Interspeech 2018
2-6 September 2018, Hyderabad

Automatic detection of multi-speaker fragments with high time resolution

Kazimirova E[1], *Belyaev A*[12]

[1]Neurodatalab, USA
[2]Lomonosov MSU, Russia

e.kazimirova@neurodatalab.com, a.belyaev@neurodatalab.com

Abstract

Interruptions and simultaneous talking represent important patterns of speech behavior. However, there is a lack of approaches to their automatic detection in continuous audio data. We have developed a solution for automatic labeling of multi-speaker fragments using harmonic traces analysis. Since harmonic traces in multi-speaker intervals form an irregular pattern as opposed to the structured pattern typical for a single speaker, we resorted to computer vision methods to detect multi-speaker fragments.

A convolutional neural network was trained on synthetic material to differentiate between single-speaker and multi-speaker fragments. For evaluation of the proposed method the SSPNet Conflict Corpus with provided manual diarization was used. We also examined factors affecting algorithm performance.

The main advantages of the proposed method are calculation simplicity and high time resolution. With our approach it is possible to detect segments with minimum duration of 0.5 seconds. The proposed method demonstrates highly accurate results and may be used for speech segmentation, speaker tracking, content analysis such as conflict detection, and other practical purposes.

Index Terms: multi-speaker detection, convolutional neural network, harmonics analysis, audio segmentation, overlapped speech, interruption, conversational analysis

1. Introduction

The detection and analysis of multi-speaker intervals in continuous audio is often necessary for improving the accuracy of speaker tracking [1], speaker recognition [2] and automatic speech recognition [3]. For these purposes the detection of multi-speaker intervals is performed in speaker-dependent environment, hence the dependency between the accuracy of multi-speaker interval detection and the performance of speaker identification algorithm. Besides, a sufficient amount of data for each speaker is needed.

However, sometimes the research goal does not require speaker identification. In conflict detection [4], for example, the number of interruptions and overlaps can be informative without any information about who were the interruptee and the interrupter. Even if speaker diarization is required, we can assume that it can be done independently of interruptions and overlaps detection [5]. Using two different specialized algorithms for these two tasks may provide a better accuracy for each of them.

Despite all the motivation, there are only few solutions for overlapped speech detection. The existing approaches rely on Gaussian mixture modeling [6], an HMM-based segmenter [5], pyknogram analysis [7], and LSTM [8]. The highest accuracy was reported in [9] with F-score equal to 0.8 for 500 ms intervals, but the results related to artificially mixed recordings and male speakers only. Thus, it can be said that although the performance of the existing solutions is pretty decent, there are still considerable possibilities for improvement.

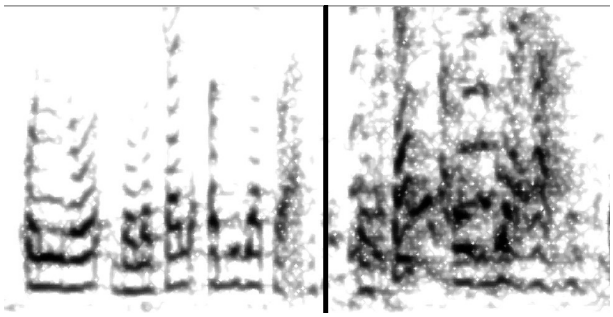

Figure 1: *Spectrogram fragment. The vertical line separates a single-speaker fragment (on the left) and a multi-speaker fragment (on the right)*

Our approach relies on combining harmonic traces analysis with computer vision methods, namely a convolutional neural network (CNN). We aimed to achieve high time-resolution along with high accuracy without any prior information about the speakers.

The remainder of the paper is organized as follows. Section 2 describes the proposed method and the evaluation procedure. Section 3 reports how present work is related to earlier works in the field. The discussion is presented in Section 4. Section 5 concludes the paper and briefly describes the future work.

