Speaker Diarization of Overlapping Speech based on Silence Distribution in Meeting Recordings

Sree Harsha Yella[1,2], *Fabio Valente*[1]

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

Abstract

Speaker diarization of meetings can be significantly improved by overlap handling. Several previous works have explored the use of different features such as spectral, spatial and energy for overlap detection. This paper proposes a method to estimate probabilities of speech and overlap classes at a segment level which are later incorporated into an HMM/GMM baseline system. The estimation is motivated by the observation that significant portion of overlaps in spontaneous conversations take place where the amount of silence is less, e.g., during speaker changes. Experiments on the AMI corpus reveal that the probability of occurrence of overlap in a segment is inversely proportional to the amount of silence in it. Whenever this information is combined with acoustic information from MFCC features in an HMM/GMM overlap detector, improvements are verified in terms of F-measure. Furthermore the paper investigates the use of exclusion and labelling strategies based on such detector for handling overlap in diarization reporting F-measure improvements from 0.29 to 0.43 in case of exclusion and from 0.15 to 0.22 in case of labelling. Consequently speaker diarization error is reduced by 8% relative compared to the baseline based solely on acoustic information.

Index Terms: speaker diarization, meeting recordings, diarization error, spontaneous overlap speech.

1. Introduction

Speaker diarization is the task of inferring "who spoke when" in an audio recording. When diarization is performed on spontaneous conversations such as meeting room recordings, significant amount of errors are due to speech from simultaneous speakers (overlap speech) [1, 2, 3]. Studies on meeting corpora have shown that significant proportion of speech is overlapped and thus diarization and ASR in spontaneous conversations have to deal with overlaps in an effective manner [4] to avoid high errors. Speaker diarization studies have also shown that effective handling of overlap speech can largely reduce the diarization error [5] and several recent works have dedicated considerable effort to this problem. In [6], authors explored various features such as energy and short-term spectral features (MFCC) for overlap detection. In [7, 8], authors investigated the use of spatial features estimated from time delay of arrival (TDOA) of speech using multiple distant microphones. Recently, the use of prosodic features [9] has shown improvements over MFCC. All the above methods use features that are frame level estimates and do not incorporate information from long term context into the detection system.

Studies on meeting conversations have shown that overlaps are more likely to occur at some specific locations such as turn exchanges and back-channels [10] and 73% of overlaps occur at end of speaker turns [10]. This paper proposes a method to estimate the probability of overlap speech in a conversation (a meeting recording) based on a longer context than a frame (at segment level) and incorporate those estimates into a baseline overlap detection system to improve its performance. The method makes use of the relation between single speaker speech, silence and overlap within a segment and is based on the observation that a significant portion of overlaps occurs in regions with small amount of silence, e.g., speaker turn changes. On the other hand, parts of the conversation with monologues contain little overlaps and also contain more silence due to speaker pauses. An example supporting this observations is explained in Figure 1. Therefore, we hypothesize that presence of low amount of silence in a segment is an indicator of presence of overlap within that segment. As silence is easier to detect compared to overlap speech, silence statistics can be used to estimate probability of overlap within the segment.

We verify this hypothesis on meetings from AMI corpus and show that the proposed method improves overlap detection and consequently speaker diarization. Rest of the paper is organized as follows, Section 2 presents briefly state-of-the-art baseline speaker diarization, overlap detection systems and the overlap handling methods. 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. Speaker diarization and overlap

The diarization process starts with speech activity detection (SAD) based on HMM/GMM system described in [11]. After this the speech segments detected are uniformly segmented and agglomeratively clustered until stopping criterion is met. The diarization output assigns each speech segment to a unique cluster (speaker) in the output (see [12] for details). The system is evaluated according to the Diarization Error Rate (DER) which is the sum of speech/non-speech error and speaker error. Speech/non-speech error 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 [13].

