CONVOLUTIONAL NEURAL NETWORK FOR SPEAKER CHANGE DETECTION IN TELEPHONE SPEAKER DIARIZATION SYSTEM

Marek Hrúz and Zbyněk Zajíc

University of West Bohemia
Faculty of Applied Sciences
NTIS - New Technologies for the Information Society
Univerzitní 8, 306 14 Plzeň, Czech Republic

ABSTRACT

The aim of this paper is to propose a speaker change detection technique based on Convolutional Neural Netvork (CNN) and evaluate its contribution to the performance of a speaker diarization system for telephone conversations. For the comparison we used an i-vector based speaker diarization system. The baseline speaker change detection uses Generalized Likelihood Ratio (GLR) metric. Experiments were conducted on the English part of the CallHome corpus. Our proposed CNN speaker change detection outperformed the GLR approach, reducing the Equal Error Rate relatively by 46 %. The final results on speaker diarization system indicate that the use of speaker change detection based on CNN is beneficial with relative improvement of diarization error rate by 28 %.

Index Terms— Convolutional Neural Network, Speaker Change Detection, Speaker Diarizatoin, Generalized Likelihood Ratio

1. INTRODUCTION

Speaker Change Detection (SCD) is the problem of finding precise boundaries, where a speaker is changing. The standard approach to SCD consists of applying a pair of sliding windows on the signal and computing the distance between their contents. Speaker changes are then found at the boundaries between the two windows where the distance achieves a significant local maximum. An example of this approach can be found in [1]. Commonly used distance metrics include the Bayesian Information Criterion (BIC), Generalized Likelihood Ratio (GLR) and Kullback-Leibler divergence. Other approach uses Deep Neural Networks (DNN) [2] to find the speaker change. In such DNN system the input of the network is a set of precomputed features. These features can have a huge impact on the success rate of the whole system.

In this paper we apply Convolutional Neural Networks (CNN) to the problem of SCD. CNNs are very successful in the field of image recognition. They were introduced in [3] and later redesigned to cope with a large dataset of images and image classes [4]. In the first layer of the CNN a set of image filters is learned. The responses of these filters are propagated through several convolutional layers and connected into dense layers. These latter layers contain an underlaying semantic information in the sense of a metric space. Closer points in this space are semantically closer. The main motivation for using a CNN is that we want to make the decision directly on the observed signal rather than on some precomputed features. The input to the CNN is a spectrogram - essentially a 2D signal similar to an image. The CNN should be able to learn important features by itself in the first layers and decide whether a speaker change is present in the observed data or not. Our intent is to introduce a CNN into the diarization pipeline as a speaker change detector.

The most common approach to Speaker Diarization (SD) consists of the segmentation of the input signal, followed by the merging of the segments into clusters corresponding to the individual speakers [1, 5]. The alternative is to combine the segmentation and clustering steps into a single iterative process [6, 7]. The segments should be long enough to allow the extraction of speaker identifying information, while limiting the risk of a speaker change being present within the segment. Our goal is to determine whether the SCD approach based on the CNN offers any improvement under such conditions with comparison to the GLR metric. For this purpose, we implement an i-vector [8, 9] based speaker diarization system [10] based on [5, 11].

2. CONVOLUTIONAL NEURAL NETWORK FOR SPEAKER CHANGE DETECTION

In our previous work we have performed first experiments with CNN for SCD [12]. Even though the CNN outperformed the BIC system based on Linear Frequency Cepstral Coeffi-

The work was supported by the Grant Agency of the Czech Republic project No. P103/12/G084. Access to computing and storage facilities owned by parties and projects contributing to the National Grid Infrastructure MetaCentrum, provided under the programme "Projects of Large Research, Development, and Innovations Infrastructures" (CESNET LM2015042), is greatly appreciated.

