A comparative study of bottom-up and top-down approaches to speaker diarization

Nicholas Evans, Simon Bozonnet, Dong Wang, Corinne Fredouille, Raphaël Troncy

A comparative study of bottom-up and top-down approaches to speaker diarization

Nicholas Evans, *Member, IEEE,* Simon Bozonnet, *Student Member, IEEE,* Dong Wang, *Associate Member, IEEE,* Corinne Fredouille and Raphaël Troncy

***Abstract*—This paper presents a theoretical framework to analyze the relative merits of the two most general, dominant approaches to speaker diarization involving bottom-up and top-down hierarchical clustering. We present an original qualitative comparison which argues how the two approaches are likely to exhibit different behavior in speaker inventory optimization and model training: bottom-up approaches will capture comparatively purer models and will thus be more sensitive to nuisance variation such as that related to the speech content; top-down approaches, in contrast, will produce less discriminative speaker models but, importantly, models which are potentially better normalized against nuisance variation. We report experiments conducted on two standard, single-channel NIST RT evaluation datasets which validate our hypotheses. Results show that competitive performance can be achieved with both bottom-up and top-down approaches (average DERs of 21% and 22%), and that neither approach is superior. Speaker purification, which aims to improve speaker discrimination, gives more consistent improvements with the top-down system than with the bottom-up system (average DERs of 19% and 25%), thereby confirming that the top-down system is less discriminative and that the bottom-up system is less stable. Finally, we report a new combination strategy that exploits the merits of the two approaches. Combination delivers an average DER of 17% and confirms the intrinsic complementary of the two approaches.**

***Index Terms*—speaker diarization, segmentation, clustering, rich transcription**

I. Introduction

THE ever-expanding volume of available audio and multimedia data has elevated technologies related to content indexing and structuring to the forefront of research. Speaker diarization [1], [2], commonly referred to as the 'who spoke when?' task, is one such example. Speaker diarization involves identifying the number of speakers within an acoustic stream and the labeling of intervals when each speaker is active. Stemming partly from the internationally competitive Rich Transcription (RT) evaluations [3] administered by the National Institute for Standards and Technology (NIST) in the U.S., speaker diarization has emerged as a prominent, core enabling technology in the wider speech processing research community.

A general speaker diarization system schematic is illustrated in Fig. 1. The first system elements involve noise reduction and beamforming, with the latter only being applied to obtain a single pseudo channel when multiple input channels are available. Following feature extraction speech activity detection is then normally performed to remove non-speech segments before the core stage of the general speaker diarization system which involves segmentation and clustering.

Whilst there are examples that do not match this dichotomy, two general approaches to segmentation and clustering have come to prominence through the official NIST RT evaluations and now dominate the literature. They involve bottom-up and top-down approaches to hierarchical clustering [2]. All the speaker diarization systems submitted to the NIST RT evaluations fit into one of these two categories which are the focus throughout this paper. Bottom-up systems are initialized with a large number of clusters which are gradually merged whereas the top-down systems are initialized with a single cluster before more are introduced through cluster splitting. Both processes are iterative and are repeated until the optimal number of speakers is reached. The bottom-up approach is an example of agglomerative hierarchical clustering whereas the top-down approach is an example of divisive hierarchical clustering.

The bottom-up approach is by far the most popular and systems based on this approach have consistently achieved the best levels of performance in the NIST RT evaluations, e.g. [4], [5], although top-down systems also achieve respectable results [6]. While some have reported that bottom-up approaches are more robust than their top-down counterparts [1] our own work [7] shows that the two approaches give comparable results, with neither being consistently superior to the other. Purification techniques which aim to 'purify' clusters of speech from all but the dominant speaker, are reported by many to give significant and consistent improvements with bottom-up approaches [8], [9], [10]. Our experience, however, shows that performance can sometimes deteriorate when purification is applied to bottom-up strategies but that it leads to consistent improvements in top-down systems [7]. These observations led us to investigate the two diarization approaches more thoroughly and to study their relative merits.

In this paper, we present an original theoretical framework for speaker diarization and use it to compare the bottom-up and top-down approaches to speaker diarization. The study shows that the two clustering approaches are similarly effective in searching for the optimal number of speakers but behave differently in discriminating between individual speakers and in normalizing unwanted acoustic variation, i.e. that which does not pertain to different speakers. This can make top-down

N. Evans, S. Bozonnet, D. Wang and R. Troncy are with the Department of Multimedia Communications, EURECOM, France, e-mails: {evans, bozonnet, wangd, troncy}@eurecom.fr.

C. Fredouille is with the University of Avignon, CERI/LIA, France, email: corinne.fredouille@univ-avignon.fr.

Manuscript received September 1, 2010; revised January 31, 2011.

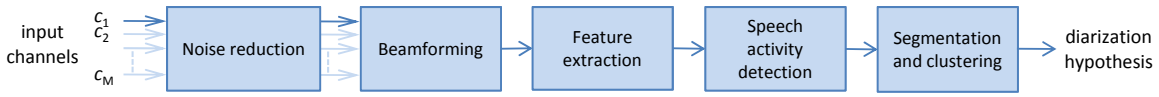

Fig. 1. An overview of a typical speaker diarization system with one or multiple input channels.

systems more stable but less discriminative, and vice versa for bottom-up systems. We also explain why purification works well with top-down approaches but why it can degrade results when applied to bottom-up systems. Finally, the study leads to a combined approach to speaker diarization which exploits the benefits of both bottom-up and top-down approaches.

The remainder of this paper is organized as follows. In Section II we present a theoretical framework for bottom-up and top-down hierarchical clustering approaches to speaker diarization. This includes a formal definition of the task and an analysis of the challenges that must be addressed by practical speaker diarization systems. The generalized bottom-up and top-down approaches are reviewed and compared on a qualitative basis in Section III. In Section IV, we describe our own specific bottom-up and top-down experimental systems and approaches to purification and system combination. Results are reported in Section V before conclusions and some thoughts for future work are presented in Section VI.

II. Speaker diarization: a theoretical framework

In this section we propose a theoretical framework for the speaker diarization task. Although it is not the only possible approach, the formulation presented is representative of state-of-the-art technologies based on probabilistic modeling. All the assumptions made in theory development are consistent with modern speaker diarization systems that have been entered into the official NIST RT evaluations [3].

Based on the probabilistic framework, we analyze the main challenges that must be addressed in related practical systems. This analysis leads naturally to the two principal approaches to speaker diarization, namely the bottom-up and top-down approaches that are studied and compared later in this paper.

