# THE LIA-EURECOM RT'09 SPEAKER DIARIZATION SYSTEM: ENHANCEMENTS IN SPEAKER MODELLING AND CLUSTER PURIFICATION

*Simon Bozonnet and Nicholas W. D. Evans*

EURECOM
BP193, F-06904 Sophia Antipolis Cedex, France
{bozonnet, evans}@eurecom.fr

*Corinne Fredouille*

University of Avignon, LIA/CERI
BP1228, F-84911 Avignon Cedex 9, France
corinne.fredouille@univ-avignon.fr

**ABSTRACT**

There are two approaches to speaker diarization. They are bottom-up and top-down. Our work on top-down systems show that they can deliver competitive results compared to bottom-up systems and that they are extremely computationally efficient, but also that they are particularly prone to poor model initialisation and cluster impurities. In this paper we present enhancements to our state-of-the-art, top-down approach to speaker diarization that deliver improved stability across three different datasets composed of conference meetings from five standard NIST RT evaluations. We report an improved approach to speaker modelling which, despite having greater chances for cluster impurities, delivers a 35% relative improvement in DER for the MDM condition. We also describe new work to incorporate cluster purification into a top-down system which delivers relative improvements of 44% over the baseline system without compromising computational efficiency.

***Index Terms*—** Speaker diarization, speaker segmentation, speaker clustering, cluster purification, DER, MDM, SDM

## 1. INTRODUCTION

Speaker diarization, commonly referred to as the 'who spoke when?' task, involves the detection of speaker turns within an audio document (segmentation) and the grouping together of all same-speaker segments (clustering). Much progress has been made in the field over recent years partly spearheaded by the NIST Rich Transcription (RT) evaluations [1] in the proceedings of which are found two general approaches: top-down and bottom-up. The bottom-up approach is by far the most common. Very few systems, such as the LIA's evolutive hidden Markov model (E-HMM) system, are based on top-down approaches.

The bottom-up, hierarchical, agglomerative clustering approach trains a number of models $N$ (which exceeds the predicted number of speakers) and aims to successively merge and reduce the number of models until there remains only one for each speaker. In contrast, the top-down approach first models the audio show with a single speaker model and aims to successively add new speaker models until the full number of speakers are deemed to be accounted for.

Even though the best performing systems over recent years have all been bottom-up approaches we believe that the top-down approach is not without significant merit: first, results on the NIST RT'09 dataset show that the top-down approach gives very reasonable performance on the multiple distant microphone (MDM) condition (even though we did not use estimates of inter-channel delay as features) and that it gives extremely competitive results on the single distant microphone (SDM) condition; second, top-down approaches such as the E-HMM system are significantly less computationally demanding than bottom-up approaches and third, there is as-yet-untapped potential to reduce model impurities through cluster purification.

The contribution in this paper is two-fold. First we report new enhancements to the top-down, E-HMM speaker diarization system for conference meeting data that result in improved speaker modelling, and hence better diarization performance. Second, we present new work undertaken since the most recent NIST RT'09 evaluation that deliver additional significant improvements in performance through cluster purification.

Given their dominance in the literature, previous attempts at cluster purification have generally focused only on bottom-up diarization systems. Small improvements in LIA's top-down E-HMM diarization system using purification were reported in [2] but the module was later removed as subsequent developments in model initialisation and speaker modelling rendered the improvements negligible. This paper argues why cluster purification should nonetheless have particular potential in a top-down approach and presents a novel attempt at integrating a new approach to cluster purification first proposed in [3] in our system to speaker diarization. Whilst performance is not as good as that of the best performing system (also bottom-up) the gap in performance between state-of-the-art top-down and bottom-up systems is significantly reduced while retaining computational efficiency.

The remainder of this paper is organised as follows. Section 2 describes new E-HMM system enhancements and justifies our efforts to incorporate cluster purification into a top-down speaker diarization system. The purification algorithm used in this work is described in Section 3 and our experimental work is presented in Section 4. Finally, our conclusions are presented in Section 5.

## 2. THE E-HMM SPEAKER DIARIZATION SYSTEM

Details of the top-down E-HMM speaker diarization system, developed using the freely available open source ALIZE toolkit [4], have been published previously in [2, 5] and a full description of our most recent system is available in [6]. Accordingly only a brief system overview is reported here. Highlighted are system enhancements implemented for our submission to the NIST RT'09 speaker diarization evaluation. We also explain the motivation to apply cluster purification in a top-down speaker diarization system.