Previous works [5, 1, 2] have shown that overlap speech regions degrade speaker diarization in two ways. When overlap segments are included in the agglomerative clustering, GMM models are corrupted thus producing an increase in speaker er-

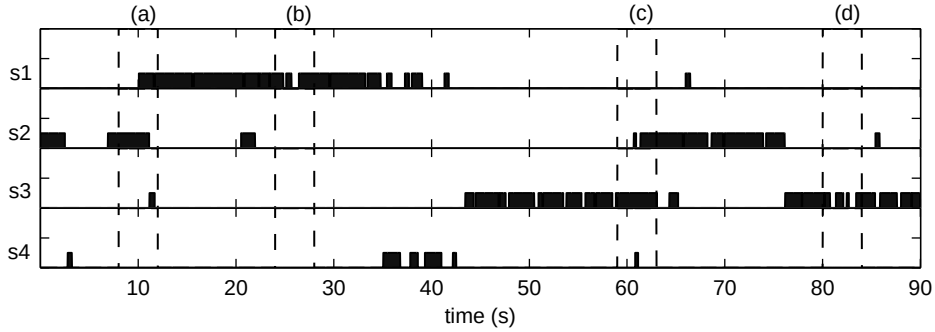

Figure 1: *Speaker vocalizations from a snippet of multi-party conversation. The fixed length segments (a) and (c) are in regions of speaker change and contain overlap whereas segments (b) and (d) contain single speaker speech. It can be observed that duration of silence within segments a,c is significantly less when compared to that in b,d.*

ror. Furthermore, as conventional diarization systems output a single speaker for each time instant, whenever overlap regions are scored, an increase in the missed speech error is verified. Overlap handling is addressed performing three steps: an initial detection, followed by exclusion and labelling tasks described below.

2.1. Baseline overlap detection and handling system

Overlap detection is typically obtained using an HMM/GMM system with two states, one representing speech class (speech from a single speaker) and the other representing the overlap class [6, 9] (speech from multiple speakers). 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 (see [6, 9]) which affect DER differently. The optimal value of the penalty is obtained by tuning on a separate data set. This system will be referred as baseline overlap detector from here after. In summary, let V denote the sequence of single-speaker speech, overlap-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:

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

Prior probabilities of single-speaker speech and overlap-speech are represented in Equation (1) by the term $P(V)$. In the AMI corpus, approximately 18% of speech is overlapped. This value is an average over the entire corpus and obviously 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) [4].

Once overlap speech is detected, two strategies have been proposed to handle it and are referred as overlap exclusion and overlap labelling [5].

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 [5, 6, 9].

Overlap labelling: In this case, 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 [5], while in the second, they are assigned to two speaker with highest posterior probability in these regions [6]. Overlap labeling reduces the missed speech error [5, 6, 9].

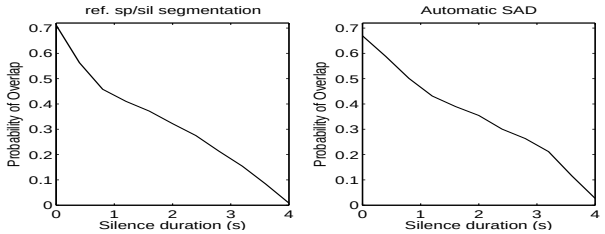

Figure 2: *Probability of overlap based on silence duration obtained using ground truth speech/sil segmentation and automatic SAD output.*

3. Overlap detection based on silence distribution

As previously described, the overlap detection starts with a speech/silence segmentation followed by a single-speaker speech/overlap detection. Under the rationale that, the statistics of silence, single speaker speech and overlap during a conversation are related to one another(see [14]), this work investigates how the amount of single-speaker speech and overlap relates to the amount of silence in a segment. The study is carried on two disjoint subsets of AMI meeting corpus one for training and the other for testing. The ground-truth segmentation is obtained by force-aligning the manual segmentation.

