A DENOISING AUTOENCODER FOR SPEAKER RECOGNITION. RESULTS ON THE MCE 2018 CHALLENGE

Roberto Font

Biometric Vox S.L.

ABSTRACT

We propose a Denoising Autoencoder (DAE) for speaker recognition, trained to map each individual ivector to the mean of all ivectors belonging to that particular speaker. The aim of this DAE is to compensate for inter-session variability and increase the discriminative power of the ivectors prior to PLDA scoring. We test the proposed approach on the MCE 2018 1st Multi-target speaker detection and identification Challenge Evaluation. This evaluation presents a call-center fraud detection scenario: given a speech segment, detect if it belongs to any of the speakers in a blacklist. We show that our DAE system consistently outperforms the usual LDA + PLDA pipeline, achieving a Top-S EER of 4.33% and Top-1 EER of 6.11% on the evaluation set, which represents a 45.6% 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

Since the introduction of total variability modeling and ivectors [1] as a fixed and low dimensional representation of speech segments, the GMM-ivector, and more recently DNN-ivector [2, 3] paradigm, followed by a discriminative backend, has become the de-facto standard in speaker recognition. This backend usually consists on Linear Discriminat Analysis (LDA) to project ivectors to a lower dimension while increasing their discriminative power, length-normalization, and Probabilistic Linear Discriminant Analysis (PLDA) to account for inter-session variability [4, 5, 6]. More recently, alternative representations, such as x-vectors [7, 8], have been proposed. However, LDA followed by PLDA remains the state-of-the-art backend.

We propose the use of a Denoising Autoencoder (DAE) [9] to increase the discriminative power of ivectors and compensate for inter-session variability. An autoencoder is a neural network architecture that learns an internal representation that allows it to reconstruct its inputs. A denoising autoencoder is a particular type of autoencoder that learns to reconstruct a "clean" version of its inputs. In our case, the DAE takes as input an ivector and tries to map it to the mean of all the ivectors of that particular speaker. To this end, the DAE is trained to maximize the cosine distance between its output and the mean ivector for that speaker. Our proposed backend consists of: length normalization, DAE transformation and PLDA scoring.

We test this approach in the MCE 2018 1st Multi-target speaker detection and identification Challenge Evaluation. The task for the MCE 2018 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, i.e. detecting if the segment belongs to any of the blacklist speakers; and Top-1 detection, i.e. detecting which specific blacklist speaker (if any) is speaking in the segment. We refer the reader to [10] for a detailed description of the challenge.

In this paper, we describe in detail our submission for the challenge and show that the proposed DAE + PLDA backend outperforms the conventional LDA + PLDA approach. Our best system achieves a Top-S EER of 4.33% and Top-1 EER of 6.11% on the evaluation set, which represents a 45.6% error reduction with respect to the baseline system provided by the organizers. We have released source code for DAE training and testing[1].

Previous work has proposed a number of alternatives to LDA [11, 12] to account for the multimodal, non-Gaussian distribution of ivectors. Our approach differs from these alternatives in the sense that it is not designed to replace LDA but to attack the problem from a different angle. As we show in Section 4, in fact, both techniques can be combined by using LDA in the DAE-transformed ivectors or by transforming LDA-projected ivectors.

The use of denoising autoencoders for speaker recognition has been previously proposed for tasks such as denoising ivectors [13] or domain adaptation [14]. In [15, 16] an approach similar to ours is proposed. First, a Restricted Boltzmann Machine (RBM) is trained and then a DAE is fine-tuned. In contrast, our approach is much simpler. We show that even the simplest DAE can outperform the traditional LDA-PLDA backend.

The rest of the paper is organized as follows. Section 2

[1] http://github.com/BiometricVox/DAE_SpeakerID

provides an overview of the MCE 2018 evaluation, Section 3 describes both the baseline LDA-PLDA system and the proposed DAE-PLDA system and Section 4 presents the results. Finally, conclusions are drawn in Section 5.