### 2.1. System overview

Our top-down system is composed of three stages: (i) speech activity detection (SAD), (ii) speaker segmentation and clustering and (iii) normalisation and resegmentation, in addition to some preprocessing such as Wiener filter noise reduction [7] and beamforming [8], where multiple microphones are available.

Following the first stage SAD (the previous pre-segmentation stage used in our RT'07 system [5] has now been removed), the second stage speaker segmentation and clustering is initialised with a root speaker model $L_0$ which is trained using all of the speech segments available. New models, each one characterising a single speaker, are then introduced iteratively, and are trained with a relevant segment from $L_0$ with several embedded iterations of decoding and adaptation [6]. Finally, a resegmentation is applied, during which speakers with too few data are removed from the model.

### 2.2. Enhancements to speaker modelling

In contrast to our previous system, during the first pass segmentation and clustering stage speaker models are now trained using expectation maximisation (EM) instead of being obtained through maximum a posteriori (MAP) adaptation of a background model (though the second-pass resegmentation still uses MAP adaptation). Speaker models now have 16 components instead of 128 and are now initialised on the longest available segments (cf. maximum likelihood criterion previously) that are greater than 6 seconds in length (cf. 3 seconds). In the final stage a further resegmentation is applied but this time using feature normalisation. Again, full details are available in [6].

### 2.3. Cluster purity

As detailed later in Section 4, these modifications lead to significant improvements in diarization performance. However, despite numerous efforts over recent years, the E-HMM system seems to be particularly prone to poor speaker model initialisation. In contrast to bottom-up approaches, new speaker models are trained on relatively small speech segments and, in addition, there is always the possibility that (i) the segments do not contain representative or sufficient speech, or that (ii) they contain speech from more than one speaker. While the modifications described above lead to models now being trained on segments of greater length there is an increased chance for impurities, i.e. data from more than one speaker. In consequence, the models may not reliably attract all other segments from the same speaker during subsequent Viterbi decoding. The accuracy of the segmentation and clustering stage has a strong impact on overall diarization performance and in this paper we aim to enhance it through purification.

Since the sequential adding of speaker models effectively reduces the pool of segments, assigned to $L_0$, on which new speaker models may be trained (both for initialisation and for subsequent decoding/adaptation), there is significant potential to reduce cluster impurities through top-down approaches. To illustrate, suppose that a new speaker $L_1$ is added to the E-HMM and that, during the subsequent decoding/adaptation loop, the resulting speaker model does reliably attract, from those currently assigned to the root model $L_0$, all the other segments which correspond to $L_1$. The root model should then contain few segments corresponding to speaker $L_1$. When the next speaker $L_2$ is added then there is a reduced chance that its model will be contaminated with segments corresponding to $L_1$. However, there is also the possibility that the newly trained model $L_1$ also contains segments from other speakers. Thus we can envisage two approaches to purification. The first should aim to further improve speaker modelling so that, during decoding, newly added speaker models more reliably attract the speaker's corresponding segments from those assigned to the root model $L_0$. The second strategy should aim to purify newly trained models by reassigning data which is deemed to correspond to another speaker. In the following we describe how this is achieved through cluster purification applied after segmentation and clustering.

## 3. CLUSTER PURIFICATION

Purification is not a new idea and several different purification approaches have been reported, e.g. [9]. In contrast to this previous work using bottom-up systems we here seek to demonstrate the potential for cluster purification specifically in *top-down* approaches. We describe our implementation of a new approach to cluster purification [3] that was first proposed by IIR-NTU researchers at the NIST RT'09 evaluations [1]. We begin with a brief description of the original algorithm and then describe the modifications we have made in order to exploit its potential in our top-down system.

The diarization system presented in [3] for the processing of single channel meeting shows (SDM condition) is initialised with 30 homogeneous clusters of uniform length and a 4-component GMM is trained on the data in each cluster. Each cluster is then split into segments of 500ms in length and the top 25% of segments which best fit the GMM are identified and marked as classified. The remaining 75% of worst-fitting segments are then gradually reassigned to their closest GMMs, $K$ segments at a time (the value of $K$ is not published in [3]), with iterative Viterbi decoding and adaptation until all segments are classified. Whilst in [3] this algorithm is referred to as sequential initialisation, it clearly performs a cluster purification role. The same algorithm was presented with modifications in [10] for purification purposes and for the processing of multiple microphone meeting shows (MDM condition).

