Speaker Diarization: A Review of Recent Research

Xavier Anguera Miro, *Member, IEEE*, Simon Bozonnet, *Student Member, IEEE*, Nicholas Evans, *Member, IEEE*,
Corinne Fredouille, Gerald Friedland, *Member, IEEE*, and Oriol Vinyals

***Abstract*—Speaker diarization is the task of determining “who spoke when?” in an audio or video recording that contains an unknown amount of speech and also an unknown number of speakers. Initially, it was proposed as a research topic related to automatic speech recognition, where speaker diarization serves as an upstream processing step. Over recent years, however, speaker diarization has become an important key technology for many tasks, such as navigation, retrieval, or higher level inference on audio data. Accordingly, many important improvements in accuracy and robustness have been reported in journals and conferences in the area. The application domains, from broadcast news, to lectures and meetings, vary greatly and pose different problems, such as having access to multiple microphones and multimodal information or overlapping speech. The most recent review of existing technology dates back to 2006 and focuses on the broadcast news domain. In this paper, we review the current state-of-the-art, focusing on research developed since 2006 that relates predominantly to speaker diarization for conference meetings. Finally, we present an analysis of speaker diarization performance as reported through the NIST Rich Transcription evaluations on meeting data and identify important areas for future research.**

***Index Terms*—Meetings, rich transcription, speaker diarization.**

I. Introduction

SPEAKER diarization has emerged as an increasingly important and dedicated domain of speech research. Whereas speaker and speech recognition involve, respectively, the recognition of a person’s identity or the transcription of their speech, speaker diarization relates to the problem of determining “who spoke when?.” More formally this requires the unsupervised identification of each speaker within an audio stream and the intervals during which each speaker is active.

Speaker diarization has utility in a majority of applications related to audio and/or video document processing, such as information retrieval for example. Indeed, it is often the case that audio and/or video recordings contain more than one active speaker. This is the case for telephone conversations (for example stemming from call centers), broadcast news, debates, shows, movies, meetings, domain-specific videos (such as surgery operations for instance), or even lecture or conference recordings including multiple speakers or questions/answers sessions. In all such cases, it can be advantageous to automatically determine the number of speakers involved in addition to the periods when each speaker is active. Clear examples of applications for speaker diarization algorithms include speech and speaker indexing, document content structuring, speaker recognition (in the presence of multiple or competing speakers), to help in speech-to-text transcription (i.e., so-called speaker attributed speech-to-text), speech translation and, more generally, Rich Transcription (RT), a community within which the current state-of-the-art technology has been developed. The most significant effort in the Rich Transcription domain comes directly from the internationally competitive RT evaluations, sponsored by the National Institute of Standards and Technology (NIST) in the Unites States [1]. Initiated originally within the telephony domain, and subsequently in that of broadcast news, today it is in the domain of conference meetings that speaker diarization receives the most attention. Speaker diarization is thus an extremely important area of speech processing research.

An excellent review of speaker diarization research is presented in [2], although it predominantly focuses its attention to speaker diarization for broadcast news. Coupled with the transition to conference meetings, however, the state-of-the-art has advanced significantly since then. This paper presents an up-to-date review of present state-of-the-art systems and reviews the progress made in the field of speaker diarization since 2005 up until now, including the most recent NIST RT evaluation that was held in 2009. Official evaluations are an important vehicle for pushing the state-of-the-art forward as it is only with standard experimental protocols and databases that it is possible to meaningfully compare different approaches. While we also address emerging new research in speaker diarization, in this paper special emphasis is placed on established technologies within the context of the NIST RT benchmark evaluations, which has become a reliable indicator for the current state-of-the-art in speaker diarization. This paper aims at giving a concise reference overview of established approaches, both for the general reader and for those new to the field. Despite rapid gains in popularity over recent years, the field is relatively embryonic compared to the mature fields of speech and speaker recognition. There are outstanding opportunities for contributions and we hope that this paper serves to encourage others to participate.

Manuscript received August 19, 2010; revised December 03, 2010; accepted February 13, 2011. Date of current version January 13, 2012. This work was supported in part by the joint-national “Adaptable ambient living assistant” (ALIAS) project funded through the European Ambient Assisted Living (AAL) program under Agreement AAL-2009-2-049 and in part 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. The work of X. Anguera Miro was supported in part by the Torres Quevedo Spanish program. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Sadaoki Furui.

X. Anguera Miro is with the Multimedia Research Group, Telefonica Research, 08021 Barcelona, Spain (e-mail: xanguera@tid.es).

S. Bozonnet and N. Evans are with the Multimedia Communications Department, EURECOM, 06904 Sophia Antipolis Cedex, France (e-mail: bozonnet@eurecom.fr).

C. Fredouille is with the University of Avignon, CERI/LIA, F-84911 Avignon Cedex 9, France (e-mail: corinne.fredouille@univ-avignon.fr).