2. CHALLENGE OVERVIEW

The MCE 2018 data have been generated from real call center user-agent telephone conversations. Instead of raw audio data, organizers processed the original data and provided 600-dimensional ivectors. This way, no special signal processing knowledge was needed to enter the evaluation. The details on the training of this ivector system can be found in [10].

The challenge data was distributed to the participants divided into three separate subsets: training, development and evaluation. The training and development portions were labeled with speaker identity, while the evaluation set was unlabeled. The composition of the different subsets is summarized in Table1.

Table 1. Summary of the different subsets

	# speakers	# utterances
Training blacklist	3.631	10.893
Training background	5.000	30.952
Development blacklist	3.631	3.631
Development background	5.000	5.000
Evaluation	?	16.017

Two complementary tasks were considered: Top-S detection and Top-1 detection. For Top-S detection, the system must decide whether 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 [10]. For both tasks, the performance metric is Equal Error Rate (EER).

During the development of our system, we focused on the Top-1 detection task. We found 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.

3. SYSTEM DESCRIPTION

3.1. LDA–PLDA system

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

- Ivectors are projected to unit length (length-normalized).
- LDA is used to project the ivectors 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 considered PLDA variant is the two-covariance model [17]. 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.

3.2. DAE System

The proposed system consists of a neural network that has as input an ivector and as output a vector with the same dimension. During training, the target is the mean of all ivectors from that speaker, and we try to maximize the cosine distance (or minimize cosine proximity) between the output and the target. During our experiments we found that cosine proximity worked better than Mean Squared Error for this task.

Our proposed architecture, shown in Figure 1, has a single hidden layer with 2000 units and tanh activation, and an output layer with dimension 600 and linear activation.

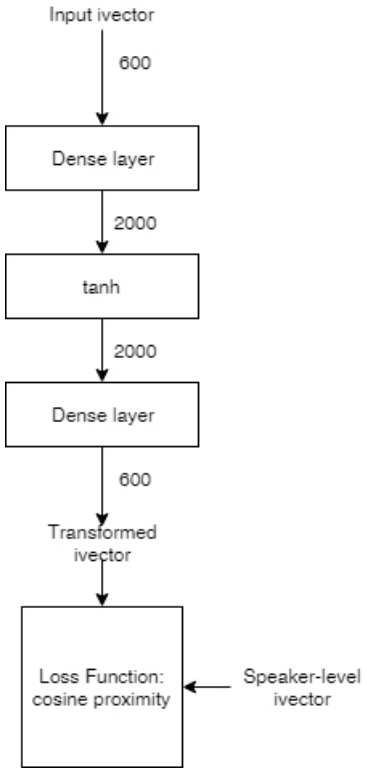

Fig. 1. Proposed DAE architecture

Additionally to this simple DAE, we test a possible extension as proposed in [13]. We extend the DAE to try to dis-

criminate between speakers. The architecture of this discriminative DAE is shown in Figure 2. We extend the DAE with two additional hidden layers of sizes 2000 and 1000 respectively and a new output layer that makes a prediction about the speaker identity. During training, the loss is a linear combination of cosine proximity for the intermediate output and cross-entropy for the final output. The rationale behind this extension regards with the computation of more discriminative transformations at the intermediate output. However, as shown in Section 4, this extension did not bring any additional improvement over the simple DAE.

