An Integrated Top-Down/Bottom-Up Approach To Speaker Diarization

Simon Bozonnet[1], *Nicholas Evans*[1], *Corinne Fredouille*[2], *Dong Wang*[1] *and Raphaël Troncy*[1]

[1]EURECOM, BP193, F-06904 Sophia Antipolis Cedex, France
[2]University of Avignon, LIA/CERI, BP1228, F-84911 Avignon Cedex 9, France
{bozonnet,evans,wang,troncy}@eurecom.fr, corinne.fredouille@univ-avignon.fr

Abstract

Most speaker diarization systems fit into one of two categories: bottom-up or top-down. Bottom-up systems are the most popular but can sometimes suffer from instability from merging and stopping criteria difficulties. Top-down systems deliver competitive results but are particularly prone to poor model initialization which often leads to large variations in performance. This paper presents a new integrated bottom-up/top-down approach to speaker diarization which aims to harness the strengths of each system and thus to improve performance and stability. In contrast to previous work, here the two systems are fused at the heart of the segmentation and clustering stage. Experimental results show improvements in speaker diarization performance for both meeting and TV-show domain data indicating increased intra and inter-domain stability. On the TV-show data in particular, an average relative improvement of 32% DER is obtained.

Index Terms: speaker diarization, speaker segmentation, speaker clustering, system combination, SDM

1. Introduction

Speaker diarization relates to the problem of determining 'who spoke when' within an audio document. More precisely it involves segmenting the audio content and grouping together same-speaker segments via unsupervised identification. Much progress has been made in the domain over the last few years stemming somewhat from the international Rich Transcription (RT) evaluations spearheaded by the National Institute of Standards and Technology (NIST) in the US [1].

Most state-of-the-art speaker diarization systems fit into one of two categories: bottom-up or top-down. The bottom-up approach, commonly referred to as hierarchical agglomerative clustering, is by far the most popular. These systems have consistently produced the best performance in NIST evaluations [1]. The bottom-up approach is initialized with a number of models that exceeds the predicted number of speakers. Clusters are then successively merged until there remains only one for each speaker. In contrast, the top-down approach is initialized with a single speaker model. New models are added successively until the full number of speakers is reached.

Each of the approaches has its own strengths and weaknesses. Bottom-up systems can sometimes suffer from poor robustness in the stopping criterion [2] while top-down systems often suffer from poor initialization [3]. Hence both approaches can lead to unstable performance. It is thus of interest to integrate the two approaches to harness the merits of each system to improve stability and performance.

A 'piped' approach was reported in [4] where the output of a bottom-up system is applied to the input of a top-down system. Also reported in [4] is a 'fused' system where each approach is applied separately and the output labels are combined before a second resegmentation is applied. There are also many examples in the literature where systems are combined at the feature level, e.g. [5]. Some other examples include [6] where one system based on an agglomerative Information Bottleneck (aIB) approach is combined with a sequential Information Bottleneck (sIB) approach. Finally in [7] two different hierarchical clustering systems are coupled and used sequentially.

Only few of theses works, such as [6] and [7], involve truly integrated approaches. The contribution in this paper is a new top-down scenario within which a bottom-up system is integrated in order to provide for more robust initialization, and hence improved, more stable performance. The novelty lies in how the two approaches are combined. In contrast to previous work our approach fuses top-down and bottom-up approaches at the heart of the clustering and segmentation stage.

The remainder of this paper is organized as follows. Section 2 describes the two approaches to speaker diarization and their strengths and weaknesses. The new integrated system is described in Section 3. Our experimental work to evaluate the new system is presented in Section 4 and finally our conclusions are presented in Section 5.

2. Approaches to Diarization

Bottom-up and top-down systems differ in the way that each is initialized. Bottom-up systems are initialized with a large number of clusters which are gradually merged while top-down systems are initialized with a single cluster before more are introduced through cluster splitting. In this section we briefly describe the top-down and bottom-up approaches to speaker diarization that we aim to integrate.

2.1. Top-down

Developed by LIA using the freely available open source ALIZE toolkit [8], our top-down system is based upon an evolutive hidden Markov model (E-HMM) approach [9] to speaker diarization where states correspond to speakers and transitions between states correspond to speaker turns. Speakers are modeled with Gaussian mixture models (GMMs). Details of the E-HMM approach have been published previously and, with the exception of a recently introduced purification stage [3], the system is exactly the same as that used for LIA-EURECOM's submission [10] to the NIST RT'09 evaluation [1].

