Asymmetric and trial-dependent modeling: the contribution of LIA to SdSV Challenge Task 2

Pierre-Michel Bousquet, Mickael Rouvier

LIA - Avignon University

first.lastname@univ-avignon.fr

Abstract

The SdSv challenge Task 2 provided an opportunity to assess efficiency and robustness of modern text-independent speaker verification systems. But it also made it possible to test new approaches, capable of taking into account the main issues of this challenge (duration, language, ...). This paper describes the contributions of our laboratory to the speaker recognition field. These contributions highlight two other challenges in addition to short-duration and language: the mismatch between enrollment and test data and the one between subsets of the evaluation trial dataset. The proposed approaches experimentally show their relevance and efficiency on the SdSv evaluation, and could be of interest in many real-life applications.

1. Introduction

The Short-duration Speaker Verification Task 2 evaluation is a text-independent speaker recognition evaluation, based on the recently released DeepMine dataset [1, 2]. This dataset is comprised of various duration utterances (with a significant proportion of less than 10 seconds) recorded by Persian native persons, some of them in English. The evaluation proposes to test and improve speaker recognition methods on speech data with varying degree of phonetic overlap between the enrollment and test utterances [3]. Robustness of speaker embeddings extracted from deep neural networks (DNN) to short-duration utterances and efficiency of the domain adaptation techniques (as Persian language is unknown to the usual speech databases) can be seen as the main objectives of this challenge. The fairly wide DeepMine development dataset provided for this challenge, which is speaker-labeled, allows to better fit model to data, even if the availability of some English speeches spoken by Persian native persons is lacking.

The task of language domain adaptation is usually addressed during the back-end procedure. Several methods have been proposed, unsupervised [4, 5, 6], or supervised [7, 8, 9] when in-domain labeled data are available. For SdSv, the availability of a relatively large size and labeled in-domain dataset makes it possible to also consider language pre-adaptation inside the supervised learning of a DNN-based feature extractor. Section 2.2 details our proposed approach of DNN Persian-refinement.

Table 1: *Data provided by SdSv for speaker verification.*

enrollment	test
1 to 29 utterances (average of 7) giving a net speech duration from 3 to 120 seconds	1 utterance (95% less than 5 seconds)

The challenge focuses on short-duration and cross-lingual speaker recognition but it also has a particularity, which is often overlooked in the speaker recognition field: Table 1 shows that the characteristics of the speech material provided for enrollment and for test are different enough to assume a mismatch between the distribution of their vector representations. It would also be of benefit to take into account such a mismatch. Moreover, mixing, in a unique evaluation, trials with a small or large enrollment sample and, also, test utterances in Persian or English can limit efficiency of a unique modeling. Designing specific back-end models for dealing with trial mismatch could be of interest. Section 3 explains how we hit on all these points.

2. Front-end feature extraction

2.1. Initial DNN learning

The system used in SdSV Challenge is based on x-vector/PLDA. Our x-vector system is built based on the Kaldi recipe [10], but with some modifications. Voxceleb2 [11] and Librispeech [12] sets are combined to generate the training set for the x-vector extractor.

The following data augmentation methods are used in this paper. Apart from the four augmentation methods used in [10], we also include audio compression randomly picked between ogg, mp3 and flac codec, high-pass filtering randomly picked in [1000Hz;3000Hz] and low-pass filtering randomly picked in [500Hz;1500Hz]. Finally, the training data consist of 8-fold augmentation that combines clean data with 7 copies of augmented data.

During the training part the utterances are further cut into segments of 2s for the neural network training. 60-dimensional filter banks (Fbanks) are used for the x-vector system, with an energy-based Voice Activity Detector (VAD) to remove silence. A short-time cepstral mean subtraction is applied over a 3-second sliding window.

Table 2 presents the Extended-TDNN architecture used. In addition to this architecture, we proposed to increase the dimension of each layer to 1024 only for the frame-level. Except the layer 9 which is used as an expansion layer and is fixed to 3000 dimension. The embeddings are extracted after the first dense layer with a dimensionality of 512. The neural network is trained for 9 epochs using natural-gradient stochastic gradient descent and minibatch size of 128.