A. Task definition

Speaker diarization can be defined as an optimization task on the space of speakers given the audio stream that is under evaluation. We first assume that non-speech segments have been removed from the acoustic stream and that features are extracted such that the remaining speech information is represented by a stream of acoustic features O. Letting S represent a speaker sequence and G a segmentation of the audio stream by S, then the task of speaker diarization can be formally defined as follows:

$$(\tilde{S}, \tilde{G}) = \underset{S,G}{\operatorname{argmax}} P(S, G|O), \tag{1}$$

where $\tilde{S}$ and $\tilde{G}$ represent respectively the optimized speaker sequence and segmentation, i.e. who (S) spoke when (G). We can factorize (1) into a posterior probability by applying the Bayesian rule:

$$\begin{aligned}(\tilde{S}, \tilde{G}) &= \underset{S,G}{\operatorname{argmax}} \frac{P(S, G)P(O|S, G)}{P(O)} \\ &= \underset{S,G}{\operatorname{argmax}} P(S, G)P(O|S, G),\end{aligned} \tag{2}$$

where $P(O)$ is suppressed since it is independent of S and G. (2) shows that two models are required in order to solve the optimization task: acoustic models which describe the acoustic attributes of each speaker, constituting $P(O|S, G)$, and speaker turn models which describes the probability of a turn between speakers with a given segmentation, constituting $P(S, G)$.

Usually the acoustic models are implemented as Gaussian mixture models (GMMs). Letting S_i denote the i-th speaker in S, and O_i the corresponding speech segment according to G, we have:

$$P(O|S, G) = \prod_i P(O_i|\lambda_{S_i}, G), \tag{3}$$

where λ_{S_i} denotes the GMM speaker model for speaker S_i.

By applying various different assumptions one can obtain different forms of the speaker turn model. For example, if we assume that the speaker labels either side of the turn are irrelevant and take only the utterance duration into account then we have the following duration model:

$$P(S, G) = P(G), \tag{4}$$

where $P(G)$ can be modeled with a normal or Poisson distribution for example. Alternatively, and as is common in practice, one may assume a uniform distribution and thus omit the turn model entirely. Substituting (3) and (4) into (2) we obtain:

$$(\tilde{S}, \tilde{G}) = \underset{S,G}{\operatorname{argmax}} P(G) \prod_i P(O_i|\lambda_{S_i}, G), \tag{5}$$

which provides a full solution to the speaker diarization problem.

B. Challenges

In practice, the implementation of a practical speaker diarization system is rather more complex than may first appear from the basic framework presented above. The first challenge involves the optimization of the speaker sequence S in (5). This is not straightforward since the inventory of S is unknown, i.e. we do not know how many speakers N there are within the acoustic stream. This means that it is not possible to optimize the speaker sequence S without a jointly-optimized speaker inventory. Second, although we suppose that a set of acoustic models can reliably represent the acoustical characteristics of the speakers, the speech signal O is rather

complex. Whilst the acoustic models depend fundamentally on the speaker, they also depend on a number of other nuisance factors such as the linguistic content, for example the words or phones pronounced, which are not related specifically to the speaker. In the following we assume for simplicity that the major nuisance variation relates only to the phone class of uttered speech, which we denote as Q, though other acoustic classes are also valid. Due to its significant effect on the speech signal, Q should appear in the solutions and must be addressed appropriately.

To formulate a solution which addresses these two challenges, we first introduce the speaker inventory Δ, and let $\Gamma(\Delta)$ represent all possible speaker sequences. Returning to (2) we derive the solution as follows:

$$\begin{aligned}(\tilde{S},\tilde{G},\tilde{\Delta}) &= \underset{S,G,\Delta:S\in\Gamma(\Delta)}{\operatorname{argmax}} P(S,G|O)\\ &= \underset{S,G,\Delta:S\in\Gamma(\Delta)}{\operatorname{argmax}} P(S,G)\sum_{Q} P(O,Q|S,G)\\ &= \underset{S,G,\Delta:S\in\Gamma(\Delta)}{\operatorname{argmax}} P(S,G)\sum_{Q} P(O|S,G,Q)P(Q|S)\\ &= \underset{S,G,\Delta:S\in\Gamma(\Delta)}{\operatorname{argmax}} P(S,G)\sum_{Q} P(O|S,G,Q)P(Q), \quad (6)\end{aligned}$$

where Q is naturally independent of G and we have further assumed it to be independent of S. The solution reveals two important issues that any practical speaker diarization system must address. First, the speaker inventory Δ must be optimized together, not only with the speaker sequence S, but also the segmentation G. There is no analytical solution for Δ and so a trial-and-error search is typically conducted. This search can be either from a smaller inventory to a larger inventory, or from a larger inventory to a smaller inventory. These strategies correspond respectively to the top-down and bottom-up approaches to speaker diarization. Secondly, when comparing (2) and (6), we see that:

$$P(O|S,G) = \sum_{Q} P(O|S,G,Q)P(Q). \quad (7)$$

This means that in the optimization task one should either use a phone-independent model $P(O|S,G)$ and apply (2), or a phone-dependent model $P(O|S,G,Q)$ with prior knowledge of $P(Q)$ and apply (6). Due to its simplicity and effectiveness, most speaker diarization systems nowadays adopt the former approach. For such a system $P(O|S,G)$ must be trained with speech material containing all possible phones, otherwise Q will be not marginalized. In other words, for a phone-independent system, acoustic speaker models must be *normalized* across phones Q to ensure that the resulting model is phone-independent, otherwise optimization according to (2) will be suboptimal.

In summary, a practical diarization system should incorporate an effective search strategy to optimize the speaker inventory Δ, and a set of well-trained speaker models to infer the speaker sequence S and segmentation G. Ideally, the models should be most discriminative for speakers and fully normalized across phones. From this perspective, the direction in which the optimal speaker inventory is searched for (bottom-up or top-down) is inconsequential. Searching from either direction will in any case arrive at the optimal inventory[1]. However, the merging (bottom-up) or splitting (top-down) operations in the search process are likely to impact upon the discriminative power and phone-normalization of the intermediate and final speaker models. Therefore, the two approaches will exhibit different behaviors and relative strengths and shortcomings in practice. This is the starting point of our analysis for these two approaches.

III. Approaches to speaker diarization

In this section we review the general bottom-up and top-down approaches to speaker diarization. The two approaches constitute the segmentation and clustering component in Fig. 1 and encapsulate the trial-and-error search for an optimal speaker inventory Δ and thus the optimization of S and G. The presentation in this section relates to the most general of bottom-up and top-down approaches to hierarchical clustering and, for the most part, it expressly avoids any relation to specific systems. Details of our own implementations are presented in Section IV.

Following the presentation of the two different approaches we outline some additional assumptions which infer new hypotheses related to their relative merits and shortcomings. They are discussed in this section in a purely qualitative context. In Section V the aim is then to compare the behavior of our specific implementations to the hypotheses presented here.

