A MULTI-TARGET SPEAKER DETECTION AND IDENTIFICATION SYSTEM BASED ON COMBINATION OF PLDA AND DNN

Nikša Jakovljević, Tijana Delić, Dragiša Mišković, Simona Đurić and Tatjana Lončar Turukalo

VINS [38] University of Novi Sad, Faculty of Technical Sciences

ABSTRACT

This paper describes a multi-target speaker detection and identification system based on combination of probabilistic linear discriminant analysis (PLDA) and deep neural network (DNN). PLDA returns log-likelihoods for each blacklist speaker that unknown i-vector is generated by given speaker. DNN returns the same likelihoods as PLDA, but in comparison to the PLDA, one can find that it better separates blacklist and background speakers than particular blacklist speakers. For that reason, score from the DNN is used to decide whether the unknow i-vector belongs to the background speaker, and if it does, corresponding PLDA score is decreased by minimum absolute value of the all PLDA scores. The proposed composite system on the MCE 2018 development set achieves 0.86 % and 7.14 % Top-S and Top-1 EER, respectively.

Index Terms— DNN, GPLDA, identification, i-vector, speakers

1. INTRODUCTION

On a number of standard datasets, the combination of the i-vectors and probabilistic linear discriminant analysis (PLDA) achieved state-of-the-art performance [1-3]. Therefore, we decided to implement an PLDA system, and in case if it would achieve better performance than baseline, we would treat it as a baseline in our experiments.

Our focus was on the development of a method based on deep neural networks (DNN), since in the last few years such systems were successfully applied to different classification tasks [4-8]. The first approach was based on assumption that DNN can map i-vectors into new vector space where the vectors corresponding to different speakers would be more separable than in the original i-vector space. To achieve this, the Siamese network architecture [5] was chosen as an appropriate architecture (see Fig. 1). The cosine distance was kept as a distance measure in the target space, because it is viewed as good measure in the original space. The network was trained to decide whether two input i-vectors belong to the same speaker or not. Both, minimum square distance and minimum cross entropy were chosen as the objective function. Unfortunately, Top-S and Top-1 scores obtained

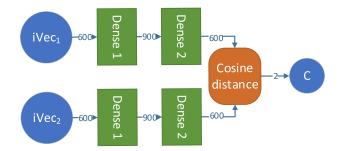


Figure 1: Siamese network architecture with 2 full connected layers (Dense 1/2).

with this approach was higher than the baseline, thus it will be excluded for further consideration.

The second approach was also based on Siamese network, but the objective function was changed. The single subnetwork (in Fig.1 it contains layers Dense 1 and 2) was trained to minimize cosine distance between output vector and mean i-vector representing speaker in the training set for given input i-vector. Our intention was to infer DNN to reduce noised caused by channel. The performance of this system is very close to the baseline performance.

The last approach was simple feedforward DNN which map input i-vectors into probabilities that each of the blacklist speakers generate it. Although Top-1 EER of this system was significantly higher than Top-1 EER of the baseline system, Top-S EER was almost 2 times smaller, thus we decided to exploit this fact to improve our PLDA system.

2. SYSTEM DESCRIPTION

A. Probabilistic Linear Discriminant Analysis

In this work we applied a modified PLDA model used in [8] where each i-vector w_r can be decomposed as:

$$w_r = m + \Phi \beta + \epsilon_r \tag{1}$$

where: m is the global offset, the columns of Φ provide a basis for the speaker-specific subspace and ϵ_r is a residual term modeling channel variability and approximation error. The model parameters $\{m, \Phi, \Sigma\}$ are estimated using expectation maximization (EM) algorithm on the training data. Since PLDA training uses random initialization, 10

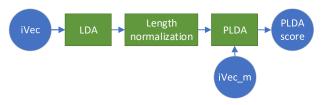


Figure 2: Simplified block diagram of PLDA scoring.

independent runs were made for each set of model parameters.

