The Influence of Speech Activity Detection and Overlap on Speaker Diarization for Meeting Room Recordings

Corinne Fredouille[1] *and Nicholas Evans*[1,2]

[1]LIA, University of Avignon (France), [2]University of Wales Swansea (UK)

(corinne.fredouille,nicholas.evans)@univ-avignon.fr

Abstract

This paper addresses the problem of speaker diarization in the specific context of meeting room recordings which often involve a high degree of spontaneous speech with large overlapped speech segments, speaker noise (laughs, whispers, coughs, etc.) and very short speaker turns. A large variability in signal quality has brought an additional level of complexity. This paper investigates the effects of speech activity detection and overlapped speech through speaker diarization experiments conducted on the NIST RT'05 and RT'06 data sets. Results indicate that our system is highly sensitive to the shape of the initial segmentation and that, perhaps surprisingly, perfect references can even degrade performance. Finally we propose a direction for future research to incorporate confidence values according to acoustic attributes in order to unify what is currently a somewhat disjointed approach to speaker diarization.

Index Terms: speaker diarization, meeting room, speech activity detection, overlapped speech.

1. Introduction

The speaker diarization task is an especially important contribution to the overall Rich Transcription (RT) paradigm, as evidenced by the RT evaluation campaigns administered by the National Institute of Standards and Technology (NIST). Also known as "Who spoke When", the speaker diarization task consists in detecting the speaker turns within an audio document (segmentation task) and in grouping together all the segments belonging to the same speaker (clustering task). Applied initially to conversational telephone speech and subsequently to broadcast news, the current focus is on meeting room recordings, a task which poses a number of new challenges. Meeting room recordings often involve a greater degree of spontaneous speech with large overlapped speech segments, speaker noise (laughs, whispers, coughs, etc.) and very short speaker turns. Due to the availability of many different recording devices and room layouts, a large variability in signal quality has brought an additional level of complexity to the speaker diarization task and more generally to the RT domain. This paper addresses the two most immediate and specific difficulties associated with meeting room recordings: non-speech event detection and overlapped speech.

Speech activity detection (SAD) is an intrinsic process of speaker diarization systems. The complexity of SAD tends to increase when applied to the meeting room environment due to the inherent variability in signal quality. As dictated by standard NIST protocols the performance of SAD systems is assessed in terms of the missed speaker error rate (speakers labelled as non-speech events) and the false alarm speaker error rate (non-speech events labelled as speakers) [1]. False alarms may influence the speaker diarization process since mis-classified non-speech events will contaminate the process as potential speakers. On the other hand, an under-zealous SAD process which identifies fewer and smaller speech segments may not yield sufficient data to reliably train each speaker's model and in turn may also adversely affect the behaviour of the speaker diarization system.

Overlapped speech is one of the main characteristics of spontaneous speech, and consequently of meeting room recordings. Very few researchers have studied overlapped speech, especially in meeting room environments [2, 3]. This is changing rapidly however as, for the first time, the most recent NIST RT'07 evaluation campaign contains an overlapped speech detection task. Since current state-of-the-art automatic systems are not currently capable of detecting overlapped speech, they are systematically involved in the speaker diarization process in the same manner as single-speaker segments. Therefore, it can be assumed that multi-speaker segments may disturb the statistical modelling processes involved in classical speaker diarization for both the speaker turn detection and clustering tasks.

This paper presents an original assessment on the impact of speech activity detection and overlapping speaker segments on a state-of-the-art speaker diarization system. The remainder of this paper is organised as follows: Section 2 presents the baseline E-HMM based speaker diarization system that was used for the experimental work reported in this paper. Section 3 defines the experimental protocol (corpora and evaluation metric). Sections 4 and 5 report experiments which assess the effects of non-speech event detection and overlapped speech. Finally, our conclusions are presented in Section 7.