G. Friedland and O. Vinyals are with the International Computer Science Institute (ICSI), Berkeley, CA 94704 USA (e-mail: fractor@icsi.berkeley.edu; evans@eurecom.fr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TASL.2011.2125954

Section II presents a brief history of speaker diarization research and the transition to the conference meeting domain. We describe the main differences between broadcast news and conference meetings and present a high-level overview of current approaches to speaker diarization. In Section III, we present a more detailed description of the main algorithms that are common to many speaker diarization systems, including those recently introduced to make use of information coming from multiple microphones, namely delay-and-sum beamforming. Section IV presents some of the most recent work in the field including efforts to handle multimodal information and overlapping speech. We also discuss the use of features based on inter-channel delay and prosodics and also attempts to combine speaker diarization systems. In Section V, we present an overview of the current status in speaker diarization research. We describe the NIST RT evaluations, the different datasets and the performance achieved by state-of-the-art systems. We also identify the remaining problems and highlight potential solutions in the context of current work. Finally, our conclusions are presented in Section VI.

II. Speaker Diarization

Over recent years, the scientific community has developed research on speaker diarization in a number of different domains, with the focus usually being dictated by funded research projects. From early work with telephony data, broadcast news (BN) became the main focus of research towards the late 1990s and early 2000s and the use of speaker diarization was aimed at automatically annotating TV and radio transmissions that are broadcast daily all over the world. Annotations included automatic speech transcription and meta data labeling, including speaker diarization. Interest in the meeting domain grew extensively from 2002, with the launch of several related research projects including the European Union (EU) Multimodal Meeting Manager (M4) project, the Swiss Interactive Multimodal Information Management (IM2) project, the EU Augmented Multi-party Interaction (AMI) project, subsequently continued through the EU Augmented Multi-party Interaction with Distant Access (AMIDA) project and, and finally, the EU Computers in the Human Interaction Loop (CHIL) project. All these projects addressed the research and development of multimodal technologies dedicated to the enhancement of human-to-human communications (notably in distant access) by automatically extracting meeting content, making the information available to meeting participants, or for archiving purposes.

These technologies have to meet challenging demands such as content indexing, linking and/or summarization of on-going or archived meetings, the inclusion of both verbal and nonverbal human communication (people movements, emotions, interactions with others, etc.). This is achieved by exploiting several synchronized data streams, such as audio, video and textual information (agenda, discussion papers, slides, etc.), that are able to capture different kinds of information that are useful for the structuring and analysis of meeting content. Speaker diarization plays an important role in the analysis of meeting data since it allows for such content to be structured in speaker turns, to which linguistic content and other metadata can be added (such as the dominant speakers, the level of interactions, or emotions).

Undertaking benchmarking evaluations has proven to be an extremely productive means for estimating and comparing algorithm performance and for verifying genuine technological advances. Speaker diarization is no exception and, since 2002, the US National Institute for Standards and Technology (NIST) has organized official speaker diarization evaluations[1] involving broadcast news (BN) and, more recently, meeting data. These evaluations have crucially contributed to bringing researchers together and to stimulating new ideas to advance the state-of-the-art. While other contrastive sub-domains such as lecture meetings and coffee breaks have also been considered, the conference meeting scenario has been the primary focus of the NIST RT evaluations since 2004. The meeting scenario is often referred to as "speech recognition complete," i.e., a scenario in which all of the problems that arise in any speech recognition can be encountered in this domain. Conference meetings thus pose a number of new challenges to speaker diarization that typically were less relevant in earlier research.

A. *Broadcast News Versus Conference Meetings*

With the change of focus of the NIST RT evaluations from BN to meetings diarization algorithms had to be adapted according to the differences in the nature of the data. First, BN speech data is usually acquired using boom or lapel microphones with some recordings being made in the studio and others in the field. Conversely, meetings are usually recorded using desktop or far-field microphones (single microphones or microphone arrays) which are more convenient for users than head-mounted or lapel microphones.[2] As a result, the signal-to-noise ratio is generally better for BN data than it is for meeting recordings. Additionally, differences between meeting room configurations and microphone placement lead to variations in recording quality, including background noise, reverberation and variable speech levels (depending on the distance between speakers and microphones).

Second, BN speech is often read or at least prepared in advance while meeting speech tends to be more spontaneous in nature and contains more overlapping speech. Although BN recordings can contain speech that is overlapped with music, laughter, or applause (far less common for conference meeting data), in general, the detection of acoustic events and speakers tends to be more challenging for conference meeting data than for BN data.

Finally, the number of speakers is usually larger in BN but speaker turns occur less frequently than they do in conference meeting data, resulting in BN having a longer average speaker turn length. An extensive analysis of BN characteristics is reported in [3] and a comparison of BN and conference meeting data can be found in [4].

[1] Speaker diarization was evaluated prior to 2002 through NIST Speaker Recognition (SR) evaluation campaigns (focusing on telephone speech) and not within the RT evaluation campaigns.