The PLDA verification score – that represents loglikelihood ratio for the hypothesis test corresponding to that both i-vectors were or were not generated by the same speaker, can be calculated as:

$$s_{PLDA} = \widetilde{w}_i^T Q \widetilde{w}_i + \widetilde{w}_m^T Q \widetilde{w}_m + 2 \widetilde{w}_i^T P \widetilde{w}_m$$
 (2)

with:

$$\begin{split} \vec{W} &= W - m \\ Q &= \Sigma_{tot}^{-1} - (\Sigma_{tot} - \Sigma_{ac} \Sigma_{tot}^{-1} \Sigma_{ac})^{-1} \\ P &= \Sigma_{tot}^{-1} \Sigma_{ac} (\Sigma_{tot} - \Sigma_{ac} \Sigma_{tot}^{-1} \Sigma_{ac})^{-1} \\ \Sigma_{ac} &= \Phi \Phi^T \\ \Sigma_{tot} &= \Sigma_{ac} + \Sigma \end{split}$$
 There are strong empirical evidences of non-Gaussian

There are strong empirical evidences of non-Gaussian behavior of speaker and channel effects, which can be compensated by length normalization of input vectors [8], thus vector length normalization precedes PLDA scoring.

In many cases dimensionality reduction improves results, thus we decided to apply linear discriminant analysis (LDA) before PLDA [10]. Since we were not convinced that LDA would lead to better performance, it was optional step in this study. In Fig. 2 we show a simplified block diagram of ivector PLDA scoring.

To estimate parameters of LDA and PLDA matrices only training set was used, but the number of retained dimensions was tuned on a development set. The optimal number of retained dimension was 400, and the number of iterations in EM estimation of PLDA matrices 40.

Top-S and Top-1 EERs on development set are 2.15 % and 7.24 %, respectively.

B. Deep Neural Network

The used model is simple feedforward neural network as it shown in Fig 3. Inputs are original i-vectors, while the outputs are one hot vectors of dimension 3631, since it is the number of speakers from blacklist. The first idea was to have one-hot vector for each speaker, but it would be extremely high dimensional and unnecessary because the task is not to identify speakers who are not from blacklist, and also it would be problem when those speakers are changed. The next idea was to put one additional output, in order to have unique one-hot vector for all speakers that are not from blacklist. This would lead to overfitting of that particular output since there would be a few thousand times more inputs corresponding to exactly that output in comparison to available samples that correspond to other possible outputs.



Figure 3: Simplified block diagram of DNN scoring.

The most logical approach was to put all zeros as the target when input corresponds to speaker that are not from blacklist, but the used cost function expects to have 1 as a sum of outputs, thus the only possibility was to set all outputs to 1/3631 when the input speaker is not from the blacklist. That way all outputs will have the same chance to learn when its output should be 1 or some low value.

The model consists of: input layer of dimension 600 (as it is the length of one i-vector), 2 hidden layers of dimension 1024, and output layer of dimension 3631 (see Fig. 3). Dimension of 1024 for hidden layers is chosen as number higher than 600 and smaller than 3631 which is at the same time power of 2. As activation function, tangent hyperbolic function is used for input and hidden layers, while for the output layer linear function is chosen. Since the task is classification in multiple classes, used cost function is cross entropy with softmax. The optimizer used is stochastic gradient descent with momentum of 0.9 and with regularization parameter L2 = 0.00001. During the training, one sample represents one batch. Learning rate was 0.01 and training lasts 20 epochs when cost function of training set got into saturation. Everything is implemented in CNTK framework [11].

Top-S and Top-1 EERs on development set are 0.84 % and 16.08 %, respectively. Although Top-S error is very good with both false negative rate and false positive rate less than 1%, but the top-1 error is not satisfying. It can be concluded that trained network is able to classify speakers between blacklist and background, but not to identify blacklist speakers accurate enough.