2.2. Front-end language adaptation

In order to adapt the x-vector system to a new language, we use the neural network trained on Voxceleb2 and Librispeech corpus as pre-trained model. Then, we propose to freeze on pre-trained model all pre-pooling TDNN layers and re-train the other layers on DeepMine corpus (using 8-fold augmentation). The neural network is trained only with 1 epoch and minibatch size of 128 (we observe in the leaderboard that more epochs do

Table 2: *Topology of the Extended-TDNN x-vector architecture.*

Layer	Layer type	Context	Size
1	TDNN-ReLU	t-2:t+2	1024
2	Dense-ReLU	t	1024
3	TDNN-ReLU	t-2, t, t+2	1024
4	Dense-ReLU	t	1024
5	TDNN-ReLU	t-3, t, t+3	1024
6	Dense-ReLU	t	1024
7	TDNN-ReLU	t-4, t, t+4	1024
8	Dense-ReLU	t	1024
9	Dense-ReLU	t	3000
10	Pooling (mean+stddev)	t	6000
11	Dense(Embedding)-ReLU	t	512
12	Dense-ReLU	t	512
13	Dense-Softmax	t	Nb spks

not improve results). The resulting "Persian-refined" DNN better combines the rich information of the wide but out-of-domain initial training set and adequacy to the target language.

3. Back-end asymmetric modeling

3.1. Four-covariance model

As explained in the introduction and observed in Table 1, it can be assumed that the distributions of the target speaker model and of the test x-vector are sufficiently distinct to require two PLDA modelings. Introduced in [13] for mismatch of duration between enrollment and test data, the four-covariance model (4-cov) is an asymmetric modeling, which allows to compute two distinct PLDA models, here one for enrollment data and the other for test data, then to fit a probabilistic relation between them in order to compute a LLR-score, despite the mismatch.

For SdSv challenge, we choose as target speaker model the length-normalized average of the enrollment sample, since this approach has proved to be efficient and robust [14]. Let denote by $\mathbf{w}_1$ a vector of type 1 (here of the latter type, computed on an enrollment sample as described in column 1 of Table 1) and similarly $\mathbf{w}_2$ of type 2 (here a test vector as described in column 2 of Table 1). The Gaussian PLDA model [15] for type i, $i = 1$ or 2, assumes that:

$$\begin{aligned} \mathbf{w}_i &= \mu_i + \mathbf{\Phi}_i y_i + \varepsilon_i \\ y_i &\sim \mathcal{N}(0, \mathbf{I}) \\ \varepsilon_i &\sim \mathcal{N}(0, \mathbf{\Gamma_i}) \end{aligned} \tag{1}$$

where $\mathcal{N}$ denotes the normal pdf, $\mathbf{I}$ is the identity matrix, μ_i a global offset and the latent variable y_i is only dependent on the speaker and statistically independent of the residual term ε_i. The 4-cov modeling assumes a linear relation between the two PLDA models by their speaker factors:

$$y_2 = \mathbf{A} y_1 + \eta \tag{2}$$

$$\eta \sim \mathcal{N}(0, \mathbf{M}) \tag{3}$$

To estimate the matricial parameters $\mathbf{A}$ and $\mathbf{M}$, the point estimate of training speaker factors y_i is computed using the expectation given by the PLDA E.M. algorithm:

$$y_i = \left(n_s \mathbf{\Phi}_i^t \mathbf{\Gamma}_i^{-1} \mathbf{\Phi}_i + \mathbf{I}\right)^{-1} \mathbf{\Phi}_i^t \mathbf{\Gamma}_i^{-1} \sum_{k=1}^{n_s} (w_k - \mu_i) \tag{4}$$

where w_k denotes the k^{th} of n_s examples for the speaker s. Then, a multivariate regression is carried out, which minimizes the least square error. Denoting by $\mathbf{Y}_i$ the row-matrix of the y_i the closed-form expressions of $\mathbf{A}$ and $\mathbf{M}$ are:

