IMPROVED OVERLAP SPEECH DIARIZATION OF MEETING RECORDINGS USING LONG-TERM CONVERSATIONAL FEATURES

Sree Harsha Yella[1,2] and Hervé Bourlard[1,2]

[1] Idiap Research Institute, CH-1920 Martigny, Switzerland
[2] Ecole Polytechnique Fédérale de Lausanne, Lausanne, Switzerland
sree.yella@idiap.ch, herve.bourlard@idiap.ch

ABSTRACT

Overlapping speech is a source of significant errors in speaker diarization of spontaneous meeting recordings. Recent works on speaker diarization have attempted to solve the problem of overlap detection using classifiers trained on acoustic and spatial features. This paper proposes a method to improve the short-term spectral feature based overlap detector by incorporating information from long-term conversational features in the form of speaker change statistics. The statistics are obtained at segment level(around few seconds) from the output of a diarization system. The approach is motivated by the observation that segments containing more speaker changes are more probable to have more overlaps. Experiments on AMI meeting corpus reveal that the number of overlaps in a segment follows a Poisson distribution whose rate is directly proportional to the number of speaker changes in the segment. When this information is combined with acoustic information in an HMM/GMM overlap detector, improvements are verified in terms of F-measure and consequently, diarization error (DER) is reduced by 5% relative to the baseline overlap detector.

Index Terms— speaker diarization, spontaneous conversations, meetings, spontaneous overlapping speech.

1. INTRODUCTION

Speaker diarization, the task of inferring "who spoke when" in an audio recording has evolved over the years from broadcast news domain to spontaneous meeting recordings [1, 2]. Spontaneous conversations such as meetings have significant proportion of speech from simultaneous speakers (overlaps) which creates difficulties for automatic systems processing such data [3]. Studies on speaker diarization have shown that overlaps are one of the significant source of the errors [4, 5, 1, 6]. Presence of overlaps in clustering corrupts the models as they contain speech from multiple speakers and, this increases the clustering error. Also, since a typical diarization system hypothesizes only a single speaker at each instant of a recording, in the case of overlaps, this results in missed speech error for at least one speaker. Motivated by these factors, a mechanism to handle overlaps in meetings has been proposed in [7]. Several recent works on speaker diarization have explored various features and methods for detecting overlaps. In [8], authors explored various features such as energy and short-term spectral features (MFCC) for overlap detection. In [9, 10], authors investigated the use of spatial features estimated from time delay of arrival (TDOA) of speech using multiple distant microphones. Also, the use of prosodic features [11] has shown improvements over MFCC. Recently, convolutive non-negative sparse coding based methods have been explored for overlap detection with encouraging results [12, 13, 14].

All the above mentioned works are based on acoustic information and do not exploit higher level information in conversations which carries useful cues for overlap detection. Studies on meeting conversations have shown that overlaps are more likely to occur at some specific locations such as turn exchanges and back-channels [15] and 73% of overlaps occur at end of speaker turns [15]. In [16], authors have tried to address this issue by incorporating information from speech/silence statistics at segment level to improve overlap detection. Experiments on AMI meetings [17] have revealed that probability of occurrence of overlap in a segment is inversely proportional to the amount of silence in the segment [16] and incorporating this information into the feature based classifier has improved its performance and consequently reduced diarization error.

Current work performs a similar study to that of the work done in [16], but explores the usefulness of conversational features such as speaker change statistics to predict overlap in a segment. The statistics are computed from a long-term segment with a length of few seconds. The approach is motivated from our observation that segments containing more speaker changes tend to have more overlaps. Speaker changes have been used previously to deal with overlaps. In [18], a two-pass diarization system was used, where speech around speaker changes obtained from first pass was used to train an overlap model which was used in the second pass to handle overlaps. In the current work, instead of training an overlap model, we use speaker change statistics in a segment obtained from first pass of diarization to estimate the probability of overlap in that segment and these probabilities are incorporated into the baseline detector as prior probabilities. We perform experiments on meetings from AMI corpus and show that the proposed method improves overlap detection and consequently speaker diarization. The rest of the paper is organized as follows, section 2 presents briefly state-of-the-art baseline speaker diarization and overlap detection systems. Section 3 describes the proposed method for estimating the probability of single speaker speech and overlap; furthermore it proposes a way of incorporating them into baseline overlap detector. Section 4 describes the experimental results on overlap detection and speaker diarization and section 5 concludes the paper.

