The LIA-EURECOM RT‘09 Speaker Diarization System

Corinne Fredouille[1], Simon Bozonnet[2] and Nicholas Evans[2,3]

[1] LIA-University of Avignon, BP1228, 84911 Avignon Cedex 9, France
[2] EURECOM, BP193, F-06904 Sophia Antipolis Cedex, France
[3] School of Engineering, Swansea University, Singleton Park, Swansea, SA2 8PP, UK
corinne.fredouille@univ-avignon.fr, {bozonnet,evans}@eurecom.fr

Abstract. This paper presents LIA-EURECOM’s joint submission to the NIST Rich Transcription 2009 (RT‘09) speaker diarization evaluation. We describe a number of modifications to our previous system which involve beamforming for the multiple distant microphone (MDM) condition and also significant enhancements to the speaker segmentation stage of the core speaker diarization system. These modifications lead to improvements in both speech activity detection (MDM only) and also to overall diarization performance. We present experimental results on a development set of 23 shows and the RT‘07 dataset, which was used for validation. Experimental results on the latter show a relative improvement in DER of 27% is achieved with our new system on the MDM condition. Similar experiments on the RT‘09 dataset show a relative improvement in DER of 35%. Our results for the MDM condition compare reasonably well with those of others even if, other than for beamforming, we did not use any delay features. Results for the single distant microphone condition (SDM) compare especially well with others’ work and highlight the merit of our top-down, evolutive hidden Markov model (E-HMM) approach to speaker diarization.

1 Introduction

This paper describes and assesses a number of modifications made to our previous submission [1] to the NIST Rich Transcription 2007 (RT‘07) speaker diarization evaluation [2] and our new system that was used for LIA-EURECOM’s joint submission to the most recent RT‘09 evaluation [3]. The changes involve the use of delay and sum beamforming for the multiple distant microphone (MDM) condition and significant changes to the speaker segmentation algorithm. Our RT‘07 system [1] used a pre-segmentation stage which aimed to identify speaker-homogeneous segments with which to initialise speaker models. The resulting segments were relatively small and with such limited data we were restricted to using MAP adaptation of a background model to train speaker models. For our RT‘09 system we removed the pre-segmentation stage favouring, instead, larger segments, though with greater potential for impurities (multiple speakers), and EM training. The modifications show significant improvements on both the RT‘07 and RT‘09 datasets and for both MDM and SDM conditions.

LIA-EURECOM officially entered only the MDM condition of the RT'09 evaluation. However, we have not undertaken significant work to optimise the beamforming frontend and, despite some recent work to improve diarization performance using delay features [4], we did not yet succeed in integrating these with the acoustic features. Success using fused delay and acoustic features has been reported previously [5–7] and from the results and system descriptions of the RT'09 evaluation, it seems that the best performing systems all use some form of delay features for the required MDM evaluation condition. Thus, in order to give a more meaningful comparison of our core clustering system with those of others, in this paper we present both MDM and SDM results, even if the latter were not published officially by NIST. It should be noted, however, that the only difference between our SDM system and our MDM system is the absence of beamforming. The acoustic features, the diarization system and all configuration parameters are absolutely identical.

The remainder of this paper is organised as follows. Section 2 describes our speaker diarization system and highlights the changes made to the system that we used for submission to the NIST RT'07 evaluation. Section 3 presents our development, validation and evaluation experimental work and results before our conclusions are presented in Section 4.

2 Speaker Diarization System

LIA-EURECOM's NIST RT'09 submission is based upon LIA's speaker diarization system [1,8–10] . It is an evolutive hidden Markov model (E-HMM) approach developed using the freely available, open source ALIZE toolkit [11]. There are three main steps involved. They are:

- speech activity detection (SAD),
- speaker segmentation and clustering, and
- normalisation and resegmentation,

in addition to some pre-processing to accommodate multiple channels where appropriate.