[2] Meeting databases recorded for research purposes usually contain head-mounted and lapel microphone recordings for ground-truth creation purposes only.

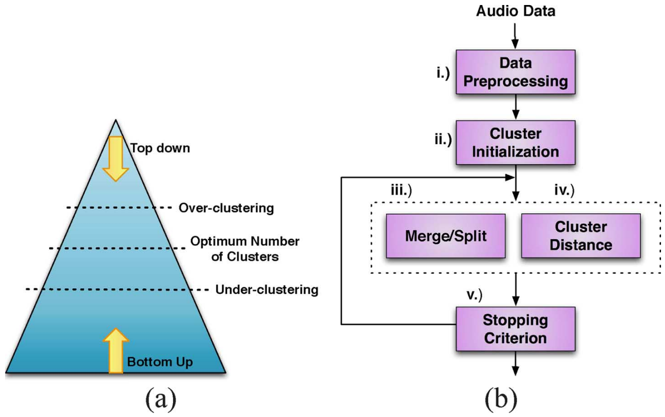

Fig. 1. General Diarization system. (a) Alternative clustering schemas. (b) General speaker diarization architecture.

B. Main Approaches

Most of present state-of-the-art speaker diarization systems fit into one of two categories: the bottom-up and the top-down approaches, as illustrated in Fig. 1(a). The top-down approach is initialized with very few clusters (usually one) whereas the bottom-up approach is initialized with many clusters (usually more clusters than expected speakers). In both cases the aim is to iteratively converge towards an optimum number of clusters. If the final number is higher than the optimum then the system is said to under-cluster. If it is lower it is said to over-cluster. Both bottom-up and top-down approaches are generally based on hidden Markov models (HMMs) where each state is a Gaussian mixture model (GMM) and corresponds to a speaker. Transitions between states correspond to speaker turns. In this section, we briefly outline the standard bottom-up and top-down approaches as well as two recently proposed alternatives: one based on information theory; and a second one based on a non parametric Bayesian approach. Although these new approaches have not been reported previously in the context of official NIST RT evaluations they have shown strong potential on NIST RT evaluation datasets and are thus included here. Additionally, some other works propose sequential single-pass segmentation and clustering approaches [5]–[7], although their performance tends to fall short of the state-of-the-art.

1) Bottom-Up Approach: The bottom-up approach is by far the most common in the literature. Also known as agglomerative hierarchical clustering (AHC or AGHC), the bottom-up approach trains a number of clusters or models and aims at successively merging and reducing the number of clusters until only one remains for each speaker. Various initializations have been studied and, whereas some have investigated k-means clustering, many systems use a uniform initialization, where the audio stream is divided into a number of equal length abutted segments. This simpler approach generally leads to equivalent performance [8]. In all cases the audio stream is initially over-segmented into a number of segments which exceeds the anticipated maximum number of speakers. The bottom-up approach then iteratively selects closely matching clusters to merge, hence reducing the number of clusters by one upon each iteration. Clusters are generally modeled with a GMM and, upon merging, a single new GMM is trained on the data that was previously assigned to the two individual clusters. Standard distance metrics, such as those described in Section III-C, are used to identify the closest clusters. A reassignment of frames to clusters is usually performed after each cluster merging, via Viterbi realignment for example, and the whole process is repeated iteratively, until some stopping criterion is reached, upon which there should remain only one cluster for each detected speaker. Possible stopping criteria include thresholded approaches such as the Bayesian Information Criterion (BIC) [9], Kullback–Leibler (KL)-based metrics [10], the generalized likelihood ratio (GLR) [11] or the recently proposed T_s metric [12]. Bottom-up systems submitted to the NIST RT evaluations [9], [13] have performed consistently well.

2) Top-Down Approach: In contrast with the previous approach, the top-down approach first models the entire audio stream with a single speaker model and successively adds new models to it until the full number of speakers are deemed to be accounted for. A single GMM model is trained on all the speech segments available, all of which are marked as unlabeled. Using some selection procedure to identify suitable training data from the non-labeled segments, new speaker models are iteratively added to the model one-by-one, with interleaved Viterbi realignment and adaptation. Segments attributed to any one of these new models are marked as labeled. Stopping criteria similar to those employed in bottom-up systems may be used to terminate the process or it can continue until no more relevant unlabeled segments with which to train new speaker models remain. Top-down approaches are far less popular than their bottom-up counterparts. Some examples include [14]–[16]. While they are generally out-performed by the best bottom-up systems, top-down approaches have performed consistently and respectably well against the broader field of other bottom-up entries. Top-down approaches are also extremely computationally efficient and can be improved through cluster purification [17].

