INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Detecting overlapped speech on short timeframes using deep learning

Valentin Andrei, Horia Cucu, Corneliu Burileanu

University "Politehnica" of Bucharest

valentin.m.andrei@gmail.com, horia.cucu@upb.ro, corneliu.burileanu@upb.ro

Abstract

The intent of this work is to demonstrate how deep learning techniques can be successfully used to detect overlapped speech on independent short timeframes. A secondary objective is to provide an understanding on how the duration of the signal frame influences the accuracy of the method. We trained a deep neural network with heterogeneous layers and obtained close to 80% inference accuracy on frames going as low as 25 milliseconds. The proposed system provides higher detection quality than existing work and can predict overlapped speech with up to 3 simultaneous speakers. The method exposes low response latency and does not require a high amount of computing power.

Index Terms: overlapped speech, auditory scene analysis, deep neural networks, deep learning.

1. Introduction

Overlapped speech is an effect that adds challenge to existing state-of-the-art speech analysis systems like speaker diarization and speech recognition. Being able to accurately tag each signal frame as overlapped or non-overlapped speech adds value but can also be a start point for more focused applications. Overlapped speech detection can be included in the computational auditory scene analysis (CASA) greater research field and can have interesting applications in forensics, blind source separation, ambient assisted living, smart surveillance or automated data mining.

In the last decade, there were several attempts to improve robustness of speech analysis systems to overlapped speech. In [1] – one of the first studies on this topic – the authors describe a classifier with three possible outputs: non-speech, overlapped speech and single source speech, designed to improve speaker diarization error. The system is based on HMM-GMM models, uses well established features like MFCC, RMS, LPC coefficients and improves diarization error with 7.4%. The method described in [2] goes up to 94% diarization accuracy when deciding between one speaker and two simultaneous speakers. However, the overlapped speech classification quality is not described. In [3], an SVM classifier with MFCC, spectral energy, voice quality metrics and prosody as features, is being used to improve speaker diarization but similarly, no separate assessment of overlapped speech detection accuracy is made.

In [4], identification of overlapped speech is done for in-vehicle safety enhancement. The system can classify only single source speech and two overlapping speakers and showcases that the error can be improved by not training the models with silence frames, as done in [1]. The method uses artificially mixed recordings starting from the TIMIT database, described in [5], and achieves an F-Score of almost 0.7. This is the highest claimed accuracy we were able to find and we think this is enabled by the precise annotation of the training and testing dataset. In [7], the authors use the method described in [4] for word-count estimation designed for automated data mining. Other studies focused on detecting overlapped speech were presented in [6], [8] and [9]. They use various classification and feature engineering techniques but the claimed error is high, fact that is motivated by the difficulty of perfectly annotating the signal frames on the selection of datasets that the authors used. One of the first studies that uses deep learning for detecting overlapped speech is presented in [10]. The method uses long-short term memory learning structures and an interesting combination of features like spectral flux and kurtosis along established features like MFCCs, LPCs, in-band energy, etc. The claimed *error* is 76% and the authors explain this is because of the unbalanced ratio between the number of overlapping and non-overlapped speech samples.

After reviewing the existing literature, we can argument that overlapped speech detection approaches can be extended to provide information like understanding how the signal frame length affects the classification accuracy or if it is possible to use the same techniques when more than two speakers are active. Also, we lack a standardized dataset focused solely on evaluating the overlapped speech classification quality. Since the amount of computing power has seen a great increase in the last years, researchers can focus more on techniques like deep learning for improving classification. In our paper, we aim to address several aspects mentioned in this paragraph.

In [11] – [13] we presented an algorithm for determining the number of active speakers in competing speaker environments. We presented two studies realized with more than 30 volunteers each, that helped us compare the accuracy of our method with human selective auditory attention capabilities. In the mentioned research, we focused on recordings where up to 10 speakers are simultaneously talking. We achieved an accuracy of approximately 75% for frames longer than 3.5 seconds. The increased frame duration can limit the adoption of the method in practical applications.

The current study focuses only on detecting overlapped speech in the presence of maximum 3 simultaneous speakers. We identified this as a reasonable practical threshold and is also similar with what human listeners can track efficiently as shown in [11] and [12]. Our most important goal is to be able to improve classification for short frames of up to 25 ms.