In both approaches presented below the aim is to model each of the N true speakers with a single GMM. Speaker turns are represented by transitions between models thus forming an ergodic hidden Markov model (HMM) in which each state represents a speaker and where all states are fully connected. The difference between the bottom-up and top-down approaches lies in where the trial-and-error search starts from and how an optimal set of GMM speaker models is derived.

A. Bottom-up

The bottom-up approach is often referred to as agglomerative hierarchical clustering (AHC). The procedure is illustrated to the left of Fig. 2 which shows how clustering begins with a larger speaker inventory (bottom) before similar clusters are merged to obtain a smaller, more optimal size (top). Only a single iteration is illustrated in Fig. 2 and in this example the process stops when two clusters are obtained. The resulting diarization hypotheses are illustrated in the left column with the corresponding ergodic HMMs in the middle column.

The search procedure starts with a model initialization which involves over-segmenting and under-clustering the acoustic stream into a larger number of clusters than the assumed number of true speakers, and training a GMM model on the acoustic data in each cluster. Various approaches can

[1]We assume that the number of speakers is known approximately so that the bottom-up approach is initialized with more clusters than true speakers in order to avoid the risk of over-clustering.

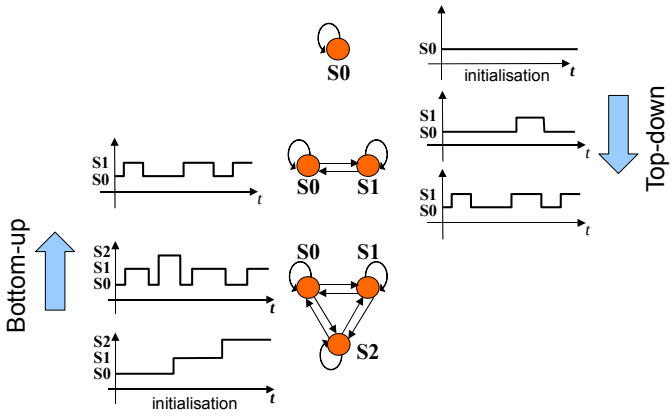

Fig. 2. An illustration of the bottom-up (left) and top-down (right) approaches to speaker diarization. Here there are $N = 2$ true speakers and the bottom-up approach is initialized with 3 clusters.

be applied to formulate the initial segmentation but linear segmentation is commonly used [11]. In the artificial example illustrated in Fig. 2 we assume that there are not more than $N = 2$ true speakers. Initialization produces 3 GMMs which are connected to form a 3-state, ergodic HMM[2]. Using the initial HMM/GMMs the acoustic stream is re-segmented by Viterbi realignment before the models are refined according to the new segmentation. New models are re-estimated with an expectation maximization (EM) procedure which converges to a stable segmentation and a set of locally optimized GMMs after several iterations. Next, the new models are examined in a pairwise fashion, and the two most similar are merged together to form a new GMM. Various distance metrics can be used to estimate model similarity and hence control merging. The most popular approaches involve the Bayesian information criterion (BIC) [12] and its variants. After a number of iterations of realignment and re-estimation a new, stable diarization hypothesis is obtained before the next cycle of model merging is considered. This process is repeated until an optimal speaker inventory (2 speakers in Fig. 2) is obtained according to some stopping criteria, which may also be based upon a BIC criterion. In each iteration, the number of speakers is reduced by one, and the speaker sequence (S) and segmentation (G) are optimized according to (1).

The bottom-up approach is the most popular and has achieved general success in the NIST RT evaluations [13], [14], [15], [16], [17] in addition to data from other domains [18], [19]. Nonetheless, some authors report that instabilities related to initialization [20], model merging and the sensitivity of the stopping criterion [21] might degrade its performance.

B. Top-down

The top-down approach operates from a smaller speaker inventory to a larger speaker inventory and is a form of divisive hierarchical clustering (DHC). In the example shown to the right of Fig. 2, the approach starts with a single general speaker model and constructs an optimal speaker inventory by introducing new speakers one-by-one.

Initialization involves the training of a general speaker model, denoted by $S0$, with all the available acoustic data. A new speaker model, denoted by $S1$, is then introduced and trained with some appropriate data from the general speaker model. Various approaches may be used to select the segment but the single largest segment identified from the speech activity detection (SAD) output has proved to give the most consistent performance [6]. As with the bottom-up approach, several iterations of Viterbi realignment and EM training are applied to iteratively refine the model, until a stable segmentation is obtained. New speakers are then added in the same way by the repeated splitting of existing models followed by several iterations of Viterbi realignment and EM training. The process continues until the optimal speaker inventory is obtained according to some stopping criteria, e.g. when there is no longer sufficient data with which to introduce a new speaker or when an upper limit on the size of the speaker inventory is reached. This process is illustrated in Fig. 2, where the process starts with a single general speaker model and stops with an inventory containing $N = 2$ speakers.

The top-down approach to speaker diarization is less popular than its bottom-up counterpart but has nonetheless been shown to give competitive performance in NIST RT evaluations [6], [22]. Compared to the bottom-up approach, which reduces the number of models at each iteration through cluster merging, the top-down approach increases the number at each iteration through cluster splitting. In the artificial example illustrated in Fig. 2 the diarization hypothesis obtained with the two approaches is the same and thus the differences between the two approaches may seem insignificant. In practice, however, they cause distinctly different behavior in terms of diarization performance and system stability, as we now discuss.

C. A qualitative comparison

The bottom-up and top-down approaches to speaker diarization are fundamentally opposing strategies. The bottom-up approach is a specific-to-general strategy whereas the top-down approach is general-to-specific. The latter will produce more reliably trained models as relatively more data are available for training. However, the models are likely to be less discriminative until sufficient speakers and their data are liberated to form distinct speaker models. The bottom-up approach, in contrast, is initialized with a larger number of models and is there more likely to discover specific speakers earlier in the process, however the models may be weakly trained until sufficient clusters are merged.

The two approaches thus have their own strengths and weaknesses and are therefore likely to exhibit different behavior and results. In the following we discuss some particular characteristics in further detail with the aim of better illuminating their potential merits.

1) Discrimination and purification: A particular advantage of the bottom-up approach rests in the fact that it is likely to capture comparatively purer models. Whilst they may correspond to a single speaker, they may also correspond

[2]In practice the number of initial clusters would be much greater than the assumed number of true speakers.

to some other acoustic unit, for example a particular phone class. This is particularly true when short-term cepstral-based features are used, though recent work with prosodic features has potential to encourage convergence specifically toward speakers [23]. In contrast, since it initially trains only a small number of models using relatively larger quantities of data, the top-down approach effectively normalizes phone classes, but it also normalizes speakers at the same time. To achieve the best discriminative power *across speakers*, a purification step becomes essential for both approaches: for the bottom-up approach, it is necessary to purify the resulting models of interference from phone variation, whereas for the top-down approach it is necessary to purify the resulting models of data from other speakers. Purifying phones involves phone recognition which is usually rather costly; purifying speakers, however, is much easier with some straightforward assumptions. We have achieved significant improvements in diarization performance using purification in our top-down approach. This recent work is presented in Section IV-D.

