Recurrent Neural Network Based Speaker Change Detection from Text Transcription Applied in Telephone Speaker Diarization System ⋆

Zbyněk Zajíc[0000−0002−4153−6560], Daniel Soutner[0000−0002−8899−8260], Marek Hrúz[0000−0002−7851−9879], Luděk Müller[0000−0002−6581−6348], and Vlasta Radová[0000−0002−3258−8430]

University of West Bohemia, Faculty of Applied Sciences,
NTIS - New Technologies for the Information Society
and Department of Cybernetics
Univerzitní 8, 306 14 Plzeň, Czech Republic
{zzajic, dsoutner, mhruz, muller, radova}@ntis.zcu.cz

Abstract. In this paper, we propose a speaker change detection system based on lexical information from the transcribed speech. For this purpose, we applied a recurrent neural network to decide if there is an end of an utterance at the end of a spoken word. Our motivation is to use the transcription of the conversation as an additional feature for a speaker diarization system to refine the segmentation step to achieve better accuracy of the whole diarization system. We compare the proposed speaker change detection system based on transcription (text) with our previous system based on information from spectrogram (audio) and combine these two modalities to improve the results of diarization. We cut the conversation into segments according to the detected changes and represent them by an i-vector. We conducted experiments on the English part of the CallHome corpus. The results indicate improvement in speaker change detection (by 0.5 % relatively) and also in speaker diarization (by 1 % relatively) when both modalities are used.

Keywords: Recurrent Neural Network, Convolutional Neural Network, Speaker Change Detection, Speaker Diarization, I-vector

1 Introduction

The problem of Speaker Diarization (SD) is defined as a task of categorizing speakers in an unlabeled conversation. The Speaker Change Detection (SCD) is often applied to the signal to obtain segments which ideally contain a speech of a single speaker [1]. The telephone speech is a particular case where the speaker turns can be extremely short with negligible between-turn pauses and frequent overlaps. SD systems for telephone conversations often omit the SCD process and use a simple constant length window segmentation of speech [3]. In our previous papers [10], [11], we introduced the SD

⋆ This research was supported by the Ministry of Culture Czech Republic, project No. DG16P02B009.

system with SCD based on Convolutional Neural Network (CNN) for segmentation of the acoustic signal. This SD system is based on i-vectors [12] that represent speech segments, as introduced in [13]. The i-vectors are clustered in order to determine which parts of the signal were produced by the same speaker and then the feature-wise resegmentation based on Gaussian Mixture Models is applied.

In all SD systems mentioned above, only the audio information is used to find the speaker change in the conversation. In this work we aimed to use the lexical information contained in the transcription of the conversation, which is a neglected modality in the SCD/SD task: The work [14] investigates whether the statistical information on the speaker sequence derived from their roles (using speaker roles n-gram language model) can be used in speaker diarization of meeting recordings. Using Automatic Speech Recognition (ASR) system transcription for diarization of a telephone conversation was used in [15] where only speech and non-speech regions were classified.

We can see the lexical information as an additive modality compared to the acoustic data. Also, both the SCD based on the linguistic and acoustic information can be combined to improve the accuracy of the SD system. A similar approach was recently published in [31].

2 Segmentation

2.1 Oracle Segmentation

We implemented oracle segmentation as in [16] for the purpose of comparison: the conversations are split according to the reference transcripts, each individual speaker turn from the transcript becomes a single segment.

2.2 CNN based SCD on spectrogram

In our previous work [17], we introduced the CNN for SCD task. We trained the CNN as a regressor on spectrograms of the acoustic signal with a reference information L about the existing speaker changes, where L can be seen as a fuzzy labeling [10] with a triangular shape around the labeled speaker change time points produced by humans. The main idea behind it is to model the uncertainty of the annotation. The speaker changes are identified as peaks in the network's output signal P using non-maximum suppression with a suitable window size. The detected peaks are then thresholded to remove insignificant local maxima. We consider the signal between two detected speaker changes as one segment. The minimum duration of one segment is limited to one second, shorter parts are not used for clustering, and the decision about the speaker in them is waiting for the resegmentation step. This condition is made to avoid clustering segments containing an insignificant amount of data from the speaker to be modeled as an i-vector. It is also possible to use this system for SCD (with small modification) in the online SD system [18].