2. Speaker diarization system

This work is concerned with the multiple distant microphones located on the meeting room tables (MDM task of the RT'05 and RT'06 evaluation plans [1]). For all experimental work reported in this paper, the different channels are processed very simply by summing related signals in order to yield a unique virtual channel which is used in all subsequent stages. No interchannel delay compensation is performed.

The LIA speaker diarization system was developed using the open source ALIZE speaker recognition toolkit [4] and is composed of 4 main steps:

- Speech/non-speech detection
- Pre-segmentation
- Speaker segmentation and clustering
- Post-normalisation and resegmentation

2.1. Speech/non-speech detection

The speech activity detection (SAD) algorithm employs feature vectors composed of 12 un-normalised Linear Frequency Cepstrum Coefficients (LFCCs) plus energy augmented by their first and second derivatives. It utilises an iterative process based on a Viterbi decoding and model adaptation applied to a two-state HMM, where the two states represent speech and non-speech events respectively and are initialised with a 32-component GMM model trained on separate data using an EM/ML algorithm. State transition probabilities are fixed to 0.5. Finally, some duration rules are applied in order to refine the speech/non-speech segmentation yielded by the iterative process.

2.2. Pre-segmentation

The pre-segmentation phase aims to provide an approximate speaker turn labelling to initialise and speed-up the subsequent segmentation and clustering stages. Now the signal is characterised by 20 LFCCs, computed every 10ms using a 20ms window. The cepstral features are augmented by energy but no feature normalisation is applied at this stage. A classical generalised likelihood ratio (GLR) criterion-based speaker turn detection is applied to two consecutive 0.5-second-long windows with a 0.05 second shift (single diagonal matrix Gaussian components). Once speaker turns are detected, a clustering process is applied in order to group together successive segments that are deemed to be sufficiently similar according to a thresholded GLR criterion.

2.3. Speaker segmentation and clustering

This step is the core of the LIA speaker diarization system. It relies on a one-step segmentation and clustering algorithm in the form of an evolutive hidden Markov model (E-HMM) [5]: each E-HMM state aims to characterise a single speaker and the transitions represent the speaker turns.
This process, still based on 20 LFCCs plus energy coefficients, can be defined as follows:
1. Initialisation: The HMM has only one state, called L_0. A world model with 128 Gaussian components is trained on the entire audio show. The segmentation process is initialised with the segmentation outputs issued from the pre-segmentation stage which are utilised for the selection process.
2. Speaker addition: a minimum 3-second-long candidate segment is selected among all the segments belonging to L_0 according to a likelihood maximisation criterion. The selected segment is attributed to L_x and is used to estimate the associated GMM model.
3. Adaptation/decoding loop: The objective is to detect all segments belonging to the new speaker L_x. All speaker models are re-estimated through an adaptation process according to the current segmentation. A Viterbi decoding pass, involving the entire HMM, is performed in order to obtain a new segmentation. This adaptation/decoding loop is re-iterated while some significant changes are observed on the speaker segmentation between two successive iterations.
4. Speaker model validation and stop criterion: The current segmentation is analysed in order to decide if the new added speaker L_x is relevant, according to some heuristical rules on the duration of the segments corresponding to speaker L_x. The stop criterion is reached if there are no more minimum 3-second-long candidate segments available in L_0 which may be used to add a new speaker; otherwise, the process goes back to step 2.

The segmentation stage is followed by a resegmentation process, which aims to refine the boundaries and to delete irrelevant speakers (e.g. speakers with too short speech segments). This stage is based on the third step of the segmentation process only: an HMM is generated from the segmentation and the iterative adaptation/decoding loop is launched. Here, an external world model, trained on microphone-recorded speech, is used for the speaker model adaptation.
The resegmentation stage does not utilise the pre-segmentation output. Indeed, all the boundaries (except speech/non-speech boundaries) and segment labels are re-examined during this process.

