Neural Network Speaker Descriptor in Speaker Diarization of Telephone Speech

Zbyněk Zajíc[1], Jan Zelinka[1,2] and Luděk Müller[1,2]

University of West Bohemia, Faculty of Applied Sciences,
[1]NTIS - New Technologies for the Information Society and [2]Dept. of Cybernetics,
Univerzitní 8, 306 14 Plzeň, Czech Republic, www.zcu.cz
zzajic@ntis.zcu.cz, zelinka@kky.zcu.cz, muller@ntis.zcu.cz

Abstract. In this paper, we have been investigating an approach to a speaker representation for a diarization system that clusters short telephone conversation segments (produced by the same speaker). The proposed approach applies a neural-network-based descriptor that replaces a usual i-vector descriptor in the state-of-the-art diarization systems. The comparison of these two techniques was done on the English part of the CallHome corpus. The final results indicate the superiority of the i-vector's approach although our proposed descriptor brings an additive information. Thus, the combined descriptor represents a speaker in a segment for diarization purpose with lower diarization error (almost 20 % relative improvement compared with only i-vector application).

Keywords: Neural network, Speaker diarization, i-Vector

1 Introduction

For a majority of speech processing tasks is convenient to work with a signal containing only one voice. In the real world, this condition is very difficult to fulfill. So, the Speaker Diarization (SD) system is necessary to determining "Who spoke when" without any prior information about the number of speakers and their identities. The process of diarization divides an input signal and merges these segments into clusters corresponding to individual speakers [23, 24]. Another approach (not used in this paper) combines the segmentation and the clustering step [6, 28].

The main problem in SD is how to describe the segments of the signal for subsequent clustering. Ideally, each segment consists of only one speaker. In recent years, i-vector approach has gained popularity in the Speaker Verification (SV) task [4, 8] as well as in SD [9, 35].

For many years, Neural Networks (NNs) have been successfully used in the field of speech recognition generally [7], and nowadays NNs are used extensively also in SD systems: in the segmentation task [13, 15] or in the clustering process [14, 20]. In paper [25], NNs are adopted to replace unsupervised Universal Background Model (UBM) for an accumulation of statistics in the i-vector generation process.

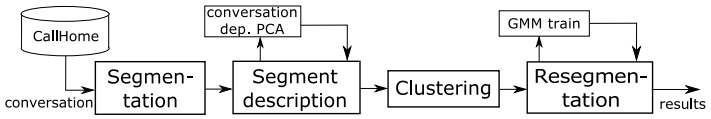

Fig. 1. The schema of the diarization system

In this paper, we propose the NN-based descriptor as a representation of a speaker in an acoustics data segment. Similar approaches using NN were adopted to the representation of the speaker for the SD task in [31, 30], where the NN is used for replacing the Mel Frequency Cepstral Coefficients (MFCCs) features or very recently in [2], where the triplet loss paradigm was used for training the NN descriptor with extremely short speech turn. This speaker representation must be closer to the representations of the segments containing a speech of the same speaker than to the other segments. In the spontaneous conversation, the continuous speech of one speaker (one segment) could by very short, i.e. much less than ten seconds [35]. Our proposed NN-based descriptor creates his own features and accumulates the statistics from very short speech segments (appearing in the telephone speech diarization task).

2 Diarization System

Our SD system consists of four modules (see Figure 1). These are described in detail in the following subsections.

2.1 Segmentation

The input conversation (audio signal) is divided into short segments. The duration of the segments should be long enough to represent the contained speaker and simultaneously to avoid the risk of a speaker change being present within the segment, as may happen in longer segments. Usually, the Speaker Change Detection (SCD) that allows obtaining segments with only one speaker is applied for this purpose [23, 1]. However, in a spontaneous telephone conversation containing obstacles as very short speaker turns and frequent overlapping speech, diarization systems often omit the SCD process and use a simple constant length window segmentation of speech [24, 26]. This two principles of segmentation (constant length window and SCD based on GLR distances or Convolutional NN) is compared in our papers [32, 15]. The results of both these approaches are very similar. Thus, only the segmentation with constant window length is applied in this work.

