NEW IMPLEMENTATIONS OF THE E-HMM-BASED SYSTEM FOR SPEAKER DIARIZATION IN MEETING ROOMS

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

[1]LIA, University of Avignon (France), [2]Swansea University (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. Some new enhancements to the E-HMM-based speaker diarization system are reported. These involve a different approach to speaker modelling utilising EM/ML-based training rather than MAP adaptation as in our previous work. Using the new system we investigate the effects of speech activity detection through speaker diarization experiments conducted on 23 meetings extracted from the NIST/RT evaluation campaign datasets. We propose a new approach, which assigns confidence values according to the type of information carried by the signal and incorporates these values directly into the speaker diarization system. Experimental results show that, perhaps surprisingly, the non-speech segments do not systematically affect the robustness of the speaker diarization system, and more precisely the speaker model training process.

Index Terms— speaker diarization, meeting rooms, confidence values, speaker recognition

1. INTRODUCTION

The speaker diarization task, also known as the "Who spoke When?" task, aims to detect the speaker turns within an audio document (segmentation task) and to group together all the segments belonging to the same speaker (clustering task). Involved as a main task in the Rich Transcription evaluation campaigns administered by the National Institute of Standards and Technology (NIST), for the last few years speaker diarization research has focused on meeting room recordings, now considered to be the most challenging task. Meeting room recordings often involve a high degree of spontaneous speech with large overlapping 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. As a necessary step for speaker diarization, speech activity detection is also challenging in this specific context, due to the inherent variability in signal quality. Recently particular emphasis has been placed on the detection of overlapping speech, one of the main characteristics of spontaneous speech, and consequently of meeting room recordings. There is, however, little published research on this topic [1, 2, 3], probably somewhat due to the difficulty of the task.
Recently the authors proposed an experimental framework [3], aimed at assessing the impact of speech activity detection and overlapping speaker segments on a state-of-the-art speaker diarization system. Experiments outlined interesting behaviours of the speaker diarization system, demonstrating greater sensitivity to the "shape" (in terms of number and length of segments) of the initial segmentation (speech/non-speech segmentation or cleaned of overlapping speech), than to the quality (even with a perfect, errorless segmentation). The authors also outlined the difficulties in supporting one interpretation regarding the effects of speech activity detection and overlapping speech on speaker diarization performance. The paper concluded by proposing an alternative approach which should utilise confidence values according to the type of information (i.e. speech/non-speech), and incorporate these values directly into the speaker diarization system.
This paper reports such an alternative approach and presents a new investigation into the impact of speech activity detection on speaker diarization. The paper is organised as follows: Section 2 describes the speaker diarization system and modifications involving EM/ML-based speaker training. Section 3 presents the proposed confidence value-based approach. Section 4 defines the experimental protocol followed by the experimental results given in Section 5. Finally, Section 6 draws some conclusions and proposes some future work.

2. SPEAKER DIARIZATION SYSTEM

Whilst still based on the Evolutive-HMM (E-HMM), the LIA speaker diarization system employed in this paper differs from that utilised in [2, 3]. The main variations lie (i) in the training algorithm involved with speaker modelling for the segmentation step, and (ii) on the related selection approach, which has been simplified as described below. The diarization system, developed using the open source ALIZE speaker recognition toolkit [4], involves 3 main steps:
- speech activity detection (SAD),
- speaker segmentation and clustering, and
- resegmentation,

in addition to some preprocessing to accommodate multiple channels.

2.1. Multi-channel handling

The speaker diarization task involved in this paper relates to multiple distant microphones located on meeting room tables (MDM task of the NIST/RT evaluation plans [5]). To deal with this task, a single virtual channel is formed using the BeamformIt 2.0 toolkit[1] with a 500 ms analysis window and a 250 ms frame rate.

2.2. Speech Activity Detection

The speech activity detection (SAD) algorithm employs feature vectors composed of 12 un-normalised Linear Frequency Cepstrum Co-

Nicholas Evans is now with Institut Eurécom, France

[1]Available at: http://www.icsi.berkeley.edu/xanguera/beamformit

efficients (LFCCs) plus energy augmented by their first and second derivatives. It utilises an iterative process based on Viterbi decoding and model adaptation applied to a two-state HMM, where the two states represent speech and non-speech events and are initialised with a 32-component GMM trained on separate data using an EM/ML algorithm. State transition probabilities are fixed to 0.5. Finally, duration rules are applied in order to refine the speech/non-speech segmentation yielded by the iterative process.