In summary, the top-down system is composed of five stages. Speech Activity Detection (SAD) is performed first to identify the speech segments which are used for the second stage segmentation and clustering. The new purification step follows before the fourth stage resegmentation step which aims to remove irrelevant speakers. The fifth stage involves feature normalization and a final resegmentation.

The main weakness of the top-down approach lies in how new speakers are introduced to the E-HMM model. First, a general GMM model $L0$ is learned from all the available speech

segments with an expectation maximization (EM) algorithm. New speaker models are iteratively introduced using a single segment of speech from the pool of segments assigned to $L0$ according to some criterion, e.g. taking the longest segment. As described in [3] there is an inherent trade-off between segment purity and data quantity. If the selected segment is too large then the chances of it containing data from more than a single speaker is increased. If the segment is too small then there is a greater chance that it is pure (contains speech from a single speaker only), however there may not be sufficient data with which to train a reliable model. In the top-down approach, the quality of the initialization of one speaker model will affect the initialization of other speaker models introduced subsequently and, as a consequence, performance can vary greatly depending on the quality of the initialization. It is for this reason that purification has proved to be so effective in top-down approaches to speaker diarization [3]. Their susceptibility to poor speaker model initialization is the main weakness of top-down systems. We seek here to avoid it by harnessing the merits of a bottom-up system through an integrated top-down/bottom-up approach.

2.2. Bottom-up

The bottom-up system used for all experiments reported here is our own implementation of a system developed by I2R for their entry to the most recent NIST RT'09 evaluation [1]. The approach is composed of 3 main stages: a SAD stage, an initialization with sequential EM, and finally an agglomerative hierarchical clustering. Full details can be found in [11]. Our implementation of the system was developed using the same ALIZE toolkit that was developed and used by LIA to implement the E-HMM top-down speaker diarization system described above.

Bottom-up systems do not suffer from the same initialization problems as top-down systems. Initialization is generally based on a linear under-clustering of the data where the audio document is simply segmented into a number of equal length segments. Clusters are progressively re-trained through several steps of alignment and model adaptation, before merging is performed to reduce their number. In contrast to other bottom-up approaches e.g. [12], in the system described in [11] models are tuned gradually with sequential EM. A fraction of the least likely data in each cluster are unclassified and are then sequentially reassigned with embedded adaptation. This acts to purify the clusters as they are formed and provides for robust initialization. However, bottom-up systems can sometimes suffer from instabilities related to the merging and stopping criteria which can lead to situations of over- or under-clustering in the final segmentation hypothesis [2].

3. Integrated Approach

Given the drawbacks of each system, it is of interest to combine them in an integrated approach. We propose a new system whose skeleton is based upon the LIA-EURECOM top-down system, described in Section 2.1, but where each speaker model is trained by following an integrated bottom-up approach with sequential EM training, as presented in Section 2.2.

As presented in Section 2.1, and as illustrated in Figure 1, the first step involves the learning of a general model $L0$ which is tuned by EM using all the available speech segments. Then initialization with sequential EM as described in [11] is applied using all of the speech data assigned to model $L0$. However instead of splitting the data uniformly into 30 clusters as presented in [11], the speech segments assigned to model $L0$ are divided linearly into 30-second sub-clusters (3 in Figure 1 labeled A, B, and C). Our experiments show that this approach gives better results. Then the steps described in [11] are performed 10 times

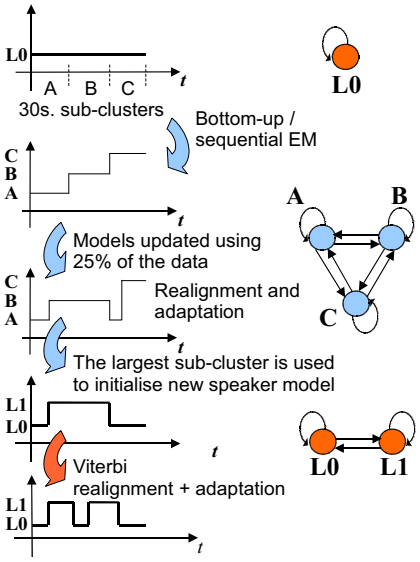

Figure 1: The integrated approach.

on the resulting sub-clusters: 25% of the data which best fits the corresponding model are considered as classified whereas other data are unlabelled. The models are updated using only the classified data and a decoding is performed where only a fraction of the newly classified data are reassigned to their nearest sub-clusters. Several steps of Viterbi realignment and adaptation are performed until all the data are classified. As illustrated in Figure 1 the sub-cluster which is assigned the greatest amount of speech data is used to introduce a new speaker $L1$ into the E-HMM system. The data in all other sub-clusters are assigned back to $L0$. Several iterations of Viterbi decoding and adaptation are performed with the E-HMM until the system is stable.

