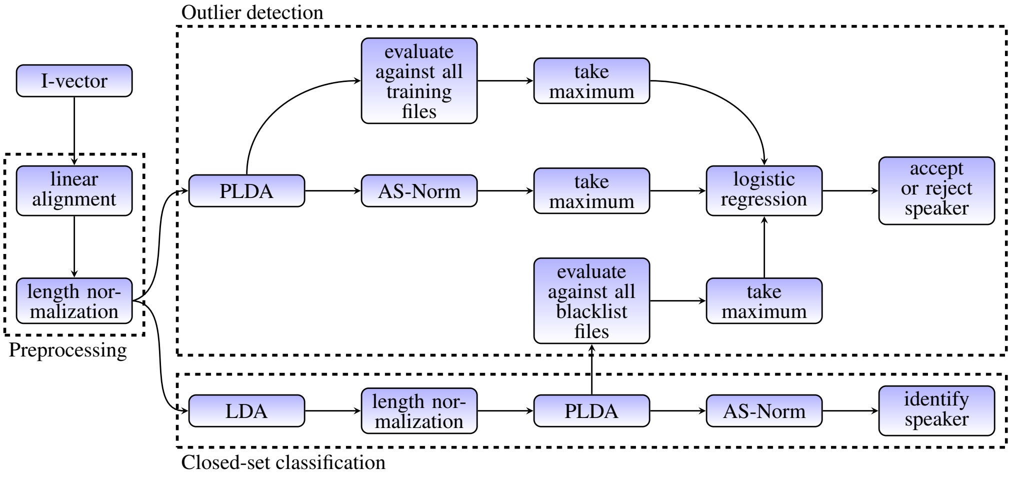

Figure 2: Structure of our proposed speaker identification system.

small number of samples per blacklist speaker. A similar behavior has been reported in [19] where a denoising autoencoder with a single hidden layer instead of multiple ones is used for preprocessing the i-vectors. We also experimented with adding this autoencoder to our system, but, surprisingly, this degraded the performance instead of improving it.

2.4.2. Ensembling strategy

Although the closed-set classification chain and PLDA-based outlier detection model are sufficient for open-set classification, a bit of performance can be gained by ensembling additional scores. The general idea is that utterances, which are very different from anything a model has been trained with, are more likely to be an outlier (i.e. belong to a background speaker) than being uttered by a blacklist speaker. Therefore, any given utterance has been evaluated against all available training files (both, background and blacklist) using the PLDA-based outlier detection model and against all blacklist files using the closed-set PLDA model. After that, the maximum of each of these scores and the regular outlier detection scores where taken and the three maxima were combined via logistic regression whose objective it is to determine whether an utterance is an outlier or not. Note, that using such a logistic regression model is a way of calibrating the scores such that they are suited for being compared to a threshold.

Training a logistic regression model for score-based fusion requires realistic scores similar to those obtained with the final test data [35]. To this end, we trained all models except the logistic regression model using the training data only and evaluated them with the blacklist and background files of the development dataset to obtain meaningful scores. These scores are then used to train the logistic regression model with balanced class weights and a SAGA solver with L2-regularization as implemented in Scikit-learn [21]. After that, all models but the logistic regression model are retrained using both the training and development datasets for final evaluations.

3. Experimental evaluations

3.1. MCE dataset

All experimental evaluations have been conducted on the dataset of the MCE challenge [13]. This has the following advantages: First, the dataset is specifically designed for open-set speaker identification and freely available. Thus, all results can easily be reproduced. Second, we can compare our results to those obtained by other systems, namely the baseline system [16] and the two top-performing systems [19, 18] of the challenge.

The MCE dataset [13] consists of $D = 600$ dimensional i-vectors belonging to $M = 3631$ blacklist speakers and an unknown number of non-blacklist speakers, which are extracted from customer-agent call-center conversations. To train the i-vector extractor, 13000 hours of unlabeled speech have been used. Each of the blacklist speakers has exactly 3 occurrences within the training dataset, which contains 41845 i-vectors in total. The development set consists of 8631 i-vectors and the test set consists of 16017. In both sets, each blacklist speaker is present exactly once. But note that for the test set this is not a priori knowledge and thus should not be used when developing or training the system.