3) Other Approaches: A recent alternative approach, though also bottom-up in nature, is inspired from rate-distortion theory and is based on an information-theoretic framework [18]. It is completely non parametric and its results have been shown to be comparable to those of state-of-the-art parametric systems, with significant savings in computation. Clustering is based on mutual information, which measures the mutual dependence of two variables [19]. Only a single global GMM is tuned for the full audio stream, and mutual information is computed in a new space of relevance variables defined by the GMM components. The approach aims at minimizing the loss of mutual information between successive clusterings while preserving as much information as possible from the original dataset. Two suitable methods have been reported: the agglomerative information bottleneck (aIB) [18] and the sequential information bottleneck (sIB) [19]. Even if this new system does not lead to better performance than parametric approaches, results comparable to state-of-the-art GMM systems are reported and are achieved with great savings in computation.

Alternatively, Bayesian machine learning became popular by the end of the 1990s and has recently been used for speaker diarization. The key component of Bayesian inference is that it does not aim at estimating the parameters of a system (i.e., to perform point estimates), but rather the parameters of their related distribution (hyperparameters). This allows for avoiding any premature hard decision in the diarization problem and for automatically regulating the system with the observations (e.g., the complexity of the model is data dependent). However, the computation of posterior distributions often requires intractable integrals and, as a result, the statistics community has developed approximate inference methods. Monte Carlo Markov chains (MCMCs) were first used [20] to provide a systematic approach to the computation of distributions via sampling, enabling the deployment of Bayesian methods. However, sampling methods are generally slow and prohibitive when the amount of data is large, and they require to be run several times as the chains may get stuck and not converge in a practical number of iterations.

Another alternative approach, known as Variational Bayes, has been popular since 1993 [21], [22] and aims at providing a deterministic approximation of the distributions. It enables an inference problem to be converted to an optimization problem by approximating the intractable distribution with a tractable approximation obtained by minimizing the Kullback–Leibler divergence between them. In [23] a Variational Bayes-EM algorithm is used to learn a GMM speaker model and optimize a change detection process and the merging criterion. In [24], variational Bayes is combined successfully with eigenvoice modeling, described in [25], for the speaker diarization of telephone conversations. However, these systems still consider classical Viterbi decoding for the classification and differ from the nonparametric Bayesian systems introduced in Section IV-F.

Finally, the recently proposed speaker binary keys [26] have been successfully applied to speaker diarization in meetings [27] with similar performance to state-of-the-art systems but also with considerable computational savings (running in around 0.1 times real-time). Speaker binary keys are small binary vectors computed from the acoustic data using a universal background model (UBM)-like model. Once they are computed all processing tasks take place in the binary domain. Other works in speaker diarization concerned with speed include [28], [29] which achieve faster than real-time processing through the use of several processing tricks applied to a standard bottom-up approach ([28]) or by parallelizing most of the processing in a GPU unit ([29]). The need for efficient diarization systems is emphasized when processing very large databases or when using diarization as a preprocessing step to other speech algorithms.

III. Main Algorithms

Fig. 1(b) shows a block diagram of the generic modules which make up most speaker diarization systems. The data preprocessing step (Fig. 1(b)-i) tends to be somewhat domain specific. For meeting data, preprocessing usually involves noise reduction (such as Wiener filtering for example), multi-channel acoustic beamforming (see Section III-A), the parameterization of speech data into acoustic features (such as MFCC, PLP, etc.) and the detection of speech segments with a speech activity detection algorithm (see Section III-B). Cluster initialization (Fig. 1(b)-ii) depends on the approach to diarization, i.e., the choice of an initial set of clusters in bottom-up clustering [8], [13], [30] (see Section III-C) or a single segment in top-down clustering [15], [16]. Next, in Fig. 1(b)-iii/iv, a distance between clusters and a split/merging mechanism (see Section III-D) is used to iteratively merge clusters [13], [31] or to introduce new ones [16]. Optionally, data purification algorithms can be used to make clusters more discriminant [13], [17], [32]. Finally, as illustrated in Fig. 1(b)-v, stopping criteria are used to determine when the optimum number of clusters has been reached [33], [34].

A. Acoustic Beamforming

The application of speaker diarization to the meeting domain triggered the need for dealing with multiple microphones which are often used to record the same meeting from different locations in the room [35]–[37]. The microphones can have different characteristics: wall-mounted microphones (intended for speaker localization), lapel microphones, desktop microphones positioned on the meeting room table or microphone arrays. The use of different microphone combinations as well as differences in microphone quality called for new approaches to speaker diarization with multiple channels.

The multiple distant microphone (MDM) condition was introduced in the NIST RT'04 (Spring) evaluation. A variety of algorithms have been proposed to extend mono-channel diarization systems to handle multiple channels. One option, proposed in [38], is to perform speaker diarization on each channel independently and then to merge the individual outputs. In order to do so, a two axis merging algorithm is used which considers the longest detected speaker segments in each channel and iterates over the segmentation output. In the same year, a late-stage fusion approach was also proposed [39]. In it, speaker segmentation is performed separately in all channels and diarization is applied only taking into account the channel whose speech segments have the best signal-to-noise ratio (SNR). Subsequent approaches investigated preprocessing to combine the acoustic signals to obtain a single channel which could then be processed by a regular mono-channel diarization system. In [40], the multiple channels are combined with a simple weighted sum according to their SNR. Though straightforward to implement, it does not take into account the time difference of arrival between each microphone channel and might easily lead to a decrease in performance.

