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

The state-of-the-art in automatic speaker recognition is undergoing a shift from using a i-vector/PLDA framework towards Deep Learning (DL) approaches. In this work, we investigate the efficacy of DL on speaker identification and especially Convolutional neural Networks (CNN). We report modest proof-of-concept experiments designed to evaluate potential. We carry out our experiments on a subset of Librispeech database. The obtained results suggest that the proposed approach merits further investigation. Avenues for future research are described and have potential to deliver improvements in the achieved performance.

Keywords: Deep Learning, CNN, end-to-end, speaker identification

1. Introduction

Deep learning (DL) has shown remarkable success in numerous speech tasks, including speech [1] and speaker recognition [2, 3]. In speaker recognition, DL techniques have been explored in different ways: for feature extraction [4], as an alternative to i-vectors within a probabilistic linear discriminant analysis (PLDA) framework [5], for the learning of posteriors in a joint factor analysis framework [6] or for PLDA backend scoring [7]. A common characteristic to all of these works is the use of DL techniques as a means of replacing a specific element of a more complex toolchain. In This work DL is applied in a so-called end-to-end approach using Convolutional Neural Networks (CNN). The motivation to use CNN is inspired by the recent successes of CNN in many computer vision applications. Accordingly, the objective of the present work is not to outperform the existing state of the art, but more specifically to investigate the longer term potential of the idea.
Section 2 introduces the proposed architecture and its application to speaker identification. Experiments and results are described in Section 3. Conclusions are presented in Section 4.

2. Approach

CNNs are exceptionally good at capturing high level features in spatial domain and have demonstrated unparalleled success in computer vision related tasks. One natural advantage of using CNN is that it's invariant against translations of the variations in frequencies, which are common observed across speaker with different pitch due to their age or gender.

The proposed end-to-end CNN-based speaker identification system is depicted in Figure 1. It consists of three convolutional layers, where the second and third ones have max pooling. After the convolutions, two densely connected layers were added with a a final one with the softmax function to obtain the output. The convolution layers as well as the first fully connected one were normalized using Batch Normalization (BN). ReLU was used for all activation functions. Training used the ADAM optimizer [9].

To fully take advantage of deep learning, it would be natural to directly feed the network with the lowest possible signal representation (e.g., raw audio samples), avoiding any kind of pre-computed intermediate representations. However, for our proposed end-to-end model, raw audios were converted to Mel Frequency Cepstral Coefficients (MFCC) in order to speed up the training process. In this context, 13-dimensional MFCCs were then derived from each utterance with a frame length of 25ms. First and second order derivatives were appended to static MFCCs giving features of 39 coefficients. Then, the features are input to the CNN model.

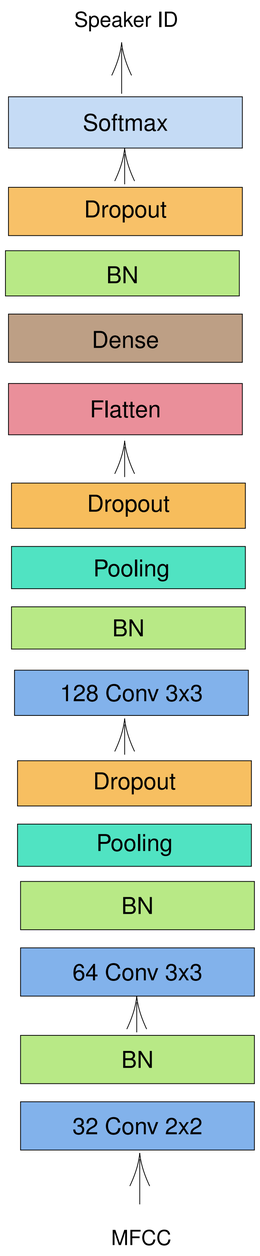

Figure 1: *The proposed CNN architecture*

3. Experiments

This section describes experiments which aim to test the potential of the end-to-end speaker identification system described in section 2.

3.1 Data

In order to investigate the performance of the end-to-end system, speaker identification experiments were carried out on the "train-clean-100" subset (251 speakers) of Librispeech database [8]. The dataset was further split into two sets: a training set containing 80% of the available data of each speaker and a testing set that comprises the remaining 20% of the data. In our experiments, no validation set was used.

3.2 Speaker Identification experiments

The input batch size was set to 32 and the model was trained for up to 50 epochs. The early stopping criteria was used with the final model being the one with low validation loss. Fig. 2 shows the results of both the training and testing sets.

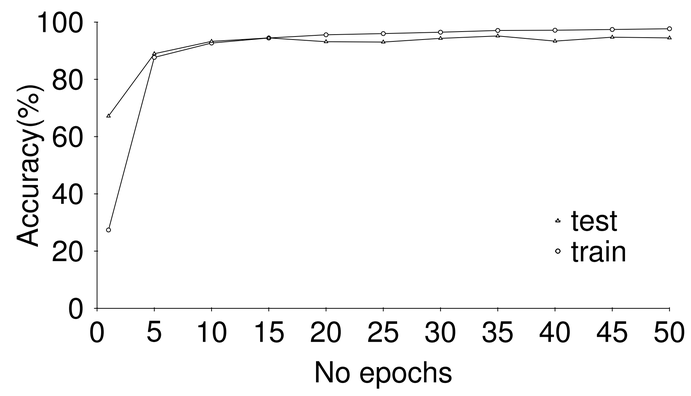

Figure 2: *251-person Speaker Identification. Shown are the training and validation sets accuracy.*

4. Conclusion

In this work, 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 94% for a subset of 251 speakers from Librispeech database.
These findings suggest that the end-to-end approach merits further attention and experimentation with larger datasets. Future works will investigate the use of raw audio signals as input to the network rather than hand-crafted features such as MFCCs. Furthermore, we would like to extend the model presented here to deal with noisy recordings also. We also note that the networks may be further optimized, for example, by exploring triplet-loss training.

References

[1] D. Yu and L. Deng. Automatic Speech Recognition - A Deep Learning Approach. *Springer*, 2015.

[2] M. McLaren, Y. Lei, and L. Ferrer. Advances in deep neural network approaches to speaker recognition. *ICASSP, 2015*.

[3] F. Richardson, D. Reynolds, and N. Dehak. Deep neural network approaches to speaker and language recognition. *IEEE Signal Processing Letters, vol. 22, no. 10, pp. 1671–1675, 2015.*

[4] L. Li, Y. Chen, Y. Shi, Z. Tang, and D. Wang. Deep speaker feature learning for text-independent speaker verification. *INTERSPEECH, 2017, pp. 1542-1546.*

[5] D. Snyder, D. Garcia-Romero, D. Povey, and S. Khudanpur. Deep neural network embeddings for text-independent speaker verification. *INTERSPEECH, 2017, pp. 999-1003.*

[6] S. Dey, S. Madikeri, M. Ferras, and P. Motlicek. Deep neural network based posteriors for text-dependent speaker verification. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5050-5054.*

[7] H.-S. Lee, Y.-D. Lu, C.-C. Hsu, Y. Tsao, H.-M. Wang., and S.-K. Jeng. Discriminative autoencorders for speaker verification. *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on, 2017.*

[8] P. Vassil, G. Chen, D. Povey, and S. Khudanpur. Librispeech: an ASR corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5206-5210. IEEE, 2015.*

[9] D. P. Kingma, J. L Ba. Adam: a method for stochastic optimization. *ICLR, 2015.*