As outlined below there are three main differences to the system used for LIA's submission to the RT'07 evaluations [1]. They lie in (i) the pre-processing; (ii) the selection approach which determines which data are used to initialise each speaker model, and (iii) the training algorithm used for speaker modelling in the segmentation stage. In particular, the pre-segmentation stage, which was added for the RT'07 system, has now been removed.

2.1 Pre-processing and multi-channel handling

All audio files are first treated with Wiener filter noise reduction [12]. Then, for the MDM condition only, a single virtual channel for each show is created using the BeamformIt v2 toolkit [13, 14] with a 500ms analysis window and a 250ms frame rate. This latter stage is not necessary for the SDM condition. This is the only difference between the diarization systems used for the MDM and SDM experiments that are reported in this paper.

2.2 Speech activity detection

The speech activity detection (SAD) algorithm is the same as that used for the RT‘07 evaluation. In brief, it employs feature vectors composed of 12 un-normalised Linear Frequency Cepstrum Coefficients (LFCCs) computed every 10ms using a 20ms window. They are augmented with energy and first and second derivatives which results in a feature vector of 39 coefficients. The iterative SAD process is based on Viterbi decoding and model adaptation applied to a two-state HMM. The two states represent speech and non-speech events and are each initialised with a 32-component GMM model trained on appropriate, 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

The speaker segmentation and clustering stage now works directly on the SAD output; the pre-segmentation stage used for the RT’07 system has now been removed. The segmentation and clustering process relies on a one-step segmentation and clustering algorithm in the form of an evolutive hidden Markov model (E-HMM). Each E-HMM state aims to characterise a single speaker and the transitions represent the speaker turns. All possible changes between speakers are authorized. Here the signal is characterised by 20 un-normalised LFCCs, computed every 10ms using a 20ms window. The cepstral features are augmented by energy resulting in a feature vector of 21 coefficients.

The process for each audio show is as follows:

1. Initialization: The E-HMM has only one state, L0. A world model of 16 Gaussian components (cf. 128 for the RT’07 system) is trained by EM on all of the speech data. An iterative process is then started where a new speaker is added at each iteration.
2. A new speaker Lx is added to the E-HMM: The longest segment (cf. maximum likelihood criterion for RT’07 system) with a minimum duration of 6 seconds (cf. 3s for RT’07 system) is selected among all of the segments currently assigned to L0. The selected segment is attributed to Lx and is used to estimate a new GMM with EM training (cf. MAP adaption for RT’07 system.)
3. Adaptation/Decoding loop: The objective is to detect all segments belonging to the new speaker Lx. All speaker models are re-estimated through an adaptation process, according to the current segmentation (EM Algorithm) and a new segmentation is obtained via Viterbi decoding. This adaptation/decoding loop is repeated 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 analyzed in order to decide if the new added speaker Lx is relevant, according to some heuristical rules on the total duration assigned to speaker Lx. The stop criterion is reached if there are no more minimum 6 second long segments available in L0 with which to add a new speaker, otherwise the process goes back to step 2.

2.4 Resegmentation

The segmentation stage is followed by a resegmentation step, used to refine the segmentation outputs and to remove irrelevant speakers (e.g. speakers with too few segments). A new 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.

2.5 Normalization and resegmentation

The last step consists in reapplying a resegmentation but using a different parameterization using data normalization. Here, 16 LFCCs, energy, and their first derivatives, extracted every 10 ms using a 20ms window, make up a feature vector of 34 coefficients. The parameter vectors are normalized, segment-by-segment, to fit a zero-mean and unity-variance distribution.

