DA-IICT Submission for SdSV Challenge

Divyesh G. Rajpura, Madhu R. Kamble, Hemant A. Patil

Speech Research Lab
Dhirubhai Ambani Institute of Information and Communication Technology (DA-IICT)
Gandhinagar, Gujarat, India.

{divyesh_rajpura, madhu_kamble, hemant_patil} @daiict.ac.in

Abstract

Automatic Speaker Verification (ASV) has gained much attention by speech community in recent years. With the recent technological advancements, such as voice assistants and smart devices, effective voice biometric systems have become an essential requirement. In the past few years, Deep Neural Network (DNN)-based approaches have shown significant improvement in the different task of speech processing, including ASV. However, short-duration and far-field speaker verification is still a challenging task. In this regard, Short-duration Speaker Verification (SdSV) challenge is organized. In this study, we make use of Generalized end-to-end (GE2E) loss to learn speaker discriminative embeddings, namely, *d*-vector. In contrast to triplet loss, GE2E does not require sample selection, which makes it computationally efficient. For the text-independent task of SdSV challenge, GE2E-based *d*-vector gave 22% of EER and minDCF of 0.8929.

Index Terms: Short-duration Speaker Verification, Deep Neural Network, *d*-vector.

1. Introduction

Automatic Speaker Verification (ASV) is the task of verifying the claimed identity of the speaker. Recent technological advances, such as smart devices and voice assistants, have introduced many research challenges in ASV. In the past few years, ASV is considered as an embedding learning task, where the aim is to map the utterance from higher-dimensional space to a low-dimensional space. Embedding learning is very crucial as it should learn embeddings such that same-speaker utterances are nearer and between-speaker utterances are far from each other in embedding space. Moreover, it should also be able to generate distinct embedding for unknown speakers.

Factor Analysis (FA)-based identity vector (i.e., i-vector)[1], and Deep Neural Network (DNN)-based x-vector [2] is state-of-the-art in ASV. However, these approaches require separate back-end for scoring as it is not directly optimized for scoring in embedding space. One of the major scoring backed used in the above approaches is Probabilistic Linear Discriminant Analysis (PLDA). Therefore, the success of these approaches is highly dependent on the scoring metric. Moreover, it is still difficult for PLDA to handle short utterances.

To address the issue of separate back-end for scoring, Deep Metric Learning (DML) has been widely used, which aims to optimize embeddings with the scoring metric in embedding space during embedding learning. These approaches allow us to use classical scoring methods, such as cosine similarity. Recently, there has been much work done to incorporate the DML with embedding learning, such as *d*-vector [3] and deep speaker [4].

We have used DNN-based embedding, specifically, *d*-vector to solve the challenging problem of Text-Independent Speaker Verification as part of Short-duration Speaker Verification Challenge during INTERSPEECH 2020 [5].

2. The *d*-vector Approach

The *d*-vector was originally, proposed in [3]. The main contribution reported by study in [3] is Generalized End-to-End Loss (GE2E), which makes it possible to jointly learn the embedding with DML to make end-to-end system without separate back-end for scoring. The more detail about GE2E is given in Section 3. Let us assume we have P number of speakers $s_1, s_2, ..., s_P$ and C number of utterances per speaker $u_{i1}, u_{i2}, ..., u_{iC}$ where $i = 1, 2, ..., P$ in training set. For each training step, we select M number of speakers and N number of utterances per speaker to form a batch, where $M < P$ and $N < C$. We fed these batches to Bi-directional Long Short Term Memory (Bi-LSTM) network. Bi-LSTM is followed by a Fully Connected (FC) layer, which is used to apply an additional transformation. We use the last hidden layer of LSTM as input to the FC layer as it contains essential information from all previous sequences. Finally, the embedding vector, i.e., *d*-vector, is defined as L_2-normalization of network output.

3. Generalized End-to-End Loss (GE2E)

GE2E includes DML with embedding learning to alleviate the need of separate backend for scoring. It is based on a very similar concept to Triplet Loss. In contrast to triplet loss, GE2E does not require the sampling of positive and negative samples during training, which makes it more computationally efficient. Given a batch consist of M number of speakers, and N number of utterances per speaker, step-by-step process for calculating GE2E loss for utterance u_{ji} is as follows:

- Compute model c_j of speaker j by calculating mean of embeddings of utterance u_{ji} of speaker, j, i.e.,

$$c_j = \frac{1}{N} \sum_{n=1}^{N} u_{jn}. \tag{1}$$

  As suggested in [3], excluding the utterance u_{ji} when computing model of c_j of speaker j for utterance i can help improve the result.