C. Fusion of PLDA and DNN

Since the DNN can classify speakers to blacklist and non-blacklist, and PLDA is very good at identifying speakers, we decided to merge those two systems. Finally, our system uses DNN to determine what speakers are from the blacklist and PLDA to calculate scores (distances between speakers). Then, all scores for speakers that DNN determined as non-blacklist, are decreased by penalty p, i.e.:

$$score = \begin{cases} s_{PLDA} & s_{DNN} > \theta \\ s_{PLDA} - p & s_{DNN} \le \theta \end{cases}$$
 (4)

where: s_{PLDA} is score obtained by PLDA, s_{DNN} is score obtained by DNN and θ is threshold.

Since the target output for DNN in case of background speakers was $1/3631 = 2.8 \cdot 10^{-4}$, and we expected unseen blacklist vectors we set the value θ on 10^{-2} . The value of penalty p was set on 300, since it is close to s_{PLDA} range in training set. Top-S and Top-1 EERs on development set are 0.86 % and 7.14 %, respectively.

Table 1: Results on development set.

Submission	Fixed	Open
Primary	Top-S: 0.86% Top-1: 7.14% (257)	/
Contrastive 1	/	/
Contrastive 2	/	/

3. ACKNOWLEDGEMENT

This work was supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia, TR 32035 and III 47020. We acknowledge the support of AlfaNum Company from Novi Sad, Serbia, for providing us computation resources.

4. REFERENCES

- [1] Johan Rohdin, Anna Silnova, Mireia Diez, Oldrich Plchot, Pavel Matejka, and Lukas Burget, "End-to-end DNN Based Speaker Recognition Inspired by i-vector and PLDA," ArXiv e-prints arXiv:1710.02369v2 2018.
- [2] Md Hafizur Rahman, Ahilan Kanagasundaram, Ivan Himawan, David Dean, and Sridha Sridharan, "Improving PLDA speaker verification performance using domain mismatch compensation techniques," Computer Speech & Language, vol. 47, pp. 240-258, 2018.
- [3] Alam Jahangir, Patrick Kenny, Gautam Bhattacharya and Marcel Kockmann, "Speaker verification under adverse conditions using i-vector adaptation and neural networks," in *Proc. Interspeech*, 2017, pp. 3732–3736.
- [4] Geoffrey Hinton, Li Deng, Dong Yu, Abdel Rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, George Dahl, and Brian Kingsbury,

- "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, November 2012.
- [5] Weibo Liu, Zidong Wang, Xiaohui Liu, Nianyin Zeng, Yurong Liu, and Fuad E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.
- [5] Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov, "Siamese neural networks for one-shot image recognition," in ICML Deep Learning Workshop, 2015, vol. 2.
- [6] Fred Richardson, Douglas Reynolds, and Najim Dehak, "Deep neural network approaches to speaker and language recognition," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1671–1675, 2015.
- [7] Zhanxiong Wang, Keke He, Yanwei Fu, Rui Feng, Yu-Gang Jiang, and Xiangyang Xue, "Multi-task deep neural network for joint face recognition and facial attribute prediction," in *Proc. Int. Conf. on Multimedia Retrieval*, Bucharest, Romania, 2017, pp. 365–374.
- [8] Pavel Matejka, Ondrej Glembek, Ondrej Novotny, Oltrich Plchot, Franišek Grezl, Lukas Burget, and Jan Honza Cernocky, "Analysis of DNN approaches to speaker identification," in Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5100–5104.
- [9] Daniel Garcia-Romero and Carol Y. Espy-Wilson, "Analysis of i-vector length mormalization in speaker recognition systems," in *Proc. Interspeech* 2011, Florence, Italy, August 27-31, 2011, pp. 249–252.
- [10] Pavel Matejka, Ondrej Glembek, Fabio Castaldo, Md Jahangir Alam, Oltdirch Plchot, Patrick Kenny, Lukaš Burget, and Jan Černocky, "Full-covariance UBM and heavy-tailed PLDA in ivector speaker verification," in *Proc. ICASSP*, 2011, pp. 4828– 4831.
- [11] Frank Seide and Amit Agarwal, "CNTK: Microsoft's opensource deep-learning toolkit," in *Proc. Int. Conf. on Knowledge Discovery and Data Mining (KDD '16)*, New York, NY, USA, 2016, pp. 2135-2135.