3 Evaluation

This section presents our development work and evaluation results. The previous RT evaluation showed that, even if our system gave reasonable results, performance was quite unstable across different meetings. This year, we used a larger development set and kept in reserve a separate dataset for validation. We used data from the RT‘04, ‘05 and ‘06 datasets to create a broad development set of 23 shows and this dataset was used to optimise all system parameters. Diarization results on the development set are presented in Section 3.1. In order to confirm the improvements on unseen data we processed the RT‘07 dataset with our new system prior to our submission to the RT‘09 evaluation. These results, together with a comparison of our RT‘07 and RT‘09 system results are presented in Section 3.2. Finally our evaluation results are presented in Section 3.3.

Ideally we would illustrate the improvement obtained with each of the modifications to our previous system, described in Section 2, on their own. However, the modifications are highly integrated and it is either not meaningful or possible to do this. For example, the turn detection and clustering algorithms used in our RT‘07 system produced a pre-segmentation hypothesis which was used to identify relatively small segments to initialise speaker models. The idea here was to identify speaker-homogeneous segments. However, shorter segments do not give sufficient data to use EM training and, instead, speaker models were trained using a background model and MAP adaptation. Only by removing the turn detection and clustering algorithms, i.e. by using the speech activity detection output to identify segments for speaker model training, are sufficiently large segments identified so that EM training proves beneficial. Therefore, in this paper, we present a straight forward comparison of our RT‘07 and RT‘09 systems. To avoid confusion with similarly named datasets these systems are from hereon referred to simply as our *old* and *new* systems.

3.1 Development set results

Our development set results are illustrated in Table 1. Results are illustrated both with and without scoring overlap regions for each show (first column). Averages results are presented in the final row. In accordance with NIST protocols everywhere in this paper we refer only to the results where overlap regions are scored. Results where overlap regions are not scored are provided to facilitate comparisons to previously published work and to observe the penalty incurred by not addressing overlap. The second and third columns of Table 1 illustrate the speech activity detection (SAD) performance. The average SAD performance is 4.0% and 2.8% for missed speech (MissSp) and false alarm speech (FA) respectively. The fourth column illustrates the speaker error (SpkErr) and shows a range of 0.2% to 37% and an average of 11%. The final column shows the overall diarization error rate (DER) where the range is 0.8% to 41% and the average is 18%. These results show that the SAD performance is relatively stable across the whole dataset but that our system is relatively unstable in terms of speaker error. Nonetheless the average performance is better than that obtained for our previous system as reported in [1].

3.2 Validation set results

In order to validate the modifications on unseen data we processed the RT‘07 dataset with the new system and compared the results to those obtained with our old system. Results are illustrated in Table 2 for both MDM and SDM conditions. The results obtained using our old system are exactly the same as those published in [1]. As illustrated in the 3rd and 4th columns of Table 2 there were small changes in the SAD performance for the MDM condition between the old and new systems. In our old system multiple channels were accommodated simply by adding the channels together without any delay correction. In our new system we used the BeamformIt toolkit [13,14] as described in Section 2 and this accounts for the difference in SAD performance for the MDM condition. Since there is no beamforming for the SDM condition here the SAD performance is the same for both old and new systems. As might be expected the SAD performance

Table 1. Missed speech (MissSp), false alarm speech (FA), speaker error (SpkErr) and overall diarization error rate (DER) for the development dataset. In all cases results are illustrated with/without scoring overlapping segments.