2.4. Post-normalisation and resegmentation

As reported in the literature [6] and drawing upon the speaker recognition domain, this last step consists in applying data normalisation. The resegmentation phase, described in the previous section, is repeated, but with a different parameterisation and now with data normalisation. Here the feature vector, comprising 16 LFCCs, energy, and their first derivatives, are normalised on a segment-by-segment basis to fit a zero-mean and unity-variance distribution. This segment-based normalisation relies on the output segmentation issued from the first resegmentation phase. The application of such a normalisation technique at the segmental level facilitates the estimation of the mean and variance on speaker-homogeneous data (compared with an estimate on the overall audio file involving many speakers).

3. Experimental protocol

3.1. Meeting corpora

The experiments reported in this paper were conducted on two different databases: the RT'05 and RT'06 data sets of the 2005 and 2006 NIST RT evaluation campaigns respectively (conference sub-domain) [1]. These data sets include between 8 and 10 meeting excerpts of about 12 minutes each, recorded at 4 or 5 different sites. The number of meeting participants varies from 4 to 9. In the same manner, rooms are equipped differently, involving various kinds of acquisition/recording devices. Lapel, head and table microphones and microphone arrays may be available. Therefore, a signal file is provided for each microphone located in a meeting room.
In this paper, the focus is made on the table microphones, for which the number of microphone channels varies from 2 to 7.

3.2. Performance measurement

The performance of the speaker diarization system is expressed in terms of the Diarization Error Rate (DER in %) which measures, in combination, both the quality of the speech activity detection (through the missed speaker error rate, denoted Mis and the false alarm speaker error rate, denoted FA) and the speaker diarization system quality (through the speaker error rate, denoted Spk). The DER is a standardised error metric as detailed in [1]. For all experiments, the DER is computed over all speech, excluding the overlapped speech (as specified by NIST until the RT'06 evaluation) although this overlapped speech has been taken into account in the speaker diarization process.
It should also be noted that the speaker diarization references that were used for all work reported in this paper are, due to their earlier availability, the ICSI/SRI forced alignment references [7] and not the manual transcription files proposed by NIST.

4. Non-speech effects

This section aims to investigate the impact of speech activity detection on the speaker diarization process. This impact will be evaluated through different configurations:
• (1-Ref) by using the reference segmentation, in which all the non-speech segments have been removed;
• (2-FA/0) by mixing the reference and automatic speech/non-

speech segmentations (issued from the automatic speech activity detection, described in section 2.1) in order to remove all the non-speech segments mis-classified as speech by the automatic process (false alarm speaker error rate fixed to 0%);
• (3-Mis/0) by mixing the reference and automatic speech/non-speech segmentations in order to remove all the speech segments mis-classified as non-speech by the automatic process (missed speaker error rate fixed to 0%);
• (4-Auto) by using the automatic speech/non-speech segmentation.

Table 1 presents the results of the speaker diarization process, for both the RT'05 (top) and RT'06 (bottom) data sets for these different configurations. Performance is provided for each individual meetings as well as for the overall sets (last row).
Even with perfect speech activity detection (no missed or false alarm error rates), the best performance (indicated in boldface) in terms of speaker error rate (Spk) is achieved with 33% of files when compared with the other configurations. In contrast, with automatic speech/non-speech detection, now with missed and false alarm error rates, the best performance is achieved with 39% of files (3 additional files give performances very close to the best). Secondly, the comparison of the (2-FA/0) and (4-Auto) configurations shows that in most cases, performance is quite similar (for 5 files) or better with automatic segmentation (for 8 files). The comparison of the (3-Mis/0) and (4-Auto) configurations leads to a similar distribution (8 similar and 6 better for automatic segmentation). Thirdly, the comparison of the (2-FA/0) and (3-Mis/0) configurations highlights that the (2-FA/0) configuration exhibits higher speaker error rates or quite similar to those of the (3-Mis/0) configuration (78% of files). Finally, regarding the (4-Auto) configuration, 67% of files have a similar or better speaker error rate than the (2-FA/0) and (3-Mis/0) configurations.
These different observations show that the speaker diarization system may be very sensitive to the speech/non-speech segmentation. Indeed, even if the speech/non-speech segmentation is perfect (as with the reference segmentation), the "shape" of the segmentation in terms of segment quantity and length can affect the speaker diarization system behaviour. Regarding for instance the RT'05 data set, the (1-Ref) configuration gives around 3000 speech segments, (2-FA/0) 2000 speech segments, (3-Mis/0) 730 speech segments and the (4-Auto) 670 speech segments. These discrepancies can explain why more accurate segmentations do not lead, in most cases, to a significant decrease in the speaker error rate: averaged DERs for the RT'05 and RT'06 data sets are similar or worse than compared with the automatic speech/non-speech segmentation.