We have found it necessary to modify this approach in order to bring its potential to the E-HMM system. In our system purification is applied after segmentation and clustering which produces a number of clusters (generally only a few more than the true number of speakers) each of which, ideally, corresponds to a single speaker. Of course there remains the distinct potential for impurities and our experiments on development data have shown that speaker clusters are typically between 50% and 95% pure.

Thus, in contrast to the bottom-up approach, where the initial clustering is generally random and uniform, our cluster purification algorithm operates on clusters which should already contain a dominant speaker. The original algorithm was intended for clusters of relatively lower initial purity and we have found that the same algorithm applied directly to our system can, in some cases, reduce cluster purity. However, the algorithm brings impovement in performance with the following modifications. First, we increased the model complexity to 16 components and second, we increased the amount of segments kept during each iteration to 55% (both determined empirically). These modifications are perhaps consistent with intuition given the initial cluster purity.

## 4. EXPERIMENTAL WORK

The experiments presented here aim to demonstrate the improvement in diarization performance obtained with modifications to speaker modelling that were presented in Section 2. Also reported are new experiments to assess the performance of the cluster purification algorithm described in Section 3.

| System | Dev. Set | Valid. Set | Eval. Set |
|---|---|---|---|
| RT'07 | 24.8/22.5 | 24.2/21.5 | 36.3/32.0 |
| RT'09 | 17.8/14.9 | 17.7/14.3 | 23.5/18.5 |
| RT'09+Pur. | 16.0/13.0 | 17.9/14.6 | 20.3/15.2 |

**Table 1**. A comparison of diarization performance on the MDM condition and three different datasets: development (Dev. Set), validation (Valid. Set) evaluation (Eval. Set). Results reported for three different systems: the system used for our RT'07 submission, that used for our RT'09 submission and the same system using cluster purification (RT'09+Pur.). Results illustrated with/without scoring overlapping speech.

### 4.1. Datasets and metrics

We report experiments on a development dataset comprising meeting shows from the NIST RT'04, '05 and '06 datasets (23 shows in total). This set alone was used to optimise both our baseline system and the purification algorithm. In addition we present results on a separate validation set, namely the NIST RT'07 dataset (8 shows). Finally, to confirm improvements in performance on unseen data we present results on the most recent NIST RT'09 evaluation dataset (7 shows). The use of strictly standard datasets and experimental protocols allows the direct comparison of results to our own previously published work [6] and to those of others [1].

In accordance with NIST evaluations we focus on the MDM condition. However, we do not use inter-channel delay features and so, in order to give a more meaningful assessment of our core diarization system, independently of beamforming performance and fused delay features, we also report results on the SDM condition. Except for the front-end beamforming the systems used for MDM and SDM conditions are otherwise identical. Diarization performance is assessed in terms of the standard diarization error rate (DER) using official NIST scoring tools. Finally, since our current system is not capable of detecting overlapping speech segments we report DERs both with (as per NIST standards) and without scoring overlapping speech.

### 4.2. Speaker modelling

Table 1 illustrates a comparison of speaker diarization performance for the MDM condition using the three different system variations (1st column) and the three different datasets (columns 2 to 4). For the newer RT'09 system (using EM speaker modelling) results of 17.8% (Dev. Set), 17.7% (Valid. Set) and 23.5% (Eval. Set) compare favourably with those of 24.8% (Dev. Set), 24.2% (Valid. Set) and 36.3% (Eval. Set) obtained using the previous RT'07 system (using MAP adaptation speaker modelling). For the Eval. Set these improvements amount to a relative improvement of 35% where overlapping speech is scored and 42% where overlapping speech is ignored. Speaker modelling through EM is thus shown to deliver much better performance than speaker modelling through MAP adapation even if the amount of data used for initialisation is still relatively small compared to that usually used by bottom-up systems. However, even if the results show more consistent and stable average performance across the three different datasets the DERs for individual audio shows in the Eval. Set were found to vary between 6% and 52%. It is the hypothesis under investigation here that this is caused by the use of impure segments for model initialisation.

| System | Dev. Set | Valid. Set | Eval. Set |
|---|---|---|---|
| RT'09 | 80/68/95 | 79/67/92 | 72/41/87 |
| RT'09+Pur. | 83/64/95 | 82/71/92 | 78/54/96 |

**Table 2**. Average/minimum/maximum cluster purities (%Pur) without (RT'09) and with (RT'09+Pur.) purification for the Dev. Set, Valid. Set and the Eval. Set. Results for MDM condition.