$$\begin{aligned} \mathbf{A} &= \mathbf{Y}_2^t \mathbf{Y}_1 \left(\mathbf{Y}_1^t \mathbf{Y}_1\right)^{-1} \\ \mathbf{M} &= cov\left(y_2 - \mathbf{A} y_1\right) \end{aligned} \tag{5}$$

where $cov()$ is the covariance matrix. A straightforward computation shows that the LLR score between two vectors $\mathbf{w}_1, \mathbf{w}_2$ of type 1 and 2 can be expressed in a simple form (simpler than in the original paper) as:

$$s(\mathbf{w}_1, \mathbf{w}_2) = -\frac{1}{2} \begin{pmatrix} \mathbf{w}_1 - \mu_1 \\ \mathbf{w}_2 - \mu_2 \end{pmatrix}^t \mathcal{M} \begin{pmatrix} \mathbf{w}_1 - \mu_1 \\ \mathbf{w}_2 - \mu_2 \end{pmatrix} \tag{6}$$

up to a constant, where

$$\begin{aligned} \mathcal{M} = &\begin{pmatrix} \mathbf{\Phi}_1 \mathbf{\Phi}_1^t + \mathbf{\Gamma}_1 & \mathbf{\Phi}_1 \mathbf{\Phi}_1^t \mathbf{A}^t \\ \mathbf{A} \mathbf{\Phi}_1 \mathbf{\Phi}_1^t & \mathbf{A} \mathbf{\Phi}_1 \mathbf{\Phi}_1^t \mathbf{A}^t + \mathbf{\Gamma}_2 + \mathbf{M} \end{pmatrix}^{-1} \\ &- \begin{pmatrix} \mathbf{\Phi}_1 \mathbf{\Phi}_1^t + \mathbf{\Gamma}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{\Phi}_2 \mathbf{\Phi}_2^t + \mathbf{\Gamma}_2 \end{pmatrix}^{-1} \end{aligned} \tag{7}$$

3.2. Specific score normalization

Taking benefit of the score normalization to enhance performance requires adapting the usual S-normalization to the specific case of an asymmetric model: the impostor cohorts are dependent on the type of data and the order of pairwise vectors to score must be respected. Given a trial between enrollment-based and test vectors $\mathbf{w}_e$ and $\mathbf{w}_t$, score-normalization is performed on score $s(\mathbf{w}_e, \mathbf{w}_t)$ such that:

$$\begin{aligned} \widehat{s}(\mathbf{w}_e, \mathbf{w}_t) = &\frac{1}{2} \frac{s(\mathbf{w}_e, \mathbf{w}_t) - \mu\left(s(\mathbf{w}_e, \mathbf{\Omega}_t)\right)}{\sigma\left(s(\mathbf{w}_e, \mathbf{\Omega}_t)\right)} \\ &+ \frac{1}{2} \frac{s(\mathbf{w}_e, \mathbf{w}_t) - \mu\left(s(\mathbf{\Omega}_e, \mathbf{w}_t)\right)}{\sigma\left(s(\mathbf{\Omega}_e, \mathbf{w}_t)\right)} \end{aligned} \tag{8}$$

where $\mathbf{\Omega}_e, \mathbf{\Omega}_t$ are cohort impostors, specific to enrollment and test, and μ, σ are the mean and standard deviation functions, possibly computed on the top scores only.

3.3. Trial-dependent models

Table 3: *Percentages of trials in the evaluation trial set, depending on the target speaker model (how many enrollment segments are available ?) and on the test language.*

	test language		
enrollment #segs	Persian	English	Total
< 5	36%	38%	74%
$\geqslant 5$	4%	22%	26%
	40%	60%	

Table 3 details the proportion of trials in the evaluation set, depending on the size of the speaker enrollment sample and on the language of test. The 4-cov model allows to fit PLDA models to each of these enrollment-test cases. Table 4 shows the different training sets used for PLDA, depending on the trial. We apply the 4-cov model to each type of mismatch: (average of sample of various size and duration)/(one short duration

utterance in Persian or English). The language of the test segments is estimated by a speech detector. For test utterances in English, PLDA is interpolated as proposed in [7], using our English training database. Let us note that the x-vectors of this database are extracted from our Persian-refined neural network, hence partially adapted to Persian language.