3.2. A comparison of the individual outlier detectors

We will now evaluate and compare all individual models. This has the following two purposes: First, we can justify our choice of models and determine which single model is most successful in detecting outliers and second, the results can be compared later to those obtained with the whole ensemble. For all evaluations, we utilized the development set as additional training data when training the system for evaluations with the test set.

The top-S EERs obtained with the individual models are depicted in Fig. 3. First, one can observe that the top-S EERs on the development set are all relatively small whereas on the test set, all EERs and the difference between them are generally higher. This seems only natural, since all models have been designed and fine-tuned in order to perform well on the development set. But interestingly, the cosine similarity based model requires no training and also performs much worse on the test set, which consists of 16017 files. Hence, it seems that the results obtained on the smaller development set (3631 + 5000 files) are less reliable and probably a bit too optimistic. As the cosine distance can be understood as a measure of similarity, this also indicates that the non-blacklist samples of the development set are “more different” from the blacklist speakers than the ones of the test set.

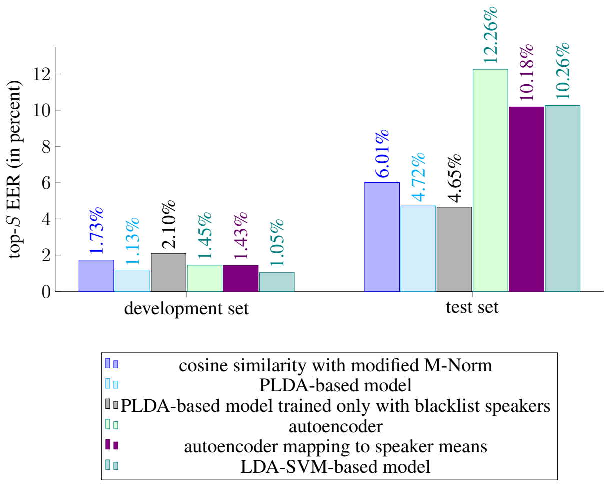

Figure 3: Comparison of top-S Equal Error Rates obtained with individual outlier detectors without using linear alignment.

Another observation to be made is that the LDA-SVM-based model and the autoencoders have much higher top-S EERs than the simple cosine similarity-based model. As stated before, it is well-known that naively training a discriminative classifier with samples of non-blacklist speakers does not work well. Thus, the only surprise is that we were able to achieve a very low top-S EER on the development set when using the SVM for that exact purpose. Furthermore, both autoencoders do not have enough samples per blacklist speaker in order to perform well on both datasets.

In contrast to the previously discussed models, the PLDA-based outlier detector yields significant performance gains over using the cosine similarity. Since PLDA is the state-of-the-art backend in speaker identification with i-vectors, this is not too surprising. What can be found interesting is that training with samples of non-blacklist speakers actually improves performance on the development set although these samples do not sufficiently cover the whole non-blacklist speaker space. However, having more training data at hand results in more accurate estimates of the PLDA parameters, and thus, leads to better results. Interestingly, the top-S EERs obtained with the test set are about the same for both PLDA models. Still, when considering the performances on both sets it seems to be beneficial to train with samples of non-blacklist speakers, which are also used for AS-Norm and thus needed anyway.

3.3. The effects of linear alignment

As shown in Fig. 4, almost all top-S and top-1 EERs are decreasing on both the development and the test set when applying linear alignment. Especially when using cosine similarity the benefits are immense. This makes linear alignment a highly beneficial preprocessing technique. The only exception is the top-S EER obtained on the development set with the PLDA-based model, which is about the same with and without applying linear alignment.

Linear alignment strongly resembles the effect of LDA or within-class covariance normalization (WCCN) [36, 37] by minimizing intra-class variability. Because of this, LDA and WCCN have also been evaluated as preprocessing techniques to be able to compare all approaches in terms of performance. There are two observations to be made: First of all, applying LDA massively degrades the performance. The reason is that not only intra-class variability is minimized but also the inter-class variability leading to unwanted effects when trying to detect outliers. Second, applying WCCN to preprocess the i-vectors also improves the performance but not as much as linear alignment.