2) Normalization and initialization: Theoretically, the EM algorithm ensures that both the bottom-up and top-down approaches will converge to a local maximum of the objective function for a fixed size Δ. If the differences between speakers is the dominant influence in the acoustic space then we can safely assume that the local maximum represents an optimal diarization on speakers, as opposed to any other acoustic class. In this case, initial models are not predominantly important, and thus both bottom-up and top-down approaches will tend to provide similar diarization results. However, in addition to the speaker the acoustic signal bears a significant influence from the linguistic contents, and more specifically the phones. Therefore, the local maximums of the objective function may correspond to phones Q instead of speakers S if the speaker models are not well normalized, i.e. Q is not fully marginalized. This analysis highlights a major advantage of the top-down approach to speaker diarization: by drawing new speakers from a potentially well-normalized background model, newly introduced speaker models are potentially more reliable than those generated by linear initialization and model merging in the bottom-up approach.

An interesting point derived from the above analysis is that the bottom-up and top-down approaches, which possess distinct properties in terms of model reliability and discrimination, are likely to result in different local maximums of the objective function, suggesting that their combination may thus provide for more reliable diarization. Previous work would seem to support this observation [24]. We report our recent work on system combination in Section IV-E.

IV. Experimental systems

Given a probabilistic approach and stated assumptions the framework presented in Section II led to the two hierarchical clustering approaches to speaker diarization. They are described in their most general form in Section III and it is the aim of the work presented later in this paper to validate the hypothesized characteristics presented therein. Whilst this work aims to compare the two approaches in a general manner it is impossible that the experimental validation entirely avoids any dependency on specific system implementations. Described in this section are the two different systems that we implemented in order to undertake the experimental work reported in this paper. The two systems are as general as is possible and are based on speaker diarization systems that have achieved state-of-the-art performance in the NIST RT evaluations.

TABLE I
SAD performance on the RT‘07 and RT‘09 datasets.

Dataset	FA	Miss	Total
RT‘07	4.7	1.1	5.8
RT‘09	7.2	1.8	9.0

Both baseline systems comprise a common SAD component and a segmentation and clustering component. In both cases the latter is a two-stage strategy involving an EM based segmentation and clustering process to generate an initial, coarse diarization and then a maximum a posteriori (MAP) based re-segmentation with feature normalization to refine the diarization hypothesis. Also described here is our approach to speaker purification and system combination. Both bottom-up and top-down speaker diarization systems were implemented with the ALIZE toolkit [25], which ensures that the comparative study better reflects core differences in the clustering and segmentation strategies instead of any nuances related to difference in estimation, decoding or adaptation algorithms.

A. Speech activity detection

SAD is a fundamental pre-processing step in all speaker diarization systems and aims to remove non-speech segments from the audio stream so that downstream speaker segmentation and clustering concentrates only on segments containing speech; furthermore, it provides an effective initialization for the top-down approach as we will discuss shortly. Our SAD system follows standard noise suppression [26] and is a simple model-based approach involving the alignment of the acoustic data to a two-state HMM in which the two states represent speech and non-speech data respectively. A large amount of speech and non-speech data from a separate development dataset, mostly from the AMI conference meeting corpus [27], are used to train the two 32-component GMMs with an EM-based algorithm. An ergodic HMM is formed by connecting the two GMMs with transition probabilities of 0.5. Key to good SAD performance is the sequential application of Viterbi realignment and model re-estimation which are applied iteratively to ensure that the models adjust to the prevailing ambient conditions. To ensure a realistic segmentation, some heuristic rules are applied to prohibit rapid transitions between speech and non-speech states. Table I illustrates SAD performance for the RT‘07 and RT‘09 datasets in terms of average false alarm (FA) and missed (Miss) speech rates. The fourth column is the addition of the FA and Miss rates and indicates overall SAD performance. Scores of 5.8% and 9.0% on the two datasets respectively show that, despite its simplicity, this approach performs well compared to other systems submitted to the NIST RT evaluations.

B. Bottom-up

Except for a novel progressive training approach to model initialization, which was proposed in [28] and referred to as sequential EM, the first stage EM-based segmentation and clustering process is a conventional AHC approach as described in Section III-A. GMM speaker models contain 4 components but, in an otherwise standard AHC approach, they are initially trained using only a small fraction of the available data before several steps of re-estimation are performed with an increasing amount of data at each step. This process is repeated until all the data are used in the final training cycle. New speaker models with 16 components are then estimated and used for the remaining merging steps. In our experience progressive training can lead to significant improvements in performance over a conventional AHC system. Cluster merging is controlled with the modified Information Change Rate (ICR) criterion [21] and continues until the stopping criterion is met. In contrast to the exemplary, state-of-the-art system presented in [28], we find that the T_s stopping criterion gives better results than the Rho criterion as used in [28], [29].

The second, MAP-based re-segmentation stage is common to both bottom-up and top-down approaches. The diarization hypothesis from the EM stage is used to train a new model for each speaker through the MAP adaptation of a generic background model which is trained on a large amount of external data. Speaker models now contain 128 components and are more complex than for the first EM-based segmentation and clustering stage. With the initial coarse segmentation from the EM-stage, the MAP-based adaptation tends to deliver more reliable performance. As in the EM stage, several iterations of Viterbi realignment and adaptation are applied to obtain a stable diarization hypothesis. Speaker clusters with too few speech data (less than 8 seconds) are removed. A final stage of re-segmentation is then applied in exactly the same way but with features that are normalized to have zero mean and unity variance, i.e. cepstral mean and variance normalization. We acknowledge that this setup potentially detracts from the independence of the assessment reported here since bottom-up and top-down segmentation and clustering algorithms are arguably used to initialize re-segmentation rather than being used for speaker diarization directly. In all of our experiments, however, MAP-based re-segmentation leads to consistent improvements in performance for both systems and as such its application has minimal impact on the comparative aspects of the study and does not alter the conclusions reported here.

C. Top-down

The top-down system is a DHC approach according to the general procedure described in Section III-B. It is based on the evolutive hidden Markov model (E-HMM) that was originally proposed in [30]. The current system has evolved significantly from the original work and, with significant improvements to speaker modeling, the system was used for LIA-EURECOM's submission to the most recent NIST RT'09 evaluation. As is the case for the bottom-up system there are two stages. The first generates an initial, coarse diarization hypothesis through an EM stage which is refined through a second MAP stage. The latter is identical to that used in the bottom-up system.