For a better understanding, we detail one case of Table 4. The last row corresponds to trials with more than 5 examples for enrollment and a test utterance in English:

- the PLDA training dataset for model 1 of 4-cov model (the one for enrollment) is made up of length-normalized averages of 12 vectors lasting more than 7.5 seconds, extracted from utterances of the DeepMine development set [1]).
- the PLDA training datasets for model 2 of 4-cov model (the one for test) are comprised of utterances lasting less than 5 seconds, from (i) the same DeepMine development set, (ii) our adapted English development set. The resulting model for test interpolates the last two sub-models (i) and (ii) [7].

As the final score file to submit mixes four scoring formulas, the scores are calibrated by using development trial datasets, specific to the four cases of Table 4 and all based on DeepMine development data.

Table 4: *Datasets for trial-dependent model training. L-average means the length-normalized average of the enrollment sample.*

trial:		4-covariance model	
enrollment #segs	test language	model 1 for enrollment	model 2 for test
< 5	Persian	3 vectors L-average < 5 sec.	< 5 sec.
< 5	English	3 vectors L-average < 5 sec.	< 5 sec. & English-dev
$\geqslant 5$	Persian	12 vectors L-average $\geqslant 7.5$sec.	< 5 sec.
$\geqslant 5$	English	12 vectors L-average $\geqslant 7.5$sec.	< 5 sec. & English-dev

4. Experiments and results

For acoustic features MFCC are extracted by using Kaldi toolkit [16] with 23 cepstral coefficients and log-energy, a cepstral mean normalization being applied with a window size of 3 seconds. Voice Activity Detection removes silence and low energy speech segments. The simple energy-based VAD uses the C0 component of the acoustic feature.

Table 5 provides results of our contributions, in terms of EER and minDCF, as reported in the SdSv Task 2 leaderboard. The first system (initial) trains the DNN and all the back-end transformations by using only the out-of-domain database, described in section 2.1. The second system benefits from the DeepMine in-domain development set provided by the SdSv organizers. It is used to refine the DNN learning by using the additional training stage described in section 2.2, then for learning all the back-end transformations (centering, whitening, length-normalization and PLDA) instead of the initial database. Let us note that, hence, no adaptation of out-of-domain data to Persian language is carried out during the back-end process to enhance modeling. The third system additionally leverages an out-of-domain development set for back-end trainings. This dataset is extracted from the one used for the first learning step of the DNN extractor and adapted by using fDA [5], an unsupervised domain adaptation method, similar to CORAL [4], which takes into account the residual components. The resulting adapted set then allows interpolation between out-of-domain and in-domain PLDA models [7]. As expected, systems employing the in-domain development set during front-end and back-end learning outperform the initial submission. It is worth noting that including the adapted out-of-domain development set into the PLDA modelings (row 3) significantly increases performance, but only in terms of minimal DCF.

Table 5: *Results of the different contributions to the SdSv evaluation.*

	EER%	minDCF
Initial	7.38	0.3682
With DeepMine dev. set	4.41	0.2103
+ out-of-domain adapted set	4.42	0.1823
4cov-model	3.28	0.1554
+ specific S-norm	3.15	0.1427
+ trial-dependent models	2.88	0.1261

The fourth system applies the four-covariance model. Let us note that this system does not use the adapted out-of-domain dataset during the back-end trainings. For the enrollment model, the training speaker models are the length-normalized averages of 15 examples (3 original segments + 12 data augmented) and, for the test model, only training segments of less than 5 seconds are selected. The gain of performance involved by this method is significant, both in terms of EER and minDCF, even without the help of the wide out-of-domain development set.

The following system adds to the latter the specific score-normalization proposed in section 3.2, with 400 top-scores. The resulting gain of performance shows that the normalization of score is compatible with an asymmetric model.