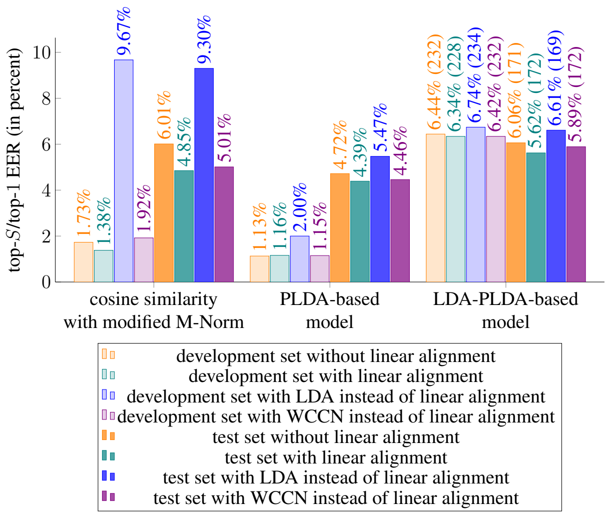

Figure 4: Comparison of top-S Equal Error Rates obtained without linear alignment, with LDA instead of linear alignment and with linear alignment. For the LDA-PLDA-based model, top-1 Equal Error Rates obtained with the proposed ensemble are shown instead.

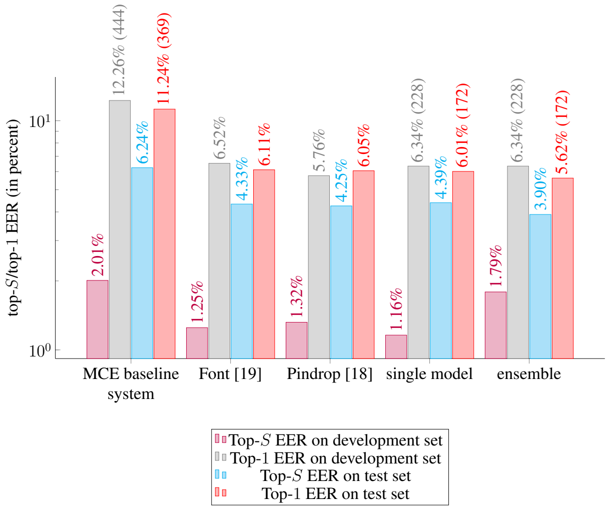

Figure 5: Comparison of Equal Error Rates obtained with different speaker identification systems. Numbers in brackets denote absolute number of confusion errors.

3.4. Evaluating the performance of our proposed system

In Fig. 5, we compared the EERs obtained with our system to the ones of the baseline system and the two top-performing systems of the MCE challenge [18, 19]. More precisely, our system is evaluated in two configurations: using the PLDA-based outlier detector only (right of center) and using the ensemble of outlier detectors (far right). First of all, both versions of our system as well as Font [19] and Pindrop [18], perform much better than the baseline system. In fact, the top-S EER is 37.5% lower and the top-1 EER is 50% lower when comparing the baseline system to our ensembled version. In addition, our single model-based system yields about the same performance as Font [19] and Pindrop [18]. The only exception, is the development set top-1 EER obtained by Pindrop, which is significantly lower.

In contrast to that, our ensembled system has a much lower top-S and top-1 EER on the test set than both our single model-based system, Font [19] and Pindrop [18]. However, in this case ensembling multiple scores also massively degrades the top-S EER on the development set when comparing it to the single model. A possible explanation is that the very low top-S EER of 1.16% obtained with the PLDA-based outlier detector alone is hard to preserve when fusing with scores, which, by themselves, lead to much worse results. But since the corresponding top-1 EER stays the same and both EERs drastically improve on the test dataset, this at least indicates that the scores contain complementary information. Thus, ensembling the scores the way we proposed still seems to be highly beneficial.