2. BASELINE SYSTEMS

2.1. Baseline speaker diarization system

Speaker diarization system used in the current work is based on a non-parametric bottom-up agglomerative framework [19]. The diarization output assigns each speech segment to a unique cluster (speaker) in the output. The system is evaluated using the metric known as the Diarization Error Rate (DER) which is the sum of speech/non-speech error and speaker error. Speech/non-speech er-

ror is the sum of miss and false alarm errors. Speaker errors are clustering errors happening whenever speech segments of a speaker are attributed to a different one. This metric has been used in several NIST Rich Transcription evaluation campaigns [2].

2.2. Baseline overlap detection system

Overlap detection is typically done using an HMM/GMM system with two states, one representing speech class (speech from a single speaker) and the other representing the overlap class (speech from multiple speakers) [8, 11]. The emission probabilities of the states are modelled by GMMs with diagonal covariance trained using 12 dimensional MFCC features and energy along with deltas. The features are mean and variance normalized. A minimum duration constraint is imposed on each HMM state. Furthermore, an overlap insertion penalty is introduced to control the trade-off between misses and false alarms [8, 11] which affect DER differently. The optimal value of the penalty is obtained by tuning on a separate data set. This system will be referred to as baseline overlap detector from here after.

Let V denote the sequence of single-speaker, overlapping speech states and X denote the sequence of acoustic features; the baseline overlap classifier infers the most probable sequence of states by Viterbi decoding as:

$$V^* = \arg\max_V P(V|X) = \arg\max_V P(X|V)P(V) \qquad (1)$$

The prior probability $P(V)$ can significantly change from one recording to another, as well as within the same recording (for instance presentations and monologues contain less overlap than discussions) [3]. But, in the baseline overlap detector, the prior probability of a class is fixed within and across the meetings, which is obtained based on the proportion of samples observed in each class in the training data. Current work tries to address this issue by estimating the prior probabilities from conversational features such as speaker change statistics from a long-term context.

3. OVERLAP DETECTION BY CONVERSATIONAL FEATURES

Studies on conversational analysis have shown that overlaps tend to occur more often at some specific parts of conversations than the remaining parts [15]. Especially it was shown that significant proportion of the overlaps occur during speaker turn changes [15]. Motivated by these studies, the current work analyzes the relationship between the occurrence of overlap in a segment and the number of speaker changes in the segment. Specifically, the study hypothesizes that overlap probability in a segment is directly proportional to the number of speaker changes in the segment i.e, segments containing more speaker changes are highly probable to have more number of overlaps than those having less changes. To verify this hypothesis, experiments are conducted on AMI meeting corpus [17] which contains multi-party meeting recordings collected in multiple meeting room environments. The corpus was divided into two halves of train and test set for the experiments. In the first experiment, the distribution of the number of overlaps is analyzed for different number of speaker changes in a segment of given length. In the current study, an occurrence of overlap is defined as a contiguous segment of overlapped speech surrounded by single-speaker speech or silence regions and the number of overlaps is obtained by counting such occurrences. Let sc, nov denote the variables indicating number of speaker changes and overlaps respectively in a segment

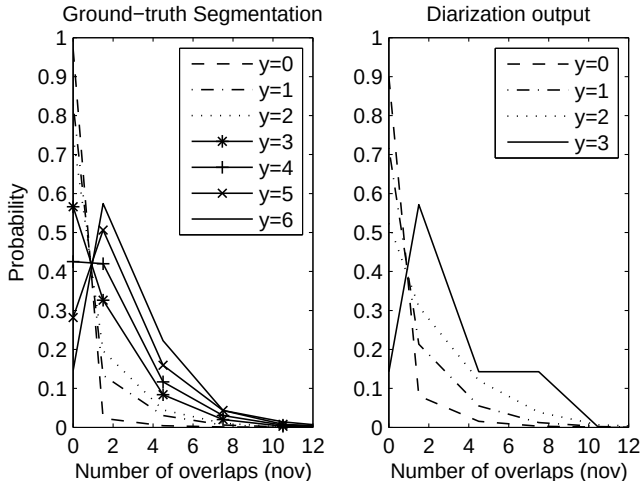

Fig. 1. *Probability distributions of number of occurrences of overlaps (nov) for different number of speaker changes (y) obtained using ground truth segmentation and diarization output.*