### 4.3. Purification

The fourth line of Table 1 (RT'09+Pur.) illustrates the results after purification is applied after segmentation and clustering. The purification algorithm has a small effect on the Dev. Set and leads to a relative improvement of 10% (16.0% cf. 17.8%) over the RT'09 system. Results are almost identical on the Valid. Set but are improved on the Eval. Set. Here results of 23.5% without purification and 20.3% with purification correspond to a relative improvement of 14% (18% without scoring overlapping speech) and 44% over the RT'07 baseline. Thus the purification algorithm gives as good or better results and helps to stabalise the results across the three datasets, though it is of interest to understand why the algorithm performs significantly better on the Eval. Set than on the dataset on which it was optimised.

To help explain this behaviour we measured the cluster purity statistics before and after purification. For this we introduce an additional metric (%Pur) specifically designed to assess the performance of the purification algorithm. Among all of the data assigned to any one cluster we simply determine the percentage of data that corresponds to the most dominant speaker, as determined according to reference transcriptions. The %Pur metric is the average purity for all speaker models after segmentation and clustering and performance is gauged by comparing %Pur before and after purification. Note that the DER is not appropriate for assessing purity as it penalises the case where there are more models than speakers - this is generally the case with our algorithm (the later resegmentation stage aims to reduce their number). Thereafter the final DER metric is the most suitable and is that used everywhere else in this paper.

Table 2 illustrates the purity for all three datasets both without and with purification (lines 2 and 3, RT'09 system and RT'09+Pur. respectively). The average, minimum and maximum cluster purity are shown in each case for the three different datasets. The results show that in all cases the average cluster purity increases after purification. Of particular note, is the general increase in the minimum cluster purity (with the exception of the Dev. Set), whereas the maximum purity only changes for the Eval. Set. Note that the lowest purities before purification (average, minimum and maximum) all correspond to the Eval. Set and also that the biggest improvement in minimum purity (54% cf. 41%) is also achieved on the Eval. Set. This goes someway to explain the behaviour noted above but it is nonetheless of interest to see the improvement in purity across the individual shows.

Figure 1 illustrates the %Pur metrics before and after purification (solid and dashed profiles respectively) for each of the 7 files in the RT'09 dataset (horizontal axis). The results show that where the initial models are already of high purity (e.g. the first and third shows) then the purification algorithm has little effect. However, when the initial clusters are of relatively poor purity (e.g. the fifth show) then purification leads to a marked improvement. For this particular show the cluster purity increases from 56% to 67% with purification. With few exceptions this behaviour is typical of that across the other datasets. Since the initial cluster purities are particularly bad for the Eval. Set (illustrated in Table 2) it is thus of

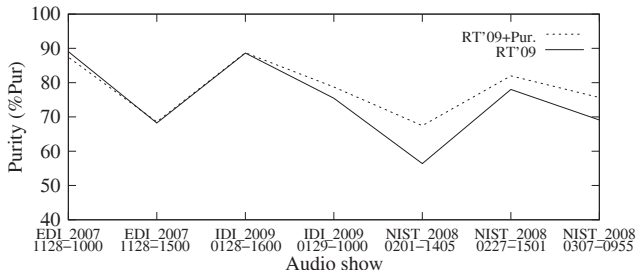


**Figure 1**. %Pur metrics for the NIST RT'09 dataset (MDM condition) before and after purification (solid and dashed profiles respectively).

no surprise that the effect of purification is greatest here. Even so, we note that other researchers have found that this dataset was more 'difficult' compared to previous datasets and the performance of our new system is also slightly inferior to that on the Dev. Set and Valid. Set even if the purification system reduces the difference.