The last system applies the trial-dependent 4-cov modeling described in section 3.3 and Table 4. The gain of performance confirms the heterogeneity between the trial partitions listed in Table 4 and the ability of the 4-cov model to handle such type of mismatch.

Relevance and efficiency of our various contributions are clearly demonstrated. The final system takes full account of the challenges of SdsV Task 2: short-duration utterances and adaptation to new language, reported in terms of performance in the first rows of Table 5, then mismatch between enrollment and test distributions or trial partitions, reported in the last rows.

5. Conclusions

The SdSv challenge made it possible to test and compare the efficiency of DNN based systems to deal with short-duration utterances. Data augmentation could also contribute to better fit these data, which are known to be very varied. The task of language adaptation was usually tackled during the back-end process. For the SdSv challenge, the availability of a sizable in-

domain labeled dataset allowed to extend this task to the DNN supervised learning stage.

Our contribution highlights the concern of mismatch between enrollment and test speech material, in terms of quantity of information. The proposed four-covariance model applies a specific asymmetric modeling, which focuses on a type of mismatch. It reveals the benefit of refining the back-end modeling to take into account this issue. Moreover, this model allows for better fit of specificities, here the relative heterogeneity of the evaluation trials.

The last system of Table 5 was our final submission for this challenge. The good ranking obtained with a system using a single front-end feature extractor shows that a system including all these contributions is able to compete with fusions of systems based on distinct DNN architectures and configurations.

6. Acknowledgements

This research was supported by the ANR agency (Agence Nationale de la Recherche), on the RoboVox project (ANR-18-CE33-0014).

7. References

[1] H. Zeinali, H. Sameti, and T. Stafylakis, "DeepMine speech processing database: Text-dependent and independent speaker verification and speech recognition in Persian and English." in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2018, pp. 386–392.

[2] H. Zeinali, L. Burget, and J. Cernocky, "A multi purpose and large scale speech corpus in Persian and English for speaker and speech recognition: the DeepMine database," in *Proc. ASRU 2019 The 2019 IEEE Automatic Speech Recognition and Understanding Workshop*, 2019.

[3] K. A. Zeinali, Hossein nad Lee, J. Alam, and L. Burget, "Short-duration speaker verification (SdSV) challenge 2020: the challenge evaluation plan." arXiv preprint arXiv:1912.06311, Tech. Rep., 2020.

[4] J. Alam, G. Bhattacharya, and P. Kenny, "Speaker verification in mismatched conditions with frustratingly easy domain adaptation," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2018.

[5] P.-M. Bousquet and M. Rouvier, "On Robustness of Unsupervised Domain Adaptation for Speaker Recognition," in *Proc. Interspeech 2019*, 2019, pp. 2958–2962. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-1524

[6] K. A. Lee, Q. Wang, and T. Koshinaka, "The CORAL+ algorithm for unsupervised domain adaptation of PLDA," *CoRR*, vol. abs/1812.10260, 2018. [Online]. Available: http://arxiv.org/abs/1812.10260

[7] D. Garcia-Romero and A. McCree, "Supervised domain adaptation for i-vector based speaker recognition," in *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 2014, pp. 4047–4051.

[8] H. Aronowitz, "Inter dataset variability compensation for speaker recognition," *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4002–4006, 2014.

[9] J. A. V. López and E. Lleida, "Bayesian adaptation of PLDA based speaker recognition to domains with scarce development data," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2012.

[10] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[11] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in *Proc. Interspeech 2018*, 2018, pp. 1086–1090.

[12] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[13] P.-M. Bousquet and M. Rouvier, "Duration mismatch compensation using four-covariance model and deep neural network for speaker verification," in *Proc. Interspeech 2017*, 2017, pp. 1547–1551. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-93

[14] P. Rajan, A. Afanasyev, V. Hautamäki, and T. Kinnunen, "From single to multiple enrollment i-vectors: Practical PLDA scoring variants for speaker verification," *Digit. Signal Process.*, vol. 31, pp. 93–101, 2014.

[15] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *International Conference on Computer Vision*. IEEE, 2007, pp. 1–8.

[16] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *IEEE Workshop on Automatic Speech Recognition & Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.