5. Overlapped speech effects

Overlapped speech is a significant characteristic of meeting room recordings as is illustrated in Table 2 in which the second column shows the percentage of overlap per file for the RT'05 (top) and RT'06 (bottom) data sets. The overlap ranges from 1 to 25%.
The motivation of the experiments reported here is to assess the effects of overlapped speech on the speaker diarization process. The effects are empirically evaluated by comparing diarization performance while:
• applying the baseline system described in section 2, on the entire meeting file;
• cleaning automatic speech/non-speech segmentations of overlapped speech (issued from references) and applying the baseline speaker diarization system.

Table 1: *Speaker diarization performance for the RT'05 and RT'06 data sets, expressed in terms of Missed speaker (Mis), False Alarm (FA) and speaker error rates (Spk) according to different configurations: (1-Ref) reference segmentation w/o non-speech segments, (2-FA/0) automatic speech/non-speech segmentation with false alarm error corrected, (3-Mis/0) automatic speech/non-speech segmentation with missed errors corrected (4-Auto) automatic speech/non-speech segmentation.*

	1-Ref	2-FA/0		3-Mis/0		4-Auto		
Test RT'05	Spk	Mis	Spk	FA	Spk	Mis	FA	Spk
AMI_20041210	**0.7**	0.6	1.1	0.9	2.3	0.6	0.9	1.3
AMI_20050204	16.7	1.3	21.1	1.0	**10.4**	1.3	1.0	34.6
CMU_20050228	21.4	5.2	7.6	1.0	**4.5**	5.2	1.0	6.2
CMU_20050301	**7.6**	0.6	13.5	1.9	13.7	0.6	1.9	13.8
ICSI_20010531	**12.8**	4.3	29.5	3.2	16.2	4.3	3.2	13.5
ICSI_20011113	35.3	1.1	49.0	2.9	41.6	1.1	2.9	**32.3**
NIST_20050412	22.6	0.0	30.5	4.4	2.6	0.0	4.4	**2.1**
NIST_20050427	**6.0**	0.3	8.9	6.5	6.1	0.3	6.5	7.3
VT_20050304	18.4	0.4	31.6	1.2	20.8	0.4	1.2	**8.9**
VT_20050318	**12.8**	2.5	17.0	2.3	18.3	2.5	2.3	26.0
Overall'05	15.1	1.6	22.5	2.5	**13.6**	1.6	2.5	14.0

	1-Ref	2-FA/0		3-Mis/0		4-Auto		
Test RT'06	Spk	Mis	Spk	FA	Spk	Mis	FA	Spk
CMU_20050912	4.5	0.1	**3.4**	8.1	10.1	0.1	8.1	11.3
CMU_20050914	7.5	0.7	4.9	3.6	53.9	0.7	3.6	**4.2**
EDI_20050216	23.8	1.6	**15.1**	1.6	40.2	1.6	1.6	22.6
EDI_20050218	**5.9**	1.0	10.5	2.7	11.6	1.0	2.7	10.7
NIST_20051024	9.4	0.5	22.4	2.0	9.4	0.5	2.0	**9.3**
NIST_20051102	11.4	0.2	**10.6**	3.9	15.0	0.2	3.9	22.9
VT_20050623	13.2	0.4	13.8	8.0	3.6	0.4	8.0	**3.3**
VT_20051027	29.7	1.5	27.7	3.0	15.3	1.5	3.0	**11.0**
Overall'06	12.9	0.7	13.5	4.0	19.7	0.7	4.0	**12.2**