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) [6]. Each E-HMM state aims to characterise a single speaker and the transitions represent the speaker turns. Here the signal is characterised by 20 LFCCs, computed every 10 ms using a 20 ms window. The cepstral features are augmented by energy but no feature normalisation is applied.

As detailed in [3], the segmentation process begins by initializing the HMM with only one state representing the entire audio show. An iterative process is then started where a new speaker is added at each iteration. Successive Viterbi decoding and speaker model training loops attribute speech segments to the different speakers involved in the E-HMM. This iterative process is performed until a stop criterion is reached, which here is based on the ability, or not, for a new speaker to be added to the E-HMM.

Two main modifications have been introduced to the system compared with previous work. The first change relates to the iterative process involving the Viterbi decoding and the speaker model training. The speaker model adaptation techniques utilised in previous work have been replaced with EM/ML (Expectation - Maximization / Maximum Likelihood) based speaker model training. Speaker model training is widely used by speaker diarization systems, especially in the first speaker segmentation and clustering steps, when distance-based or criterion-based measurements are involved. Here, GMM-based speaker models are utilised, with 16 Gaussian components (diagonal covariance matrix) for all the HMM states, except for the last one for which only 8 Gaussian components are estimated. The difference in the number of Gaussian components aims to balance the different amounts of data attributed to the last speaker compared with the others. If sufficient data is available to estimate GMM speaker models, EM/ML is assumed to be more reliable than speaker model adaptation techniques (such as MAP adaptation) especially in the case of multi-speaker segments. Indeed, for the segmentation step, it is usual to process segments involving one or more speakers (due to the initialisation steps, classification errors, overlapping speech, etc.). Adaption with speech segments involving multiple speakers may significantly distort the resulting speaker model and consequently the overall speaker diarization process. The effect is even more pronounced when a strong adaptation weight is used for the speaker data (because of the small amount of available speaker data) compared with that used for the a priori data.

The second modification is directly linked to the use of the EM/-ML algorithm for speaker model training in the segmentation process. As mentioned above, sufficient data must be available for speaker model estimation with the EM/ML algorithm. For any newly detected speaker the amount of training data may be particularly low at the beginning of the iterative training process as well as for subsequent iterations, depending on the degree of detected speaker activity. In previous work, different strategies have been proposed to select initial speech segments used to add a new speaker: a maximum likelihood criterion [6], a criterion based on the maximum likelihood ratio [2], and a maximum likelihood criterion coupled with a pre-processing turn detection and local clustering [3]. Whilst the constraints are relaxed with successive Viterbi decoding and speaker model training iterations, all selection strategies were constrained to utilise fixed-size segments (3 or 6 seconds depending on the strategies) in order to limit multi-speaker segments during model initialisation. However, such limited segment sizes are not appropriate for the EM/ML algorithm thus a new strategy has been introduced and here we select the largest speech segment available (with a minimum size fixed to 6 seconds). This strategy is quite simple, but rather different from the previous ones. Here no effort is made to control the number of speakers present in the selected segments since their duration may vary greatly. This approach is inspired by other speaker diarization systems such as the ICSI system [7] for which speaker models are trained on large segment clusters, resulting in very good performance. Such large initial segments allow the use of the EM/ML algorithm to estimate speaker models, even for the first iteration of the process.

2.4. Resegmentation

The segmentation stage is followed by a resegmentation step, used to refine the segmentation outputs. An HMM is generated from the segmentation output and an iterative speaker model training/Viterbi decoding loop is launched. In contrast to the segmentation stage, here MAP adaptation (coupled with a generic speech model) replaces the EM/ML algorithm for speaker model estimation since the segmentation step provides an initial distribution of speech segments among the different speakers detected. For the resegmentation process, all the boundaries (except speech/non-speech boundaries) and segment labels are re-examined.