This process is repeated in exactly the same way to add additional speakers to the E-HMM until there is no longer sufficient data assigned to $L0$ with which to create a new speaker model. Thus in this approach we harness the better initialization provided by the bottom-up approach to initialize each new speaker model in the top-down approach. In contrast to previous work the two systems are thus fused at the heart of the clustering and segmentation stage.

4. Experiments and results

The experiments presented in this section aim to demonstrate the improvement in performance and stability of the new integrated diarization system. We describe the experimental conditions and datasets before our results.

4.1. Systems

In the following we present a performance comparison of top-down, bottom-up and integrated approaches. In order to make the comparison fair and meaningful, and to focus the analysis on speaker segmentation and clustering performance only (the assessment metric also reflects SAD performance), the same SAD was used for all experiments reported here. In addition, a normalization and resegmentation stage, as described in [13], are applied for each system since they bring consistent benefits. Finally, all results are presented with/without the purification step introduced in [3] in order to determine whether or not additional benefits are obtained over those brought by the integrated system. Whether or not improvements are obtained with purifica-

System	Dev. Set	RT07	RT09	GE
Top-down	22.7/20.0	18.3/15.0	26.0/21.5	40.4/36.0
Top-down+Pur.	21.1/18.3	17.8/14.4	21.1/16.0	38.5/33.9
Bottom-Up	21.7/18.9	23.8/20.8	19.1/13.5	33.7/29.0
Bottom-Up+Pur.	21.6/18.8	22.7/19.6	27.0/21.8	33.9/29.1
Integrated System	17.3/14.3	16.5/13.0	23.8/18.6	30.9/26.3
Integrated System+Pur.	16.2/13.2	16.4/12.9	23.5/18.2	28.4/23.2

Table 1: % Speaker diarization performance in terms of DER with/without scoring overlapping speech. Results illustrated without and with (+Pur.) purification for the Dev. Set and the RT'07, RT'09 and GE datasets.

tion also reflects initialization robustness.

Where we refer to a bottom-up system, all results reported in this paper reflect those obtained with our own implementation of I2R's system which departs somewhat from the original system [11] as described above. In addition, whereas the modified Information Change Rate (ICR) merging criterion based on information theory was used [2, 11], we prefer the T_s stopping criterion to the Rho criterion presented in [11, 14] since it leads to better performance in our implementation.

4.2. Datasets

Each system is assessed using 4 different datasets: 3 NIST RT meeting datasets and 1 TV-show dataset that we used in previous work [15]. All the meeting data involves single distant microphone (SDM) data, so only a single channel is used for each show. To optimize each system we used a development dataset comprising 23 meeting shows from the NIST RT'04, '05 and '06 datasets. Evaluation is performed on the NIST RT'07 and RT'09 datasets using the same system without modification. To assess the stability of the system to totally new data, it was then applied, without modification, to the TV-show data. This dataset is composed of 19 hours of data from the 'Grand Echiquier' (GE) corpus which comprises over 50 French-language TV-talk-show programs from the 1970-80s [15]. The GE database has been used previously among both national and European multimedia research projects, e.g. The European K-space Network of Excellence [16]. In all cases, speaker diarization performance is measured in terms of the diarization error rate (DER).

4.3. Performance

Results for the 4 different datasets are presented in Table 1 where the diarization error rate (DER) is given with/without the scoring of overlapping speech. Since none of the systems assessed in this paper is capable of detecting or labelling overlapping speech, we refer to scores where overlapping speech is ignored. Rows 2 and 3 of Table 1 present results obtained with our baseline system [13] and the same system with the recently added purification step [3]. The purification step delivers a significant increase in performance for the RT'09 dataset (from 21.5% to 16.0%) and smaller improvements on other datasets. The differences in theses scores reflect impurities caused through poor initialization and the subsequent improvement obtained with purification. Results for other systems are also given with/without the purification step to evaluate initialization performance and purity.

