MagmaNet: Ensemble of 1D Convolutional Deep Neural Networks for Speaker Recognition in Hungarian

Attila Gróf[1], Annamária Kovács[2,3], Anna Moró[1], Miklós Gábriel Tulics[2] and Máté Ákos Tündik[2]

[1]Dept. of Telecommunications and Media Informatics, Budapest University of Technology and Economics, [2]Budapest University of Technology and Economics, [3]Research Centre for Natural Sciences, Hungarian Academy of Sciences, Hungary

Speaker recognition is one of the classical and widely documented tasks in speech technology. In short it can be said that speaker recognition is the process of automatically recognizing the person behind the voice, on the basis of information obtained from his or her speech.

Building speaker recognition systems has a large literature, especially from GMM/HMM era, which were the first pioneers in this field, but Neural Network-based solutions are gaining ground in this area as well.

Zaki and his colleagues used cascade neural networks [1] for speaker classification, others, such as Zakariya et al. used i-vectors and ANN for text-independent speaker classification [2].To achieve higher performance using different ANNs committee neural networks were proposed by Narender et al. [9]: several ANNs were trained to achieve excellent recognition, but in the final step, only the five best performing networks were fed into the (final) committee network. This way, the final decision was determined based on the majority voting of the member networks.

The authors of [11] used a deep 2D-Convolutional Neural Network (CNN), applied to the raw spectrograms. The main idea was to create convolutions that sample raw information across both frequency and time makes the spectrograms suitable for CNN analysis. The input was an image of size 513 by 107, a spectrogram of a 20ms segment of an utterance. They used a pre-trained network, AlexNet for the classification. They reached 17,1% Equal Error Rate (EER). An i-vector based system was also trained and tested on the same data, resulting in 39,7% EER.

The state-of-the-art approaches for text-independent speaker recognition are using Joint Factor Analysis (JFA) or i-vector based modeling. JFA is a powerful and widely used technique for compensating the variability caused by different channels and sessions. The total variability i-vector modeling has gained significant attention in speaker recognition due to its excellent performance, low complexity and resulting small model sizes. The authors of [10] reached 2,53% EER.

Although several ANN/DNN implementation exist in the speaker recognition field, one-dimensional convolution was not applied specifically to this problem. However, not only the architecture of the neural network is important, it is substantial to choose the proper speech features for this task as well. Much depends on the appropriate feature selection. There are well-established speech features for these tasks: Melfrequency cepstral coefficients (MFCCs) [3], linear-prediction coefficients (LPC) [4], mean Hilbert envelopes [5], and also hybrid feature-sets [6] [7].

In this research we showed that using a low number of features is sufficient to train a DNN-based system to perform this classification task with a high performance. For this reason, we used a small database containing a total of 161 sound recordings from 11 native Hungarian speakers, reading “The NorthWind and the Sun” folk-tale. We used five speakers to differentiate them, the other speakers were used as an impostor model.

MFCCs, LPCs and Linear Predictive Cepstral Coefficients (LPCC) were extracted in 30 ms windows with 10 ms overlaps, these features were used as input vectors. We implemented four neural network models: a Multi-layer Perceptron (MLP), a 1D ConvNet and a 1D Dilated ConvNet with an LSTM Layer, and an Ensemble one called ’MagmaNet’1. Our 1D Convolutional - based Neural Network Architectures can be seen on Figure 1.

The impact on the classification accuracy of the acoustic parameters was examined. In case of MLP the best classification results was obtained using only MFCC features (47%). As one of the simplest DNN architectures, the MLP classification accuracy can be considered as a baseline. Convolutional Networks yield better accuracies when MFCC and LPCC features were combined together. Both networks reached an accuracy of 76%. Manual and automatic hyperparameter optimization was performed for each network.

In hope of getting further improvement in classification accuracy, in our first attempt the two convolutional networks were merged and were simultaneously trained. We performed the same processing steps on the Ensemble model like on the other networks. First, we selected the best feature set

(MFCC) for classification, then we performed manual hyperparameter optimization obtaining accuracy results of 74% on the validation set and 68% on the test set. Our second attempt was using pre-trained 1D Convolution models, storing their weights, and applying them in a transfer learning approach.