3. CONFIDENCE VALUE-BASED APPROACH

This paper investigates an alternative approach to speaker diarization, suggested in previous work [3], to deal with non-speech and overlapping segments or, more generally, to deal with segments where the confidence in their content is low. This approach assigns confidence values depending on the type of information carried by the signal (and detected by the pre-processing), and incorporates these values directly into the speaker diarization system. Here, confidence values are estimated for each frame (even if all the frames of a segment are assigned the same confidence value) and are utilised in the speaker model training during both the segmentation and resegmentation steps described in Section 2.

Training and confidence value: Regarding the EM/ML algorithm, the following parameter reestimation formulae are used for each iteration [8] given λ, an M-component GMM, characterised by p_i, μ_i, Σ_i $(i = 1, ..., M)$, which denote the mixture weights, mean vectors and covariance matrices respectively for Gaussian component i trained on a sequence of T D-dimensional vectors $X = x_1, ..., x_T$:

$$\bar{p_i} = \frac{1}{T} \sum_{t=1}^{T} p(i|x_t, \lambda) \quad (1)$$

$$\bar{\mu_i} = \frac{\sum_{t=1}^{T} p(i|x_t, \lambda) x_t}{\sum_{t=1}^{T} p(i|x_t, \lambda)} \quad (2)$$

$$\bar{\sigma_i} = \frac{\sum_{t=1}^{T} p(i|x_t, \lambda) x_t}{\sum_{t=1}^{T} p(i|x_t, \lambda)} - \bar{\mu_i^2} \quad (3)$$

with

$$p(i|x_t, \lambda) = \frac{p_i b_i(x_t)}{\sum_{k=1}^{M} p_k b_k(x_t)} \quad (4)$$

and

$$b_i(x_t) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} exp\left[-\frac{1}{2}(x_t - \mu_i)' \Sigma_i^{-1} (x_t - \mu_i)\right] \quad (5)$$

Considering now w_t as the confidence value assigned to observation x_t, Equation 4 can be rewritten as:

$$p(i|x_t, \lambda) = \frac{w_t p_i b_i(x_t)}{\sum_{k=1}^{M} p_k b_k(x_t)} \quad (6)$$

4. EXPERIMENTAL PROTOCOL

Meeting corpora:
The experiments reported in this paper were conducted on 23 meeting files extracted from the datasets of the 2004, 2005 and 2006 NIST RT evaluation campaigns (conference sub-domain) [5]. These data sets include meeting excerpts from 10 to 18 minutes, recorded at 7 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. A signal file is provided for each microphone located in a meeting room. In this paper, the focus is made on the distant table microphones (MDM task defined by NIST).

Performance measurement:
The performance of the speaker diarization system is expressed in terms of the Diarization Error Rate (DER in %) [5], 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 (through the speaker error rate, denoted Spk). For all experiments, the DER is computed over all speech, excluding overlapping speech (as specified by NIST until the RT'06 evaluation), although overlapping speech has been taken into account in the speaker diarization process.

5. EXPERIMENTAL RESULTS

In this section, we present experiments and results based on the confidence value-based approach presented in Section 3. Different sets of experiments examine the effect of the approach, through speaker diarization performance, focusing on one type of information: non-speech segments. Diarization performance using the EM/ML algorithm and the new selection strategy is also presented.

5.1. Speaker model training

This paper proposes a new implementation of the segmentation process, based on the EM/ML algorithm for the speaker model training instead of the MAP-based adaptation technique used in previous work. Table 1 compares speaker diarization system performance for both implementations. This comparison is given individually for the segmentation and resegmentation steps. The missed and false alarm speaker error rates, common for both implementations, are also provided.

	SAD (%)		EM/ML (%)		MAP (%)	
All meetings	Mis.	FA	Spk.	DER	Spk.	DER
Segmentation step	1.2	2.8	17.6	21.5	29.9	33.8
Resegmentation step	1.2	2.8	12.2	16.1	15.3	19.2

Table 1. *Overall speaker diarization performance, in terms of % DER, for EM/ML and MAP-based implementations.*

Performance reached after the segmentation step highlights the efficiency of the EM/ML algorithm in coping with the issues relating to the E-HMM framework, compared with the MAP-based adaptation. The MAP adapted speaker models are less robust and lead to coarse segmentation outputs. However, the difference is largely decreased after the resegmentation step, where only 3% absolute difference is observed between both implementations. The EM/ML-based implementation exhibits the best performance.