Upon comparison of results for the top-down and bottom-up systems, with and without purification respectively (rows 3 and 4), we see that the top-down system gives similar results for the development set (18.3% vs. 18.9%). We observe that the top-down system gives better performance for the RT'07 dataset (14.4% vs. 20.8%) whereas for the RT'09 and GE datasets, the bottom-up system gives better performance (13.5% vs. 16.0% and 29.0% vs. 33.9%). Results for the bottom-up system with purification are presented in row 5 of Table 1. This time, no consistent improvement in performance is obtained with purification. For the RT'09 dataset in particular, performance is even worse (21.8% vs. 13.5%). These observations would seem to suggest that the clusters produced by the bottom-up system are inherently purer than those produced by the top-down system and thus no improvements in performance are obtained when purification is applied. Moreover, bottom-up systems often result in over-clustering of the data, i.e. a tendancy to produce a high number of small clusters. In this case the purification algorithm results in unreliable models trained on insufficient data. This effect is noticed with the RT'09 dataset in which there are a number of speakers with very small floor time and in these cases the performance of the bottom-up system deteriorates with purification.

Finally rows 6 and 7 of Table 1 show results for the new integrated system that is described in Section 3, once again with and without purification respectively. Referring first to results without purification and their comparison to results for the baseline system (row 2), we observe largely consistent improvements in performance. Relative improvements of 29%, 13%, 13% and 27% are obtained for the development, RT'07, RT'09 and GE datasets respectively. With added purification only small improvements in performance are obtained for the development and GE datasets (8% and 12% relative improvements respectively), indicating once again that the integrated approach succeeds in improving initialization robustness: no consistent improvement is achieved with purification. For the RT'09 dataset, however, the performance with the integrated approach is worse than performance with the top-down system with purification and the bottom-up system (16.0% and 13.5% respectively). Whilst this is disappointing we note that the RT'09 dataset has a particularly high degree of overlapping speech and very short speech segments. Other researchers have also reported difficulties with this particular dataset[1]. We also note that the decrease in performance is concentrated on only one show whereas for other shows performance largely improves or is unaffected. When this one file is removed from the scoring the difference between the performance of the integrated system and the top-down and bottom-up systems is reduced. For the integrated system the performance is 15.3% and 15.6% with and without purification respectively.

4.4. Stability

The box plots in Figure 2 depict performance and stability for each of the 4 systems: the baseline top-down system with purification, the bottom-up system both with and without purification (both are included since performance with purification is inconsistent) and the new integrated system. All plots illustrate the spread in performance across an entire dataset, first for meeting data and second for the TV-show data. The rectangular boxes show the inter-quartile range (IQR) and illustrate the intra-domain stability, while the middle line indicates the median performance. The comparison of any corresponding pair of box plots (one for meeting data, one for TV-show data) serves to illustrate the inter-domain stability.

The first two box plots illustrate performance for the baseline top-down system with purification, first for meetings and then for TV-show data. We observe that performance differs greatly between the two datasets. The third and fourth box plots

[1] from discussions with other participants at the NIST RT'09 workshop and as illustrated in the NIST workshop presentation available at http://www.itl.nist.gov/iad/mig/tests/rt/2009/workshop/RT09-SPKR-v3.pdf

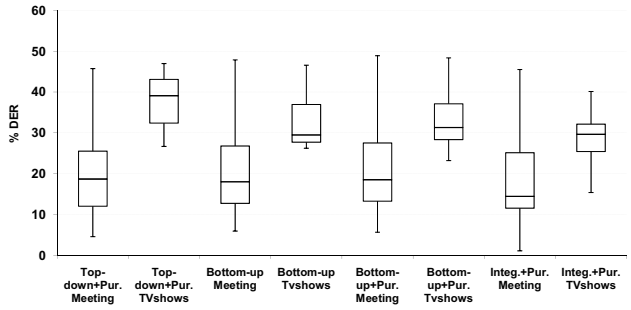

Figure 2: Box plot of the variation in DER for the four systems and both meeting (averaged across the Dev. Set, RT'07 and RT'09 datasets) and TV-show data (GE dataset).

illustrate comparative performance for the bottom-up system *without* purification whereas the fifth and sixth box plots show performance for the same system *with* purification, in both cases for meeting data and then TV-show data. For meetings the median performance is almost the same as for the top-down system with purification whereas for the TV-show data the bottom-up system achieves significantly better performance. There is negligible difference between results with or without purification which again confirms the benefit of the initialisation process in the bottom-up system.

