SdSV Challenge Technical Report: the GREAT system

Zhiyong Chen, Zongze Ren and Shugong Xu

Shanghai Institute for Advanced Communication and Data Science, Shanghai University, China

bicbrv, zongzeren, shugong@shu.edu.cn

Abstract

This paper gives a brief report on the systems developed for the SDSV20 challenge from team GREAT. There will be a paper to describe our exploration in this challenge with further details.

Index Terms: automatic speaker verification, domain adaptation

1. Introduction

The Short-duration Speaker Verification Challenge 2020(SDSV20)[1] aims to evaluate new technologies for text-dependent and text-independent speaker verification in a short duration scenario. SDSV20 Task 2 is a typical text-independent automatic speaker verification(TI-ASV) task where a domain mismatch problem exists. Participants need to train an ASV model with a sufficient amount of data from two English speech corpora which have different noise conditions, plus a tiny labeled corpus of Persian language which is in a similar condition that will be met in the testing stage.

We define the domain-mismatched problem encountered in SDSV20 as a supervised few-shot multi-domain adaptation ASV(SFDA-ASV) problem. Since we found this problem was less well addressed in the previous studies, we have tried to develop a domain adaptation method that works well in this scenario during the competition period.

Partly inspired by the end-to-end architecture of[2], we have successfully developed our method to solve the imbalanced multi-domain adaptation problems including SFDA-ASV, which is concise and consistently effective. Our major models went through the real-condition leaderboard evaluation of SDSV20, and the proposed methods work consistently better comparing with the baseline.

2. Supervised Few-shot Adversarial Multi-domain Adaptation

Our proposed adversarial learning architecture is shown in Figure 1. Which we use X-vector with large margin discriminative loss as the end-to-end ASV backbone architecture for its superior performance. We use AM-Softmax Cross-Entropy as the speaker loss in default. Standard X-vector[3] is used. The core idea of domain adaptation method for solving the SFDA-ASV problem is to design a robust and effective adversarial loss and explicitly consider the extreme domain-data imbalance problem. We manage to achieve this goal by using our proposed reweighted-balance adversarial domain adaptation(RW-B-ALDA).

3. Datasets and systems

The SDSV20 designated Voxceleb1&2[4], Librispeech-clean[5] and SDSV20 development-set(devset) were used to conduct all our experiments. Voxceleb1&2 contains audio clips from Youtube, which is noisy and mainly in English. Librispeech-clean is a clean corpus of read English speech. SDSV20 devset contains selected parts of DeepMine[6], which is a clean Persian corpus.

3.1. Evaluation data

Four trial-sets were used for evaluation. (1)Vox: The Voxceleb1 testing-set, (2)Libri: Librispeech-dev-100 section, (3)SDSV-DEV: 100 speakers from SDSV20 devset and (4)SDSV-LB: the SDSV20 leaderboard post-evaluation designated set. Default trial lists were used for Vox and SDSV-LB, the other two trial lists were composed by ourselves similarly to conduct our preliminary experiments.

3.2. Training data

All speakers(7323 speaks) in Voxcelev1&2 training-set, the maximum speakers from SDSV20 devset we can utilize(488 speaks) and all speakers from Librispeech-clean(1172 speaks) were used to run the evaluation. Every speaker was equally sampled 1500 times. Background noises and reverberations were used to augment original speeches. 30-dimensions MFCC acoustics feature was extracted as the input to the deep neural network. Note all data were preprocessed similarly as Kaldi[7].

3.3. Training and testing details

We used SGD as optimizer for all experiments. For all preliminary studies, step decay learning rate schedule from 0.01 to 0.0001 was used. Cosine annealing learning rate schedule[8] was used for all experiments in comprehensive evaluation. Cosine similarity scoring was used for all our systems. SDSV20 designated evaluation metrics[1] were adopted. Systems were fully implemented with Python & Pytorch including the preprocessing and backend processes.