Section 2 describes the approach used for building the training and inference datasets and emphasizes the importance of being able to accurately tag training samples. In section 3 we describe the feature engineering techniques and the motivation for selecting each feature. Section 4 focuses on the deep learning model building and training methods. In section 5 and 6 we present the results and discuss some of the most important findings of this study.

http://dx.doi.org/10.21437/Interspeech.2017-188

2. Training and inference datasets

In [1], [9], [10] the authors opted to *train* their systems with the Augmented Multi-Party Interaction Corpus, known as the AMI Meeting Corpus ([14]). However, AMI Corpus was designed for improving meeting interaction and therefore lacks some important properties that serve the purpose of this paper. One of them is that there is no precise tagging of overlapped speech intervals. We tried to extract individual sources using blind speech separation but due to the placement of the microphones, we were not able to get reasonable quality. The voice activity detectors detected speech even for background voices and this made annotation difficult. Another challenging property was the noticeable room reverberation.

Because of this we decided to train our models with artificially mixed recordings where we can mark precisely the overlapped speech periods. A similar approach was used in [2] and [8] which resulted in higher accuracy and generalization performance. Romanian language was used for both training and inference datasets but we do not expect the language selection to influence the findings described in this paper.

2.1. Source mixing and tagging

Figure 1 illustrates how the mixing and the selection of frames was realized. Each single speaker source was passed through a voice activity detector – [15]. With the voice activity masks, we were able to control that the entire frame represents overlapped speech. For example, frame 1 was selected as overlapped speech. Frames 2, 3, 4 were dropped because not all speakers have shown voice activity throughout the *entire* frame duration. Frame 5 was selected as non-overlapped speech since there was only one speaker active. A normalization relatively to signal power was done.

All the silence frames were ignored. We did not use them as they have numerical properties that make converging more difficult and we need the model to focus on detecting subtle differences between the non-overlap and overlapped speech samples. In practice, a modern voice activity detector would be able to filter out silence frames with high accuracy.

2.2. Training dataset

We used only male speakers because we predicted that combining male and female voices would have improved artificially the accuracy. Recordings were collected from 10 speakers in a soundproof room with minimized reverberation. We used a noise cancelling microphone and a sampling frequency of 44.1 kHz. The speakers were instructed to read a paper on a random topic at moderate pace. The same recordings were used in the studies presented in [11] – [13].

In post-processing, we extracted frames with more than 99% speech activity and randomly combined them in mixtures with up to 3 simultaneous speakers. The number of non-overlap speech samples is equal to the number of overlapped-speech inputs as in training we want to avoid biasing the classifiers towards a certain class. By randomly combining frames from the recordings we can create a sufficient number of samples – we used 100000 for training.

2.3. Inference dataset

Existing studies estimate the quality of overlapped speech detection by measuring its impact over the accuracy of a

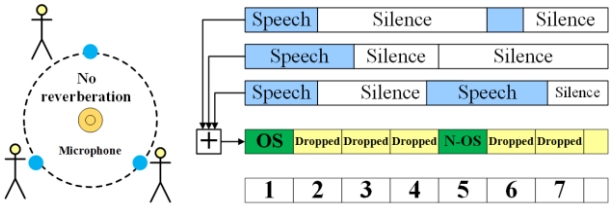

Figure 1: *Source mixing approach*

system with different purpose, like speaker diarization as in [1] and [6], speech analysis for assisted driving as presented in [4] or conflict detection described in [16] and [19].

We would like to evaluate speech overlap detection independently of other factors. We could not find a public speech corpus where recorded mixtures and individual speaker sources are available. This is mandatory for perfect annotation of overlapped speech periods and to ensure the classification accuracy is being measured correctly. Due to that, we used artificial mixed recordings. We used a new cohort of 5 male speakers that were not involved in recording the training dataset. Acquisition was done in identical conditions. We generated approximately 20000 samples for inference.

3. Features

A wide selection of features was used in prior work. Most often, we encounter the usage of MFCC like in [1], [4], [7], [8]. Other studies used spectral energy, loudness, in band energy, spectral flux, spectral kurtosis or voice derived prosody elements ([1] – [10]).

It is commonly accepted that deep learning models can yield satisfactory accuracy even when using unprocessed input like the time-based signal or frequency spectrum. In our paper, we selected the spectrum as the "raw" input and combined it with a set of processed features that during the experiments, demonstrated to improve detection accuracy.

3.1. Feature selection

It is important to note that even though we recorded at 44.1 kHz, we discovered that above 16 kHz the accuracy of the model stagnates, and therefore we resampled the inputs using 16 kHz. We used the set of features enumerated below. When the achieved accuracy was not satisfactory, we used also squared features. In some cases, this could lead to "overtraining" effects but we avoided this by increasing the size of the validation dataset used in each iteration of training.