2.2 Segment Description

After a recording is segmented, a speaker representation is computed for each segment. For this purpose, the i-vectors representation of the speaker in the

acoustics data borrowed from SV is used in recent SD systems [35, 24]. The i-vector representation can handle relatively short speaker utterances. For each conversation segment the supervector of statistics is accumulated [33] and subsequently, the i-vector is extracted from this supervector. For the i-vector extraction, the Factor Analysis (FA) approach [17, 18] (or extended Joint Factor Analysis (JFA) [16] to handle more sessions of each speaker) is used for dimensionality reduction of the supervector of statistics. In Section 3 our proposed approach to segment description based on NN is described.

When segment representation computation works perfectly, a representation is closer to the segments which contain the same voice than to the segments which contain some other voice, i.e. the representation makes clusters. But the low amount of speaker's data in the segments disturbes this presumption. Because of the differences among all conversations (and the similarity inside one conversation), we also compute a conversation dependent Principal Component Analysis (PCA) transformation, which further reduces the dimensionality of the i-vector. The dimension of the PCA latent space is dependent on the parameter p, the ratio of eigenvalue mass [27].

2.3 Clustering and Resegmentation

The segments representations are clustered in order to determine which segments are produced by the same speaker. Since the homogeneity of one segment can not be ensured, it is convenient to refine the final diarization by resegmentation based on a smaller unit then segments. The system iteratively performs the resegmentation applying a Gaussian Mixture Model (GMM) representation of each cluster and redistributing of the whole conversation frame by frame according to the likelihood of the GMMs.

3 Neural Network Descriptor

In order to verificate the speaker or to resolve another similar task such as our main problem, special statistics that describe relevant speaker are computed from a recording. A recurrent NN could be employed to compute such statistics. But, because a recurrent NN training has some issues (especially it has high computation demands), a standard feed-forward NN was used. A similar system for SV task was introduced in our paper [34].

Naturally, a standard feed-forward NN gives exactly the same number of vectors as it has on its input. To compute one single vector of speaker statistics, an average of all vectors was computed. All parts of a recording are not equally relevant. In particular, parts where is no activity of speakers vocal tract are certainly not relevant at all. Furthermore, for one speaker statistics are relevant another parts than for another speaker statistics. Therefore, instead of a simple average, a weighted average where each statistic has own series of weights was computed. The weights could be computed as means of separated NNs. But

in this case, some information could be surely computed redundantly. To prevent this redundancy, one single NN with two output layers was trained instead training of two separated NN.

Our speaker descriptor computes the square of the euclidian distance between two vectors of speaker statistics (i.e. results of the weighted average) and then a sigmoid function is applied. The resultant metric range is obviously in the interval between zero and one. In the training process, an inclination of the sigmoid function was fixed. Only a bias have been trained by means of backpropagation in the same way as all other parameters.

No part of the descriptor is trained separately. Naturally, targets in the training process were ones (for matching pairs of recordings) and zeros (otherwise). The used criterion was modified mean square error. The modification lies in different weights for different types of errors. The criterion ε is given by the following equation

$$\varepsilon = \sum_i w_i (y_i - t_i)^2, \tag{1}$$

where y_i denotes i-th output, t_i denotes i-th target, $w_i = 100$ when $t_i = 1$ and $y_i < 0.5$ or $t_i = 0$ and $y_i > 0.5$. Otherwise, $w_i = 1$. This approach emphasizes errors which lead to a classification error.

As we found in our preliminary experiments, using weighted average brings one serious risk. This risk is a collapse of training algorithm. In such collapse, weights choose only one or very small number of feature vectors to compute statistics. These statistics are nearly irrelevant then. To prevent the training process from these collapsing, two systems with tied parameters were trained. The first one with the plain averaging and the second with the weighted averaging. The first one has been deviating the second one from collapsing. A schema of the resultant speaker descriptor is displayed in Figure 2.