2. Multi-speaker detection algorithm

2.1. Spectrogram

Spectrogram was computed in the frequency range from 0 to 1500 Hz using 80 ms Hanning window with a 80% overlap between windows. This frequency range was chosen because harmonic traces could be seen most clearly in it. To enhance harmonic visibility, we plotted the spectrogram in reversed gray colors with color level from -50 dB to 0 dB with a 1 dB step. An example of the spectrogram is presented on Figure 1. The harmonics are clearly distinguishable on the spectrogram. For a single speaker their traces are parallel (left half of the picture). When two or more speakers pronounce different tone phonemes simultaneously, the traces cross and the pattern becomes irregular losing its apparent parallel structure (right half of the picture).

2.2. Convolutional Neural Network

The architecture of the designed convolutional neural network was relatively simple. It was composed of three convolutional

10.21437/Interspeech.2018-1878

blocks with pooling and batch normalization. Logistic regression with cross-entropy as a loss function was used in our model.

We formed a training dataset using the audio from TED Talks [10] with 100 male and 100 female voices. The forming of the set involved the following steps:

- Removing unvoiced speech and pauses from the audios with the Praat software [11].
- Creating multi-speaker fragments by mixing the tracks of two different speakers.
- Extracting non-overlapped five-second intervals.
- Computing a spectrogram for every five-second interval as described above. In total, 1584 pictures were obtained, 792 for each of the two classes (single- and multi-speaker).
- Finally, the pictures were obtained for every 500 ms with a 10 ms step. They represented a sum of three spectrum channels - a) log-normal spectrum on channel 1; b) log-normal spectrum with histogram equalization on channel 2; c) log-normal spectrum with Contrast Limited Adaptive Histogram Equalization (CLAHE) on channel 3. See more below.

All frames (356,400 for each of the two classes) were divided into a training set and a test set in the ratio 3:1. The number of training epochs was 70. The initial weights were assigned using Xavier initialization with magnitude of 3. We used SGD optimizer with learning rate of 0.005. The accuracy on the training sample reached $99.87 \pm 0.03\%$, while on the test sample it was equal to $94.64 \pm 0.03\%$. The output of CNN was a score of multi-speaker presence ranging from 0 to 1 for every 500 ms frame in the audio with a 10 ms step.

2.3. Evaluation procedure

As we achieved significant test accuracy on the TED Talks dataset, we aimed to evaluate final model performance on the new data in order to test its applicability across different datasets. This step allows us to predict whether this model could potentially be used for practical purposes in the wild.

We evaluated our algorithm on the SSPNet Conflict Corpus [4] that contained 1430 clips (30 seconds each) from political debates supplemented by manual speaker diarization for each audio which we took as ground truth. This dataset seemed to be suitable for our purposes because political debates contain natural overlapped speech and interruptions.

First, we tested the performance of the proposed algorithm with and without the implementation of histogram equalization methods (simple histogram equalization and CLAHE, as described in the previous paragraph). We considered a manually labeled multi-speaker interval to be detected if at least 50% of it was covered by automatically detected intervals (ADIs). Otherwise, it was tagged as missed. ADIs that did not intersect with any of the manually labeled intervals were tagged as false alarms. The equal error rate values (EER,%) are presented in Table 1. These results testify to the fact that using both histogram equalization methods is beneficial for our purpose.

Second, we performed some additional processing of the output. It was necessary for specifying the results, because for the evaluation procedure spectrogram pictures were obtained for raw audios without the removal of pauses and unvoiced speech. The general scheme of the evaluation procedure is presented in Figure 2. The processing included the following steps.

Table 1: *Equal error rates of multi-speaker interval detection with different spectrum channels combinations.*

Channels	**EER, %**
ch1	19.8
ch1+ch2	17.51
ch1+ch2+ch3	**13.85**

- Subsequent ADIs with gaps of no more than 0.15 s between them were merged together, see 1.1 on Figure 2.
- ADIs where the percentage of unvoiced sounds and/or pauses (UnvP) exceeded the threshold were eliminated, see 1.2 on Figure 2. In the training sample we only had voiced sounds, so the presence of other sounds could have caused detection mistakes. For that reason we assumed that the elimination of unvoiced intervals and pauses would increase the performance. This step involved using the Praat voiced-unvoiced annotation with standard parameters.
- Intervals shorter than minimum allowed length (MinL) were eliminated, see 1.3 on Figure 2.