Table 2 illustrates the performance of the speaker diarization system according to these two configurations. Three main trends can be highlighted: (1) cases for which the baseline system using the entire file outperforms the case where the automatic speech/non-speech segmentations are cleaned of overlapped speech (in boldface), (2) cases for which the use of the segmentations cleaned of overlapped speech outperforms the baseline, (3) cases for which the difference is not considered as significant (in italic). Results indicate that, perhaps surprisingly, across the two data sets a performance loss is observed for only 1/3 of files when overlapped speech is involved in the speaker diarization process (case 2 above).
In particular *CMU_20050228* (RT'05, top) exhibits 23% overlapped speech yet the baseline error rate of 6% rises to 26% when overlapped speech segments are removed. Similar behaviour may be observed with other meetings, which exhibit more than 20% overlapped speech. In these cases overlapped speech does not appear to significantly affect the speaker diarization process and the removal of overlapped speech leads to a significant performance loss.
However, some files exhibit a lower speaker error rate when overlapped segments are removed. In particular for *EDI_20050218* (RT'06, bottom) the removal of 13% overlapped speech results in a speaker error rate of 11% dropping to 3%. Here, confusion between speakers has been significantly reduced by cleaning the segmentation of multi-speaker segments. Moreover, the system detects the right number of speakers (the baseline system erroneously detects two additional speakers).
Regarding the average performance, DERs for the RT'05 and RT'06 data sets increase from 14% to 16% and from 12% to

Table 2: *Speaker diarization performance for the RT'05 and RT'06 data sets expressed in terms of Missed speaker (Mis), False Alarm (FA) and speaker error rates (Spk) according to: (1) baseline system and (2) baseline system cleaned of overlapped speech segments.*

		Baseline			Baseline cleaned of overlapped speech		
Test RT'05	Overlap Rate(%)	Mis	FA	Spk	Mis	FA	Spk
AMI_20041210	2.7	0.6	0.9	*1.3*	0.6	0.9	*0.8*
AMI_20050204	9.5	1.3	1.0	34.6	1.3	1.0	31.8
CMU_20050228	22.8	5.2	1.0	**6.2**	5.2	1.0	**25.8**
CMU_20050301	12.8	0.6	1.9	*13.8*	0.6	1.9	*13.9*
ICSI_20010531	7.8	4.3	3.2	**13.5**	5.3	3.2	**26.7**
ICSI_20011113	18.7	1.1	2.9	32.3	1.1	2.9	30.0
NIST_20050412	20.9	0.0	4.4	**2.1**	0.0	4.4	**9.9**
NIST_20050427	8.5	0.3	6.5	7.3	0.3	6.5	4.3
VT_20050304	1.0	0.4	1.2	**8.9**	0.4	1.2	**9.9**
VT_20050318	8.5	2.5	2.3	26.0	2.5	2.3	10.0
Overall'05	11.0	1.6	2.5	**14.0**	1.6	2.5	**16.0**

		Baseline			Baseline cleaned of overlapped speech		
Test RT'06	Overlap Rate(%)	Mis	FA	Spk	Mis	FA	Spk
CMU_20050912	25.6	0.1	8.1	**11.3**	0.1	8.1	**17.1**
CMU_20050914	24.0	0.7	3.6	**4.2**	0.7	3.6	**12.8**
EDI_20050216	10.3	1.6	1.6	**22.6**	1.6	1.6	**38.7**
EDI_20050218	12.5	1.0	2.7	10.7	1.0	2.7	2.8
NIST_20051024	15.8	0.5	2.0	*9.3*	0.5	2.0	*8.8*
NIST_20051102	11.3	0.2	3.9	22.9	0.2	3.9	19.4
VT_20050623	19.0	0.4	8.0	*3.3*	0.4	8.0	*3.8*
VT_20051027	4.7	1.5	3.0	**11.0**	1.5	3.0	**16.8**
Overall'06	15.5	0.7	4.0	**12.2**	0.7	4.0	**15.1**

15% respectively when overlapped speech segments are discarded. The difference in performance is, perhaps surprisingly, small and would suggest that the diarization system is at least somewhat robust to overlapped speech segments.