The mentioned NN computes the speaker statistics from features vectors. The standard feature extraction methods such as Linear/Mel Frequency Cepstral Coefficients (LFCCs/MFCCs) might lose a lot of information about speaker identity. Hence, another NN-based feature extraction method was applied. An input of an NN for feature extraction is the absolute spectrum that is very close to the raw signal. The layers used in the described NN-based feature extraction method were not trained to make an LFCCs approximator, but all layers in the whole system for speaker description were randomly initialized and trained simultaneously after the initialization. For testing purposes, only one part of the trained NN was used. The output y is considered to by the vector describing the speaker used in SD system.

The delta and the delta-delta coefficient computation is likely beneficial. However, mean or even variance normalization could be inappropriate in the case of speaker verification. Thus, the original features were not replaced with a normalization but were joined together with delta and delta-delta coefficients, mean normalization (MN), variance normalization (VN) with new delta and delta-delta coefficients into a new larger feature vector. Moreover, splicing that makes long-temporal-feature vectors was applied. The resultant number of features is

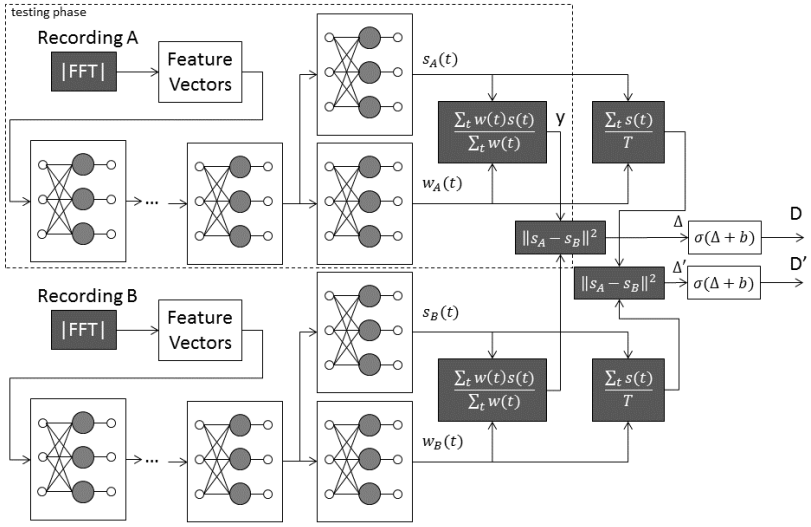

Fig. 2. The neural-network-based speaker verification system. After the feature extraction, the statistics ($\mathbf{y}$) are accumulated and in the case of the training process two different decisions (D and D') about the similarity between recording A and B are made. For the testing purpose, only $\mathbf{y}$ is used as a speaker descriptor (speaker vector)

too high. Hence, the last fully connected layer was applied to reduce the feature dimension. The NN-based feature vector computation is shown in Figure 3.

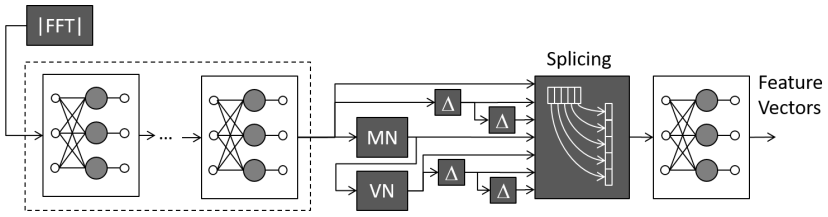

Fig. 3. The neural-network-based feature vectors extraction

The difference between our approach and the one introduced in [2, 10] lies in the fact that we using these DNN features instead of precomputed MFCCs. Meanwhile in [31, 30], the DNN features are used with MFCCs as a stream in an GMM/HMM diarization model.