Initialization involves the training of a single, 16-component root speaker model $S0$ with an EM algorithm. New speakers are added to the model one-by-one by training new models with an EM algorithm using the single longest segment of speech that is assigned to $S0$ at any iteration. Following the addition of each speaker several iterations of Viterbi realignment and model re-estimation are used to refine the speaker models and diarization hypothesis. The quantity of data assigned to any new speaker must be sufficient to indicate a significant speaker; newly added speaker models that are assigned less than 8 seconds of data are rejected. In this case the system reverts to the previous hypothesis and the next largest segment is used to add a new speaker. This process continues until no more segments of greater than 6 seconds in length remain assigned to the root model, $S0$.

D. Purification

Purification is a data filtering technique; the central idea is to remove noisy data so that models are trained on data that is indicative of the target class only and not of unwanted variation. Purification techniques have been extensively studied within bottom-up approaches to speaker diarization [8] but there is comparatively very little work in the context of top-down approaches.

Our purification algorithm is based on the original work in [28] and is very similar to the progressive training approach described in Section IV-B. The algorithm operates between the first EM stage and the second MAP re-segmentation stage. Each hypothesized cluster is split into sub-segments of 500ms in length. The 55% of segments which best fit the corresponding GMM model are then used to estimate new models with EM training. The process is repeated ten times where at each iteration 45% of the data with the smallest likelihood are always discarded before being reassigned to their nearest cluster with Viterbi decoding. The diarization hypothesis obtained in the final iteration is then used in the final MAP-based resegmentation stage described above. Note that since purification is applied before the second stage MAP-based resegmentation it can influence the number of clusters in the final diarization hypothesis.

E. Combination

As outlined above the bottom-up and top-down clustering strategies are likely to produce different diarization outputs and it is thus of interest to combine their outputs. We hypothesize that for both approaches, some models may reliably represent specific, individual speakers, whereas others may be relatively unreliable. They may correspond to multiple speakers or to local maxima of the objective function which are not related to differences between speakers but to some other acoustic phenomena. If it is possible to identify reliable models then better diarization performance may be achieved by re-clustering the data assigned to the unreliable models.

A number of combination approaches have been proposed previously, at the clustering stage [24], [31] or at the output

stage [32], [33], [34]. Better performance is usually obtained but, with the exception of [35], none of the previous work considered the combination of both bottom-up and top-down system outputs without further re-segmentation. In our work, to leverage the respective merits of both the bottom-up and top-down approaches, we treat the top-down output as a base segmentation and apply the bottom-up output to purify it. Specifically, for each cluster contained in the top-down system output C_i a cluster contained in the bottom-up system C_n is chosen as a matching cluster if (i) they share a sufficient proportion of frames and (ii) among all other clusters contained in the bottom-up system C_n is the closest to C_i, where the inter-cluster distance is measured in terms of ICR. Each matched cluster pair is accepted as a reliable speaker and is retrained with only those frames that are common to both C_i and C_n. This set of reliable, matching clusters is denoted Ξ. All unreliable, or unmatched clusters are then compared to Ξ in order to identify additional reliable clusters, as follows:

$$\Xi \leftarrow C_m \tag{8}$$

if

$$\ell(C_m, \Xi) = \max_k \ell(C_k, \Xi) \quad C_k \notin \Xi \tag{9}$$

and

$$\ell(C_m, \Xi) > \theta \tag{10}$$

where θ is a tunable threshold, and where ℓ is the minimum ICR distance defined by:

$$\ell(C_k, \Xi) = \min_t ICR(C_k, C_t) \quad C_k \notin \Xi, C_t \in \Xi. \tag{11}$$

Additionally there is no significant overlap between C_m and any of the clusters in set Ξ. This procedure is conducted iteratively until no further reliable clusters remain. For each new added cluster, the 50% best-fitting frames (according to likelihood) are used to re-estimate a new speaker model. In contrast to previous work [35] the outputs of *both* the bottom-up and top-down systems are utilised in order to select frames for re-estimating new speaker models in the case of un-matched clusters. This acts to purify the speaker models. Further purification is achieved by training models using only the best fitting data and thus better speaker diarization performance is expected. This approach can be regarded as an extension to the work in [33], [35] which accepts matched clusters only.

V. Results and Discussion

In this section we present our experimental work. Whilst results give significant insight into the behavior of the two approaches we acknowledge that they pertain to the specific systems described in Section IV. Even though the systems are largely standard the observations should not be considered to be absolutely general.

We first introduce the standard evaluation protocols and performance metrics that are used in the NIST RT evaluations, and then the different datasets that were used for development and evaluation. We then report speaker diarization experiments for the baseline systems and the same systems with purification. Finally we analyze the results in terms of phone normalization and cluster purity.

A. Protocols and metrics

The NIST RT evaluations [3] have an instrumental role in assessing the state-of-the-art and in providing standard evaluation protocols, performance metrics and common datasets. Each evaluation involves various experimental conditions involving different microphone configurations. In order to assess the core segmentation and clustering components independently from beamforming [36], [37] or integrated inter channel delay features [38], [39], all experiments reported here involve the single distance microphone (SDM) condition; we expect that the observations and conclusions apply equally well to the core, multiple distant microphone (MDM) condition.

To evaluate the performance of a speaker diarization system, NIST defines a time-based metric known as the diarization error rate (DER). This is calculated as the fraction of speaker time that is not correctly attributed based on the optimal mapping between speakers in the reference and those hypothesized by the speaker diarization system. The DER is formally defined as:

$$DER = \frac{\sum_i \{D_i^R \cdot (\max(N_i^R, N_i^S) - N_i^C)\}}{\sum_i \{D_i^R \cdot N_i^R\}} \tag{12}$$

where D_i^R denotes the duration of the i-th reference segment, and where N_i^R and N_i^S are respectively the number of speakers according to the reference or the number of speakers hypothesized by the diarization system. N_i^C is the number of speakers that are correctly matched by the diarization system. Note that with overlapping speech, both N^R and N^S can be larger than one. Our speaker diarization systems are not capable of detecting overlapping speech, and thus N^S is either zero or one. While NIST defines protocols to evaluate performance with or without the scoring of overlapping speech, the primary metric *includes* overlap. Consequently all results discussed in the text involve the scoring of overlapping speech. Corresponding results where overlapping speech is not scored are included in the tables for comparative purposes only.

B. Datasets

Following the protocols and metrics discussed above, our experimental systems were optimized on a development dataset of 23 meetings from the NIST RT'04, '05 and '06 evaluation datasets. Performance was then assessed on the independent RT'07 and RT'09 evaluation datasets. There is no overlap between development and evaluation datasets and in all cases no prior knowledge is available except an approximate idea of the number of speakers. This is used solely in the case of the bottom-up system and only so that the system is initialized with a number of clusters that exceeds the maximum number of true speakers. In all cases we report only results obtained on the evaluation datasets.

