An end-to-end deep learning architecture for speaker identification

Sara SEKKATE, Mohammed KHALIL, and Abdellah ADIB
Team Networks, Telecoms & Multimedia
LIM@II-FSTM, B.P. 146
Mohammedia 20650, Morocco
sarasekkate@gmail.com, medkhalil87@gmail.com, adib@fstm.ac.ma

Abstract

In this paper, a complete end-to-end speaker identification system is presented. The use of ResCNN architecture as well as triplet-loss training based on cosine similarity are investigated. Experiments on two different databases demonstrate the effectiveness of the deep learning speaker identification based architecture.

1 Introduction

Speaker Identification (SI) is the task of determining who is talking from a set of known speakers. State-of-the-art SI systems are based on i-vectors [1] with a probabilistic linear discriminant (PLDA) backend for classification. Due to its great success in speech processing, Deep Learning (DL) has gained extensive interest of its application in recent years. Some of the proposed applications make use of DL techniques in the i-vector extraction process [2, 3] or applied on i-vectors as a backend [4, 5]. This paper focuses on another possible use of DL which consists of representing the speaker characteristics with a single low dimensional vector, rather than the baseline i-vector algorithm. These vectors are often referred to as speaker embeddings. We aim to investigate an end-to-end SI system based on DL that has drawn more attention recently [6]. First, we derive log-mel filterbank features that are input to a residual Convolutional Neural Network (ResCNN). For the end-to-end objective training function, triplet-loss training is used. To measure the similarity within trials, cosine distance is utilized on speaker embeddings. Finally, the end-to-end system is evaluated on SI task using two different datasets. The rest of this paper is organized as follows: Section 2 presents the end-to-end SI system. The performance of the presented system is evaluated, and the results are discussed in Section 3. Finally, Section 4 gives a conclusion of the paper.

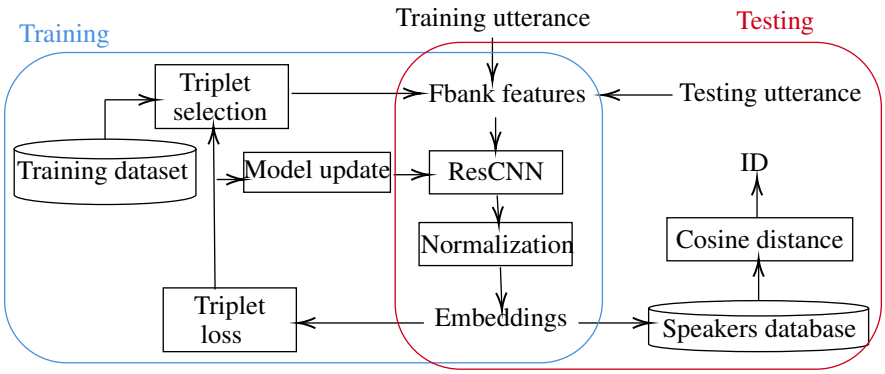

Figure 1: The structure of the end-to-end SI system

2 The system description

Figure 1 depicts the structure of the end-to-end SI system. It consists of extracting log-mel filterbank features and a deep ResCNN architecture followed by length normalization. Finally, the triplet loss layer operates on pairs of embeddings, by maximizing the cosine similarities of embedding pairs from the same speaker, and minimizing those from different speakers.

Layer	struct	stride	dimension	no. param
conv64-s	5×5,64	2×2	2048	1.6K
res64	$[(3 \times 3,64) \times 2] \times 3$	1×1	2048	36.9K $\times$ 6
conv128-s	5×5,128	2×2	2048	204K
res128	$[(3 \times 3,128) \times 2] \times 3$	1×1	2048	147K $\times$ 6
conv256-s	5×5,256	2×2	2048	819K
res256	$[(3 \times 3,256) \times 2] \times 3$	1×1	2048	590K $\times$ 6
conv512-s	5×5,512	2×2	2048	3.2M
res512	$[(3 \times 3,512) \times 2] \times 3$	1×1	2048	2.3M
Average	-	-	2048	0
Affine		-	512	1M
ln	-	-	512	0
triplet	-	-	512	0
Total				24M

Table 1: Architecture of the ResCNN model. “Average” and "ln" stand for the temporal pooling and the length normalization layers, respectively.

3 Experimental setup and results

In order to investigate the performance of the end-to-end system, speaker identification experiments were carried out on two datasets: TIMIT [7] and Librispeech [8]. In each corpus, speech data was divided into two sets. For TIMIT database, 430 speakers were used for training and 200 speakers were reserved for testing. For Librispeech, the train-clean-100 subset (251 speakers) was used for training while test-clean (40 speakers) subset was used for testing. Subsequently, Voice Activity Detection (VAD) was performed on all speech samples to remove silent segments [9]. 64-dimensional log-mel filterbank features were then derived from each utterance with a frame length of 25ms. The output is further input to the ResCNN model described in Table 1. The obtained results on the two databases are shown in Table 2. Both corpora yield similar results, we achieve an accuracy of 98% corresponding to 18 misidentified speakers.

Dataset	Accuracy
TIMIT	98.5%
LibriSpeech	98.6%

Table 2: Accuracy of the presented end-to-end SI system

4 Conclusion

In this paper, we have investigated whether it is feasible to identify speakers using an end-to-end deep learning architecture which directly maps the utterance to an identification result. Experimental results effectively showed that the presented approach demonstrates a promising new direction for SI applications by achieving an accuracy of 98% for both TIMIT and LibriSpeech databases.

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*, 19(4):788–798, May 2011.

[2] Y. Lei, N. Scheffer, L. Ferrer, and M. A McLaren. novel scheme for speaker recognition using a phonetically-aware deep neural network. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2014.

[3] Y. Liu, Y. Qian, N. Chen, T. Fu, Y. Zhang, and K. Yu. Deep feature for text-dependent speaker verification. *Speech Communication*, 73:1 – 13, 2015.

[4] J. Villalba, N. Brümmer, and N. Dehak. Tied variational autoencoder backends for i-vector speaker recognition. In *INTERSPEECH*, 2017.

[5] O. Kudashev, S. Novoselov, T. Pekhovsky, K. Simonchik, and G. Lavrentyeva. Usage of dnn in speaker recognition: Advantages and problems. In *Advances in Neural Networks – ISNN 2016*, pages 82–91, Cham, 2016. Springer International Publishing.

[6] R. Ji, X. Cai, and X. Bo. An end-to-end text-independent speaker identification system on short utterances. In *INTERSPEECH*, pages 3628–3632, 09 2018.

[7] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. Darpa timit acoustic phonetic continuous speech corpus cdrom. 1993.

[8] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, South Brisbane, Queensland, Australia*, pages 5206–5210, April 2015.

[9] J. Ramirez, J. M. Gorriz, and J. C. Segura. Voice activity detection. fundamentals and speech recognition system robustness. In Michael Grimm and Kristian Kroschel, editors, *Robust Speech*, chapter 1. IntechOpen, Rijeka, 2007.