4. Conclusions and future work

In this paper, we presented a freely available open-set speaker identification system for i-vectors and theoretically motivated its structure. It consists of an effective preprocessing technique, called linear alignment, in combination with an ensemble of outlier detectors and a fairly standard closed-set speaker identification chain. In experimental evaluations, it was shown that linear alignment is highly beneficial. Furthermore, multiple models for outlier detection have been presented and evaluated, which justifies the choice of models used in our system. In conclusion, our open-set speaker identification system greatly outperforms the baseline system of the MCE 2018 challenge. More concretely, its top-S EER is 37.5% and the top-1 EER is 50% lower than the corresponding ones obtained with the baseline system. Moreover, the obtained performance is even better than all previously published results on the same dataset we are aware of.

In the near future, we plan to examine the following augmentations of the presented system, possibly leading to an even better performance: Since LDA plays a fundamental role in closed-set identification, it may be beneficial to replace it with Deep LDA [38], which is reported to have a better performance than LDA. An additional improvement of the closed-set classification chain may be accomplished by using discriminative PLDA with SVMs as presented in [39, 40]. Furthermore, additional experiments carried out on larger datasets with available audio files are necessary to evaluate the presented system with other speaker embeddings as for example x-vectors [15].

5. References

[1] Walter J. Scheirer, Anderson Rocha, Archana Sapkota, and Terrance E. Boult, “Towards open set recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, July 2013.

[2] Walter J. Scheirer, Lalit P. Jain, and Terrance E. Boult, “Probability models for open set recognition,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 11, pp. 2317–2324, 2014.

[3] A. M. Ariyaeeinia, J. Fortuna, P. Sivakumaran, and A. Malegaonkar, “Verification effectiveness in open-set speaker identification,” *IEE Proceedings-Vision, Image and Signal Processing*, vol. 153, no. 5, pp. 618–624, 2006.

[4] Wonkyung Park, Jae C. Oh, Misty K. Blowers, and Matt B. Wolf, “An open-set speaker identification system using genetic learning classifier system,” in *8th annual conference on Genetic and evolutionary computation*. ACM, 2006, pp. 1597–1598.

[5] Amit Malegaonkar and Aladdin Ariyaeeinia, “Performance evaluation in open-set speaker identification,” in *European workshop on biometrics and identity management*. Springer, 2011, pp. 106–112.

[6] Chao Gao, Guruprasad Saikumar, Amit Srivastava, and Premkumar Natarajan, “Open-set speaker identification in broadcast news,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 5280–5283.

[7] Seyed Omid Sadjadi, Timothée Kheyrkhah, Audrey Tong, Craig S Greenberg, Douglas A Reynolds, Elliot Singer, Lisa P Mason, and Jaime Hernandez-Cordero, “The 2016 NIST speaker recognition evaluation.,” in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 1353–1357.

[8] J. Fortuna, Perasiriyan Sivakumaran, Aladdin M. Ariyaeeinia, and Amit Malegaonkar, “Relative effectiveness of score normalisation methods in open-set speaker identification,” in *ODYSSEY - 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 *9th European Conference on Speech Communication and Technology (INTERSPEECH)*, 2005.

[10] Yaniv Zigel and Moshe Wasserblat, “How to deal with multiple-targets in speaker identification systems?,” in *ODYSSEY - The Speaker and Language Recognition Workshop*. IEEE, 2006, pp. 160–166.

[11] David A. van Leeuwen, Niko Brümmer, and Albert Swart, “A comparison of linear and non-linear calibrations for speaker recognition,” in *ODYSSEY - The Speaker and Language Recognition Workshop*, 2014.

[12] Niko Brümmer and Daniel Garcia-Romero, “Generative modelling for unsupervised score calibration,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1680–1684.

[13] 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,” in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 356–360.

[14] Najim Dehak, Patrick J. Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[15] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[16] Elliot Singer and Douglas A. Reynolds, “Analysis of multitarget detection for speaker and language recognition,” in *ODYSSEY - The Speaker and Language Recognition Workshop*, 2004, number 4, pp. 301–308.

[17] Simon J.D. Prince and James H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *11th International Conference on Computer Vision (ICCV)*. IEEE, 2007, pp. 1751–1758.

[18] Elie Khoury, Khaled Lakhdhar, Andrew Vaughan, Ganesh Sivaraman, and Parav Nagarsheth, “Pindrop labs’ submission to the first multi-target speaker detection and identification challenge,” in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 1502–1505.