It should be noted that the experimental work reported in this paper relates specifically to meeting domain data. However, the hypothesis and observations are expected to be

TABLE II
DERS WITH (OV) AND WITHOUT (NOV) THE SCORING OF OVERLAPPING SPEECH, WITH AND WITHOUT PURIFICATION.

System	RT'07		RT'09	
	OV	NOV	OV	NOV
Bottom-up	23.8	20.8	19.1	13.5
Bottom-up + Pur.	22.7	19.6	27.0	21.8
Top-down	18.3	15.0	26.0	21.5
Top-down + Pur.	17.8	14.4	21.1	16.0
Combined	16.1	12.8	17.8	12.3

general and apply to different data so long as the assumptions still hold. A review of data characteristics and their effect on speaker diarization has been reported previously [2], [40] and thus it is not discussed further in this article. An assessment of our speaker diarization systems on television chat show data was reported in previous work [7]. Given that the same assumptions also apply in this context it is of no surprise that behavior was observed to be consistent to that reported later in this paper and indicates a certain level of generality to different data.

C. Diarization performance

Speaker diarization performance using a bottom-up approach is illustrated on row 3 of Table II in which results are presented with (OV) and without (NOV) the scoring of overlapping speech. DER scores of 23.8% and 19.1% are obtained on the RT'07 and RT'09 datasets respectively. Of note is the large difference in performance with and without the scoring of overlapping speech for the RT'09 dataset. This is due to the high degree of overlapping speech in this dataset (13.6% for RT'09 cf. 7.6% for RT'07) which is well known to have a significant impact on the performance of state-of-the-art speaker diarization systems [41].

Speaker diarization performance using a top-down approach is illustrated on row 5 of Table II. DERs of 18.3% and 26.0% are obtained on the RT'07 and RT'09 datasets respectively and thus indicate an inconsistency in the comparative performance of top-down and bottom-up approaches: the top-down system gives superior performance for the RT'07 dataset whereas the bottom-up system is superior for the RT'09 dataset. The hypothesis is that factors unrelated to differences between speakers lead to unstable performance. This hypothesis is discussed further in Section V-D. First though, we report the impact of purification on both system outputs.

The performance of the bottom-up system with purification is illustrated on row 4 of Table II. DERs of 22.7% and 27.0% show that, while there is a small improvement over the baseline bottom-up system for the RT'07 dataset, there is a marked degradation in performance for the RT'09 dataset. The performance of the top-down system with purification is illustrated on row 6. DERs of 17.8% and 21.1% show a consistent improvement over the baseline top-down system. This suggests that, although purification may provide performance improvement for both the bottom-up and top-down systems, it is rather unstable with the bottom-up system and can lead to a degradation in performance in some cases. Comparatively, the top-down system achieves stable and consistent performance gains with purification, which supports our conjecture that (i) clusters identified by top-down systems are less discriminative and thus require purification, and (ii) those identified by bottom-up systems are less well normalized and that performance cannot always be improved through purification.

Upon comparison of results for the best bottom-up and top-down systems, we observe an inconsistency in performance. With purification, the top-down system outperforms the bottom-up system for the RT'07 dataset whereas it gives poorer results for the RT'09 dataset. This lends further support to the idea of combining the outputs of the two different systems. Diarization results with the combined system are illustrated on row 7 of Table II. They correspond to the combination of the outputs of the baseline bottom-up system and the top-down system with purification. DERs of 16.1% and 17.8% for the RT'07 and RT'09 datasets respectively show improved performance over both single systems. The combination strategy is thus successful in exploiting the merits of each approach.

D. Phone normalization

In this section we aim to account for the inconsistencies in system performance outlined above. According to the arguments presented in Section III-C bottom-up approaches are relatively more likely than top-down approaches to convergence to sub-optimal local maxima of (3). These are likely to correspond to nuisance variation and, whilst other acoustic classes are also relevant, we hypothesize here that the phones uttered are among the most significant competing influences in the acoustic space.

To help confirm this, or otherwise, we measured the difference in the phone distribution between each pair of clusters in the diarization hypothesis. The phone distribution is computed as the fraction of speech time attributed to each phone and thus requires a phone-level reference to determine the phone class of each frame. This was accomplished by a forced alignment of the phone transcription of each word in the reference annotation to the corresponding speech. The phone distribution of each cluster is used to calculate the average inter-cluster distance D as follows:

$$D = \binom{N}{2}^{-1} \sum_{n=1}^{N} \sum_{m=n+1}^{N} D_{\text{KL2}}(C_n||C_m),$$

where N is the size of the speaker inventory Δ, i.e. the number of clusters, and where the binomial coefficient $\binom{N}{2}$ is the number of unique cluster pairs. $D_{\text{KL2}}(C_n||C_m)$ is the symmetrical Kullback-Leibler (KL) distance between the phone distributions for clusters C_n and C_m, defined as:

$$D_{\text{KL2}}(C_n||C_m) = \frac{1}{2}\Big(D_{\text{KL}}(C_n||C_m) + D_{\text{KL}}(C_m||C_n)\Big)$$

where $D_{\text{KL}}(C_n||C_m)$ is the KL divergence of C_n from C_m. We note that the symmetrical KL metric has been used for the segmentation and clustering of broadcast news [42].

TABLE III
INTER-CLUSTER PHONE DISTRIBUTION DISTANCES.

System	Mean		Variance	
	RT'07	RT'09	RT'07	RT'09
Bottom-up	0.17	0.14	0.167	0.013
Bottom-up + Pur.	0.13	0.12	0.017	0.005
Top-down	0.11	0.10	0.006	0.004
Top-down + Pur.	0.07	0.08	0.001	0.002
Combined	0.07	0.07	0.001	0.001

TABLE IV
AVERAGE CLUSTER PURITY AND NUMBER OF CLUSTERS.

System	Cluster Purity (%)		No. Clusters	
	RT'07	RT'09	RT'07	RT'09
Bottom-up	80.6	79.2	7.0	7.0
Bottom-up + Pur.	82.2	75.2	5.8	6.9
Top-down	81.8	79.1	5.0	6.0
Top-down + Pur.	84.1	81.4	4.8	5.3
Combined	81.7	81.6	4.4	4.6
Ground-truth	100.0	100.0	4.4	5.4

In the case where clusters are well normalized against phone variation then the average inter-cluster distance is expected to be small, since the clusters should have similar phone distributions. Significant differences between distributions, however, indicate poor phone normalization and possibly a sub-optimal local maximum of (3). This latter case might reflect a higher degree of convergence toward phones, or other acoustic classes, rather than toward speakers.