of given length, $n^s(sc = y)$ denotes the number of segments of length s seconds which contain y number of speaker changes and, $n^s(nov = k, sc = y)$ denotes the number of segments containing k number of overlaps and y speaker changes. Then, the probability of having k number of overlaps in a segment of length s seconds conditioned on the fact that it contains y number of speaker changes $P^s(nov = k|sc = y)$ can be obtained as,

$$P^s(nov = k|sc = y) = \frac{n^s(nov = k, sc = y)}{n^s(sc = y)} \qquad (2)$$

Fig. 1 shows the distribution ($P^s(nov|sc = y)$) of the number of overlaps in segments of length five seconds ($s = 5$) for different number of speaker changes (y). The speaker changes are obtained from the ground truth speaker segmentation and diarization output for left and right subplots respectively. It can be observed from Fig. 1 that, as the number of speaker changes increases, the probability of occurrence of more overlaps also increases. Also, it can be observed that distributions, $P^s(nov|sc = y)$ for different y seem to follow a Poisson distribution with a rate that is directly proportional to the number of speaker changes (y). Speaker changes in the diarization output are fewer when compared to ground truth speaker segmentation due to constraints and errors introduced by the automatic system. Nevertheless, similar phenomenon can be observed for distributions estimated from diarization output also. Fig. 1 supports our hypothesis that segments containing more speaker changes contain more overlaps. This information can be useful when incorporated into the baseline overlap detector which is based on acoustic features, as it does not contain evidence from the conversational patterns in the meetings.

Motivated from the empirical distributions in the Fig. 1, we model the probability of number of occurrences of overlaps in a given segment by a Poisson distribution whose rate depends on the number of speaker changes in the segment i.e.,

$$P^s(nov = k|y) = \frac{(\lambda_y^s)^k e^{-\lambda_y^s}}{k!} \qquad (3)$$

where, the rate parameter λ_y^s is a maximum likelihood estimate from the training set of meetings which is simply the mean of the number

of occurrences of overlaps in segments of length s seconds which contain y number of speaker changes. After estimating the set of rate parameters λ_y^s for different values of y, the probability of occurrence of overlap in a segment conditioned on the number of speaker changes in the segment can be obtained as,

$$P^s(ov|y) = 1 - P^s(nov = 0|y) \quad (4)$$

$$= 1 - e^{-\lambda_y^s} \quad (5)$$

and, the probability of single speaker speech can be obtained as,

$$P^s(sp|y) = 1 - P^s(ov|y) \quad (6)$$

$$= e^{-\lambda_y^s}. \quad (7)$$

To compute these statistics for the whole recording, the segment is progressively shifted by one frame at each step and probabilities $P^s(ov_i|y_i)$ and $P^s(sp_i|y_i)$ are estimated $\forall i$ where $i \in \{1 \dots N\}$ and N is the total number of frames in the file and y_i is the number of speaker changes in the segment centered around frame i. This process is depicted in Figure 2.

To verify how the estimated probabilities $P^s(ov_i|y_i), P^s(sp_i|y_i)$ generalize on test set of meetings different from those used for estimating the rate parameters, cross entropy between these estimated probabilities and test distribution was computed. The probabilities for the test distribution are obtained for each frame $i \in \{1 \dots N\}$ as follows, $P^t(ov_i) = 1, P^t(sp_i) = 0$ if the frame i is overlapped ($i \in \{OV\}$) and $P^t(sp_i) = 1, P^t(ov_i) = 0$ if the frame i is single-speaker speech ($i \in \{SP\}$). The knowledge of whether a frame i belongs to the set $\{OV\}$ or $\{SP\}$ is obtained from the ground-truth segmentation of the test set of meetings. The cross entropy between the test distribution and the estimated distribution is computed as follows.

$$C(s) = -\frac{1}{L}\left(\sum_{i\in\{OV\}} log(P^s(ov_i|y_i)) + \sum_{j\in\{SP\}} log(P^s(sp_j|y_j))\right) \quad (8)$$

where, L is the total number of frames used in the computation. To eliminate the bias in the measure due to uneven number of samples present in single-speaker speech and overlap classes, the cross entropy measure is computed by considering equal number of samples from each class.