Since the NIST RT'05 evaluation, the most common approach to multi-channel speaker diarization involves acoustic beamforming as initially proposed in [41] and described in detail in [42]. Many RT participants use the free and open-source acoustic beamforming toolkit known as BeamformIt [43] which consists of an enhanced delay-and-sum algorithm to correct misalignments due to the time-delay-of-arrival (TDOA) of speech to each microphone. Speech data can be optionally preprocessed using Wiener filtering [44] to attenuate noise using, for example, [45]. A reference channel is selected and the other channels are appropriately aligned and combined with a standard delay-and-sum algorithm. The contribution made by each signal channel to the output is then dynamically weighted according to its SNR or by using a cross-correlation-based metric. Various additional algorithms are available in the BeamformIt toolkit to select the optimum reference channel and to stabilize the TDOA values between channels before the signals are summed. Finally, the TDOA estimates themselves are made available as outputs and have been used successfully to improve diarization, as explained in Section IV-A. Note that, although there are other algorithms that can provide better beamforming results for some cases, delay-and-sum beamforming is the most reliable one when no information on the location or nature of each microphone is known *a priori*. Among alternative beamforming algorithms we find maximum likelihood (ML) [46] or generalized sidelobe canceller (GSC) [47] which adaptively find the optimum parameters, and minimum variance distortionless response (MVDR) [48] when prior information on ambient noise is available. All of these have higher computational requirements and, in the case of the adaptive algorithms, there is the danger of converging to inaccurate parameters, especially when processing microphones of different types.

B. Speech Activity Detection

Speech activity detection (SAD) involves the labeling of speech and nonspeech segments. SAD can have a significant impact on speaker diarization performance for two reasons. The first stems directly from the standard speaker diarization performance metric, namely the diarization error rate (DER), which takes into account both the false alarm and missed speaker error rates (see Section VI-A for more details on evaluation metrics); poor SAD performance will therefore lead to an increased DER. The second follows from the fact that nonspeech segments can disturb the speaker diarization process, and more specifically the acoustic models involved in the process [49]. Indeed, the inclusion of non-speech segments in speaker modelling leads to less discriminant models and thus increased difficulties in segmentation. Consequently, a good compromise between missed and false alarm speech error rates has to be found to enhance the quality of the following speaker diarization process.

SAD is a fundamental task in almost all fields of speech processing (coding, enhancement, and recognition) and many different approaches and studies have been reported in the literature [50]. Initial approaches for diarization tried to solve speech activity detection on the fly, i.e., by having a nonspeech cluster be a by-product of the diarization. However, it became evident that better results are obtained using a dedicated speech/nonspeech detector as preprocessing step. In the context of meetings nonspeech segments may include silence, but also ambient noise such as paper shuffling, door knocks or non-lexical noise such as breathing, coughing, and laughing, among other background noises. Therefore, highly variable energy levels can be observed in the nonspeech parts of the signal. Moreover, differences in microphones or room configurations may result in variable SNRs from one meeting to another. Thus, SAD is far from being trivial in this context and typical techniques based on feature extraction (energy, spectrum divergence between speech and background noise, and pitch estimation) combined with a threshold-based decision have proven to be relatively ineffective.

Model-based approaches tend to have better performances and rely on a two-class detector, with models pre-trained with external speech and nonspeech data [6], [41], [49], [51], [52]. Speech and nonspeech models may optionally be adapted to specific meeting conditions [15]. Discriminant classifiers such as linear discriminant analysis (LDA) coupled with Mel frequency cepstrum coefficients (MFCCs) [53] or support vector machines (SVMs) [54] have also been proposed in the literature. The main drawback of model-based approaches is their reliance on external data for the training of speech and nonspeech models which makes them less robust to changes in acoustic conditions. Hybrid approaches have been proposed as a potential solution. In most cases, an energy-based detection is first applied in order to label a limited amount of speech and nonspeech data for which there is high confidence in the classification. In a second step, the labeled data are used to train meeting-specific speech and nonspeech models, which are subsequently used in a model-based detector to obtain the final speech/nonspeech segmentation [9], [55]–[57]. Finally, [58] combines a model-based with a 4-Hz modulation energy-based detector. Interestingly, instead of being applied as a preprocessing stage, in this system SAD is incorporated into the speaker diarization process.