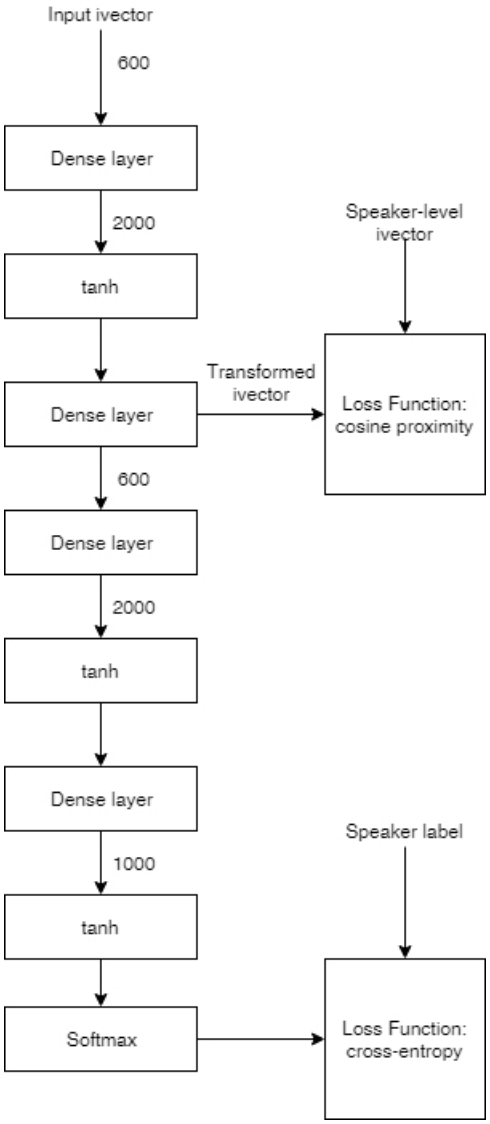

Fig. 2. Discriminative DAE architecture

We train both neural networks with Adam optimizer, a learning rate of 0.001 and a batch size of 128 for 5 epochs using only the background training set. For that, we used Keras with Tensorflow backend.

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

3.3. Score normalization

Concerning score normalization, we have used symmetric normalization (S-Norm). A set of speakers, in this case the background speakers from the training set, is used to score the test or enrollment segments against each one of the speakers in this cohort. In this way, the score after S-Norm is given by

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

where S is the raw score, 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. We refer the reader to [18] for a more detailed description of the different score normalization schemes.

For the baseline LDA-PLDA system, we use $N_t = 2000$ to compute μ_t and σ_t, and $N_m = 3000$ to compute μ_m and σ_m. For the DAE-PLDA system, we use $N_t = 1000$, $N_m = 500$. These optimal values were found by performing a grid search on the development set.

4. RESULTS

The results obtained from the development set of the different systems under study are shown in Table 2. As we can see, score normalization is extremely beneficial, particularly for the DAE system. In the last two rows, we can see that combining the DAE with LDA gives reasonable results but does not provide any further improvement. We can also see that the discriminative DAE does not improve the results over the simple DAE.

Table 2. Performance on development set

System	Top-S EER [%]	Top-1 EER [%]
Baseline	2.01	12.26
LDA + PLDA	1.82	6.96
LDA + PLDA + S-Norm	1.26	6.72
DAE + PLDA	1.73	7.22
DAE + PLDA + S-Norm	1.25	**6.52**
disc. DAE + PLDA + S-Norm	1.38	6.80
DAE + LDA + PLDA + S-Norm	1.33	6.64
LDA + DAE + PLDA + S-Norm	**1.15**	6.78

To score the evaluation set, we pooled together the blacklist training and development sets and used this combined set

instead of the blacklist training set. In this way, speaker models are computed using the 4 utterances available between training and development sets. For the DAE system, as an extra regularization step, we trained an ensemble of 10 models with different initializations. Apart from this modifications, the rest of parameters are kept identical.

Results are shown in Table 3. Again, the DAE system achieves lower error rates than the LDA-PLDA system. It is worth noting that the system exhibits no overfitting. In fact, results on the evaluation set are appreciably better than on the development set, as expected in case of no overfitting, since now the speaker models are computed using 4 utterances instead of 3.

Finally, we experimented with score-level fusion of both system using logistic regression, but we could not get any significant improvement. Our primary submission to the challenge consisted on the DAE-PLDA system as primary submission and the LDA-PLDA system as contrastive submission.