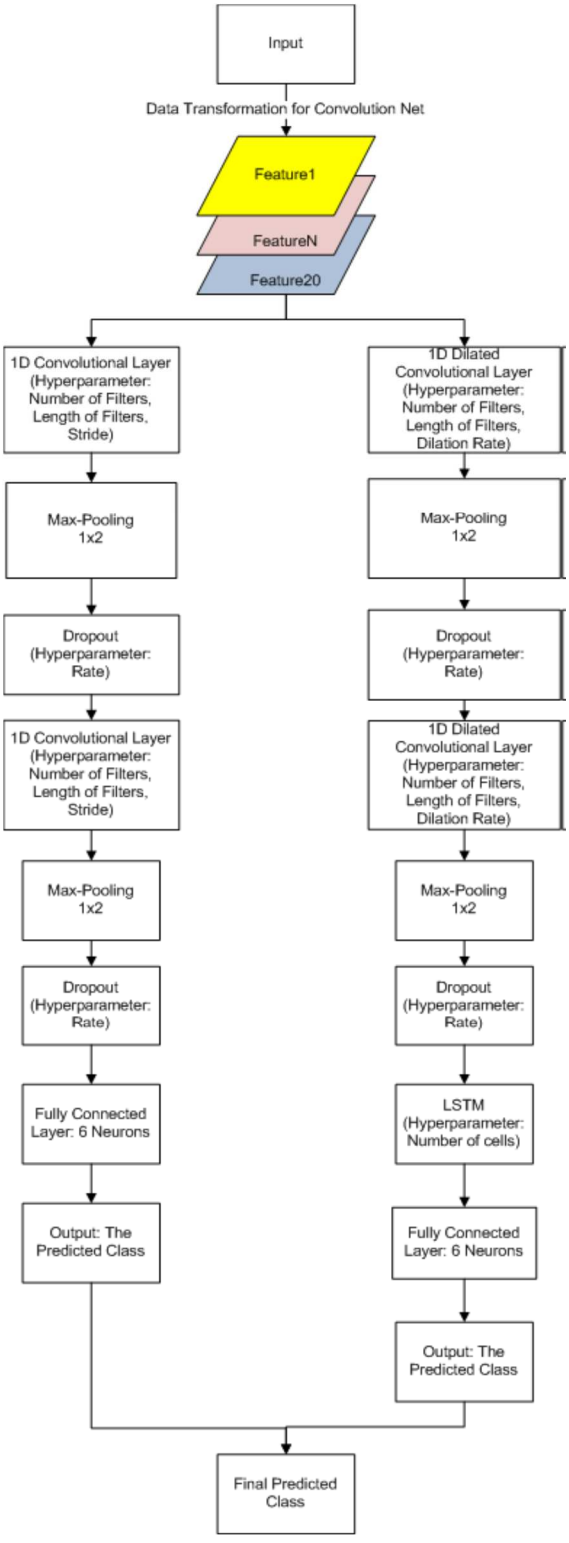

Figure 1. The MagmaNet components: 1. 1D ConvNet:on the left, 2. 1D Dilated
ConvNet with LSTM: on the right. Ensemble Model: together

This 'MagmaNet' Ensemble model reached 78% accuracy on the test set, which is a quite good achievement for this classification task, on this limited database. We can show comparable results in accuracy with [12], who investigated speaker identification in TV Broadcast data. Although the data is different, the task is similar.

We have some ideas for future improvement: we would like to examine various number of features, so take more effort to feature selection. We are also aware of the limitation of our research, thus we are going to switch to a public database containing more recordings. All in all, our results confirmed the plausibility of using 1D Convolution based DNN as a means for implementing a valuable speaker recognition solution.

References

[1] M. Zaki, A. Ghalwash, A. Elkouny, 1996. Speaker recognition system using a cascade neural network, Int. J. Neural Syst. 7. 203–212.

[2] Zakariya Qawaqneh, Arafat Abu Mallouh, Buket D. Barkana 2017. Deep neural network framework and transformed MFCCs for speaker's age and gender classification, Knowledge-Based Systems, Volume 115.

[3] P. Mermelstein 1976. "Distance measures for speech recognition, psychological and instrumental," in Pattern Recognition and Artificial Intelligence, C. H. Chen, Ed., pp. 374-388. Academic, New York.

[4] S.B. Davis, and P. Mermelstein, 1980. "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences," in IEEE Transactions on Acoustics, Speech, and Signal Processing, 28(4), 357–366.

[5] Seyed Omid Sadjadi, John H.L. Hansen 2015. Mean Hilbert envelope coefficients (MHEC) for robust speaker and language identification, Speech Communication, Volume 72. 138–148.

[6] Leandro D. Vignolo, S.R. Mahadeva Prasanna, Samarendra Dandapat, H. Leonardo Rufiner, Diego H. Milone 2016. Feature optimisation for stress recognition in speech, Pattern Recognition Letters, Volume 84. 1–7.

[7] Ji-Won Cho, Hyung-Min Park 2016. Independent vector analysis followed by HMM-based feature enhancement for robust speech recognition, Signal Processing, Volume 120. 200–208.

[8] S. K. Singh, Supervisor: Prof P. C. Pandey: Features and Techniques for speaker recognition

[9] Reddy N.P., Buch, O., 2003. Speaker verification using committee neural networks, Computer Meth. Programming in Biomed., Vol. 72. 109–115.

[10] Ming Li, Andreas Tsiartas, Maarten Van Segbroeck and Shrikanth S. Narayanan 2014.

Speaker verification using simplified and supervised i-vector modeling.

[11] Lior Uzan, Lior Wolf: I Know That Voice: Identifying the Voice Actor Behind the Voice

[12] Mateusz Budnik, Laurent Besacier, Ali Khodabakhsh, Cenk Demiroglu. 2016. Deep complementary features for speaker identification in TV broadcast data. Odyssey Workshop 2016, Jun 2016, Bilbao, Spain. Odyssey.