We also tested the dependency of the algorithm performance from MinL and UnvP. Figure 3 shows the detection error tradeoff (DET) curves for different parameter sets. In Table 2 EERs for different parameter values are listed.

Table 2: *Equal error rates of multi-speaker interval detection with different parameters.*

MinL, s	**UnvP, %**	**EER, %**
0.5	0%	13.75
0.5	50%	12.65
0.5	66%	**11.31**
0.7	0%	12.69
0.7	50%	11.87
0.7	66%	**11.30**
1.0	0%	14.11
1.0	50%	12.66
1.0	66%	12.18

The best performance was achieved when the minimal allowed length of the detected interval was set at 0.5 s or 0.7 s, whereby such intervals where unvoiced sounds constituted more than 66% were eliminated.

We also examined the way in which the performance of our algorithm depends on the speakers' genders. Table 3 contains the EER values for the cases where two men, two women, or a man and a woman speak simultaneously.

Table 3: *Equal error rates of multi-speaker interval detection for different gender mixes.*

Speakers' Genders	**EER, %**
male + male	14.94
female + female	**13.65**
male + female	14.83

After having examined the parameter dependency, we applied our algorithm to the SPPNet Conflict Corpus with a 0.8

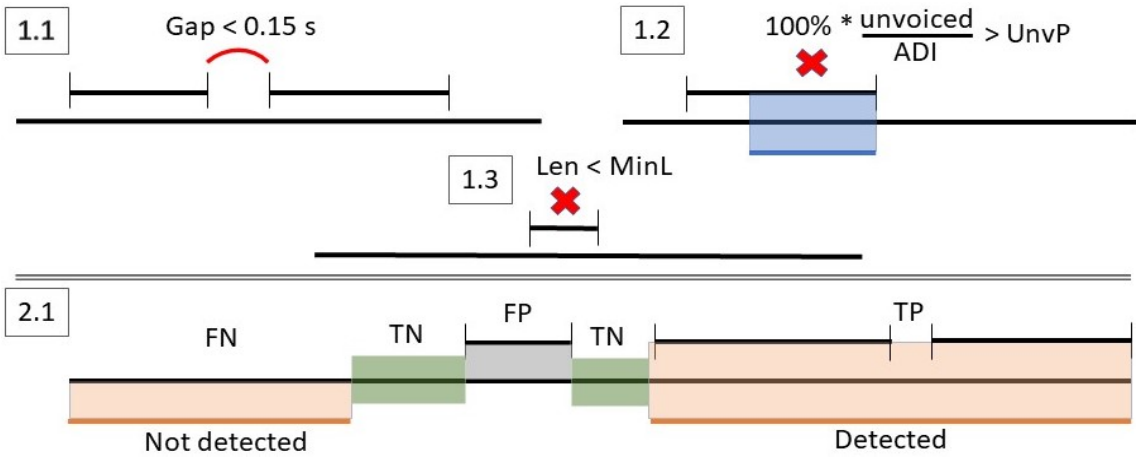

Figure 2: *Scheme of the proposed algorithm evaluation procedure. Automatically detected intervals (ADI) are presented in black, unvoiced intervals are presented in blue, manually detected multi-speaker intervals - in orange. The crosses mark ADIs which were eliminated. 1.1 - 1.3 - additional processing of ADIs (merging intervals, eliminating unvoiced and short intervals). 2.1 - comparing automatic and manual segmentation. FN - false negative, TN - true negative, FP - false positive, TP - true positive.*