2.3 RNN based SCD on lexical information

From the global point of view, a change of a speaker mainly occurs when the speaker ended a word as opposed to in the middle of pronunciation. The probability of change

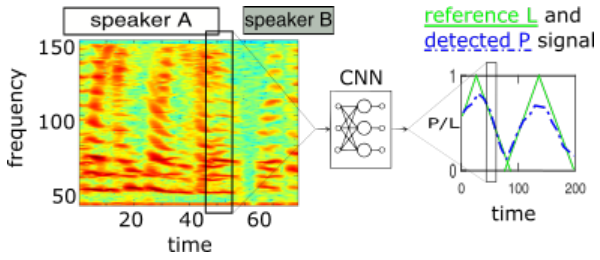

Fig. 1. The input speech as spectrogram is processed by the CNN into the output probability of change P (the dashed line). The reference speaker change L for the CNN training is depicted also (the solid line).

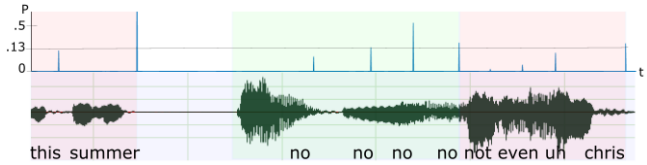

Fig. 2. The output of the RNN based SCD on lexical information. The probability P of the speaker/utterance change in time.

is even higher when he/she finished a sentence. This is the reason, why we decided to acquire extra information about the speaker change from text transcriptions using detection of utterance endings. This process might produce over-segmentation of the conversation. Although this means the coverage measure of the results will be lower, the purity of the segments will be high. Nevertheless, the over-segmentation of the conversation is not such a crucial problem, because our goal is to make the whole diarization process more accurate (not just SCD). If the segments are long enough to represent the speaker by an i-vector, the segmentation step of the SD system will assign the proper speakers to the segments. That's why we deduce that to find the end of an utterance is a reasonable requirement for segmentation.

We conducted two experiments. First with the reference transcriptions that were force-aligned with an acoustic model and the second with the recognized text from the ASR system. We followed this procedure: Obtain aligned text with time stamps (force aligned or from the ASR) from the recordings. Train a language model as Recurrent Neural Network [19] with Long Short-Term Memory (LSTM) layers [20]. Label every word from text with lexical probability, that the next word is the end of an utterance. The output from the RNN is the probability of speaker change in time (see Figure 2).

2.4 Combination of both SCD approaches

Both the approaches to the SCD problem can be combined to refine the information about the speaker change for segmentation step of SD system. Both systems output the probability of a speaker change in time. The combined system can decide about the

speaker change considering two sources; CNN on a spectrogram (audio) and RNN on a transcription (text). The output of the combined system is also a probability of speaker change (a number between zero and one). We used a weighted sum of both speaker change probabilities $P_{comb} = w * P_{spectr} + (1 - w) * P_{transc}$ and normalized the results into an interval $\langle 0; 1 \rangle$. The value of the parameter w was found experimentally to be 0.5.

3 Segment description

To describe a segment of conversation we first construct a supervector of accumulated statistics [21] and then the i-vectors are extracted using Factor Analysis [22]. In our work [11], we introduced an approach to the statics refinement using the probability of speaker change as a weighting factor into the accumulation of statistics. We also use this approach in this paper.

4 Experiments

We designed the experiment to investigate our proposed approach to SCD from RNN on transcription compared with CNN on spectrogram and with the combined system.

4.1 Corpus

The experiment was carried out on telephone conversations from the English part of CallHome corpus [23]. We mixed the original two channels into one and we selected only two speaker conversations so that the clustering can be limited to two clusters. This subset contains 109 conversations in total each has about 10 min duration in a single telephone channel sampled at 8 kHz. For training of the CNN, we used only 35 conversations, the rest we used for testing the SD system.

4.2 System

The SD system presented in our paper [11] uses the feature extraction based on Linear Frequency Cepstral Coefficients, Hamming window of length 25 ms with 10 ms shift of the window. We employ 25 triangular filter banks which are spread linearly across the frequency spectrum, and we extract 20 LFCCs. We add delta coefficients leading to a 40-dimensional feature vector ($D_f = 40$). Instead of the voice activity detector, we worked with the reference annotation about the missed speech.