Following the observations reported above, the assumption that overlapped speech may adversely affect the speaker diarization process from a statistical viewpoint can be questioned and tends to be context-dependent.

From another point of view, cleaning segmentations of overlapped speech also has an impact on the length and number of segments, especially in the case of a large amount of overlapped speech. This may affect the speaker diarization process as reported in the previous section. Because of this sensitivity, it is quite difficult to draw relevant conclusions about the real effects of the overlapped speech on the speaker diarization process.

6. Discussion

The experiments conducted on the speech/non-speech segmentation and overlapped speech highlight the importance of the initial segmentation in the speaker diarization system. Indeed, observations reported previously show that the "shape" of the segmentation (length and number of segments) is very pertinent and can affect the speaker diarization process. If the signal has to be cleaned of irrelevant acoustic events (especially in the case of non-speech signals to limit the false alarm error rates, but also overlapped speech), this pre-processing cannot be performed without taking into account the functioning of the speaker diarization system. In this sense, we argue that performing the pre-processing and the speaker diarization system successively as two separate steps is not necessarily the optimal approach in order to maximise the benefit of the enhanced initial segmentation. As an alternative, we propose to assign confidence values according to the type of information carried by the signal (and detected by the pre-processing), and to handle these values directly in the speaker diarization system without imposing on the system a large number of segments. According to the speaker diarization process utilised in this paper, these confidence values would be involved in both the training and decoding steps. This is the direction of our future work.

7. Conclusion

This paper investigates the effects of speech activity detection and overlap on a speaker diarization process in the context of meeting room recordings. Experiments conducted on two data sets, issued from the NIST RT'05 and RT'06 evaluation campaigns, outline interesting behaviours of the speaker diarization system. First, the latter is more sensitive to the "shape" of the initial segmentation (speech/non-speech segmentation or cleaned of overlapped speech), than to its quality (even with a perfect, errorless segmentation). Therefore, it is quite difficult currently to support one interpretation regarding the experimental results for both the speech/non-speech effects and the overlapped speech. Secondly, based on the experiments, the authors propose an original approach to deal with the non-relevant acoustic events (still necessary) and their implications in the speaker diarization process. Future work will focus on this proposal.

8. Acknowledgments

The authors would like to thank the ICSI group, Xavier Anguera and Chuck Wooters especially, for kindly providing their forced alignment speaker diarization references for both the RT'05 and RT'06 data sets.

9. References

[1] NIST, "Spring 2006 (RT'06S) Rich Transcription meeting recognition evaluation plan," http://www.nist.gov/speech/tests/rt/rt2006/spring/docs/rt06s-meeting-eval-plan-V2.pdf, February 2006.

[2] K. Laskowski and T. Schultz, "Unsupervised learning of overlapped speech model parameters for multichannel speech activity detection in meetings," in *ICASSP'06*, Pittsburgh, USA, May 2006.

[3] C. Fredouille and G. Senay, "Technical improvements of the e-hmm based speaker diarization system for meeting records," in *MLMI'06*, Washington, USA, May 2006.

[4] J.-F. Bonastre, F. Wils, and S. Meignier, "ALIZE, a free toolkit for speaker recognition," in *ICASSP'05*, Philadelphia, USA, March 2005.

[5] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Special issue of Computer and Speech Language Journal, Vol. 20-(2-3)*, 2006.

[6] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, "Combining speaker identification and BIC for speaker diarization," in *EuroSpeech'05*, Lisboa, Portugal, Sept. 2005.

[7] X. Anguera, C. Wooters, and J. Hernando, "Robust speaker diarization for meetings: ICSI RT06s evaluation system," in *ICSLP'06, Pittsburgh, USA*, September 2006.