The last two box plots show performance for the new integrated system. Compared to the baseline the spread in performance with meeting data is unchanged whereas the median decreases noticeably. There is thus an overall improvement in performance, however, whilst the best score also decreases, the worst score remains unchanged. The largest improvement is achieved for the TV-show data for which significant decreases in both the IQR and median performance are observed. We also notice that the difference between the box plots for meeting and TV-show data is less for the integrated system than it is for any other system. Thus the inter-domain stability is greatly improved with the new integrated approach.

5. Conclusions

This paper reports a new integrated speaker diarization system which harnesses the merits of both top-down and bottom-up approaches through their fusion at the heart of the clustering and segmentation stage. We aim to avoid weaknesses in initialisation which often afflict top-down approaches and those sometimes associated with the merging or stopping criterion in bottom-up systems. For meeting data, improvements are observed for two out of the three datasets. With purification this corresponds to an average relative improvement of 28% DER on the development set. On the separate RT'07 evaluation dataset we obtain a relative improvement of 10% DER. Even if results are not improved for the problematic RT'09 dataset, performance remains reasonably competitive and the degradation in performance corresponds to one file only.

Speaker diarization results for a TV-show dataset shows that the new integrated system is significantly more competitive on this data than is our baseline system. Here a relative improvement of 32% is achieved and serves to show that the new integrated system also improves inter-domain stability.

A possible direction for future work involves optimising the criterion used to select the subcluster that is used in the top-down approach. This may lead to further improvements since the size of the cluster (the quantity of data assigned to it as used in this work) is not necessarily the best indicator of quality. Criteria based on the cluster variance, for example, might give better results.

6. Acknowledgments

This work was partially supported by the French Ministry of Industry (Innovative Web call) under contract 09.2.93.0966, 'Collaborative Annotation for Video Accessibility' (ACAV).

7. 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] K. Han, S. Kim, and S. Narayanan, "Strategies to improve the robustness of agglomerative hierarchical clustering under data source variation for speaker diarization," *IEEE TASLP*, vol. 16, no. 8, pp. 1590–1601, 2008.

[3] S. Bozonnet, N. W. D. Evans, and C. Fredouille, "The LIA-EURECOM RT'09 Speaker Diarization System: enhancements in speaker modelling and cluster purification," in *Proc. ICASSP'10*, Dallas, Texas, USA, March 14-19 2010.

[4] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," in *CSL, selected papers from the Speaker and Language Recognition Workshop (Odyssey'04)*, 2006, pp. 303–330.

[5] V. Gupta, P. Kenny, P. Ouellet, G. Boulianne, and P. Dumouchel, "Combining gaussianized/non-gaussianized features to improve speaker diarization of telephone conversations," in *Signal Processing letters, IEEE*, 2007, pp. 1040–1043.

[6] D. Vijayasenan, F. Valente, and H. Bourlard, "Combination of agglomerative and sequential clustering for speaker diarization," in *Proc. ICASSP*, Las Vegas, USA, 2008, pp. 4361–4364.

[7] E. El-Khoury, C. Senac, and S. Meignier, "Speaker diarization: combination of the LIUM and IRIT systems," in *Internal report*, 2008.

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

[9] S. Meignier, J.-F. Bonastre, and S. Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," in *Proc. Odyssey Speaker and Language Recognition Workshop*, Chania, Creete, June 2001, pp. 175–180.

[10] C. Fredouille, S. Bozonnet, and N. W. D. Evans, "The LIA-EURECOM RT'09 Speaker Diarization System," in *RT'09, NIST Rich Transcription Workshop*, 2009.

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

[12] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Lecture notes in Computer Science - Multimodal Technologies for Perceptionof Humans*, vol. 4625/2008. Springer, 2008, pp. 509–519.

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

[14] T. H. Nguyen, E. S. Chng, and H. Li, "T-test distance and clustering criterion for speaker diarization," in *Proc. Interspeech*, Brisbane, Australia, 2008.

[15] S. Bozonnet, F. Vallet, N. W. D. Evans, S. Essid, G. Richard, and J. Carrive, "A multimodal approach to initialisation for top-down speaker diarization of television shows," EURECOM, Sophia Antipolis," Technical Report RR-10-239, 2010.

[16] K-Space, "The European K-Space Network Of Excellence," http://www.k-space.eu/.