C. Segmentation

In the literature, the term "speaker segmentation" is sometimes used to refer to both segmentation and clustering. While some systems treat each task separately many of present state-of-the-art systems tackle them simultaneously, as described in Section III-E. In these cases the notion of strictly independent segmentation and clustering modules is less relevant. However, both modules are fundamental to the task of speaker diarization and some systems, such as that reported in [6], apply distinctly independent segmentation and clustering stages. Thus, the segmentation and clustering models are described separately here.

Speaker segmentation is core to the diarization process and aims at splitting the audio stream into speaker homogeneous segments or, alternatively, to detect changes in speakers, also known as speaker turns. The classical approach to segmentation performs a hypothesis testing using the acoustic segments in two sliding and possibly overlapping, consecutive windows. For each considered change point there are two possible hypotheses: first that both segments come from the same speaker (H_0), and thus that they can be well represented by a single model; and second that there are two different speakers (H_1), and thus that two different models are more appropriate. In practice, models are estimated from each of the speech windows and some criteria are used to determine whether they are best accounted for by two separate models (and hence two separate speakers), or by a single model (and hence the same speaker) by using an empirically determined or dynamically adapted threshold [10], [59]. This is performed across the whole audio stream and a sequence of speaker turns is extracted.

Many different distance metrics have appeared in the literature. Next, we review the dominant approaches which have been used for the NIST RT speaker diarization evaluations during the last four years. The most common approach is that of the Bayesian information criterion (BIC) and its associated ΔBIC metric [33] which has proved to be extremely popular, e.g.,[60]–[62]. The approach requires the setting of an explicit penalty term which controls the tradeoff between missed turns and those falsely detected. It is generally difficult to estimate the penalty term such that it gives stable performance across different meetings and thus new, more robust approaches have been devised. They either adapt the penalty term automatically, i.e., the modified BIC criterion [33], [63], [64], or avoid the use of a penalty term altogether by controlling model complexity [65]. BIC-based approaches are computationally demanding and some systems have been developed in order to use the BIC only in a second pass, while a statistical-based distance is used in a first pass [66]. Another BIC-variant metric, referred to as cross-BIC and introduced in [67] and [68], involves the computation of cross-likelihood: the likelihood of a first segment according to a model tuned from the second segment and vice versa. In [69], different techniques for likelihood normalization are presented and are referred to as bilateral scoring.

A popular and alternative approach to BIC-based measures is the generalized likelihood ratio (GLR), e.g.,[70], [71]. In contrast to the BIC, the GLR is a likelihood-based metric and corresponds to the ratio between the two aforementioned hypotheses, as described in [39], [72], and [73]. To adapt the criterion in order to take into account the amount of training data available in the two segments, a penalized GLR was proposed in [74].

The last of the dominant approaches is the Kullback–Leibler (KL) divergence which estimates the distance between two random distributions [75]. However, the KL divergence is asymmetric, and thus the KL2 metric, a symmetric alternative, has proved to be more popular in speaker diarization when used to characterize the similarity of two audio segments [75]–[77].

Finally, in this section we include a newly introduced distance metric that has shown promise in a speaker diarization task. The information change rate (ICR), or entropy can be used to characterize the similarity of two neighboring speech segments. The ICR determines the change in information that would be obtained by merging any two speech segments under consideration and can thus be used for speaker segmentation. Unlike the measures outlined above, the ICR similarity is not based on a model of each segment but, instead, on the distance between segments in a space of relevance variables, with maximum mutual information or minimum entropy. One suitable space comes from GMM component parameters [18]. The ICR approach is computationally efficient and, in [78], ICR is shown to be more robust to data source variation than a BIC-based distance.

D. Clustering

Whereas the segmentation step operates on adjacent windows in order to determine whether or not they correspond to the same speaker, clustering aims at identifying and grouping together same-speaker segments which can be localized anywhere in the audio stream. Ideally, there will be one cluster for each speaker. The problem of measuring segment similarity remains the same and all the distance metrics described in Section III-C may also be used for clustering, i.e., the KL distance as in [10], a modified KL2 metric as in [61], a BIC measure as in [79] or the cross likelihood ratio (CLR) as in [80] and [81].

However, with such an approach to diarization, there is no provision for splitting segments which contain more than a single speaker, and thus diarization algorithms can only work well if the initial segmentation is of sufficiently high quality. Since this is rarely the case, alternative approaches combine clustering with iterative resegmentation, hence facilitating the introduction of missing speaker turns. Most of present diarization systems thus perform segmentation and clustering simultaneously or clustering on a frame-to-cluster basis, as described in Section III-E. The general approach involves Viterbi realignment where the audio stream is resegmented based on the current clustering hypothesis before the models are retrained on the new segmentation. Several iterations are usually performed. In order to make the Viterbi decoding more stable, it is common to use a Viterbi buffer to smooth the state, cluster or speaker sequence to remove erroneously detected, brief speaker turns, as in [16]. Most state-of-the-art systems employ some variations on this particular issue.