5.2. Non-speech effects

In this set of experiments, different confidence values are attributed to the non-speech segments, varying from 0.1 to 1.0 for each experiment, whereas the confidence value assigned to speech segments is fixed to 1.0 for all experiments. Two conditions are examined: (1) an automatic speech/non-speech segmentation (issued from the automatic SAD) and (2) a manual segmentation (coming from the references).

Table 2 summarises the speaker diarization performance obtained for these experiments under both conditions for the 23 meeting files. For clarity, only the worst and the best DER (denoted "Worst weight" and "Best weight" respectively in the table), depending on the confidence value considered (given in parentheses), are presented. The DER of the baseline speaker diarization system (denoted as "No weight" since non-speech segments are discarded from the speaker diarization process) is also provided for comparison.

Regarding condition (1), Table 2 shows that the best performance (in boldface) is reached for 70% of meetings with the fixed confidence values exhibiting the best DER ("Best weight"), against only 30% with the baseline system ("No weight"). Moreover, it can be observed that the fixed confidence values vary greatly depending on the meeting (from 0.1 to 1.0). Nevertheless, regarding the "Best weight" results only, the best DER is obtained for 61% of meetings with confidence values equal to or greater than 0.5, the overall best DER being reached with the confidence value fixed to 1.0. In a similar manner, regarding the "Worst weight" results only, the worst DER is also reached for 61% of meetings with a confidence values equal to or greater than 0.5. Even though these results demonstrate quite large variability in terms of DER performance and confidence values, it is interesting to underline some specific observations:
(1) low confidence value (< 0.5) for the "Worst DER" vs high confidence value (≥ 0.5) for the "Best DER": 6 meetings fall into this category, which suggests that integrating non-speech segments into the speaker model training with a high weight can improve speaker diarization robustness.
(2) low confidence value for the "Best DER" vs high confidence value for the "Worst DER": representing 4 meetings. The significant DER differences observed in this case show that introducing non-speech segments into the training scheme can disturb the robustness of speaker models.
(3) stationary DER between the "Worst DER" and the "Best DER":

representing 5 meetings for which the DER is relatively insensitive to the confidence value across the full range considered (not reported here). In these cases the presence of non-speech segments in the training process does not affect speaker model reliability. The remaining meetings exhibit a large variability, in terms of DER, according to the different confidence values used.

Regarding condition (2), Table 2 shows that the best performance (in italics) is reached for 74% of meetings with the fixed confidence values exhibiting the best DER ("Best weight"), against 26% with the baseline system ("No weight"). Here, 52% of meetings with a confidence value equal to or greater than 0.5 achieve the best DER ("Best weight"), the overall best DER still being obtained with the confidence value fixed to 1.0. Similar behaviour is observed for the "Worst weight" case.

These results show that, perhaps surprisingly, it is not necessarily the case that non-speech segments have systematically negative effects on the speaker diarization process and more precisely on the robustness of speaker models. The impact is variable depending on the meeting. Consequently, it can be assumed that even if the speech activity detection is not perfect (though giving satisfactory performance), it does not necessarily impact upon speaker model robustness in the specific context of meeting recordings. This observation is highlighted by the $EDI\_20050216$ meeting for which a DER improvement from 48.2% (baseline) to 30.2% is reached with a confidence value of 0.5 assigned to manually labelled non-speech segments. With the $EDI\_20050218$ meeting an improvement from 13.9% to 9.9% is achieved with a confidence value fixed to 1.0 and with the $VT\_20050304$ meeting an improvement from 17.2% to 2.2% is obtained with a confidence value fixed to 0.4.

