The LIA System Description for SdSV Challenge Task 2

Pierre-Michel Bousquet, Mickael Rouvier

LIA - Avignon University

first.lastname@univ-avignon.fr

Abstract

For our submission to the SdSv challenge we have explored x-vector extractor topologies, front-end language adaptation, then back-end asymmetric and trial-dependent modeling. Our final entry to the challenge is the score of a single front-end system, in order to determine a robust and efficient speaker recognition approach.

1. Introduction

The SdSv challenge consists of recordings from various duration in Persian or English, by Persian native speakers. The major challenges for this evaluation are the improvement of neural network architecture, training algorithms for supervised language adaptation and modeling to deal with short duration utterances. We show in section 3 that a mismatch between enrollment and test data can also be taken into account, as well as between some partitions of the evaluation trial set. This paper presents the technical details of our front-end feature extractor and back-end process.

2. Front-end feature extraction

The system used in SdSV Challenge is based on x-vector/PLDA. Our x-vector system is built based on the Kaldi recipe [1], but with some modifications. Voxceleb2 [2] and Librispeech [3] 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 [1], 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, 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 1 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.

In order to adapt the x-vector system to new language, we used neural network trained on Voxceleb2 and Librispeech corpus as pre-trained model. We 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 not improve results).

Table 1: *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

3. Back-end modeling

The back-end process of LIA system is based on an asymmetric model inspired by the *four-covariance* model (*4-cov*), which we introduced in [4]. This modeling distinguishes between enrollment and test distributions, leading to asymmetric scoring formulas (i.e. $score(w_1, w_2) \neq score(w_2, w_1)$).

The 4-cov model was initially designed for mismatch of duration in [4]. The SdSv challenge could take advantage of this model for the following purpose:

- as noticed in the evaluation plan, *" The enrollment data in Task 2 consists of one to several variable-length utterances. The net speech duration for each model is roughly 3 to 120 seconds "*. Each target speaker can be modeled by using a x-vector sample, with a mean of 7 observations per speaker while the test utterance to compare is unique and of short-duration (often less than 5 sec.).
- on the other hand, a non-negligible proportion of test segments are in English (non native as the speakers are Persian native).

Table 2 details the proportion of trials, depending on the size of the speaker enrollment sample and on the language of test. It is worth noting that in the trial dataset of SdSv, when the size of the enrollment sample is lower than 5, the utterances are of short duration (less than 5 seconds) in a main proportion. While for enrollment sample size upper than 5, the duration of utterances are spanning the interval 3 to 120 seconds indicated in the plan.

The 4-cov model allows to fit PLDA models specifically to each enrollment and test distribution. Table 3 shows the differ-

Table 2: *Percentages of trials in the evaluation trial dataset, depending on the target speaker model (how many enrollment segments are available ?) and on the test language.*

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

ent systems applied, depending on the trial. We apply the 4-cov model to each type of mismatch: (mean of samples of various size and duration) vs (one short duration utterance in Persian or English). The model vector is the length-normalized average of the speaker x-vector sample.

To deal with the issue of language mismatch, we apply, for learning the PLDA model specific to short duration test segments, a two-step domain adaptation method: first, x-vectors of a wide learning database, which is out of domain in terms of language, are extracted, by using the front-end configuration described above. As this neural network is re-fined to Persian language, it partially adapts the initial data to Persian. Then the parameters of in- and out-of domain PLDA models are adapted by using a weighted interpolation [5]. The resulting PLDA parameters feed the second side of a 4-covariance model, therefore specific to test utterances in English.

To better understanding, we detail one case of Table 3. Its 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 [6]).
- 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 submodels (i) and (ii) [5].

The language of the test segments is estimated by a speech detector, then the score of each trial is the one resulting from the corresponding model, according to the detected language of its test segment.

Taking benefit of the score normalization to enhance performance required 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.