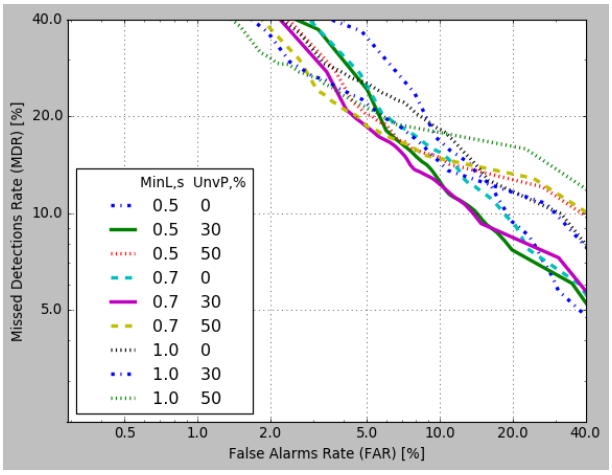

Figure 3: *DET curves for different minimum allowed lengths of ADIs (MinL, s) and for maximum allowed percentage of unvoiced sounds in ADIs (UnvP, %)*

CNN score threshold, 0.5 s and 0.7 s minimum interval lengths and 66% UnvP. The detection performance measures including F-scores and total accuracy are displayed in Table 4. Figure 4 represents an example of manual and automatic multi-speaker detection in the same audio. Manually labeled intervals are marked with the black line while ADIs are filled with gray color. In this particular example both manually labeled intervals were detected successfully.

Table 4: *Multi-speaker interval detection performance.*

MinL,s	Prec.	Rec.	F-score	Accuracy
0.5	0.64	0.83	0.71	0.90
0.7	0.71	0.78	0.75	0.92

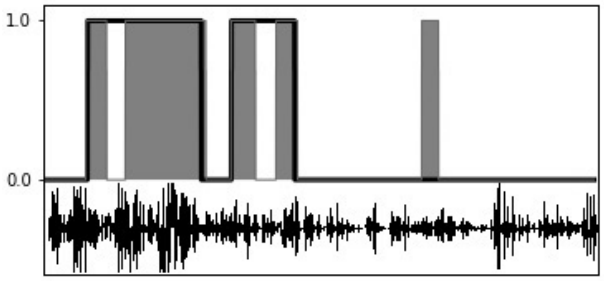

Figure 4: *Comparison of multi-speaker intervals labeled manually (black) and automatically - with the proposed algorithm (gray).*

3. Relation to prior work

As overlapped speech detection has a clear practical purpose, numerous studies have been conducted in this field. The approaches and data used in these studies are diverse.

Some works rely on multi-channel audio where it is possible to compare the signal from different channels or to perform acoustic beamforming or source separation [12, 13] combined, for example, with HMM [14]. But these methods are not applicable to one-channel data. However, an HMM-based approach was applied to one-channel data in [15], where the overlap detection error rate was equal to 78.3%, and in [16], where the researchers achieved 38% F-scores. Both of the studies were conducted using the AMI corpus. Chatlet and colleagues [17] studied how overlapping speech detection improved speaker diarization for ETAPE TV audio. They achieved a F1-measure of about 60% with overlapping speech detection systems relying on cepstral features and multi-pitch analysis.

The detection of multi-speaker fragments based on spectral harmonicity and envelope features with the use of Gaussian mixture models was presented in [7]. However, using neural networks instead of GMM turned out to be more effective. Although we have not yet examined the effect of noise level on our solution's performance, we have achieved a better performance (11.3% EER, the average SNR for the SSPNet Conflict Corpus

equals 8.4 dB) in comparison with the approximate 27% EER provided by [7]. On the SSPNet Conflict corpus [18] and [19] achieved the best conflict level detection results by implementing interruption detection in their systems.

Our approach requires fewer limitations than the majority of the existing works and may be applied to one-channel audio and an unknown number of speakers.

4. Discussion

Applying computer vision method to harmonic spectral traces allowed us to develop a solution for detecting multi-speaker intervals in continuous audio. Our method is speaker-independent and does not require any complex calculations. Based on previous researches in the area, we assumed that spectrograms should be useful for overlapped speech detection. We used histogram equalization and CLAHE to adjust image contrast. These methods, common for image enchantment [20, 21, 22], improved the algorithm performance.