Fig. 3 (left plot) shows cross entropy measure for test set of meetings from AMI meeting corpus when different segment lengths (s) are used for estimating the probabilities $P^s(nov = k|y)$. It can be observed from Fig. 3 (left plot) that the cross entropy reduces with increase in segment length (context around frame) thus indicating that the estimated probabilities are more closer to the real ones when the length of the context increases and reaches an optimum at the segment length of three seconds. We have also computed the cross entropy measure on meetings from RT09-eval set to verify whether the estimated statistics are corpus specific. The probability estimates $P(ov_i|y_i), P(sp_i|y_i)$ for RT09-eval set are estimated using the rate parameters $\{\lambda_y^s\}$ obtained from the training set of meetings in AMI corpus. Fig. 3 (right plot) plots the cross entropy measure for RT09-eval meeting set and shows similar trend to that of Fig. 3 (left plot). This shows that the estimated statistics are not specific to the AMI corpus and are generalizable to multi-party meetings in general.

Incorporating this information into the baseline HMM/GMM overlap detector described in (1) is straightforward. If $V = \{v_i\} = \{sp_i, ov_i\}$ stands for the sequence of states single-speech/overlap, $X = \{x_i\}$ the sequence of acoustic vectors and $Y = \{y_i\}$ the sequence of number of speaker changes in segment centered around

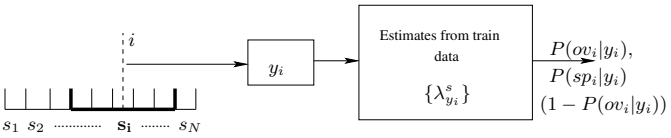

Fig. 2. *Estimation of probabilities of single-speech and overlap states for a frame i based on number of speaker changes y_i present in the segment s_i centered around the frame i.*

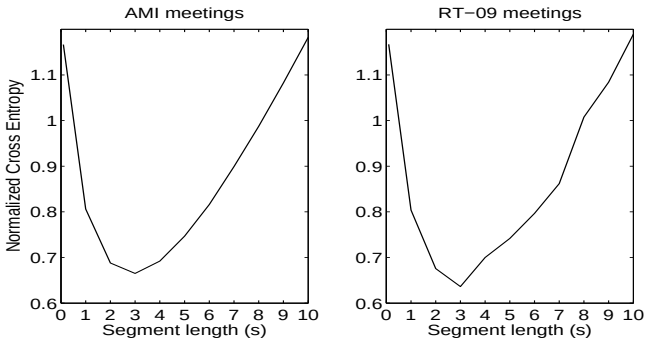

Fig. 3. *Cross entropy measure for various segment lengths on AMI meetings and RT09-eval set of meetings.*

frame i, the optimal single-speech/overlap segmentation can be obtained by Viterbi decoding as:

$$\arg\max_V P(V|X,Y) = \arg\max_V P(X|V,Y)P(V|Y)$$

$$\simeq \arg\max_V P(X|V)P(V|Y) \quad (9)$$

In (9) it is assumed that the observed features X are independent of number of speaker changes Y given the state V. In other words, the information from the acoustic features $P(X|V)$ is combined together with $P(V|Y)$ which estimates how probable an overlap is given the number of speaker changes Y in the segment. Furthermore $P(V|Y)$ is estimated from a long temporal window (three seconds) and thus includes information from long-term conversational features in the form of number of speaker changes in the window. In the current study the probabilities $P(V|Y)$ are estimated using the rate parameters obtained from the automatic diarization output.

4. EXPERIMENTS AND RESULTS

Overlap detection and speaker diarization experiments are conducted on meeting recordings in AMI meeting corpus [17]. The audio captured by multiple distant microphones is enhanced by beamforming using the *BeamformIt* toolkit [20]. Two disjoint sets of meetings for training and testing are created each consisting of 35 and 25 meetings respectively by random picking while the remaining meetings are used for estimating the rate parameters $\{\lambda_y^s\}$. Both the train and test sets contain recordings from all the meeting sites and ground truth speaker times obtained from ASR force-aligned manual transcriptions. The proposed method is compared to the baseline overlap detection system based on short-term spectral features in two tasks; overlapping speech detection and overlapping speech diarization.