We employed CNN described in [10] for segmentation based on the audio information. The input of the net is a spectrogram of speech of length 1.4 seconds, and the shift is 0.1 seconds. The CNN consists of three convolutional layers with ReLU activation functions and two fully connected layers with one output neuron. Note that for the purposes of this paper we reimplemented the network in Tensorflow [1], thus the results slightly differ from our previous work.

[1] Available on https://www.tensorflow.org

As our language model for computing lexical scores, we have chosen neural network model with two LSTM [20] layers with the size of hidden layer 640. We trained our model from Switchboard corpus [24], which is very near to our testing data. We split our data into two folds: train with 25433 utterances and development data with 10000 utterances. The vocabulary has the size of 29600 words (only from the training part of the corpus) plus the $\langle unk \rangle$ token for the unknown words. We used SGD as the optimizer. We employed dropout for regularization, and the batch size was 30 words. We evaluated our model on text data, and we achieved 72 in perplexity on development data and 70 on test data.

The ASR system setup, for automatic transcription of the data, was the same as the standard Kaldi [25] recipe s5c for Switchboard corpus; we used the "chain" model. We trained the acoustic model as Time Delayed Neural Network with seven hidden layers, each with an output of 625, the number of targets (states) was 6031. We set the inputs as MFCC features with a dimension of 40 and the i-vectors for adaptation purposes. We recognized all the recordings as one file, the Word Error Rate on tested data was $26.8\,\%$.

For the purpose of training the i-vector, we model the Universal Background Model as a Gaussian Mixture Model with 1024 components. We have set the dimension of the i-vector to 400. For clustering, we have used K-means algorithm with cosine distance to obtain the speaker clusters.

4.3 Results

The results as Purity [26] vs. Coverage [27] curve for SCD can be seen in in Figure 3 for all approaches to the segmentation, where dual evaluation metrics Purity and Coverage are used according to the work [28] to better evaluate the SCD process. The slightly modified Equal Error Rate (EER), where the Coverage and Purity have the same value, for each SCD method with the particular threshold T_{EER} can be seen in the first two columns of Table 1. The goal of the general SCD system is to get the best Purity and Coverage, but for our SD system, we want to get the best "Purity" of all segments with enough segments longer than 1 second. The 1-second threshold we set empirically as enough speech for training the i-vector to represent the speaker accurately in the segment for diarization of two-party conversation. For CallHome data with relatively long conversations (5-10 minutes), it is better for the SD system to leave some short segments out of clustering and wait for the re-segmentation step to decide about the speaker in these segments.

We use the Diarization Error Rate (DER) for the evaluation of our SD system to be comparable to other methods tested on CallHome (e.g., [29, 3]). DER has been described and used by NIST in the RT evaluations [30]. We use the standard 250 ms tolerance around the reference boundaries. DER is a combination of several types of errors (missed speech, mislabeled non-speech, incorrect speaker cluster). We assume the information about the silence in all testing recordings is available and correct. That means that our results represent only the error of incorrect speaker clusters. Contrary to a common practice in telephone speech diarization, we do not ignore overlapping segments during the evaluation. The last two columns of Table 1 shows the SD system using the

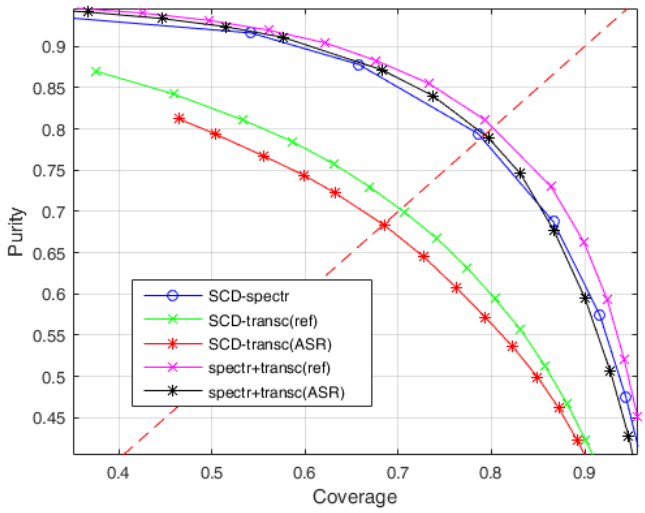

Fig. 3. Purity vs Coverage curve for SCD system with CNN on spectrogram, RNN on transcripts and Combined system.