3.1 Comparison NN vs. i-vector Approach

The NN approach to the speaker descriptor has significantly lower computational and memory demands in comparison with the i-vector extraction process (parametrization, statistics accumulation, FA application).

Both processes produce a vector describing the speaker. Thus, it is possible to combine both descriptors by simple concatenating two vectors into one final vector describing the speaker.

4 Experiments and Results

In our experiment, two described approaches to representing the speaker in each segment of conversation for clustering these segments for diarisation are examined. We compared the state-of-the-art SD system that uses the i-vector based descriptor and SD system that uses our proposed approach applying the NN-based descriptor. The result of the combination of these two principles is also given.

The experiment was carried out on telephone conversations from the English part of CallHome corpus [3] (two channels have been mixed into one), where only two speaker conversations were selected (so the clustering can be limited to two clusters), this is 77 conversation each with about 10 min duration in a single telephone channel sampled at 8 kHz.

The SD system that is the same as in our papers [15, 32] uses the feature extraction based on Linear Frequency Cepstral Coefficients (LFCCs), Hamming window of length 25 ms with 10 ms shift of the window. There are 25 triangular filter banks that are spread linearly across the frequency spectrum and 20 LFCCs were extracted. Added delta coefficients extend the feature vector to a 40-dimensional feature vector. Instead of the voice activity detector, the reference annotation about missed speech was used. For segmentation, only 2-second window with 1-second of overlap was used. The i-vector extraction system was trained using the following corpora: NIST SRE 2004, NIST SRE 2005, NIST SRE 2006 speaker recognition evaluations [19, 21, 22] and the Switchboard 1 Release 2 and Switchboard 2 Phase 3 [11, 12]. The number of Gaussians in the UBM was set to 512. The latent dimension (dimension of i-vectors) in the FA total variability space matrix in the i-vector extraction was set to 400. Finally, the dimension of the final i-vector was reduced by conversation dependent PCA with the ratio of eigenvalue mass $p = 0.5$. Since we have limited the problem to conversations with only two speakers, we applied for segments clustering only K-means algorithm with cosine distance.

For the NN training, the NIST-04,05,06 corpora were used and all recordings were cut up to 2-seconds long pieces. All pieces with too low energy (i.e. pieces which included a significant amount of silence) were excluded from the training process. The dimension of the NN input is 128, hidden layers have 1024 neurones and the dimension of the NN output is 64. The dimension of the feature vector is 40. Training process was implemented utilizing The Theano toolbox[29]

that allows complicated gradient propagations and a GPU usage with almost no effort.

For evaluation of our approach, the Diarization Error Rate (DER) was used as described by NIST in the RT evaluations [5], with 250 ms tolerance around the reference boundaries. DER combines all types of error (missed speech, mislabeled non-speech, incorrect speaker cluster). In our experiments, a correct information about the silence from the reference annotation were used and so our results represent only the error in speaker cluster. The comparison of the examined systems is shown in Table 1. The experimental results of the two approaches

Table 1. DER [%] for SD system with the i-vector speaker representation, the NN speaker representation and the combination of these two representations

descriptor	**DER** [%]
i-vector	9.59
NN	11.20
i-vector + NN	7.72

to the speaker description indicate that the proposed approach using the NN-based descriptor brings some new information about the speaker in the short segment in addition to the i-vector. The result of the NN-based approach do not overcome the result of the i-vector descriptor, but the combination gets significant improvement in DER.

5 Conclusions

In this work, our goal was to propose and investigate a novel technique to represent speaker information available in the short segment (of the conversation provided to the diarisation system) for further clustering. The NN were trained to gain a small vector (the essence of the speaker) from the short acoustics data presented to the net. The final vector representation must be as similar as possible to the representations of another segment containing a speech of the same speaker and most diverse to the others. This method of the speaker description was compared with the i-vector descriptor and both methods were tested in the speaker diarization system. The test results of these two approaches show that the i-vector approach leaded to a better performance, but the NN brings new useful information about the speaker that i-vector approach did not obtain. Hence, the combination of both these descriptors outperforms i-vector approach.