- Signal's Magnitude Frequency Spectrum
- 12 MFCC + Log Energy + 0^{th} Cepstral Coefficient
- Signal Envelope computed with Hilbert Transform
- 12^{th} Order Auto-Regressive Model Coefficients

We used the FFT transform of the signal, from 0 to 4 kHz as we considered that it contains an important quantity of unprocessed information that can be "de-cyphered" by a complex deep learning system. We did experiments with spectrogram usage though no significant improvement was observed, while we had to pay the cost of the extra compute.

MFCC were selected as a feature as it was the dominant feature set in prior work. We investigated the usage of Delta

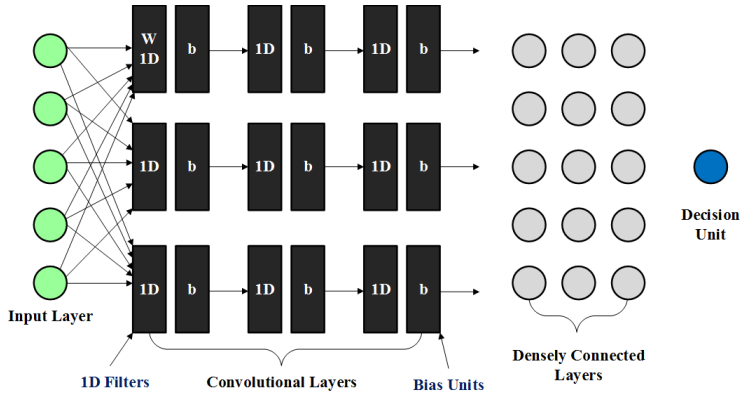

Figure 2: Neural network architecture

and Delta^2 coefficients and discovered that they do not add extra performance and impact the training process. Also, most of the derived coefficients were eliminated by running a Principal Component Analysis having the goal to select only the dominant features. We used the default parameters in computing the MFCC, as implemented in VOICEBOX.

We used the signal envelope as a feature because we expect that as the amount of overlap increases, the envelope tends to a flat shape. This was indeed a feature that improved the correct detection ratio. To compute the signal's envelope, we used the Hilbert transform.

Finally, we used per-frame 12th order AR coefficients as the 4th set of features because we observed that the numerical range of the resulting coefficients is rather generous and we speculated that this will amplify the subtle differences between overlapped and non-overlap speech frames. The coefficients were computed using 15ms windows.

3.2. Feature scaling and mean normalization

In order to speed-up the training algorithm, we used feature scaling and mean normalization. It is a well-known fact that if the numerical ranges between features differ significantly, the training algorithms tend to converge much slower. This strategy was necessary because for example MFCC values fall in a totally different range than for example the feature set derived from the signal's envelope.

In order to test inference fairly, we saved the row vectors of maximum, minimum and mean feature values computed on the training set and used them to scale and normalize for the evaluation dataset. This was done to asses even more the generalization capacity of the system.

4. Model and training strategy

As stated in the introduction of this paper, we aim to explore how does the accuracy of the overlap detection evolves with the length of the signal frame, with the ultimate goal to achieve good precision for 25ms timeframes. We will analyze 500ms, 100ms and 25ms window durations in the experiments presented in the next paragraphs.

One of the latest neural network architectures used for speech analysis is Deep Speech 2, presented in [17]. This classifier is known to improve phoneme recognition error rate over well-established frameworks like Google API, Bing Speech and Apple Dictation. We analyzed applying Deep Speech 2 for overlapped speech detection and decided that we need to simplify the architecture since we need a binary classifier that will not bring a significant impact on the performance of the application that uses it. We propose the neural network architecture illustrated in Figure 2. It encompasses convolutional layers and densely connected layers, with a final decision unit having 0.5 as threshold. All units use the sigmoid activation function. Table 1 summarizes some of the key design parameters of our proposed architecture.

Table 1: *Deep neural network design parameters*

Design Parameter	Range	Optimal
Number of convolutional layers	3-4	4
Number of 1D filters per conv. layer	5-30	20
Filter size on conv. layers	5-15	10
Number of densely connected layers	3-20	6
Units in dense layers / input size	1.1-2.0	1.5

4.1. Training strategy

We built our entire training and inference infrastructure using TensorFlow (www.tensorflow.org). Due to the high number of parameters associated with training, finding an optimal configuration is often challenging. Table 2 highlights the most important hyperparameters we used to train the model. As the parameters are interdependent, we had to do a grid search to determine the optimal values.

Stochastic Gradient Descent was selected as the model weights update method by using the *Momentum* optimizer implemented in TensorFlow and activating the *Nesterov Accelerated Gradient* presented in [18]. This is a technique that has almost become a standard in optimizing deep learning training. The learning rate was decayed across epochs to ensure convergence towards the end of the training, when the weights need to be updated in small steps.