4.1. Experiments on Overlap detection

Performances of the overlap detectors are compared in terms of Recall, Precision and F-measure. Figure 4 (a) plots the f-measures of the baseline overlap detector and the overlap detector incorporating speaker change statistics as a function of different overlap insertion penalties (OIP). It can be observed from Fig. 4 (a) that the system

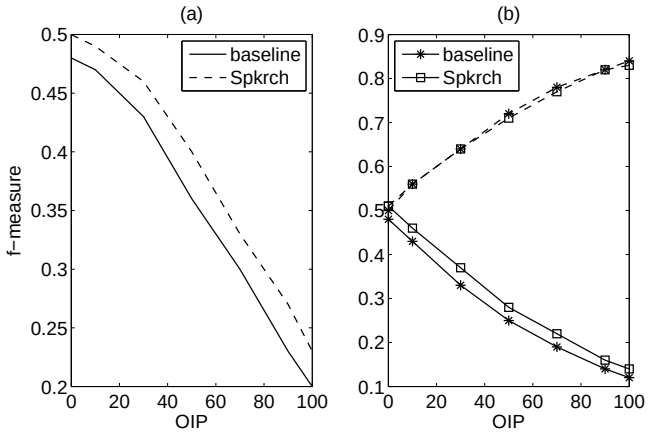

Fig. 4. *Performance of overlap detectors for different overlap insertion penalties (OIP). (a) F-measures of baseline detector, and detector based on speaker changes obtained from diarization output. (b) Precision (dashed line), Recall (solid line) for classifiers*

incorporating information about speaker changes has better performance than baseline system for all penalties. Furthermore Fig. 4 (b) plots the precision and recall for the two systems for different penalties. It can be observed that proposed approach improves the recall of the detector without any degradation in the precision which is desirable for overlap handling in speaker diarization. Similar improvements in overlap detection are observed in the case of RT09-eval meetings also.

4.2. Experiments on overlap speaker diarization

Once overlaps are detected, two strategies have been proposed in literature to handle it and are referred to as overlap exclusion and overlap labelling [7]. In overlap exclusion, prior to clustering, an overlap detection is performed and the detected segments are excluded from the clustering step in order to avoid GMM corruption. Once the final clustering is obtained, the excluded regions are assigned to a speaker by the Viterbi realignment decoder. Overlap exclusion reduces the total speaker error [7, 8, 11]. In overlap labelling, the handling happens after the diarization system is run by labelling the overlap segments with *two* speakers. This step can be performed according to two strategies: in the first one an overlap segment is assigned to the two nearest speakers in time [7], while in the second, they are assigned to two speakers with highest posterior probability in these regions [8]. Overlap labelling reduces the missed speech error [7, 8, 11].

Table 1 (first row) shows DER of 30.4 for the speaker diarization system without any overlap handling as described in [19]. Let us now compare the results obtained by the baseline overlap detector and the proposed system which exploits information from speaker changes on three tasks overlap exclusion, labelling and both. As proposed in the previous works [9, 8, 14, 16], two different operating points were chosen for exclusion and labelling tasks as they have different effects on DER. For exclusion a high recall point was chosen at the OIP of 0 and for labelling, a high precision point was chosen at the OIP of 90. The penalties were obtained by tuning on a separate development set of meetings and were kept constant for both systems. Overlap labelling for baseline and proposed method is done based on 2-nearest speaker strategy as proposed in [7]. It can be observed from Table 1 (third and fourth rows) that the proposed system has lower DER than the baseline system on all the three tasks. When both exclusion and labelling are done, the proposed method achieves about 5% relative DER reduction (from 25.5% to 24.2%). The improvement is particularly large in case of exclusion (from 26.2% to 24.6%). As the penalty term (OIP) is same for both the systems, the reduction in DERs can be attributed to the proposed incorporation of prior probability estimates from speaker change statistics.

Table 1. *DERs for various systems on test set with relative improvements over the diarization system with no overlap handling within parenthesis.*

No overlap handling		30.4	
System	Exclusion	Labelling	Both
Baseline	26.2 (13.8%)	29.7 (2.3%)	25.5 (16.1%)
Spkrch	24.6 (19.1%)	29.5 (2.9%)	24.2 (20.4%)

5. CONCLUSIONS AND FUTURE WORK

This paper proposed a method for estimating the probability of overlapping speech based on a longer context than a frame at a segment level, based on number of speaker changes in the segment. Speaker changes are obtained from automatic segmentation from the diarization output. Experiments on the AMI corpus revealed that the probability of having overlap in a segment is directly proportional to the number of speaker changes in it. Empirical distributions of number overlaps in a given segment have been shown to follow a Poisson distribution with a rate directly proportional to the number of speaker changes in the segment. The cross entropy measure revealed that probabilities estimated from a segment length of approximately three seconds minimizes the cross entropy on a separate test data set. The study also revealed that the estimated probabilities generalize to a completely different data set of meetings (RT09-eval). Furthermore, a method to include these statistics in a conventional HMM/GMM overlap detector by combining this information with acoustic features was proposed. Experimental results revealed that the proposed method outperforms the conventional overlap detector in terms of F-measure for all possible operating points. Whenever the detected overlap is used in speaker diarization for performing labelling and exclusion tasks, the DER is reduced by almost 5% relative to baseline feature based overlap detector from 25.5% to 24.2%.