Acknowledgments. This research was supported by the Ministry of Culture Czech Republic, project No.DG16P02B048.

References

1. Adami, A.G., Kajarekar, S.S., Hermansky, H.: A New Speaker Change Detection Method for Two-Speaker Segmentation. In: ICASSP. vol. 4, pp. 3908–3911 (2002)
2. Bredin, H.: TristouNet: Triplet Loss for Speaker Turn Embedding. In: ICASSP. pp. 5430–5434. New Orleans (2017)
3. Canavan, A., Graff, D., Zipperlen, G.: CALLHOME American English Speech, LDC97S42. In: LDC Catalog. Philadelphia: Linguistic Data Consortium (1997)
4. Dehak, N., Kenny, P.J., Dehak, R., Dumouchel, P., Ouellet, P.: Front-End Factor Analysis for Speaker Verification. IEEE Transactions on Audio, Speech, and Language Processing 19(4), 788–798 (2011)
5. Fiscus, J.G., Radde, N., Garofolo, J.S., Le, A., Ajot, J., Laprun, C.: The Rich Transcription 2006 Spring Meeting Recognition Evaluation. Machine Learning for Multimodal Interaction 4299, 309–322 (2006)
6. Fredouille, C., Bozonnet, S., Evans, N.: The LIA-EURECOM RT 09 Speaker Diarization System. In: NIST Rich Transcription Workshop (RT09). Melbourne, USA (2009)
7. Furui, S., Itoh, D.: Neural-Network-Based HMM Adaptation for Noisy Speech. In: ICASSP. pp. 365–368. Salt Lake City (2001)
8. Garcia-Romero, D., Espy-Wilson, C.Y.: Analysis of I-vector Length Normalization in Speaker Recognition Systems. In: Interspeech. pp. 249–252. Florence (2011)
9. Garcia-Romero, D., McCree, A., Shum, S., Brummer, N., Vaquero, C.: Unsupervised Domain Adaptation for I-Vector Speaker Recognition. In: Odyssey - Speaker and Language Recognition Workshop. pp. 260–264. Joensuu (2014)
10. Garcia-Romero, D., Snyder, D., Sell, G., Povey, D., McCree, A.: Speaker Diarization Using Deep Neural Network Embedings. In: ICASSP. pp. 4930 – 4934. New Orleans (2017)
11. Graff, D., Miller, D., Walker, K.: Switchboard-2 Phase III Audio. In: LDC Catalog. Philadelphia: Linguistic Data Consortium (1999)
12. Graff, D., Walker, K., Canavan, A.: Switchboard-2 Phase II, LDC99S79. In: LDC Catalog. Philadelphia: Linguistic Data Consortium (2002)
13. Gupta, V.: Speaker Change Point Detection Using Deep Neural Nets. In: ICASSP. pp. 4420–4424. Brisbane (2015)
14. Hershey, J.R., Chen, Z., Roux, J.L., Watanabe, S.: Deep Clustering: Discriminative Embeddings for Segmentation and Separation. In: ICASSP. pp. 31–35. Shanghai (2016)
15. Hrúz, M., Zajíc, Z.: Convolutional Neural Network for Speaker Change Detection in Telephone Speaker Diarization System. In: ICASSP. pp. 4945–4949. New Orleans (2017)
16. Kenny, P.: Joint Factor Analysis of Speaker and Session Variability: Theory and Algorithms. Tech. rep., Centre de Recherche Informatique de Montreal (2006)
17. Kenny, P., Dumouchel, P.: Experiments in Speaker Verification Using Factor Analysis Likelihood Ratios. In: Odyssey - Speaker and Language Recognition Workshop. pp. 219–226. Toledo (2004)
18. Machlica, L., Zajíc, Z.: Factor Analysis and Nuisance Attribute Projection Revisited. In: Interspeech. pp. 1570–1573. Portland (2012)
19. Martin, A., Przybocki, M.: 2004 NIST Speaker Recognition Evaluation, LDC 2006 S44. In: LDC Catalog. Philadelphia: Linguistic Data Consortium (2011)
20. Milner, R., Hain, T.: DNN-Based Speaker Clustering for Speaker Diarisation. In: Interspeech. vol. 08-12-Sept, pp. 2185–2189. San Francisco (2016)