The mean and the variance of the inter-cluster distances are presented in columns 2 and 3 of Table III for the RT'07 and RT'09 datasets respectively. For the baseline bottom-up system average inter-cluster distances of 0.17 and 0.14 are obtained. These fall to 0.13 and 0.12 with purification indicating improved normalization against phones. For the top-down system the average distances are 0.11 and 0.10. These fall to 0.07 and 0.08 with purification and are significantly better than for the bottom-up system. Reassuringly, with combination the values remain stable at 0.07 and 0.07. Columns 4 and 5 of Table III show the corresponding variances in all cases and show a consistent decrease moving down the table: reductions in the mean are accompanied by reductions in the variation. These observations suggests that on average, and as predicted, the clusters identified with the bottom-up system are indeed less well normalized against phone variation than those identified with the top-down system and that combination preserves the normalization of the top-down system.

E. Cluster purity

The observations reported above do not explain why, for the RT'09 dataset, the bottom-up system performance deteriorates with purification even though the phone normalization improves. To help explain this behavior we analyzed the average speaker purity in each system output. The cluster purity is the percentage of data in each cluster which are attributed to the most dominant speaker, as determined from the ground-truth reference. Average, time-weighted cluster purities are presented in columns 2 and 3 of Table IV. For the RT'07 dataset purification leads to marginal improvements: from 80.6% purity to 82.2% for the bottom-up system and from 81.8% to 84.1% for the top-down system. Different behavior is observed for the RT'09 dataset. Whereas purification gives an improvement from 79.1% to 81.4% for the top-down system it leads to a degradation from 79.2% to 75.2% for the bottom-up system.

Whilst a reduction in cluster purity may account for the decrease in diarization performance it is necessary to consider the number of clusters in the system output to properly interpret cluster purity and its impact on diarization performance. As explained in Section IV-D purification influences the number of identified clusters. A larger number of clusters may be associated with inherently higher purity (i.e. with a single cluster for each sample the purity is 100%) and so purity statistics alone do not fully reflect the effect of purification on diarization performance. The number of clusters detected in each system output is illustrated in columns 4 and 5 of Table IV in which the last row shows the statistics for the ground-truth reference. All systems over-estimate the number of speakers and purification always reduces the number toward the number of true speakers. When coupled with increases in average purity, then improved diarization performance should be expected. For the bottom-up system and the RT'09 dataset the decrease in the number of clusters when purification is applied is negligable, whereas the purity also decreases. This can only result in poorer diarization performance.

Turning to the combination results for the RT'07 dataset, even though the average purity decreases to 81.7% (below that for the top-down system with purification) diarization performance still improves since the number of clusters more accurately reflects the true number of speakers. For the RT'09 dataset the combined system produces clusters which are marginally more pure than any of the single systems (81.6%) even though the number of clusters decreases below the true number of speakers. Further investigation showed that the missed speakers have relatively low floor time and thus do not contribute significantly to diarization performance.

VI. CONCLUSIONS AND FUTURE WORK

This paper presents a theoretical framework for speaker diarization. It is used to compare the relative merits of the bottom-up and top-down approaches to speaker diarization. We argue that the two approaches are likely to exhibit different behavior in the face of significant variation from non-speaker-related factors and that both have the potential to benefit from purification, particularly the top-down approach. We also argue that, since the two approaches involve entirely contrasting search strategies to optimize the speaker inventory, they are likely to converge to different local maxima of the objective function and thus there is potential for them to be combined in order to improve speaker diarization performance.

These hypotheses are validated by experiments performed on two standard, single-channel NIST RT evaluation datasets. Results show that, despite the dominance in the literature of bottom-up systems, the two approaches deliver largely com-

parable performance, with neither being consistently superior. With purification consistent improvements are observed with a top-down system; with a bottom-up system, however, purification leads to inconsistent improvements and even degrades performance for one dataset. This supports our conjecture that models produced by the top-down approach tend to be less discriminative and therefore are likely to benefit from purification. Finally, the combined approach provides additional and consistent performance improvements, and demonstrates the complementarity of the bottom-up and top-down approaches. This finding highlights the importance of continuing research with both approaches to speaker diarization.

We acknowledge that the study presented in this paper is not exhaustive. Although the theoretical framework is general, it relates to a number of assumptions which lead to specific implementations. The study is therefore not absolute and does not include a number of new and emerging techniques based on different assumptions. e.g. Bayesian treatments. In addition, we focus on a particular factor, i.e. phonetic nuisance, whereas other factors may also impact on clustering and segmentation performance. In addition phonetic nuisance may affect not only the clustering and segmentation process but also other components in a typical speaker diarization system. All of these aspects require further study, even if the conclusions drawn in this paper are still widely applicable and can be migrated to other domains if the assumptions supporting the theory are still satisfied.

Finally, future work should investigate new techniques to address the respective shortcomings of the two clustering approaches. More effective purification approaches are needed to enhance the discrimination of speaker models whereas new marginalization techniques are required to attenuate nuisance variation that is unrelated to differences between speakers. This is particularly important for bottom-up systems.

ACKNOWLEDGMENT

This work was partially supported by the joint-national 'Adaptable ambient living assistant' (ALIAS) project funded through the European Ambient Assisted Living (AAL) programme, agreement number AAL-2009-2-049 and by the 'Annotation Collaborative pour l'Accessibilité Vidéo' (ACAV) project funded by the French Ministry of Industry (Innovative Web call) under contract 09.2.93.0966.

REFERENCES

[1] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE TASLP*, vol. 14, no. 5, pp. 1557–1565, 2006.

[2] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE TASLP*, 2011.

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

[5] H. Sun, B. Ma, S. Z. K. Khine, and H. Li, "Speaker diarization system for RT07 and RT09 meeting room audio," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2010)*, March 2010.

[6] 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.

[7] S. Bozonnet, N. Evans, C. Fredouille, D. Wang, and R. Troncy, "An integrated top-down/bottom-up approach to speaker diarization," in *Proc. Interspeech, to appear*, September 2010.

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

[9] ——, "Frame purification for cluster comparison in speaker diarization," in *Second Workshop on Multimodal User Authentication (MMUA)*, 2006.

[10] H. Sun, T. L. Nwe, B. Ma, and H. Li, "Speaker diarization for meeting room audio," in *Proc. Interspeech'09*, September 2009.

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

[12] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in *Proc. of DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, Virginia, USA, February 1998, pp. 127–132. [Online]. Available: http://www.nist.gov/speech/publications/darpa98/pdf/bn20.pdf

[13] C. Wooters and M. Huijbregts, "The ICSI RT07s Speaker Diarization System," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, USA, May, 2007, Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 509–519.