Consider a short segment of conversation with a duration D frames. Let us designate with $n(sl = x)$ the total number of segments which contain x frames of silence and with $n(ov, sl = x)$ the number of segments which contain x frames of silence and contain an overlap between speakers. It is possible to estimate the probability of having overlap within a segment conditioned on the amount of silence in that segment as:

$$P(ov|sl = x) = n(ov, sl = x)/n(sl = x), \qquad (2)$$

Figure 2 (left plot) shows $P(ov|sl = x)$ conditioned on the value of x for a segment of four seconds, i.e., $D = 400$ frames. It can be noticed that the probability of having overlap in a segment is inversely proportional to the amount of silence. When the amount of silence is zero, the probability of having an overlap in the segment is 0.7. In other words, it is possible to estimate the probability of having an overlap in a segment by the amount of silence in it. This information is potentially useful as speech/silence detection is a simpler task compared to single-speech/overlap detection. In order to verify

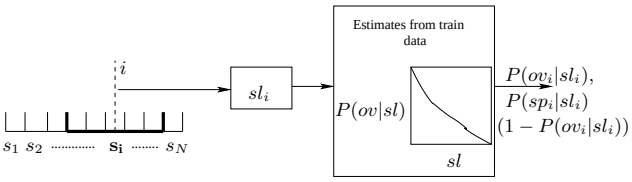

Figure 3: *Estimation of probabilities of single-speech and overlap states for a frame i based on duration of silence sl_i present in the segment s_i centered around the frame i.*

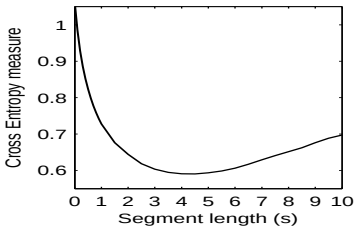

Figure 4: *Cross entropy measure for various window lengths.*

if this conclusion also holds in case of automatic speech/non-speech segmentation, the previous statistics are recomputed using the output of the automatic segmenter and plotted in Figure 2 (right plot) showing exactly similar trends. The probability of a single-speaker speech within a segment can be obtained as $P(sp|sl = x) = 1 - P(ov|sl = x)$. In order to compute these statistics for the whole recording, the segment is progressively shifted by one frame at each step and $P(ov_i|sl_i), P(sp_i|sl_i)$ are estimated $\forall i$ where $i \in \{1 \dots N\}$ and N is the total number of frames in the file. This process is depicted in Figure 3.

Let us now investigate how the statistics $P(ov|sl = x)$ and $p(sp|sl = x)$ generalize to a test set different from the one used for their estimation. In order to do this, the cross entropy between those estimates and the test distribution (P_t) obtained from ground-truth segmentation is measured. The probabilities for the test distribution are obtained for each frame $i \in \{i \dots N\}$ as follows, $P_t(ov_i) = 1, P_t(sp_i) = 0$ if the frame i is overlapped and $P_t(sp_i) = 1, P_t(ov_i) = 0$ if the frame i is single speaker speech. Then the cross entropy between the test distribution and the estimated distribution is computed as follows.

$$C = -\frac{1}{L}\left(\sum_{i\in\{OV\}} log(P(ov_i|sl_i)) + \sum_{j\in\{SP\}} log(P(sp_j|sl_j))\right)$$

where L is total number of frames used in the estimation, $\{OV\}$ is the set of frames in overlap regions and $\{SP\}$ is the set of frames in the single-speech regions. Figure 4 plots C as a function of various segment lengths D. It is important to notice that for $D = 1$, the single-speech/overlap-speech statistics reduce to the frame based statistics, and for $D > 1$, those statistics include information from longer time spans of conversation. Figure 4 reveals that segment lengths longer than one frame reduce the cross-entropy thus the statistics from the training set generalize to the test set. Furthermore the optimal segment length, i.e., the one that minimizes the cross entropy is approximatively 400 frames, i.e., 4 seconds.

Incorporating this information into the baseline HMM/GMM overlap detector described in Equation (1) is straightforward. Let us designate with $V = \{v_i\} = \{sp_i, ov_i\}$ the sequence of states single-speech/overlap, with $X = \{x_i\}$ the sequence of acoustic vectors and with $SL = \{sl_i\}$ the sequence of silence durations contained in segments centered around frame i. The optimal single-speech/overlap segmentation can be obtained by Viterbi decoding as:

$$\begin{aligned}\arg\max_V P(V|X, SL) &= \arg\max_V P(X|V, SL)P(V|SL) \\ &\doteq \arg\max_V P(X|V)P(V|SL) \quad (3)\end{aligned}$$