As the score file of trials intertwines four scorings, the scores are calibrated by using development trial datasets specific to the four cases of Table 3, all based on DeepMine development data.

All the details about the methods described here and specifically designed for the SdSv challenge (DNN Persian-refinement, asymmetric and trial-dependent modelings, score normalization) will be presented in an upcoming article.

Table 3: *Datasets for trial-dependent model training. The 4-covariance model allows to customize the model to enrollment and test materials.*

trial:		4-covariance model	
enrollment #segs	test language	model 1 for enrollment	model 2 for test
< 5	Persian	3 vectors $L2$-average < 5 sec.	< 5 sec.
< 5	English	3 vectors $L2$-average < 5 sec.	< 5 sec. & English-dev
$\geqslant 5$	Persian	12 vectors $L2$-average $\geqslant 7.5$sec.	< 5 sec.
$\geqslant 5$	English	12 vectors $L2$-average $\geqslant 7.5$sec.	< 5 sec. & English-dev

4. Results on the progress subset

4.1. During challenge period (on the progress subset)

Table 4 shows results of the successive systems submitted on the progress set (30% of the trials). For better comparison, the baseline of SdSv organizers is recalled in first row of the Table. For S-normalization, we used the top 400 segments to compute the normalization parameters of each trial.

Table 4: *Results in terms of EER and DCF of the successive systems submitted, as reported by the organizers on the progress set (30% of the trials).*

	EER%	minDCF
SdSv baseline	10.67	0.432
System		
baseline	7.71	0.3550
+ domain adaptation	5.58	0.2510
+ 4cov-model	4.21	0.1981
+ specific S-norm	4.04	0.1771
+ trial-dependent models	3.55	0.1507
+ librispeech (DNN-training)	3.01	0.1307
+ final optimization	2.89	0.1264

The baseline system is a system trained only on VoxCeleb2 using a standard PLDA (with no domain adaptation). The baseline system provides 7.71% of EER (0.355 minDCF). The domain adaptation consists in adapting the pre-trained neural network model obtained with baseline system on Deepmine corpus. The domain adaptation methods allows 5.58% of EER (0.251 minDCF). The 4cov-model, S-norm and model/trial subsets methods provide 3.55% of EER (0.151 minDCF) and, finally, the addition of Librispeech corpus in training corpus for the pre-trained neural network model, followed by a final optimization of configuration parameters, achieves 2.89% of EER (minDCF 0.126).

4.2. Post-evaluation

To better assess the benefits of each of our contribution, Table 5 shows the post-evaluation results, with systems using our overall DNN training dataset (including the LibriSpeech corpus, belatedly added during the challenge phase). Efficiency of the different stages of our final system is clearly demonstrated, leading to a competitive speaker detection accuracy. Relevance of asymmetric 4-cov modeling is also highlighted, firstly for dealing with enrollment-test mismatch, secondly to fit model to trial-subset specificity.

Table 5: *Post-evaluation results of our different contributions on the full evaluation. Unlike Table 4, all the systems are based on the same DNN learning dataset.*

	EER%	minDCF
baseline	7.38	0.3682
+ language adaptation	4.42	0.1823
+ 4cov-model	3.28	0.1554
+ specific S-norm	3.15	0.1427
+ trial-dependent models	2.88	0.1261

5. Conclusions

The SdSv challenge allowed to test robustness of text-independent speaker recognition systems and to propose new approaches. The relevance of our novelties for this evaluation was confirmed by the results we obtained. The use of asymmetric modeling, taking into account the mismatch between model speaker for enrollment and utterance of test to compare, as well as between trial partitions, has proven its efficiency. Such situations often occur in real-life applications and the SdSv evaluation provided an opportunity to demonstrate the usefulness of these approaches.

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] 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.

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

[3] 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.

[4] 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

[5] 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.

[6] 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 *Proc. Odyssey 2018 The Speaker and Language Recognition Workshop*, 2018, pp. 386–392.