21. NIST Multimodal Information Group: 2005 NIST Speaker Recognition Evaluation Training Data, LDC2011S01. In: LDC Catalog. Philadelphia: Linguistic Data Consortium (2011)
22. NIST Multimodal Information Group: 2006 NIST Speaker Recognition Evaluation Training Set, LDC2011S09. In: LDC Catalog (2011)
23. Rouvier, M., Dupuy, G., Gay, P., Khoury, E., Merlin, T., Meignier, S.: An Open-source State-of-the-art Toolbox for Broadcast News Diarization. In: Interspeech. p. 5. Lyon (2013)
24. Sell, G., Garcia-Romero, D.: Speaker Diarization with PLDA I-vector Scoring and Unsupervised Calibration. In: IEEE Spoken Language Technology Workshop. pp. 413–417. South Lake Tahoe (2014)
25. Sell, G., Garcia-Romero, D., Mccree, A.: Speaker Diarization with I-Vectors from DNN Senone Posteriors. In: Interspeech. pp. 3096–3099. Dresden (2015)
26. Senoussaoui, M., Kenny, P., Stafylakis, T., Dumouchel, P.: A Study of the Cosine Distance-Based Mean Shift for Telephone Speech Diarization. Audio, Speech and Language Processing 22(1), 217–227 (2014)
27. Shum, S., Dehak, N., Chuangsuwanich, E., Reynolds, D., Glass, J.: Exploiting Intra-Conversation Variability for Speaker Diarization. In: Interspeech. pp. 945–948. Florence (2011)
28. Shum, S.H., Dehak, N., Dehak, R., Glass, J.R.: Unsupervised Methods for Speaker Diarization: An Integrated and Iterative Approach. Audio, Speech, and Language Processing 21(10), 2015–2028 (2013)
29. Theano Development Team: Theano: A Python Framework for Fast Computation of Mathematical Expressions. arXiv e-prints abs/1605.0 (2016)
30. Wang, R., Gu, M., Li, L., Xu, M., Zheng, T.F.: Speaker Segmentation Using Deep Speaker Vectors for Fast Speaker Change Scenarios. In: ICASSP. pp. 5420–5424. New Orleans (2017)
31. Yells, S.H., Stolcke, A., Slaney, M.: Artificial Neural Network Features for Speaker Diarization. In: Proc. IEEE Spoken Language Technology Workshop. pp. 402–406. IEEE (2014)
32. Zajíc, Z., Kunešová, M., Radová, V.: Investigation of Segmentation in i-Vector Based Speaker Diarization of Telephone Speech. In: Specom. pp. 411–418. Springer International Publishing, Budapest (2016)
33. Zajíc, Z., Machlica, L., Müller, L.: Initialization of fMLLR with Sufficient Statistics from Similar Speakers. TSD 2011. Lecture Notes in Computer Science 6836, 187–194 (2011)
34. Zelinka, J., Vaněk, J., Müller, L.: Neural-Network-Based Spectrum Processing for Speech Recognition and Speaker Verification. In: Statistical Language and Speech Processing. vol. 9449, pp. 288–299. Budapest (2015)
35. Zhu, W., Pelecanos, J.: Online Speaker Diarization Using Adapted I-Vector Transforms. In: ICASSP. pp. 5045–5049. Shanghai (2016)