SCD based on spectrogram, transcription, and combination of both and the experimentally chosen threshold T (to remove insignificant local maxima in SCD system outputs) for each method.

Table 1. EER [%] for each the SCD system with particular threshold T_{EER} and DER [%] of the whole SD systems with the segmentation based on SCD. The SCD using spectrogram, reference transcription and transcription from ASR. Also, the results from combination using spectrogram with reference transcription and spectrogram with ASR transcription are reported. The experimentally chosen threshold T for segmentation is in the last column.

segmantation	**EER**	T_{EER}	**DER**	T
SCD-oracle	0.00	-	6.76	-
SCD-spectr	0.21	0.75	6.93	0.70
SCD-transc(ref)	0.30	0.17	8.07	0.17
SCD-transc(ASR)	0.32	0.08	8.62	0.12
spectr+transc(ref)	**0.20**	0.50	**6.86**	0.45
spectr+transc(ASR)	0.21	0.49	7.06	0.45

4.4 Discussion

The proposed SCD approach using lexical information from transcription performed worse than the SCD on the spectrogram. We think the main reason for this is due to the

quality of the conversation: the sentences in the telephone recordings are not always finished due to the frequent crosstalks, so the SCD based on transcription has incomplete information about the speaker change. Nevertheless, this information brings an additive knowledge about the speaker change. The combined SCD system ("spectr+transc") improved the results of the SD system. When using the transcription from ASR we obtain slightly worse results due to the accuracy of the ASR system. More sophisticated classifier using both SCD from spectrogram and transcription can be trained. However, there is a problem with the training criterion because our goal is to get better results on the SD system, not only to find the precise boundaries of the speaker changes. Also, the mistakes in the reference annotations of CallHome corpus are limiting the performance (see the result of oracle segmentation). Authors of a similar approach [31] tried to find SCD using also both acoustics and lexical information combined together and propagated thru only one LSTM neural network. Unfortunately, the evaluation of their approach was made on different data.

5 Conclusions

In this paper, we proposed a new method for SCD using the lexical information from the transcribed conversation. For this purpose, we have trained RNN with LSTM layers to evaluate the transcription of the conversation and find the speaker changes in it. This approach brings new information about the speaker change and can be used in combination with SCD method based on the audio information to improve the diarization. Our future work is to train a complex classifier to improve the speaker change detection using both modalities (text and audio).

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," in *Interspeech*, Lyon, 2013, pp. 1477–1481.
2. S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised Methods for Speaker Diarization: An Integrated and Iterative Approach," *Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, 2013.
3. G. Sell and D. Garcia-Romero, "Speaker Diarization with PLDA I-vector Scoring and Unsupervised Calibration," in *IEEE Spoken Language Technology Workshop*, South Lake Tahoe, 2014, pp. 413–417.
4. A. G. Adami, S. S. Kajarekar, and H. Hermansky, "A New Speaker Change Detection Method for Two-Speaker Segmentation," in *ICASSP*, vol. 4. Orlando: IEEE, 2002, pp. 3908–3911.
5. J. Ajmera, I. McCowan, and H. Bourlard, "Robust Speaker Change Detection," *Signal Processing Letters, IEEE*, vol. 11, pp. 649–651, 2004.
6. B. Fergani, M. Davy, and A. Houacine, "Speaker Diarization Using One-Class Support Vector Machines," *Speech Communication*, vol. 50, no. 5, pp. 355–365, 2008.
7. V. Gupta, "Speaker Change Point Detection Using Deep Neural Nets," in *ICASSP*. Brisbane: IEEE, 2015, pp. 4420–4424.
8. R. Wang, M. Gu, L. Li, M. Xu, and T. F. Zheng, "Speaker Segmentation Using Deep Speaker Vectors for Fast Speaker Change Scenarios," in *ICASSP*. New Orleans, 2017, pp. 5420–5424.