In Equation (3) it is assumed that the observed features (X) are independent of amount of silence (SL) given the state V. In other words, the information from the acoustic features $P(X|V)$ is combined together with $P(V|SL)$ which estimates how probable an overlap is given a certain amount of silence in the segment. Furthermore $P(V|SL)$ is estimated from a long temporal window (four seconds) and thus includes information from surrounding speech/non-speech estimates. $P(X|V)$ is a probability density function (a GMM) and $P(V|SL)$ is given by probabilities $P(sp_i|sl_i)$ and $P(ov_i|sl_i)$, thus a scaling factor tuned on an independent data set is introduced to bring them in comparable ranges. From here on, we will refer to the proposed method as overlap detector based on silence statistics.

4. Experiments and Results

Experiments are conducted on meeting recordings in AMI meeting corpus [15]. The corpus consists of about 100 hours of meeting recordings captured using multiple distant microphones at multiple sites. The audio signals are enhanced by beamforming using *BeamformIt* toolkit [16]. Two disjoint sets for training and testing are created each consisting of 35 and 20 meetings respectively by randomly picking while the remaining meetings are used for estimating the probabilities $P(V|SL)$. 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 differences between the baseline overlap detector and the proposed method are compared in two tasks: overlap detection and overlapping speech diarization. $P(V|SL)$ are estimated based on statistics computed using automatic speech/silence segmentation, as Figure 2 shows that the estimates are similar for both reference and automatic segmentations.

4.1. Experiments on Overlap detection

Performances of the overlap detectors are compared in terms of Recall, Precision and F-measure. Figure 5 (a) plots the f-measures of the baseline overlap detector and the overlap detector incorporating silence statistics as a function of different overlap insertion penalties (OIP). It can be observed from

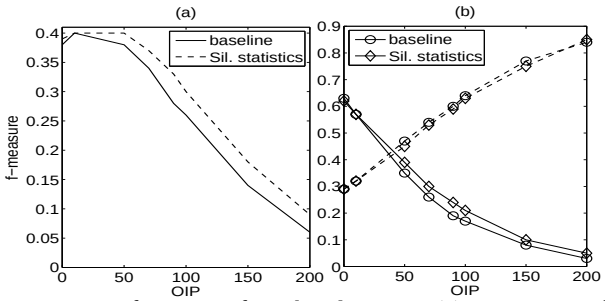

Figure 5: *Performance of overlap detectors. (a) F-measures of baseline detector, and detector based on silence statistics estimated based on automatic SAD. (b) Precision(dashed line), Recall(solid line) for classifiers*

Fig. 5(a) that the system incorporating silence statistics has better performance than baseline system for all penalties. Furthermore Fig. 5(b) plots the precision and recall for the two systems for different penalties. It can be observed that incorporation of silence statistics improves the recall.

4.2. Experiments on overlap speaker diarization

Table 1 (first line) shows DER for the speaker diarization system without any overlap handling as described in [12] which is 29.9. To get an estimate of the maximum possible improvements obtained by overlap handling, Table 1 (third line) reports the performance of labelling and exclusion methods whenever oracle overlap speech (from the reference segmentation) is used.

Let us now compare the results obtained by the baseline overlap detector and the proposed system that incorporates silence statistics on three tasks overlap exclusion, labelling and both. Overlap labelling for baseline and proposed method is done based on 2-nearest speaker strategy proposed in [5]. The improvements obtained by the baseline detector are similar to those reported in previous works [6, 9]. It can be observed from Table 1 (fourth and fifth line) 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 8% relative reduction in DER (from 26.2% to 24.3%). The improvement is particularly large in case of exclusion (from 26.8% to 25.1%), where the proposed method performs as good as the oracle.

Let us now compare the two approaches in terms of F-measure. As the operating point for overlap detectors are selected by minimizing the DER on a separate train set [9, 6], different operating points are used for exclusion and labelling. The F-measures at the operating points chosen for baseline system and the proposed method are reported in Table 2 showing improvements from 0.29 to 0.43 for the exclusion and from 0.15 to 0.22 for the labeling. As insertion penalties are same in both cases, the gain in the f-measure can be attributed to the proposed incorporation of silence statistics into the classifier.