Show	MissSp	FA	SpkErr	DER
AMI_20041210-1052	0.3/0.1	0.3/0.3	0.2/0.2	0.8/0.6
AMI_20050204-1206	1.3/0.3	2.1/2.2	5.5/5.6	9.0/8.1
CMU_20050228-1615	4.1/0.7	0.9/0.9	5.5/5.9	10.5/7.5
CMU_20050301-1415	1.5/0.0	1.7/1.8	6.0/6.0	9.2/7.8
CMU_20050912-0900	9.5/0.2	8.8/10.8	8.5/7.4	26.9/18.4
CMU_20050914-0900	7.0/0.5	6.3/7.2	6.3/6.5	19.6/14.2
EDI_20050216-1051	3.5/0.7	2.9/3.0	22.6/23.4	28.9/27.2
EDI_20050218-0900	3.1/0.4	3.1/3.3	6.4/6.5	12.5/10.2
ICSI_20000807-1000	4.3/0.1	0.4/0.5	21.5/23.0	26.1/23.5
ICSI_20010208-1430	2.8/1.1	1.1/1.1	29.8/30.4	33.7/32.7
ICSI_20010531-1030	1.9/1.1	2.7/2.8	11.5/11.7	16.2/15.6
ICSI_20011113-1100	4.2/0.1	3.4/3.7	17.6/17.5	25.3/21.3
LDC_20011116-1400	2.0/0.7	3.2/3.3	1.4/1.4	6.5/5.4
LDC_20011116-1500	6.0/0.2	1.3/1.5	9.3/9.6	16.6/11.3
NIST_20030623-1409	2.2/1.4	0.3/0.3	3.9/3.8	6.4/5.5
NIST_20030925-1517	9.4/3.3	3.6/4.2	17.0/18.6	30.1/26.0
NIST_20050427-0939	2.0/0.2	2.3/2.4	1.2/1.0	5.6/3.6
NIST_20051024-0930	3.9/0.3	1.1/1.1	6.5/6.8	11.4/8.3
NIST_20051102-1323	3.7/0.6	2.8/2.9	3.3/3.5	9.7/7.0
VT_20050304-1300	0.3/0.2	0.8/0.8	4.6/4.6	5.6/5.6
VT_20050318-1430	2.0/1.9	1.8/1.8	37.3/37.3	41.1/41.0
VT_20050623-1400	4.9/1.0	4.7/5.0	25.2/26.4	34.8/32.5
VT_20051027-1400	3.1/1.8	2.4/2.5	10.5/10.1	16.0/14.4
Average	**4.0/0.7**	**2.8/3.0**	**11.0/11.2**	**17.8/14.9**

is slightly better for the MDM conditions when compared to the SDM conditions for each system respectively.

Speaker error rates (fifth column of Table 2) of 18% and 12% for the old and new systems respectively and the MDM condition show a large improvement. This is due to the use of larger segments for model initialisation and EM training instead of MAP adaptation. However, there is no significant difference in speaker error rates between the MDM and SDM conditions for the two systems which shows that our system is only capable of utilising the additional information from multiple microphones to improve the SAD performance. This is reflected in the overall DER (final column of Table 2). DERs of 24% and 18% are illustated for the old and new systems respectively which amounts to a relative improvement of 27% for the MDM condition. Similar improvements are obtained for the SDM condition. The new diarization system was therefore used for our RT‘09 submission.

Table 2. Summary of speaker diarization performance, using our old (RT'07) and new (RT'09) diarization systems, on the MDM and SDM conditions of the RT'07 dataset.

System	Mic. Cond.	MissSp	FA	SpkErr	DER
Old	MDM	4.5/0.8	2.0/2.2	17.7/18.6	24.2/21.5
	SDM	4.7/1.1	2.1/2.3	17.7/17.9	24.5/21.3
New	MDM	4.1/0.4	1.8/1.9	11.8/11.9	17.7/14.3
	SDM	4.7/1.1	2.1/2.3	11.4/11.6	18.3/15.0

Table 3. Missed speech (MissSp), false alarm speech (FA), speaker error (SpkErr) and overall diarization error rate (DER) for the RT'09 dataset. Results with/without scoring overlapping segments.