### 4.4. SDM performance

Finally, since the performance of our MDM system is linked to beamforming performance, we now present an identical summary of results for the corresponding SDM datasets. This allows us to assess the performance of our core diarization system independently of beamforming performance. The diarization system is absolutely identical in every way except for the beamforming. Results are presented in Table 3 in an identical manner to those in Table 1. Here we see behaviour consistent with that for the MDM set. Performance is shown to improve across the three datasets and, for the Eval. Set, a DER of 21.1% with EM speaker modelling and purification compares well to 29.5% using our RT'07 system (MAP adaptation speaker modelling without purification) and 26.0% using our RT'09 system (EM speaker modelling without purification). These results correspond to relative improvements of 28% and 19% respectively. Even if the evaluation results are slightly worse than the development and validation results, the combination of EM speaker modelling and purification acts to stabilise the results across the three datasets for both MDM and SDM conditions.

### 4.5. Computational complexity

These improvements are at the expense of a small increase in computational cost. The submission criteria of the NIST RT evaluations [1] require the reporting of system efficiency in terms of a speed factor which gauges the efficiency of the system in relation to real time. Our RT'07 system with MAP speaker adaptation achieved a speed factor of 0.5. This system also included various pre-processing algorithms which greatly increased the efficiency of the speaker segmentation and clustering stage. They were, however, removed in the RT'09 system so that models are initialised on more data. This system achieved a speed factor of 1.5. The purification algorithm introduces a negligible overhead in processing time which increases the speed factor of our new system to 1.6. Compared to the speed factors of other systems published in the proceedings of the NIST RT evaluations our new system is still among the most efficient.

| System | Dev. Set | Valid. Set | Eval. Set |
|---|---|---|---|
| RT'07 | 26.4/24.1 | 24.5/21.3 | 29.5/24.7 |
| RT'09 | 22.7/20.0 | 18.3/15.0 | 26.0/21.5 |
| RT'09+Pur. | 21.1/18.3 | 17.8/14.4 | 21.1/16.0 |

**Table 3**. As for Table 1 except for the SDM condition.

## 5. CONCLUSIONS

The contributions in this paper include an improved approach to speaker modelling and a new cluster purification algorithm which, when applied to our state-of-the-art top-down speaker diarization system, collectively lead to improvements in stability and overall performance without sacrificing computational efficiency.

Enhancements to speaker modelling increase the chances of model impurities at initialisation but nonetheless result in a 35% relative improvement in DER for the MDM condition of a separate evaluation dataset when overlapping speech segments are scored. The new cluster purification delivers further relative improvements of 14% (44% when compared to the original RT'07 system) on the same data. On the SDM task the corresponding relative improvements are 28% and 19%.

These results, whilst not as good as those of the very best performing systems that were presented at the most recent NIST RT speaker diarization evaluations, greatly reduce the performance gap between state-of-the-art bottom-up and top-down systems. We believe that further, ongoing work to purify clusters as they are added will better realise the potential of top-down systems. The computational efficiency of such approaches is of particular appeal to many practical applications such as speaker indexing and other real-time applications. For these reasons it is our opinion that top-down approaches to speaker diarization warrant further attention.

## 6. REFERENCES

[1] 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.

[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] T. Nguyen et al., "The IIR-NTU Speaker Diarization Systems for RT 2009," in *RT'09, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, Florida, USA*, 2009.

[4] J.-F. Bonastre, F. Wils, and S. Meignier, "ALIZE, a free toolkit for speaker recognition," in *Proc. ICASSP'05*, Philadelphia, USA, March 2005, vol. 1, pp. 737–740.

[5] C. Fredouille and N. Evans, "The LIA RT07 speaker diarization system," in *Lecture notes in Computer Science - Multimodal Technologies for Perception of Humans*, Fiscus Stiefelhagen, Bowers, Ed. 2008, vol. 4625/2008, pp. 520–532, Springer.

[6] C. Fredouille, S. Bozonnet, and N. W. D. Evans, "The LIA-EURECOM RT'09 Speaker Diarization System," in *RT'09, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, Florida, USA*, 2009.

[7] A. Adami et al., "Qualcomm-ICSI-OGI features for ASR," in *Proc. ICSLP*, 2002, pp. 21–24.

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

[9] X. Anguera, C. Wooters, and J. Hernando, "Purity algorithms for speaker diarization of meetings data," in *Proc. ICASSP*, May 2006.

[10] H. Sun, T. L. Nwe, B. Ma, and H. Li, "Speaker diarization for meeting room audio," in *Proc. Interspeech*, September 2009.