9. L. Ten Bosch, N. Oostdijk, and J. P. De Ruiter, "Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues," in *TSD*. Brno: Springer, 2004, pp. 563–570.
10. M. Hrúz and Z. Zajíc, "Convolutional Neural Network for Speaker Change Detection in Telephone Speaker Diarization System," in *ICASSP*. New Orleans, 2017, pp. 4945–4949.
11. Z. Zajíc, M. Hrúz, and L. Müller, "Speaker Diarization Using Convolutional Neural Network for Statistics Accumulation Refinement," in *Interpeech*, Stockholm, 2017, pp. 3562–3566.
12. 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*, vol. 19, no. 4, pp. 788–798, 2011.
13. S. Shum, N. Dehak, E. Chuangsuwanich, D. Reynolds, and J. Glass, "Exploiting Intra-Conversation Variability for Speaker Diarization," in *Interspeech*, Florence, 2011, pp. 945–948.
14. F. Valente, D. Vijayasenan, and P. Motlicek, "Speaker diarization of meetings based on speaker role n-gram models," in *ICASSP*. Prague: IEEE, 2011, pp. 4416–4419.
15. S. E. Tranter, K. Yu, G. Evermann, and P. C. Woodland, "Generating and Evaluating Segmentations for Automatic Speech Recognition of Conversational Telephone Speech," in *ICASSP*. Montreal: IEEE, 2004, pp. 753–756.
16. M. Kunešová, Z. Zajíc, and V. Radová, "Experiments with Segmentation in an Online Speaker Diarization System," in *TSD*. Plzen: Springer, 2017, pp. 429–437.
17. M. Hrúz and M. Kunešová, "Convolutional Neural Network in the Task of Speaker Change Detection," in *Specom*. Budapest: Springer, 2016, pp. 191–198.
18. M. Kunešová, Z. Zajíc, and V. Radová, "Experiments with Segmentation in an Online Speaker Diarization System," in *TSD*. Prague: Springer, 2017, pp. 429–437.
19. D. Soutner and L. Müller, "Application of LSTM neural networks in language modelling," in *TSD*. Plzen: Springer, 2013, pp. 105–112.
20. S. Hochreiter and J. Urgen Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
21. Z. Zajíc, L. Machlica, and L. Müller, "Robust Adaptation Techniques Dealing with Small Amount of Data," in *TSD*, vol. 7499. Brno: Springer, 2012, pp. 418–487.
22. P. Kenny and P. Dumouchel, "Experiments in Speaker Verification Using Factor Analysis Likelihood Ratios," in *Odyssey*, Toledo, 2004, pp. 219–226.
23. A. Canavan, D. Graff, and G. Zipperlen, "CALLHOME American English Speech, LDC97S42," in *LDC Catalog*. Philadelphia: Linguistic Data Consortium, 1997.
24. J. J. Godfrey and E. Holliman, "Switchboard-1 Release 2," in *LDC Catalog*. Philadelphia: Linguistics Data Consortium, 1997.
25. Povey, Daniel, et al., "Modelos animales de dolor neuropático," in *Workshop on Automatic Speech Recognition and Understanding*, 2011, IEEE Catalog No.: CFP11SRW–USB.
26. M. Harris, X. Aubert, R. Haeb-Umbach, and P. Beyerlein, "A Study of Broadcast News Audio Stream Segmentation and Segment Clustering," in *EUROSPEECH*. Budapest, 1999, pp. 1027–1030.
27. H. Bredin, "TristouNet: Triplet Loss for Speaker Turn Embedding," in *ICASSP*. New Orleans, 2017, pp. 5430–5434.
28. H. Bredin, "pyannote.metrics: a toolkit for reproducible evaluation, diagnostic, and error analysis of speaker diarization systems," in *Interspeech*, Stockholm, 2017, pp. 3587–3591.
29. G. Sell, D. Garcia-Romero, and A. Mccree, "Speaker Diarization with I-Vectors from DNN Senone Posteriors," in *Interspeech*, Dresden, 2015, pp. 3096–3099.
30. J. G. Fiscus, N. Radde, J. S. Garofolo, A. Le, J. Ajot, and C. Laprun, "The Rich Transcription 2006 Spring Meeting Recognition Evaluation," *Machine Learning for Multimodal Interaction*, vol. 4299, pp. 309–322, 2006.
31. M. India, J. Fonollosa and J. Hernando, "LSTM neural network-based speaker segmentation using acoustic and language modelling," *Interspeech*, Stockholm, 2017 pp. 2834–2838.