[19] Roberto Font, “A denoising autoencoder for speaker recognition. results on the MCE 2018 challenge,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6016–6020.

[20] Charu Aggarwal, *Outlier Analysis*, Springer, 2nd edition, 2017.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[22] Niko Brümmer and Edward De Villiers, “The speaker partitioning problem.,” in *ODYSSEY - The Speaker and Language Recognition Workshop*, 2010, pp. 202–209.

[23] Jesús Villalba and Niko Brümmer, “Towards fully Bayesian speaker recognition: Integrating out the between-speaker covariance,” in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011, pp. 505–508.

[24] Aleksandr Sizov, Kong Aik Lee, and Tomi Kinnunen, “Unifying probabilistic linear discriminant analysis variants in biometric authentication,” in *Proc. S+SSPR*. Springer, 2014, pp. 464–475, Software available at https://sites.google.com/site/fastplda/.

[25] Pavel Matejka, Ondrej Novotnỳ, Oldrich Plchot, Lukás Burget, Mireia Díez Sánchez, and Jan Cernockỳ, “Analysis of score normalization in multilingual speaker recognition.,” in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 1567–1571.

[26] Sandro Cumani, Pier Domenico Batzu, Daniele Colibro, Claudio Vair, Pietro Laface, and Vasileios Vasilakakis, “Comparison of speaker recognition approaches for real applications,” in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011, pp. 2365–2368.

[27] Zahi N Karam, William M Campbell, and Najim Dehak, “Towards reduced false-alarms using cohorts,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4512–4515.

[28] Simon Hawkins, Hongxing He, Graham Williams, and Rohan Baxter, “Outlier detection using replicator neural networks,” in *International Conference on Data Warehousing and Knowledge Discovery*. Springer, 2002, pp. 170–180.

[29] Shivangi Mahto, Hitoshi Yamamoto, and Takafumi Koshinaka, “i-vector transformation using a novel discriminative denoising autoencoder for noise-robust speaker recognition.,” in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 3722–3726.

[30] Suwon Shon, Seongkyu Mun, Wooil Kim, and Hanseok Ko, “Autoencoder based domain adaptation for speaker recognition under insufficient channel information,” in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, vol. 2017, pp. 1014–1018.

[31] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[32] Martin Abadi et al., “Tensorflow: A system for large-scale machine learning,” *OSDI*, vol. 16, pp. 265–283, 2016.

[33] François Chollet et al., “Keras,” https://keras.io, 2015.

[34] Diederik P. Kingma and Max Welling, “Auto-encoding variational Bayes,” in *International Conference on Learning Representations (ICLR)*, 2014.

[35] Kevin Wilkinghoff, Paul M. Baggenstoss, Alessia Cornaggia-Urrigshardt, and Frank Kurth, “Robust speaker identification by fusing classification scores with a neural network,” in *13th ITG Symposium on Speech Communication*. 2018, pp. 261–265, VDE-Verlag.

[36] Andrew O. Hatch and Andreas Stolcke, “Generalized linear kernels for one-versus-all classification: Application to speaker recognition,” in *International Conference on Acoustics Speech and Signal Processing (ICASSP)*. 2006, pp. 585–588, IEEE.

[37] Andrew O. Hatch, Sachin S. Kajarekar, and Andreas Stolcke, “Within-class covariance normalization for svm-based speaker recognition,” in *9th International Conference on Spoken Language Processing (ICSLP)*. 2006, ISCA.

[38] Matthias Dorfer, Rainer Kelz, and Gerhard Widmer, “Deep linear discriminant analysis,” in *International Conference on Learning Representations (ICLR)*, 2015.

[39] Sandro Cumani, Niko Brümmer, Lukáš Burget, and Pietro Laface, “Fast discriminative speaker verification in the i-vector space,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4852–4855.

[40] Lukáš Burget, Oldřich Plchot, Sandro Cumani, Ondřej Glembek, Pavel Matějka, and Niko Brümmer, “Discriminatively trained probabilistic linear discriminant analysis for speaker verification,” in *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 4832–4835.