As part of future work, we will explore other relevant conversational features which can be easily extracted automatically and which are correlated with overlaps in spontaneous meetings. Also, we will explore novel methods to combine estimates of the prior probabilities obtained from the conversational features such as the ones proposed in [16] and the current work to exploit any complementary information captured by them.

6. ACKNOWLEDGEMENTS

The authors thank Dr. Kofi Boakye for providing the force-aligned reference segmentations. This work was funded by the Swiss National Science Foundation through SNF-RODI and SNF-IM2 grant and by the EU through FP7 SSPnet grant.

7. REFERENCES

[1] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 356–370, feb. 2012.

[2] "http://www.itl.nist.gov/iad/mig/tests/rt/," .

[3] Ozgur Cetin and Elizabeth Shriberg, "Overlap in meetings: Asr effects and analysis by dialog factors, speakers, and collection site," in *3rd Joint Workshop on Multimodal and Related Machine Learning Algorithms*, Washington DC, USA, 2006.

[4] M. Huijbregts and C. Wooters, "The blame game: Performance analysis of speaker diarization system components," in *Interspeech*, Antwerp, Belgium, 2007, pp. 1857–1860.

[5] M. Huijbregts, D.A. van Leeuwen, and C. Wooters, "Speaker diarization error analysis using oracle components," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 2, pp. 393–403, feb. 2012.

[6] Mary Tai Knox, Nikki Mirghafori, and Gerald Friedland, "Where did i go wrong?: Identifying troublesome segments for speaker diarization systems," in *Interspeech*, Portland, USA, 2012.

[7] Scott Otterson and Mari Ostendorf, "Efficient use of overlap information in speaker diarization," in *ASRU*, Kyoto, Japan, 2007.

[8] Kofi Boakye, Oriol Vinyals, and Gerald Friedland, "Improved overlapped speech handling for speaker diarization," in *Interspeech*, Florence, Italy, 2011, pp. 941–943.

[9] Martin Zelenak, Carlos Segura, and Javier Hernando, "Overlap detection for speaker diarization by fusing spectral and spatial features," in *Interspeech*, Makuhari, Japan, 2010, pp. 2302–2305.

[10] Scott Otterson, *Use of Speaker Location Features in Meeting Diarization*, Ph.D. thesis, University of Washington, Seattle, 2008.

[11] Martin Zelenak and Javier Hernando, "The detection of overlapping speech with prosodic features for speaker diarization," in *Interspeech*, Florence, Italy, 2011, pp. 1041–1043.

[12] R. Vipperla, J.T. Geiger, S. Bozonnet, Dong Wang, N. Evans, B. Schuller, and G. Rigoll, "Speech overlap detection and attribution using convolutive non-negative sparse coding," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 4181–4184.

[13] J.T. Geiger, R. Vipperla, N. Evans, B. Schuller, and G. Rigoll, "Speech overlap detection using convolutive non-negative sparse coding: New improvements and insights," in *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, 2012, pp. 340–344.

[14] Jurgen Geiger, Ravichander Vipperla, Simon Bozonnet, Nicholas Evans, Bjorn Schuller, and Gerhard Rigoll, "Convolutive non-negative sparse coding and new features for speech overlap handling in speaker diarization," in *Interspeech*, Portland, USA, 2012.

[15] Elizabeth Shriberg, Andreas Stolcke, and Don Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversation," in *Eurospeech*, Aalborg, Denmark, 2001, pp. 1359–1362.

[16] Sree Harsha Yella and Fabio Valente, "Speaker diarization of overlapping speech based on silence distribution in meeting recordings," in *Interspeech*, Portland, USA, 2012.

[17] "http://corpus.amiproject.org/," .

[18] Marijn Huijbregts, David A. van Leeuwen, and Franciska M. G. de Jong, "Speech overlap detection in a two-pass speaker diarization system," in *Interspeech*, Brighton, United Kingdom, 2009, pp. 1063–1066.

[19] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 17, no. 7, pp. 1382–1393, 2009.

[20] "http://www.xavieranguera.com/beamformit/," .