cients (LFCCs) [13] we identified weak spots of the system that we address in this paper. In the previous system only the lower 256 out of 512 frequencies were used, because we considered the higher frequencies useless for SCD. But our later experiments showed that using the full spectrum is beneficial and a smaller resolution of 256 in the frequency domain is sufficient. This yields an image (spectrogram) of sufficient quality, meaning the speech harmonics are clearly visible as seen in Figure 1. Furthermore, after observing the annotations of our training data we realized that there are time inaccuracies of the annotated speaker change points. This is due to the nature of the spontaneous telephone conversation, where the speaker changes rapidly and often one speaks over the other. We developed a new labeling strategy to cope with the uncertainty introduced by the human annotators. Instead of the previously used binary labeling we use a fuzzy labeling. The value one is assigned when the speaker change is located in the middle of the analyzed window and linearly decreases to zero as the change moves away from the middle. The value zero begins at ± 0.6 sec away from the change. Formally written the value of the label function L in time t is

$$L(t) = \max\left(0, -\frac{\min_i\left(|t - s_i|\right)}{\tau} + 1\right), \qquad (1)$$

where s_i is the time of i^{th} speaker change and $\tau = 0.6$ is the tolerance. Figure 1 depicts an example of speech and the values of the labeling.

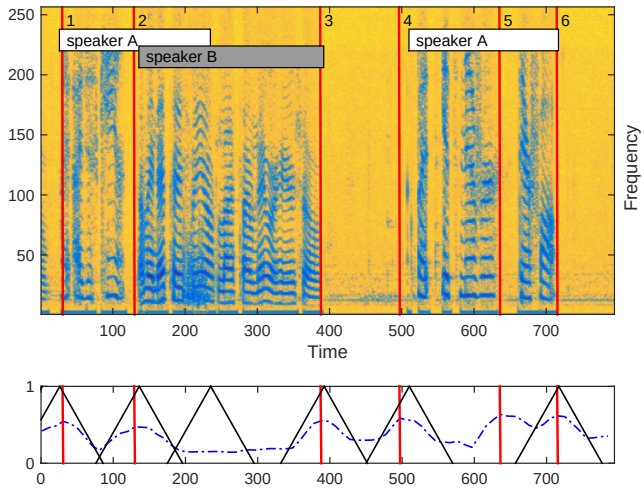

Fig. 1. An example of speech with labeling, CNN output, and detected boundaries. In the top of the image the annotation is depicted as labeled rectangles. The vertical lines are the detected speaker change points. In the lower image, the triangles represent the values of the labeling function L, the dotted line is the output of the CNN, and the red lines are the detected speaker change points.

In the figure, it can be seen that the CNN failed to detect the third labeled change (end of overlapped speech). This will add to the miss rate error. Another error is made by the net since it divided the second speech segment of speaker A with detection number 5. Although this will be evaluated as a false alarm it remains open for discussion whether it should have been annotated or not, since a silence segment is present.

Table 1. Summary of the architecture of the CNN.

Layer	Kernels	Size	Shift
Convolution	50	32 x 16	2 x 2
Max pooling		2 x 2	2 x 2
Batch Norm			
Convolution	200	4 x 4	1 x 1
Max pooling		2 x 2	2 x 2
Batch Norm			
Convolution	300	3 x 3	1 x 1
Max pooling		2 x 2	2 x 2
Batch Norm			
Fully Connected	4000		
Fully Connected	1		

2.1. Architecture of CNN

The architecture of the CNN consists of three convolutional layers with ReLU activation functions. Each convolution layer is followed by a max pooling layer and a batch normalization layer [14]. The last two layers are fully connected with sigmoid activation function. The architecture is summarized in Table 1. The shape of the filters in the first convolutional layer respects the usually rectangular shapes of the high energy speech harmonics in the spectrogram. We have doubled the size of these filters from our previous work which has slightly improved the results. This layer serves as visual features detector. The output layer consist of just one neuron with sigmoid activation function. Thus the output is limited between zero and one. It represents the likelihood of a speaker change happening in the observed window.

2.2. Training of the CNN

The CNN serves as a regressor. Given an input spectrogram it produces a number between zero and one. In other words, it tries to replicate the label function L from the training data. We use a Binary Cross Entropy loss function in the training process. It is optimized by Stochastic Gradient Descent with batch size of 64 and we change the learning rate after a fixed number of steps. When the loss function is stabilized we use RMSProp algorithm for fine tuning of the network's weights.