The key point of data preprocessing for CNN training was that we picked out only voiced sounds. Pauses and unvoiced sounds, i.e. those produced without vocal cords vibration, were removed. This enabled us to manipulate more uniform data with observable harmonics presence. However, in the course of the evaluation procedure we had to take into consideration that fragments containing unvoiced sounds might have been detected incorrectly. In order to reduce the number of such mistakes we performed additional processing of the algorithm output and removed automatically detected intervals that contained unvoiced sounds or pauses. In the present work, this step was carried out with the use of Praat voiced-unvoiced labeling. The comparison between the EER values for the algorithm performance with and without this elimination confirmed our assumption. If more than 66% of the ADI consisted of unvoiced sounds and pauses, it was likely to be a false alarm, so such ADIs had to be ignored. In the future, this additional step will either be included in the main algorithm or become unnecessary in case the training set is enriched with the corresponding data.

The output of our solution is represented by CNN scores for short intervals which may be extended to word or phrase level depending on particular research goals. In the present work we evaluated our solution's performance by comparing it with manual diarization from SPPNet Conflict Corpus. Manual diarization usually involves long fragments like words or phrases. We tried to minimize the difference in time-resolution between manually and automatically detected intervals by means of additional processing of the algorithm results. Finally, we achieved high accuracy of multi-speaker fragments detection with F-score equal to 0.75.

It should be mentioned that the difference in time-resolution between manual and automatic detection might affect false alarm rate. With manual diarization short intervals of simultaneous speech may have been ignored. This means that some of the automatically detected intervals which were tagged as false alarms may in fact be short fragments of overlapped speech (like a meaningful exclamation of the other interlocutor in the background of continuous speech). In general, the duration of simultaneous speech sufficient for labeling an interval as a multi-speaker one strictly depends on the research purpose and data, and cannot be unified.

We consider our solution to be generally language-independent, as we achieved high accuracy despite the fact that the SSPNet Conflict Corpus was in French while TED Talks dataset was in English. However, it should be taken into account that the differences in voiced/unvoiced sounds ratio or in phoneme sequences may surely affect the accuracy of the results. Thus, it might be beneficial to enrich the training set with the material in different languages.

The performance of our algorithm turned out to be slightly better in those cases where female voices were overlapping (as compared with male voices or a male and a female voices). A possible explanation may lie in the fact that the fundamental frequency of male voices is lower on average, so the frequency range up to 1500 Hz includes more harmonics for male than for female voices. The density of traces on a spectrogram is higher for male voices, therefore it is harder to distinguish between single-speaker and multi-speaker patterns, and the number of CNN mistakes increases.

5. Conclusion

We presented a solution for multi-speaker intervals detection in continuous audio. Despite the method's simplicity, the accuracy of the results was higher than that reported in other studies for non-synthesized data. We expect our method to be helpful for high time-resolution speaker diarization, speaker tracking and other audio segmentation purposes. It can also be used in speech and social behavior analysis, for example, interruption rate estimation, conflict detection, etc.

Our method provides a way of spotting multi-speaker fragments with a minimum length of 0.5 seconds by using a CNN. The implementation is based on the detection of harmonic traces' irregular structure typical of simultaneous articulation of voiced sounds by multiple speakers.

The possible future steps may be to single out those intervals where voiced sounds overlap with unvoiced ones as a separate class for CNN training and to increase the size and language diversity of the training sample. Furthermore, it would be beneficial to combine the proposed method with the detection of background noises, laughter and hesitation in order to prevent the algorithm from misinterpreting these sounds.

6. Acknowledgements

This research was supported by Neurodata Lab LLC.

7. References

[1] K. Sonmez, L. Heck, and M. Weintraub, "Speaker Tracking and Detection With Multiple Speakers," in *Proc. 6th European Conference on Speech Communication and Technology (Eurospeech '99)*, vol. 5, 1999, pp. 2219–2222.