Show	MissSp	FA	SpkErr	DER
EDI_20071128-1000	3.6/0.5	1.9/2.0	0.8/0.9	6.3/3.4
EDI_20071128-1500	7.6/0.6	5.2/6.0	27.2/30.4	40.0/37.1
IDI_20090128-1600	4.1/0.4	0.5/0.5	13.0/13.2	17.6/14.2
IDI_20090129-1000	4.6/1.1	2.7/2.9	7.8/6.8	15.1/10.9
NIST_20080201-1405	16.0/1.7	1.2/1.7	34.8/40.5	52.0/44.0
NIST_20080227-1501	8.5/0.2	0.4/0.5	18.5/19.8	27.4/20.4
NIST_20080307-0955	3.5/0.0	1.4/1.5	17.9/17.9	22.8/19.4
Average	**6.1/0.6**	**1.9/2.1**	**15.5/15.8**	**23.5/18.5**

3.3 Evaluation set results

Our official results for the MDM condition of the NIST RT'09 evaluation are presented in Table 3. With the exception of the first NIST show (sixth row) the SAD performance is reasonably consistent with that of the development set, though the average missed speech error rate is slightly higher (6.1% cf. 4.0%) and the average false alarm speech rate is slightly lower (1.9% cf. 2.8%). A speaker error rate of 16% compares reasonably well to a score of 11% for the development set but the results are again quite unstable, the range being 1% to 35%. The overall DER is 24% (cf. 18% for the development set) and shows a large variation in scores from 6% to 52%. Our preliminary investigations show that our system almost always overestimates the number of speakers which is a significant cause of the large variation in performance.

Finally we present a summary of both MDM and SDM results on the RT'09 dataset using both our old and new diarization systems. They are illustrated in Table 4. The results show a marked improvement in performance between the old and new systems. Also shown is the decreased performance for the MDM condition over the SDM condition for the old system. For the new system, where we have utilised the BeamformIt toolkit [13, 14] for beamforming, the MDM results are better than the SDM results. For the MDM condition overall DERs of 36% and 24% for the old and new systems respectively correspond to a relative

Table 4. As for Table 2 except for the RT’09 dataset.

System	Mic. Cond.	MissSp	FA	SpkErr	DER
Old	MDM	6.6/1.1	1.6/1.8	28.1/29.0	36.3/32.0
	SDM	7.2/1.8	1.2/1.4	21.0/21.5	29.5/24.7
New	MDM	6.1/0.6	1.9/2.1	15.5/15.8	23.5/18.5
	SDM	7.2/1.8	1.2/1.4	17.6/18.3	26.0/21.5

improvement of 35%. A smaller relative improvement of 12% (30% for the old system cf. 26% for the new system) is achieved on the SDM condition.

Upon a comparison of our results to those of others it is evident that our MDM system compares reasonably well. Our SDM system performs particularly well. Since the previous evaluation we have given some consideration to abandoning our top-down approach to diarization in favour of the more popular bottom-up or hierarchical, agglomerative clustering approaches [7,15]. However, our SDM results have given us cause to pursue our approach further. We have conducted some experiments with bottom-up approaches but have, so far, not been able to improve our results over those obtained with the E-HMM. This work is not yet complete but our initial findings show that the introduction of speaker models sequentially, as is the case in top-down approaches, rather than simultaneously, as is the case with bottom-up approaches, has certain merits. Most current bottom-up approaches use a uniform or linear segmentation for initialisation, e.g. [7, 15]. Whilst some attempts to improve on this, e.g. [16], have demonstrated some success there is some debate as to whether or not they bring consistent improvements. By adding speaker models sequentially, rather than simultaneously, the results of previous segmentations have the potential to assist the selection of purer segments for the initialisation of subsequent speaker models. For this reason we believe that the top-down approach might provide better potential for improved initialisation and intend to investigate this approach in the future.

4 Conclusions

This paper demonstrates the improvements made to the system used for our submission to the NIST RT‘07 speaker diarization evaluations. Our new system, which incorporates various integrated modifications to improve speaker modelling achieves a 35% relative improvement (24% cf. 36%) in DER on the MDM condition of the RT‘09 dataset. This level of performance places our system reasonably well among the other entries to the RT‘09 evaluation in terms of absolute performance, despite our not having used delay features (other than for initial delay and sum beamforming). We believe this observation highlights the strengths of our core, top-down speaker diarization algorithm. We expect performance to improve once delay features are incorporated.