4. Evaluation

Table 1 shows the results of our evaluation results on both our own experiments and the leaderboard evaluation. Note the results was recorded from the SDSV20 Post-Evaluation stage. We tested different data combination to train our model, including single domain, dual domains and triple domains scenarios. Consistent improvement for the target domain on both the devset and the leaderboard was observed in Exp 1-5 by adopting RW-B-ALDA. VoxOnly means using only Vox data and Mix-data represents intuitively mix labeled data from different domains together. Comparing Exp 2-5, only limited improvement can be achieved on the target domain by simply mixing more data from another new domain, given that Exp 4 performs even worse than Exp 3 on the leaderboard after huge amounts of data added. Only by adopting RW-B-ALDA while adding various domain data did we see a significant gain on the leaderboard, indicating speaker knowledge from different domains had been

Table 1: *Multi-domain and comprehensive evaluation results*

id	system	speaker numbers (Vox:Libri:SDSV)	Vox EER%	Vox MINC	Libri EER%	Libri MINC	SDSV-DEV(Target) EER%	SDSV-DEV(Target) MINC	SDSV-LB(Target) EER%	SDSV-LB(Target) MINC
1	VoxOnly	7323:0:0	2.29	0.23	2.51	0.16	4.80	0.39	6.18	0.256
2	MixData	7323:0:488	2.36	0.23	2.28	0.14	3.09	0.29	5.21	0.207
3	RW-B-ALDA	7323:0:488	2.33	0.24	**2.20**	**0.13**	**3.01**	**0.27**	**4.58**	**0.191**
4	MixData	7323:1172:488	2.33	0.23	**1.99**	0.10	3.03	0.27	4.76	0.192
5	RW-B-ALDA	7323:1172:488	**2.32**	0.24	**1.99**	0.11	**2.90**	**0.26**	**4.19**	**0.177**

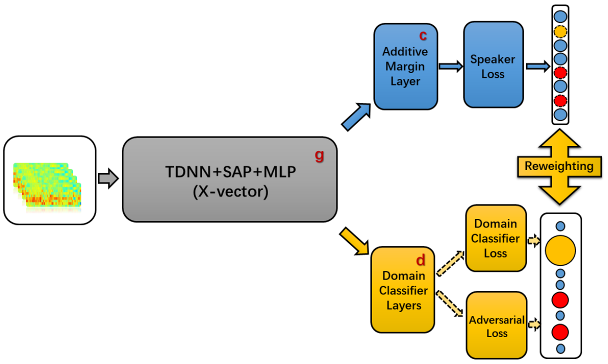

Figure 1: *Adversarial learning architecture.*

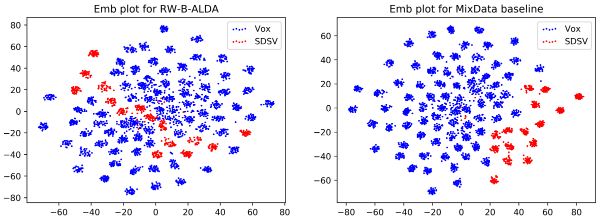

Figure 2: *Embedding plot comparison.*

mutually well transferred and robust speaker model had been learned.

By plotting the speaker embedding space with T-SNE in Figure 2 gives a strong evidence that our proposed method is effective for the SFDA-ASV problem. 90 speakers in the preliminary training-set from the two domains were used to extract their embeddings. Without RW-B-ALDA, the model overfitted to the implicit domain factors in the dataset while target domain embeddings were restricted to a corner of the whole embedding space. This unexpected effect impeded the common speaker knowledge to transfer between domains and caused degenerated generalization power of our model to perform well on multiple few-shot target domains. This problem was mitigated by using RW-B-ALDA, as a more domain invariant distribution was observed.

5. Conclusions

The core findings are that effective adversarial loss design and properly reweighting imbalanced multi-domain data are both critical to achieving the goal. Results both from our own experiments and the SDSV20 evaluation in real multi-domain conditions have proven the effectiveness of our proposed method. Our formal paper is ready to release to address this problem and our solution in detail.

6. References

[1] H. Zeinali, K. A. Lee, J. Alam, and L. Burget, "Short-duration speaker verification (sdsv) challenge 2020: the challenge evaluation plan," *arXiv preprint arXiv:1912.06311*, 2019.

[2] J. Zhou, T. Jiang, L. Li, Q. Hong, Z. Wang, and B. Xia, "Training multi-task adversarial network for extracting noise-robust speaker embedding," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6196–6200.

[3] 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.

[4] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: a large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017.

[5] 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)*. IEEE, 2015, pp. 5206–5210.

[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 *Odyssey*, 2018, pp. 386–392.

[7] D. Snyder, "Kaldi voxceleb recipe," https://github.com/kaldi-asr/kaldi/tree/master/egs/voxceleb/v2.

[8] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.