An alternative approach to clustering involves majority voting [82], [83] whereby short windows of frames are entirely assigned to the closest cluster, i.e., that which attracts the most frames during decoding. This technique leads to savings in computation but is more suited to online or live speaker diarization systems.

E. One-Step Segmentation and Clustering

Most state-of-the-art speaker diarization engines unify the segmentation and clustering tasks into one step. In these systems, segmentation and clustering are performed hand-in-hand in one loop. Such a method was initially proposed by ICSI for a bottom-up system [31] and has subsequently been adopted by many others [9], [41], [52], [84]–[86]. For top-down algorithms it was initially proposed by LIA [14] as used in their latest system [16].

In all cases the different acoustic classes are represented using HMM/GMM models. EM training or MAP adaptation is used to obtain the closest possible models given the current frame-to-model assignments, and a Viterbi algorithm is used to reassign all the data into the closest newly-created models. Such processing is sometimes performed several times for the frame assignments to stabilize. This step is useful when a class is created/eliminated so that the resulting class distribution is allowed to adapt to the data.

The one-step segmentation and clustering approach, although much slower, constitutes a clear advantage versus sequential single-pass segmentation and clustering approaches [5]–[7]. On the one hand, early errors (mostly missed speaker turns from the segmentation step) can be later corrected by the re-segmentation steps. On the other hand, most speaker segmentation algorithms use only local information to decide on a speaker change while when using speaker models and Viterbi realignment all data is taken into consideration.

When performing frame assignment using Viterbi algorithm a minimum assignment duration is usually enforced to avoid an unrealistic assignment of very small consecutive segments to different speaker models. Such minimum duration is usually made according to the estimated minimum length of any given speaker turn.

IV. Current Research Directions

In this section, we review those areas of work which are still not mature but which have the potential to improve diarization performance. We first discuss the trend in recent NIST RT evaluations to use spatial information obtained from multiple microphones, which are used by many in combination with MFCCs to improve performance. Then, we discuss the use of prosodic information which has led to promising speaker diarization results. Also addressed in this section is the "Achilles heel" of speaker diarization for meetings, which involves overlapping speech; many researchers have started to tackle the detection of overlapping speech and its correct labeling for improved diarization outputs. We then consider a recent trend towards multimodal speaker diarization including studies of multimodal, audiovisual techniques which have been successfully used for speaker diarization, at least for laboratory conditions. Finally, we consider general combination strategies that can be used to combine the output of different diarization systems. The following summarizes recent work in all of these areas.

A. *Time-Delay Features*

Estimates of inter-channel delay may be used not only for delay-and-sum beamforming of multiple microphone channels, as described in Section III-A, but also for speaker localization. If we assume that speakers do not move, or that appropriate tracking algorithms are used, then estimates of speaker location may thus be used as alternative features, which have nowadays become extremely popular. Much of the early work, e.g.,[87], requires explicit knowledge of microphone placement. However, as is the case with NIST evaluations, such *a priori* information is not always available. The first work [88] that does not rely on microphone locations led to promising results, even if error rates were considerably higher than that achieved with acoustic features. Early efforts to combine acoustic features and estimates of inter-channel delay clearly demonstrated their potential, e.g.,[89], though this work again relied upon known microphone locations.

More recent work, and specifically in the context of NIST evaluations, reports the successful combination of acoustic and inter-channel delay features [86], [90], [91] when they are combined at the weighted log-likelihood level, though optimum weights were found to vary across meetings. Better results are reported in [42] where automatic weighting based on an entropy-based metric is used for cluster comparison in a bottom-up speaker diarization system. A complete front-end for speaker diarization with multiple microphones was proposed in [42]. Here a two-step TDOA Viterbi post-processing algorithm together with a dynamic output signal weighting algorithm were shown to greatly improve speaker diarization accuracy and the robustness of inter-channel delay estimates to noise and reverberation, which commonly afflict source localization algorithms. More recently, an approach to the unsupervised discriminant analysis of inter-channel delay features was proposed in [92] and results of approximately 20% DER were reported using delay features alone.

In the most recent NIST RT evaluation, in 2009, all but one entry used estimates of inter-channel delay both for beamforming and as features. Since comparative experiments are rarely reported it is not possible to assess the contribution of delay features to diarization performance. However, those who do use delay features report significant improvements in diarization performance and the success of these systems in NIST RT evaluations would seem to support their use.