Table 3. Performance on evaluation set

System	Top-S EER [%]	Top-1 EER [%]
Baseline	6.24	11.24
LDA + PLDA	4.63	6.81
LDA + PLDA + S-Norm	4.42	6.56
DAE + PLDA	4.60	6.75
DAE + PLDA + S-Norm	**4.33**	**6.11**

5. CONCLUSIONS

We have tested a denoising autoencoder architecture to transform ivectors for speaker recognition. Results on the MCE 2018 challenge show that this simple architecture shows promise and can improve results with no appreciable overfitting. This challenge proposes an interesting open-set speaker identification task which, to date, has received little attention. It has, however, some limitations: not having access to raw-audio makes it difficult to perform effective data augmentation, and a modest (around 4) number of utterances per speaker. We hypothesize that having access to a larger number of utterances per speaker and making use of data augmentation could be beneficial to neural network-based approaches like the proposed DAE. Future work will try to validate this hypothesis by applying the proposed approach on different, larger tasks. Using x-vectors or other speaker embeddings as input, instead of ivectors, is another possible extension of the present work.

6. REFERENCES

[1] Najim Dehak, Patrick 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.

[2] Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[3] Patrick Kenny, Vishwa Gupta, Themos Stafylakis, Md. Jahangir Alam, and Pierre Ouellet, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in *Odyssey*, 2014.

[4] Patrick Kenny, "Bayesian speaker verification with heavy-tailed priors," in *Odyssey*, 2010.

[5] Mohammed Senoussaoui, Patrick Kenny, Niko Brümmer, Edward de Villiers, and Pierre Dumouchel, "Mixture of plda models in i-vector space for gender-independent speaker recognition," in *12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011.

[6] S. Cumani O. Glembek P. Matejka L Burget, O. Plchot and N. Brümmer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification," in *12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011.

[7] David Snyder, Daniel Garcia-Romero, Daniel Povey, and Sanjeev Khudanpur, "Deep neural network embeddings for text-independent speaker verification," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017.

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

[9] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th International Conference on Machine Learning*, New York, NY, USA, 2008, ICML '08, pp. 1096–1103, ACM.

[10] S. Shon, N. Dehak, D. Reynolds, and J. Glass, "MCE 2018: The 1st Multi-target Speaker Detection and Identification Challenge Evaluation (MCE) Plan, Dataset and Baseline System," *ArXiv e-prints*, July 2018.

[11] Seyed Omid Sadjadi, Jason Pelecanos, and Weizhong Zhu, "Nearest neighbor discriminant analysis for robust speaker recognition," in *15th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2014, pp. 1860–1864.

[12] Abhinav Misra, Shivesh Ranjan, and John H.L. Hansen, "Locally weighted linear discriminant analysis for robust speaker verification," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 2864–2868.

[13] Shivangi Mahto, Hitoshi Yamamoto, and Takafumi Koshinaka, "i-vector transformation using a novel discriminative denoising autoencoder for noise-robust speaker recognition," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 3722–3726.

[14] Suwon Shon, Seongkyu Mun, Wooil Kim, and Hanseok Ko, "Autoencoder based domain adaptation for speaker recognition under insufficient channel information," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 1014–1018.

[15] Timur Pekhovsky, Sergey Novoselov, Aleksei Sholohov, and Oleg Kudashev, "On autoencoders in the i-vector space for speaker recognition," in *Odyssey*, 2016.

[16] Sergey Novoselov, Timur Pekhovsky, Oleg Kudashev, Valentin Mendelev, and Alexey Prudnikov, "Non-linear plda for i-vector speaker verification," in *16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015.

[17] Niko Brümmer and Edward de Villiers, "The speaker partitioning problem," in *Odyssey*, 2010.

[18] Pavel Matejka, Ondrej Novotny, Oldrich Plchot, Lukas Burget, Mireia Diez Sánchez, and Jan Cernocky, "Analysis of score normalization in multilingual speaker recognition," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 1567–1571.