	Auto. segmentation			Ref. segmentation	
	No Weight	Worst Weight	Best Weight	Worst Weight	Best Weight
Meetings	%DER	%DER	%DER	%DER	%DER
AMI_20041210	***1.7***	3.1 (0.5)	2.6 (0.7)	3.0 (0.6)	2.0 (0.4)
AMI_20050204	16.0	17.2 (0.1)	**5.4** (0.2)	17.2 (0.1)	*5.4* (0.2)
CMU_20050228	***11.8***	18.3 (0.1)	12.5 (0.6)	16.2 (0.1)	12.7 (0.5)
CMU_20050301	**9.2**	27.4 (0.2)	9.8 (0.3)	15.5 (0.5)	*8.6 (0.1)*
CMU_20050912	21.1	21.8 (1.0)	**20.6** (0.6)	22.1 (0.8)	*18.1* (0.1)
CMU_20050914	10.5	22.1 (0.5)	**10.3** (0.1)	12.5 (0.5)	*10.3* (0.1)
EDI_20050216	48.2	52.5 (0.8)	**30.7** (0.5)	50.7 (0.1)	*30.2* (0.5)
EDI_20050218	13.9	15.3 (0.2)	**9.9** (1.0)	22.9 (0.5)	*9.9* (1.0)
ICSI_20000807	6.7	11.5 (0.1)	**4.3** (0.2)	13.2 (0.1)	*4.2* (0.9)
ICSI_20010208	32.6	35 (0.1)	**14.6** (1.0)	27.5 (0.7)	*13.3* (0.1)
ICSI_20010531	***12.6***	23.0 (0.7)	13.1 (0.3)	40.5 (0.5)	13.4 (0.2)
ICSI_20011113	***30.5***	32.5 (0.6)	31 (0.3)	39.6 (0.4)	31 (0.8)
LDC_20011116-14	5.7	5.4 (0.5)	**4.3** (0.6)	5.1 (0.6)	*4.1* (0.4)
LDC_20011116-15	11.6	29.8 (0.8)	**8.7** (0.7)	26.7 (0.2)	*9.5* (0.6)
NIST_20030623	5.8	5.9 (0.1)	**2.1** (1.0)	5.9 (0.2)	*2.1* (1.0)
NIST_20030925	***14.2***	25.0 (0.8)	16.4 (0.3)	44.1 (0.7)	17.0 (1.0)
NIST_20050427	8.2	9.1 (0.1)	**6.0** (0.6)	9.3 (0.1)	*5.3* (0.5)
NIST_20051024	8.7	8.6 (0.1)	**3.0** (1.0)	10.4 (0.1)	*3.0* (1.0)
NIST_20051102	9.3	13.5 (0.9)	**7.4** (0.1)	13.2 (0.8)	*7.3* (0.1)
VT_20050304	17.2	26 (0.9)	**2.2** (0.6)	20.6 (0.1)	*2.7* (0.4)
VT_20050318	36.0	38.1 (0.7)	**34.8** (0.1)	38.6 (0.5)	*23.3* (0.2)
VT_20050623	24.8	29.0 (0.8)	**14.5** (0.5)	36.2 (0.4)	*22.5* (0.9)
VT_20051027	***16.8***	26.3 (0.9)	18.0 (0.8)	29.3 (0.8)	18.0 (0.5)
Overall	16.1	17.4 (0.8)	**14.3** (1.0)	18.3 (0.7)	*14.3* (1.0)

Table 2. *Speaker diarization performance in terms of % DER, in the context of confidence value-based speaker training.*

6. CONCLUSIONS

This paper investigates the effects of speech activity detection on speaker diarization in the context of meeting room recordings. Together with some modifications made to our previously reported system, the paper presents an original approach, which assigns frame-based confidence values according to the type of information carried by the signal (speech, non-speech, overlapping speech) and incorporates these values into the speaker model training scheme involved in the speaker diarization process.

A large set of experiments, conducted on 23 meetings extracted from the NIST/RT evaluation campaign datasets highlight that by assigning fixed confidence values (varying from 0.1 to 1.0) to non-speech segments, high confidence values do not systematically affect speaker model robustness since for 70% of meetings, improvements in speaker diarization performance are observed with confidence values equal to or greater than 0.5, compared to the baseline system. Consequently, speaker diarization errors should not necessarily be attributed to non-speech misclassification. Further work aims to better understand this rather surprising behaviour by measuring speaker model quality along the speaker diarization process. Similar experiments will be conducted on overlapping speech segments in order to examine their effects in the same manner.

7. REFERENCES

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

[2] 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.

[3] C. Fredouille and N. Evans, "The influence of speech activity detection and overlap on the speaker diarization for meeting room recordings," in *INTERSPEECH'07, Antwerp, Belgium*, September 2007.

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

[5] 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.

[6] 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.

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

[8] D. A. Reynolds, "Speaker identification and verification using gaussian mixture speaker models," in *Speech Communication*, 1995, vol. 171-2, pp. 91–108.