[2] A. F. Martin and M. A. Przybocki, "Speaker Recognition in a Multi-Speaker Environment," *Interspeech*, pp. 787–790, 2001.

[3] O. Cetin and E. Shriberg, "Speaker Overlaps and ASR Errors in Meetings: Effects Before, During, and After the Overlap," in *2006 IEEE International Conference on Acoustics Speed and Signal Processing Proceedings*, vol. 1. IEEE, pp. I–357–I–360.

[4] S. Kim, F. Valente, M. Filippone, and A. Vinciarelli, "Predicting Continuous Conflict Perception with Bayesian Gaussian Processes," in *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, 2014.

[5] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped Speech Detection for Improved Speaker Diarization in Multiparty Meetings," in *Proc. ICASSP*, 2008, pp. 4353–4356.

[6] N. Shokouhi, A. Sathyanarayana, S. O. Sadjadi, and J. H. L. Hansen, "Overlapped-speech detection with applications to driver assessment for in-vehicle active safety systems," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2013, pp. 2834–2838.

[7] N. Shokouhi, A. Ziaei, A. Sangwan, and J. H. L. Hansen, "Robust overlapped speech detection and its application in word-count estimation for prof-life-log data," no. 978, pp. 4724–4728, 2015.

[8] J. T. Geiger, F. Eyben, B. Schuller, and G. Rigoll, "Detecting overlapping speech with long short-term memory recurrent neural networks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013, pp. 1668–1672.

[9] V. Andrei, H. Cucu, and C. Burileanu, "Detecting overlapped speech on short timeframes using deep learning," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2017-August, 2017, pp. 1198–1202.

[10] N. Pappas and A. Popescu-Belis, "Combining Content with User Preferences for TED Lecture Recommendation," *In 11th Int. Workshop on Content Based Multimedia Indexing (CBMI), Veszprem*, 2013.

[11] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, 2002.

[12] K. Laskowski and T. Schultz, "Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings," in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, vol. 1, May 2006, pp. I–I.

[13] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, Sept 2007.

[14] T. Pfau, D. P. W. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the icsi meeting recorder," in *Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on*, 2001, pp. 107–110.

[15] J. T. Geiger, F. Eyben, N. Evans, B. Schuller, and G. Rigoll, "Using linguistic information to detect overlapping speech," in *INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, August 25-29, 2013, Lyon, France*, Lyon, FRANCE, 08 2013. [Online]. Available: http://www.eurecom.fr/publication/4020

[16] K. Boakye, O. Vinyals, and G. Friedland, "Two's a crowd: improving speaker diarization by automatically identifying and excluding overlapped speech," in *INTERSPEECH*, 2008.

[17] D. Charlet, C. Barras, and J. S. Linard, "Impact of overlapping speech detection on speaker diarization for broadcast news and debates," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 7707–7711.

[18] M.-J. Caraty and C. Montacié, *Detecting Speech Interruptions for Automatic Conflict Detection*. Cham: Springer International Publishing, 2015, pp. 377–401.

[19] F. Grèzes, J. Richards, and A. Rosenberg, "Let me finish: automatic conflict detection using speaker overlap," in *INTERSPEECH*, 2013.

[20] N. Bassiou and C. Kotropoulos, "Color image histogram equalization by absolute discounting back-off," *Computer Vision and Image Understanding*, vol. 107, no. 1-2, pp. 108–122, jul 2007.

[21] S. M. Pizer, E. P. Amburn, J. D. Austin, R. Cromartie, A. Geselowitz, T. Greer, B. ter Haar Romeny, J. B. Zimmerman, and K. Zuiderveld, "Adaptive histogram equalization and its variations," *Computer Vision, Graphics, and Image Processing*, vol. 39, no. 3, pp. 355–368, sep 1987.

[22] D. Anggraeni Pitaloka, A. Wulandari, T. Basaruddin, and D. Yanti Liliana, "Enhancing CNN with Preprocessing Stage in Automatic Emotion Recognition," in *Procedia Computer Science*, vol. 116, no. 00, 2017, pp. 523–529.