3. SPEAKER DIARIZATION SYSTEM

Our speaker diarization system, described in [10], is based on the i-vectors to represent segments of speech, as introduced

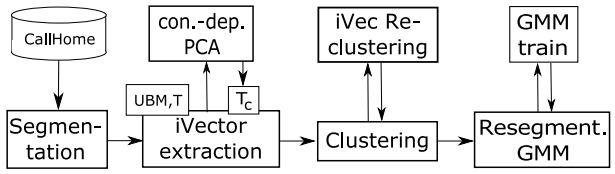

Fig. 2. Diagram of the diarization process.

in [5, 15]. The i-vectors are constructed using total variability matrix $\boldsymbol{T}$ derived from huge amount of data by Principal Component Analysis (PCA). The size of the i-vector is further reduced by conversation-dependent PCA [5] (represented by the total variability matrix $\boldsymbol{T_C}$).

The diarization process starts with the extraction of acoustic features from the conversation. Then the conversation is split into short segments. In the next step, a single i-vector is extracted from each segment and the i-vectors are clustered using K-means algorithm with cosine distance in order to determine which parts of the signal were produced by the same speaker. The system iteratively performs reclustering by computing new i-vectors from all data belonging to each cluster and assigning the segments' i-vectors to the nearest cluster i-vector (in cosine distance). Finally, the feature-wise (LFCC) iterative resegmentation by Gaussian Mixture Models (GMMs) trained on the data from each cluster to refine the final results is applied. A diagram of our diarization system can be seen in Figure 2.

3.1. Segmentation

In our system described in [10], we compared the naive approach to segmentation (constant length window) with a SCD segmentation based on GLR distances, where likely speaker changes are identified as the locations of significant local maxima of the distance function. For this purpose, we calculate the prominence of individual peaks in the distance function and select those with values exceeding a threshold. The value of the threshold is set experimentally.

In this paper, the likely speaker changes in SCD approach based on CNN (with fuzzy labeling) are identified using maximum suppression with suitable window size. This results into a set of detected peaks which are thresholded with a value of $T = 0.5$. This removes insignificant local maxima. The value of the threshold can be set to 0.5 due to the training process of the network (the output of CNN is the probability of change).

4. EXPERIMENTS

In this paper, we try to outperform classical GLR-based SCD approach used for segmentation in an i-vector based speaker diarization system by our proposed CNN-based SCD segmentation. We compared the results with a naive segmentation with constant length segments (which is frequently used in speaker diarization system for telephone speech [16], [5]). The experiment was carried out on telephone conversations from the English part of CallHome corpus [17], where only two speaker conversations were selected (so the clustering can be limited to two clusters). This is 109 conversation (32 for CNN training and 77 for testing) each with about 10 min duration in a single telephone channel sampled at 8 kHz.

The GLR-based SCD uses LFCC features with Hamming window of length 25 ms with 10 ms shift of the window. There are 25 triangular filter banks which are spread linearly across the frequency spectrum, and 20 LFCCs were extracted. Delta coefficients were added leading to a 40-dimensional feature vector. Instead of the voice activity detector, the reference annotation about missed speech was used. The length of window for GLR was set to 1.4 seconds (longer window usually used in SCD system proved to be inappropriate for telephone conversations with spontaneous speech). For CNN-based SCD the input was a spectrogram computed over a 1.4 second long window. Each column of the spectrogram is a result of a FFT into 256 frequencies with a shift of 10 ms. The absolute value of the spectrum was used.

The speaker diarization system based on i-vector uses the same LFCC feature as described above. The i-vector extraction system [18] was trained using the following corpora: NIST SRE 2004, NIST SRE 2005, NIST SRE 2006, Switchboard 1 Release 2 and Switchboard 2 Phase 3. The number of Gaussians in the Universal Background Model (UBM) was set to 1024. The latent dimension (dimension of i-vectors) in the factor analysis total variability space matrix $\boldsymbol{T}$ 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 $p = 0.5$. For naive segmentation, a 2 second window with 1 second overlap was used. For segmentation by GLR, the length of the segments was limited to 4 seconds maximum and 1 second minimum. In CNN segmentation only the 1 second minimum limitation on segment was imposed. In the reclustering and resegmentation, the maximum iteration was set to 1000 (although the convergence occurred usually much earlier). The number of GMMs components (used in resegmentation step) was depended on the amount of data. The model was trained by EM algorithm or adapted from UBM (in the case of minimal amount of data, under 20 sec).