Table 1: *DERs for various systems on test set using with relative improvements over baseline within parenthesis.*

No overlap handling		29.9	
System	Exclusion	Labelling	Both
Oracle	25.1 (16.1%)	18.9 (36.8%)	15.0 (49.8%)
Baseline	26.8 (10.4%)	29.3 (2%)	26.2 (12.3%)
Silence stats	25.1 (16.1%)	29.1 (2.7%)	24.3 (18.7%)

Table 2: *F-measures for the overlap detectors on test set at the operating points used for speaker diarization*

task	baseline	Silence Statistics
Exclusion	0.29	0.43
Labelling	0.15	0.22

5. Conclusions

Speaker diarization of spontaneous conversations like meetings is seriously affected by overlap speech. This problem has been widely addressed using signal processing approaches discarding the fact that meetings are spontaneous conversations and overlap occurs in particular moments for instance when several speakers are competing to talk at the same time. Several works have shown that during conversations silence, single-speaker speech and overlap speech are related to each other [17] and present patterns that can be modeled [14].

This paper proposed a method for estimating the probability of overlap speech based on a longer context than a frame at a segment level based on the amount of silence in the segment. As speech/silence detection is easier compared to single-speech/overlap detection, silence statistics can be used as auxiliary information during the overlap detection task.

Experiments on the AMI corpus revealed that the probability of having overlap in a segment is inversely proportional to the amount of silence in it. Cross-entropy measure revealed that silence statistics from a segment length of approximately 400 frames (4 seconds) minimizes the cross-entropy on a separate test data set. Furthermore the paper proposed a method to include these statistics in a conventional HMM/GMM overlap detector by combining this information with acoustic features.

Experiments on the AMI corpus revealed that the proposed method outperforms the conventional overlap detector in terms of F-measure for all the possible operating points. Whenever the detected overlap is used in speaker diarization for performing labelling and exclusion tasks, the DER is reduced by almost 8% relative from 26.2% to 24.3%. F-measure in overlap detection improved from 0.29 to 0.43 for the exclusion task and from 0.15 to 0.22 for the labelling task.[1]

6. References

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

[2] M. Huijbregts, D. van Leeuwen, and C. Wooters, "Speaker diarization error analysis using oracle components," *IEEE TASLP*, vol. 20, no. 2, pp. 393–403, feb. 2012.

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

[4] O. Cetin and E. 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.

[5] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *ASRU*, Kyoto, Japan, 2007.

[6] K. Boakye, O. Vinyals, and G. Friedland, "Improved overlapped speech handling for speaker diarization," in *Interspeech*, Florence, Italy, 2011, pp. 941–943.

[7] M. Zelenak, C. Segura, and J. Hernando, "Overlap detection for speaker diarization by fusing spectral and spatial features," in *Interspeech*, Makuhari, Japan, 2010, pp. 2302–2305.

[8] S. Otterson, "Use of speaker location features in meeting diarization," Ph.D. dissertation, University of Washington, Seattle, 2008.

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

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

[11] M. Huijbregts and F. de Jong, "Robust speech/non-speech classification in heterogeneous multimedia content," *Speech Communication*, vol. 53, no. 2, pp. 143–153, 2011.

[12] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE TASLP*, vol. 17, no. 7, pp. 1382–1393, 2009.

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

[14] K. Laskowski, M. Osterdorf, and T. Schultz, "Modeling vocal interaction for text-independent participant characterization in multi-party convrersation," in *9th ISCA/ACL SIGdial*, Columbus, USA, 2008, pp. 148–155.

[15] "http://corpus.amiproject.org/."

[16] "http://www.xavieranguera.com/beamformit/."

[17] O. Cetin and E. Shriberg, "Analysis of overlaps in meetings by dialog factors, hot spots, speakers, and collection site: Insights for automatic speech recognition," in *ICSLP*, Pittsburgh, USA, 2006, pp. 293–296.

[1]The authors thank Dr. Kofi Boakye for providing the force-aligned reference segmentation. 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.