[14] E. C. Koh, H. Sun, T. L. Nwe, T. H. Nguyen, B. Ma, E.-S. Chng, H. Li, and S. Rahardja, "Speaker Diarization Using Direction of Arrival Estimate and Acoustic Feature Information: The I2R-NTU Submission for the NIST RT 2007 Evaluation," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, USA, May 8-11, 2007, Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 484–496.

[15] D. A. Van Leeuwen and M. Konečný, "Progress in the AMIDA Speaker Diarization System for Meeting Data," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, USA, May, 2007, Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 475–483.

[16] J. Luque, X. Anguera, A. Temko, and J. Hernando, "Speaker diarization for conference room: The UPC RT07s evaluation system," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, USA, May, 2007, Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 543–553.

[17] X. Zhu, C. Barras, L. Lamel, and J.-L. Gauvain, "Multi-stage Speaker Diarization for Conference and Lecture Meetings," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, USA, May, 2007, Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 533–542.

[18] K. Biatov and J. Köhler, "Improvement Speaker Clustering Using Global Similarity Features," in *Interspeech 2006 - ICSLP*, 2006.

[19] M. Ben, M. Betser, F. Bimbot, and G. Gravier, "Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted gmms," in *Proc. ICSLP, Jeju Island, Korea*, 2004.

[20] D. Imseng and G. Friedland, "Tuning-Robust Initialization Methods for Speaker Diarization," *to appear in IEEE Transactions on Audio, Speech and Language Processing, available at IEEExplore*, 2010.

[21] 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.

[22] C. Fredouille and N. Evans, "The LIA RT'07 speaker diarization system," in *Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, USA, May, 2007, Revised Selected Papers*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 520–532.

[23] G. Friedland, O. Vinyals, Y. Huang, and C. Muller, "Prosodic and other long-term features for speaker diarization," *IEEE TASLP*, vol. 17, no. 5, pp. 985–993, July 2009.

[24] 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.

[25] 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.

[26] A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, "Qualcomm-ICSI-OGI features for ASR," in *Proc. ICSLP*, 2002, pp. 21–24.
[27] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska, W. Post, D. Reidsma, and P. Wellner, "The AMI meeting corpus," in *Proc. Measuring Behavior*, 2005.
[28] T. Nguyen *et al.*, "The IIR-NTU Speaker Diarization Systems for RT 2009," in *RT'09, NIST Rich Transcription Workshop*, Melbourne, Florida, USA, 2009.
[29] T. H. Nguyen, E. S. Chng, and H. Li, "T-test distance and clustering criterion for speaker diarization," in *Proc. Interspeech*, Brisbane, Australia, 2008.
[30] 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.
[31] 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.
[32] S. E. Tranter, "Two-way cluster voting to improve speaker diarisation performance," in *IEEE ASRU Workshop*, 1997, pp. 347–352.
[33] 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.
[34] M. Huijbregts, D. A. Van Leeuwen, and F. M. G. de Jong, "The majority wins: a method for combining speaker diarization systems," in *Proc. Interspeech*, 2009, pp. 924–927.
[35] S. Bozonnet, N. Evans, X. Anguera, O. Vinyals, G. Friedland, and C. Fredouille, "System output combination for improved speaker diarization," in *Proc. Interspeech, to appear*, September 2010.
[36] X. Anguera, "BeamformIt (the fast and robust acoustic beamformer)," http://www.xavieranguera.com/beamformit/.
[37] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE TASLP*, vol. 15, no. 7, pp. 2011–2021, September 2007.
[38] D. Ellis and J. C. Liu, "Speaker turn detection based on between-channels differences," in *Proc. ICASSP*, 2004.
[39] X. Anguera, C. Wooters, and J. Hernando, "Speaker diarization for multi-party meetings using acoustic fusion," in *Proc. ASRU*, Nov. 2005, pp. 426–431.
[40] N. Mirghafori and C. Wooters, "Nuts and flakes: A study of data characteristics in speaker diarization," in *Proc. ICASSP*, 2006.
[41] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, 31 2008-April 4 2008, pp. 4353–4356.
[42] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA Speech Recognition Workshop*, 1997, pp. 97–99.

Nicholas Evans was awarded M.Eng. and Ph.D. degrees from the University of Wales Swansea (UWS), UK in 1999 and 2003 respectively and was appointed as was appointed as a Lecturer in Communications in 2002. After one year at the Laboratoire Informatique d'Avignon (LIA) he joined EURECOM as an Assistant Professor in 2007 where he now heads the Speech and Audio Processing Research Group. His current research interests include speaker diarization, speaker recognition, biometrics, speech enhancement, noise compensation and echo cancellation. He is a member of ISCA, EURASIP, the IEEE and its Signal Processing Society and currently serves as an associate editor for the EURASIP Journal on Audio, Speech and Music Processing.

Simon Bozonnet obtained a Diploma in Electrical Engineering from the National Institute of Applied Sciences (INSA), Lyon, France in 2008, specializing in Signal Processing, in addition to a Master of Research in Images and Systems. He undertook his Master's thesis at the CEA (Nuclear Energy Center) in Paris, France where he worked on signal fusion and intelligent systems for source localization. In October 2008 he moved to EURECOM in Sophia Antipolis, France and joined the Multimedia Communications Department as a Ph.D. candidate. His research interests include multimedia indexing, and specifically speaker diarization.

Dong Wang received the B.Sc. and M.Sc. in computer science at Tsinghua Univ. in 1999 and 2002, and then worked for Oracle China in 2002-2004 and IBM China in 2004-2006. He joined CSTR, University of Edinburgh in 2006 as a research fellow and Ph.D. student supported by a Marie Curie fellowship, from where he received his Ph.D. in 2010. He is now working in EURECOM France as a post-doc fellow.

Corinne Fredouille was awarded a Ph.D. degree from the Laboratoire Informatique d'Avignon (LIA), University of Avignon in 2000 and was appointed as an assistant professor in 2003. Her research interests include acoustic analysis, voice quality assessment, statistical modeling, automatic speaker recognition, speaker diarization and, more recently, speech and voice disorder assessment. She has participated in several national and international speaker diarization system evaluation campaigns and has published over 15 research papers in this field. She is a member of the International Speech Communication Association (ISCA) and secretary of the French speaking communication association (AFCP).

Raphaël Troncy was awarded a Master's degree in Computer Science at the University Joseph Fourier, Grenoble, France, and a Ph.D. degree from the University of Grenoble (INRIA/INA) in 2004. He was an ERCIM Post-Doctorate Research Associate from 2004 to 2006. He was a senior researcher for CWI from 2006 to 2009 and was appointed as an assistant professor at EURECOM in 2009. He is co-chair of the W3C Incubator Group on Multimedia Semantics and Media Fragments Working Group and contributes to the Media Annotations Working Group. He is an expert in audio-visual metadata and in combining existing metadata standards (such as MPEG-7) with current Semantic Web technologies.