Table 2: *Training hyperparameters*

Parameter	Range	Optimal
Learning rate	10^{-4} – 10^{-1}	10^{-3}
Momentum	0.8 – 0.95	0.9
Batch Size	32 – 800	430
Learning rate decay rate	0.9 – 0.99	0.99
Learning rate decay epochs	10 – 50	20

Table 3: *Feature selection per targeted case*

Frame Length	FFT	MFCC	AR	Envelope	Sq. Feat.
500ms	NO	YES	NO	NO	NO
100ms	NO	YES	YES	NO	YES
25ms	YES	YES	YES	YES	YES

Table 4: *Detection performance*

Frame Length	Detection Accuracy	F-Score	Precision	Recall
500ms	80.2%	0.8	0.81	0.78
100ms	79%	0.78	0.82	0.74
25ms	74.2%	0.72	0.77	0.68

5. Results

The standard measure for assessing the quality of a binary classifier is the F-Score, accompanied by the parent metrics Precision and Recall. In existing literature, the highest F-Score – collected in circumstances similar to our experiment – was reported in [4], with a value of 0.63. We improved the F-Score with up to 26% and speculate that it can be improved even further by adding more features to the classifier.

5.1. Discussion on features' selection

We used a different selection of features depending on the frame length of the acquisitioned signal. Table 3 summarizes the selection per each type of experiment.

If we use 500ms frames, there is a higher amount of information encapsulated in the input and by using the entire feature set, we end up in having to train a classifier with about 5000 features. This is why we used only MFCC.

For 100ms frames, we started with MFCC but the results were not satisfactory so we added AR coefficients and the squared features for both sets. With this we were able to improve performance to a similar level achieved for 500ms.

As the 25ms case is the most challenging, since a 25ms frame has few information, we needed the entire feature set along with the squared features.

5.2. Discussion on detection performance

Table 4 summarizes the detection performance for each of the 3 cases. As expected, the best overlapped speech detection quality is provided by the classifier that uses 500ms windows. This is expected mainly because the amount of information contained in 500ms is high. The result is consistent with findings presented in [11] and [12] where it is shown how counting competing speakers tends to be more accurate on longer analysis durations. Using 500ms frames may not be appealing for improving speaker diarization rate or phone recognition error in the presence of multiple speakers but can be used in other types of applications like for example, crowd sensing or aggression detection as presented in [16].

Using 100ms frames may be a reasonable compromise when having to select between 500ms and 25ms. The duration is sufficient enough to achieve an F-Score of 0.78, similar to what a 500ms frame classifier can provide, and small enough to be considered for applications like speaker diarization or blind source separation. In order to improve the classification performance, we increased the feature set size.

When using 25ms frames we have to expect a lower detection quality. We obtained an F-Score of 0.72 which is still with 11% better than the highest reported score in existing prior work. This frame duration comes with the speed benefit and the possibility of having the model integrated in a vast spectrum of applications that might benefit from overlapped speech detection.

We estimate that the F-Scores reported in Table 4 are not an upper limit and admit that a more complex model can yield improved accuracy. An interesting observation is that the Precision and Recall values are close which is an indicator that the system is not biased towards a class. With respect to training time, all the proposed models were trained usually in less than 6 hours. The 500ms model was the most time consuming since it uses the highest number of features.

6. Conclusions

This work demonstrates how the usage of deep learning can improve the detection of overlapped speech. By training with input produced by 10 speakers, the system is able to generalize well, with F-Scores higher than 0.72, when presented with inference data produced by totally different speakers. This indicates that the classifier is able to filter out the data related to individual voice characteristics and focus only on properties that are related to overlapped speech.

Our study shows that a combination of MFCC, signal's frequency spectrum, the auto-regressive model coefficients and speech signal envelope can be used to successfully estimate overlapped speech. For improved accuracy, we can also use the squared features. We do not claim that this is the optimal set of features but for our experiments, they provided compelling performance.

Another important conclusion is that higher frame durations result in increased detection accuracy. By adding new features, we can also improve quality on small frames up to a reasonable level.

We admit that live recorded mixtures can be preferred over artificially mixed recordings but we have a strong reasoning for using our approach. By being able to accurately tag input frames, without the usage of blind source separation algorithms, we can greatly help the training algorithm to focus only on the relevant data subtleties. Also, by recording in a soundproof room, we were able to eliminate reverberation effects that might bias the classifier.