Results on the SDM condition show a relative improvement of 13% DER over our previous system (22% cf. 25%). This result, whilst not published officially by NIST, compares very well to those of others and further demonstrates the merit of our core diarization algorithm and thus we feel that our top-down approach to speaker diarization warrants continued attention.

Finally, our system is shown to be quite unstable across different meetings and is quite sensitive to speaker model initialisation. Our ongoing and future work will investigate improved initialisation strategies and we believe they have the potential to further improve the performance of top-down approaches to speaker diarization.

References

[1] Fredouille, C., Evans, N.: The LIA RT07 speaker diarization system. In Stiefelhagen, Bowers, F., ed.: Lecture notes in Computer Science - Multimodal Technologies for Perception of Humans. Volume 4625/2008., Springer (2008) 520–532

[2] NIST: Spring 2007 (RT'07S) Rich Transcription meeting recognition evaluation plan. http://nist.gov/speech/tests/rt/2007/docs/rt07-meeting-eval-plan-v2.pdf (2007)

[3] NIST: The NIST Rich Transcription 2009 (RT'09) evaluation. http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf (2009)

[4] Evans, N.W.D., Fredouille, C., Bonastre, J.F.: Speaker diarization using unsupervised discriminant analysis of inter-channel delay features. In: Proc. IEEE ICASSP. (2009) 4061–4064

[5] Pardo, J.M., Anguera, X., Wooters, C.: Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences. In: Proc. ICSLP. (2006)

[6] Koh, E.C.W., Sun, H., Nwe, T.L., Nguyen, T.H., Ma, B., Chng, E.S., Li, H., Rahardja, S.: Speaker diarization using direction of arrival estimate and acoustic feature information: The I2R-NTU submission for the NIST RT 2007 evaluation. In Stiefelhagen, Bowers, F., ed.: Lecture notes in Computer Science - Multimodal Technologies for Perception of Humans. Volume 4625/2008., Springer (2008) 484–496

[7] Wooters, C., Huijbregts, M.: The ICSI RT07s speaker diarization system. In Stiefelhagen, Bowers, F., ed.: Lecture notes in Computer Science - Multimodal Technologies for Perception of Humans. Volume 4625/2008., Springer (2008) 509–519

[8] Fredouille, C., Senay, G.: Technical improvements of the E-HMM based speaker diarization system for meeting records. In: MLMI'06, Washington, USA (2006)

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

[10] Fredouille, C., Evans, N.: New implementations of the E-HMM-based system for speaker diarisation in meeting rooms. In: Proc. IEEE ICASSP. (2008)

[11] Bonastre, J.F., Wils, F., Meignier, S.: ALIZE, a free toolkit for speaker recognition. In: ICASSP'05, Philadelphia, USA (2005)

[12] Adami, A., Burget, L., Dupont, S., Garudadri, H., Grezl, F., Hermansky, H., Jain, P., Kajarekar, S., Morgan, N., Sivadas, S.: Qualcomm-ICSI-OGI features for ASR. In: Proc. ICSLP. (2002) 21–24

[13] Anguera, X.: BeamformIt (the fast and robust acoustic beamformer). (http://www.xavieranguera.com/beamformit/)

[14] Anguera, X., Wooters, C., Hernando, J.: Acoustic beamforming for speaker diarization of meetings. IEEE Transactions on Audio, Speech, and Language Processing **15**(7) (2007) 2011–2021

[15] Ajmera, J., Wooters, C.: A robust speaker clustering algorithm. In: ASRU'03, US Virgin Islands, USA. (2003)

[16] Anguera, X., Wooters, C., Hernando, J.: Friends and enemies: A novel initialization for speaker diarization. In: Proc. Interspeech, ICSLP. (2006) 689–692