4.1. Speaker Change Detection Results

In this section we present the results from experiments with SCD. Each test audio file was processed by individual systems resulting into a set of speaker changes. Each speaker change is compared to the annotated changes. We consider a ± 200 ms tolerance and compute false alarm rate and miss rate of the detections. We analyzed SCD based on GLR, CNN

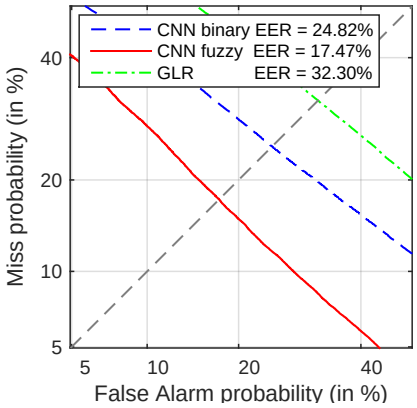

Fig. 3. DET curves for different SCD approaches based on CNN with binary labeling, CNN with fuzzy labeling and GLR metric.

with binary labeling [12], and CNN with fuzzy labeling proposed in this paper. The results are depicted in Figure 3 as DET curves [19]. The fuzzy labeling proved to be very beneficial for SCD. The Equal Error Rate (EER) [19] was reduced significantly when compared to the binary labeling, see the results in Table 2. The relative error dropped by almost 30 %.

Table 2. EER [%] for for different SCD approaches based on CNN with binary labeling, CNN with fuzzy labeling and GLR metric.

SCD	EER [%]
CNN binary	24.82
CNN fuzzy	17.47
GLR	32.30

4.2. Speaker Diarization Results

For evaluation, the Diarization Error Rate (DER) was used as described and used by NIST in the RT evaluations [20], 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 we use correct information about the silence from the reference annotation and so our results represent only the error in speaker cluster. The comparison of the examined systems is shown in Table 3.

The experimental results of SCD-based segmentation for speaker diarization task indicate, that the segmentation based on CNN lowers the DER compared to GLR metric based segmentation. The DER was reduced relatively by 28 %. However, the comparison of the results of constant length window shows that the impact of inaccurate segmentation is diminished by resegmentation process at the end of the diarization (as we have shown in [10]).

Table 3. Comparison of the system using SCD-based segmentation (CNN with fuzzy labeling and GLR) and constant length window segmentation (ConstWin). Results are given as DER [%].

segmentation	DER [%]
CNN fuzzy	9.3139
GLR	11.9797
ConstWin	9.2258

5. CONCLUSIONS

In this work, we compared two approaches to speaker change detection, GLR and our prosed CNN. The SCD segmentation methods are trying to find the precise boundaries where the speaker is changing. Result of SCD shows the superiority of the approach based on CNN over the GLR approach.

Furthermore, we investigated the effect of SCD-based segmentation (CNN and GLR) in an i-vector based speaker diarization system. We compared these results with the naive segmentation using constant length window which divides a conversation into short segments and relies on clustering and further resegmentation to refine the boundaries. The experimental results show that the CNN approach offers significantly better performance of the speaker diarization system then the GLR approach. But compared to the constant length widow, the differences are diminished by the resegmentation, which repairs the inaccurate segmentation produced by the constant length window. The effect of resegmentation is strong because there is sufficient amount of data available in each conversation for efficient training of GMM. In a task with less data, the estimation of the GMM parameters could be inaccurate even impossible. Then the speaker diarization system can not relay on resegmentation. This will be a major problem e.g. in an on-line diarization system. In our proposed approach with CNN the decision of the speaker change is made only from 1.4 second long window. No further information is needed and the result is a likelihood value in the interval $\langle 0; 1 \rangle$ which enables the use of a priori threshold.