$$c_j^{-i} = \frac{1}{N-1} \sum_{n=1, n\neq i}^{N} u_{jn}. \tag{2}$$

- Compute the similarity matrix $S_{ji,k}$ by calculating similarity of each utterance, u_{ji} with centroid of each speaker c_k in batch.

$$S_{ji,k} = \begin{cases} w * (cos(u_{ji}, c_k^{-i}) + b, & \text{if } k = j; \\ w * (cos(u_{ji}, c_k) + b, & \text{otherwise,} \end{cases} \quad (3)$$

where $1 \leq i \leq N, 1 \leq j, k \leq M$. Here, w and b are learnable parameters.

- Finally, we can use softmax loss on $S_{ji,k}$ to compute loss for each utterance u_{ji}

$$L(u_{ji}) = \log \frac{e^{u_{ji,j}}}{\sum_{k=1}^{M} e^{u_{ji,k}}}. \quad (4)$$

4. Experimental Setup

4.1. Dataset and Feature Extaction

We have used the Voxceleb1 [6] to train an embedding model. It consists of 1251 number of speakers, out of which 1201 is used for training, and 51 is used for validation. In our experiments, we use *20*-dimensional Mel-Frequency Cepstral Coefficients (MFCC) features with window length 32ms and overlap of 75%. Delta and acceleration are appended to create 60 dimension feature vectors.

4.2. Architecture Details

We have used similar architecture as in [3]. It consists of 3 Bi-directional Long Short Term Memory (BLSTM) layers with 768 hidden nodes, followed by one Fully Connected (FC) layer with 256 output nodes. The last hidden layer of BLSTM is used as input to the FC layer. The embeddings are L_2 normalized, and the dimension is set to 256. For training with GE2E loss, batches are created such that each batch consists of 8 speakers, and 8 utterances per speaker. We use Adam optimizer with an initial learning rate of 0.001. We applied the learning rate scheduler, which linearly reduces the learning rate by 0.25 at each epoch. We trained our model for 16 epochs. As suggested in [3], we keep the initial value of w to 10 and b to -5. The model was implemented in PyTorch [7] and trained using a single GeForce GTX Titan X GPU.

5. Experimental Results

We performed experiments on Task-2 (i.e., text-independent speaker verification). The result on evaluation data [8, 9] provided by challenge organizers is shown in Table 1. The results are not satisfactory. The possible reasons could be as follows (1) Language mismatch could be major reason as our training data contains only English utterances, whereas evaluation data contains both English and Persian utterances, (2) The BLSTM are known for its capability of memorizing the information over time, however, longer sequence length may discard some useful information. We enforce sequence length for batch to be number of frames in shortest utterance present in batch, which is in range of 400 to 600 frames in most of the cases, whereas in [3], it is in range of 140 to 160, (3) During inference, we give complete utterance as input to the network, which again may not be effective approach for utterances having longer sequence length.

Table 1: *Results on Task-2 Text-Independent Speaker Verification*

	EER(%)	minDCF
Baseline	10.67	0.4324
d-vector	22.61	0.8929

6. Summary & Conclusion

In this study, we evaluate GE2E loss-based embedding, namely, *d*-vector for the text-independent task of the SdSV challenge. GE2E loss uses a similar underlying concept as triplet loss, however, GE2E does not require sample selection, which makes it more computationally efficient. We are not able to achieve satisfactory results, and a key reason for that could be language mismatch as we train our model on the Voxceleb1 dataset, which contains only English utterances. In contrast, evaluation dataset contains two different languages, English and Persian. In future, the other effective DML, such as Ring loss, and CosFace can be adopted to improve performance further.

7. References

[1] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011.

[2] 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)*, Calgary, Alberta, Canada, Apr 15-20, 2018, pp. 5329–5333.

[3] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Calgary, Alberta, Canada, Apr 15-20, 2018, pp. 4879–4883.

[4] C. Li, X. Ma, B. Jiang, X. Li, X. Zhang, X. Liu, Y. Cao, A. Kannan, and Z. Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *CoRR*, vol. abs/1705.02304, 2017, {Last Accessed: 14/04/2020}. [Online]. Available: http://arxiv.org/abs/1705.02304

[5] 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.

[6] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A large-scale speaker identification dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[7] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 8026–8037. [Online]. Available: http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[8] 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*, Les Sables d'Olonne, France, Jun 26-29, 2018, pp. 386–392.

[9] 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 (ASRU) Workshop*, Sentosa, Singapore, Dec 14-18, 2019.