A likely next step is to investigate how the proposed system can improve the performance of various state of the art applications like speaker diarization, blind speech separation, phone recognition or even crowd sensing applications like aggression detection. Collecting a database with live recorded mixtures would also be required at future points but currently the logistics of acquisitioning and perfectly annotating the datasets are challenging.

7. Acknowledgements

This work has been funded by the Romanian Government through the Executive Agency for Higher Education, Research, Development and Innovation Funding (UEFISCDI), program "Partnerships in priority areas", "Collaborative Applied Research Projects", project ID: PN-II-PT-PCCA-2013-4-0789, contract number 32/2014.

8. References

[1] K. Boakye, B. Trueba-Hornero, O. Vinyals, G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings", *ICASSP 2008 – IEEE International Conference on Acoustics, Speech and Signal Processing Proceedings*, 2008

[2] W. Tsai, S. Liao, "Speaker Identification in overlapped speech", *Journal of Information Science and Engineering*, 2010, pp. 1891-1903

[3] R. Vipperla, J. T. Geiger, et. al., "Speech overlap detection and attribution using convolutive non-negative sparse coding", *ICASSP 2012 – Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2012

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

[5] Garofolo, John, et al., "TIMIT acoustic-phonetic continuous speech corpus LDC93S1", *Philadelphia: Linguistic Data Consortium*, 1993.

[6] S. H. Yella, H. Bourlard, "Overlapped speech detection using long-term conversational features for speaker diarization in meeting room conversations", *IEEE/ACM Transactions on Audio, Speech and Language Processing*, December 2014, Vol. 22, No. 12

[7] N. Shokouhi, A. Ziaei, A. Sangwan, J. H. L. Hansen, "Robust overlapped speech detection and its application in word-count estimation for prof-life-log data", *ICASSP 2015 – Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2015

[8] S. A. Chowdhury, M. Danieli, G. Riccardi, "Annotating and categorizing competition in overlap speech", *ICASSP 2015 – Proceedings of International Conference on Acoustics, Speech and Signal Processing*, 2015, pp. 5136-5121

[9] J. T. Geiger, F. Eyben, et. al., "Using linguistic information to detect overlapped speech", *INTERSPEECH 2013 – 15th Annual Conference of the International Speech Communication Association Proceedings*, 2013

[10] J. T. Geiger, F. Eyben, B. Schuller, G. Rigoll, "Detecting overlapped speech with long short-term memory recurrent neural networks", INTERSPEECH 2013 – *14th Annual Conference of the International Speech Communication Association Proceedings*, 2013

[11] V. Andrei, H. Cucu, A. Buzo, C. Burileanu, "Detecting the number of competing speakers – human selective hearing versus spectrogram distance based estimator," *INTERSPEECH 2014 – 15th Annual Conference of the International Speech Communication Association Proceedings*, 2014, pp. 467 – 470

[12] V. Andrei, H. Cucu, A. Buzo, C. Burileanu, "Counting competing speakers in a timeframe – human versus computer", *INTERSPEECH 2015 – 16th Annual Conference of the International Speech Communication Association Proceedings,* 2015, pp. 3999-4003

[13] V. Andrei, H. Cucu, A. Buzo, C. Burileanu, "Estimating competing speaker count for blind speech source separation", *SPED 2015 – Proceedings of 8th Conference on Speech Technology and Human Computer Dialogue*, 2015, pp. 152–157

[14] J. Carletta, "Announcing the AMI meeting corpus", *The ELRA Newsletter 11(1)*, January–March 2006, p. 3-5

[15] Rainer Martin, "Noise power spectral density estimation based on optimal smoothing and minimum statistics", *IEEE Trans. Speech and Audio Processing*, July 2001, Vol. 9, pp. 504-512

[16] F. Grezes, J. Richards, A. Rosenberg, "Let me finish: automatic conflict detection using speaker overlap", INTERSPEECH 2013 – *14th Annual Conference of the International Speech Communication Association Proceedings*, 2013, pp. 200–204

[17] D. Amodei, S. Ananthanarayanan, et. al. "Deep Speech 2: end-to-end speech recognition in English and Mandarin", *Proceedings of the 33rd International Conference on Machine Learning*, 2016, JMLR: W&CP Vol. 48

[18] I. Sutskever, J. Martens, G. Dahl, G. Hinton, "On the importance of initialization and momentum in deep learning", *ICML 2013 – Proceedings of International Conference on Machine Learning*, 2013, pp. 1139-1147

[19] Lefter, I., Rothkrantz, L. J. M., & Burghouts, G. J. "A comparative study on automatic audio-visual fusion for aggression detection using meta-information", *Pattern Recognition Letters*, 2013, Vol. 34(15), pp. 1953-1963.