6. REFERENCES

[1] M. Rouvier, G. Dupuy, P. Gay, E. Khoury, T. Merlin, and S. Meignier, "An open-source state-of-the-art toolbox for broadcast news diarization", Tech. rep., Idiap, 2013.

[2] V. Gupta, "Speaker change point detection using deep neural nets", in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4420–4424, 2015.

[3] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel, "Backprop-

agation Applied to Handwritten ZIP Code Recognition", in *Neural computation*, 1(4), pp. 541–551, 1989.

[4] A. Krizhevsky, I. Sutskever and G. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks" in *Advances in Neural Information Processing Systems*, pp. 1097–1105, 2012.

[5] G. Sell and D. Garcia-Romero,"Speaker Diarization with PLDA I-vector Scoring and Unsupervised Calibration", in *Proc. IEEE Spoken Language Technology Workshop*, pp. 413–417, 2014.

[6] C. Fredouille, S. Bozonnet and N. Evans, "The LIA-EURECOM RT09 Speaker Diarization System", in *Proc. NIST Rich Transcription Workshop (RT09)*, 2009.

[7] S. H. Shum, N. Dehak, R. Dehak and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach", in *Proc. IEEE Transactions on Audio, Speech, and Language Processing*, 21(10), pp. 2015–2028, 2013.

[8] N. Dehak, P. Kenny, R. Dehak, RP. Dumouchel and P. Ouellet, "Front-end factor analysis for speaker verification", in *Proc. IEEE Transactions on Audio, Speech, and Language Processing*, 19(4), pp. 788–798, 2011.

[9] D. Garcia-Romero and C.Y. Espy-Wilson, "Analysis of I-vector Length Normalization in Speaker Recognition Systems", in *Proc. Interspeech*, pp. 249–252, 2011.

[10] Z. Zajic, M. Kunesova and V. Radova, "Investigation of Segmentation in i-Vector based Speaker Diarization of Telephone Speech", in *Lecture Notes in Computer Science*, vol. 9811, pp. 411–418, 2016.

[11] M. Senoussaoui, P. Kenny, T. Stafylakis and P. Dumouchel, "A study of the cosine distance-based mean shift for telephone speech diarization", in *Proc. IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(1), pp. 217–227, 2014.

[12] M. Hruz and M. Kunesova, "Convolutional Neural Network in the Task of Speaker Change Detection", in *Lecture Notes in Computer Science*, vol. 9811, pp. 191–198, 2016.

[13] S. Chen, P. Gopalakrishnan, P.: "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion", in: *Proceedings DARPA Broadcast News Transcription and Understanding Workshop*, pp. 127–132, Vol. 8., 1998.

[14] S. Ioffe, C Szegedy: Batch normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv preprint arXiv:1502.03167*, 2015

[15] S. Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds and J. Glass, J, "Exploiting intra-conversation variability for speaker diarization", in *Proc. Interspeech*, pp. 945–948, 2011.

[16] G. Sell, D. Garcia-romero and A. Mccree, "Speaker Diarization with I-Vectors from DNN Senone Posteriors", in *Proc. Interspeech*, pp. 3096–3099, 2015.

[17] A. Canavan, D. Graff and G. Zipperlen, "CALLHOME American English Speech LDC97S42", *LDC Catalog, Philadelphia: Linguistic Data Consortium*, 1997.

[18] L. Machlica and Z. Zajic, "Factor Analysis and Nuisance Attribute Projection Revisited", in *Proc. Interspeech*, pp. 1570–1573, 2012.

[19] N. Brummer and E. de Villiers, "The BOSARIS Toolkit User Guide: Theory, Algorithms and Code for Binary Classifier Score Processing", Tech. rep., 2011.

[20] J.G. Fiscus, N. Radde, J. S. Garofolo, A. Le, J. Ajot and C. Laprun, "The Rich Transcription 2006 Spring Meeting Recognition Evaluation", in *Machine Learning for Multimodal Interaction*, 4299, pp. 309–322, 2006.