B. *Use of Prosodic Features in Diarization*

The use of prosodic features for both speaker detection and diarization is emerging as a reaction to the theoretical inconsistency derived from using MFCC features both for speaker recognition (which requires invariance against words) and speech recognition (which requires invariance against speakers) [93]. In [84], the authors present a systematic investigation of the speaker discriminability of 70 long-term features, most of them prosodic features. They provide evidence that despite the dominance of short-term cepstral features in speaker recognition, a number of long-term features can provide significant information for speaker discrimination. As already suggested in [94], the consideration of patterns derived from larger segments of speech can reveal individual characteristics of the speakers' voices as well as their speaking behavior, information which cannot be captured using a short-term, frame-based cepstral analysis. The authors use Fisher LDA as a ranking methodology and sort the 70 prosodic and long-term features by speaker discriminability. The combination of the top-ten ranked prosodic and long-term features combined with regular MFCCs leads to a 30% relative improvement in terms of DER compared to the top-performing system of the NIST RT evaluation in 2007. An extension of the work is provided in [95]. The paper presents a novel, adaptive initialization scheme that can be applied to standard bottom-up diarization algorithms. The initialization method is a combination of the recently proposed "adaptive seconds per Gaussian" (ASPG) method [96] and a new pre-clustering method in addition to a new strategy which automatically estimates an appropriate number of initial clusters based on prosodic features. It outperforms previous cluster initialization algorithms by up to 67% (relative).

C. *Overlap Detection*

A fundamental limitation of most current speaker diarization systems is that only one speaker is assigned to each segment. The presence of overlapped speech, though, is common in multiparty meetings and, consequently, presents a significant challenge to automatic systems. Specifically, in regions where more than one speaker is active, missed speech errors will be incurred and, given the high performance of some state-of-the-art systems, this can be a substantial fraction of the overall diarization error. A less direct, but also significant, effect of overlapped speech in diarization pertains to speaker clustering and modeling. Segments which contain speech from more than a single speaker should not be assigned to any individual speaker cluster nor included in any individual speaker model. Doing so adversely affects the purity of speaker models, which ultimately reduces diarization performance. Approaches to overlap detection were thoroughly assessed in [97] and [98] and, even while applied to ASR as opposed to speaker diarization, only a small number of systems actually detects overlapping speech well enough to improve error rates [99]–[101].

Initially, the authors in [102] demonstrated a theoretical improvement in diarization performance by adding a second speaker during overlap regions using a simple strategy of assigning speaker labels according to the labels of the neighboring segments, as well as by excluding overlap regions from the input to the diarization system. However, this initial study assumed ground-truth overlap detection. In [100], a real overlap detection system was developed, as well as a better heuristic that computed posterior probabilities from diarization to post process the output and include a second speaker on overlap regions. The main bottleneck of the achieved performance gain is mainly due to errors in overlap detection, and more work on enhancing its precision and recall is reported in [99] and [101]. The main approach consists of a three-state HMM-GMM system (nonspeech, nonoverlapped speech, and overlapped speech), and the best feature combination is MFCC and modulation spectrogram features [103], although comparable results were achieved with other features such as root mean squared energy, spectral flatness, or harmonic energy ratio. The reported performance of the overlap detection is 82% precision and 21% recall, and yielded a relative improvement of 11% DER. However, assuming reference overlap detection, the relative DER improvement goes up to 37%. This way, this area has potential for future research efforts.

D. Audiovisual Diarization

Reference [104] presents an empirical study to review definitions of audiovisual synchrony and examine their empirical behavior. The results provide justifications for the application of audiovisual synchrony techniques to the problem of active speaker localization in broadcast video. The authors of [105] present a multi-modal speaker localization method using a specialized satellite microphone and an omni-directional camera. Though the results seem comparable to the state-of-the-art, the solution requires specialized hardware. The work presented in [106] integrates audiovisual features for online audiovisual speaker diarization using a dynamic Bayesian network (DBN) but tests were limited to discussions with two to three people on two short test scenarios. Another use of DBN, also called factorial HMMs [107], is proposed in [108] as an audiovisual framework. The factorial HMM arises by forming a dynamic Bayesian belief network composed of several layers. Each of the layers has independent dynamics but the final observation vector depends upon the state in each of the layers. In [109], the authors demonstrate that the different shapes the mouth can take when speaking facilitate word recognition under tightly constrained test conditions (e.g., frontal position of the subject with respect to the camera while reading digits).

Common approaches to audiovisual speaker identification involve identifying lip motion from frontal faces, e.g.,[110]–[114]. Therefore, the underlying assumption is that motion from a person comes predominantly from the motion of the lower half of their face. In addition, gestural or other nonverbal behaviors associated with natural body motion during conversations are artificially suppressed, e.g., for the CUAVE database [115]. Most of the techniques involve the identification of one or two people in a single video camera only where short term synchrony of lip motion and speech are the basis for audiovisual localization. In a real scenario the subject behavior is not controlled and, consequently, the correct detection of the mouth is not always feasible. Therefore, other forms of body behavior, e.g., head gestures, which are also visible manifestations of speech [116] are used. While there has been relatively little work on using global body movements for inferring speaking status, some studies have been carried out [82], [117]–[119] that show promising initial results.

However, until the work presented in [120], approaches have never considered audiovisual diarization as a single, unsupervised joint optimization problem. The work in [120], though, relies on multiple cameras. The first paper that discusses joint audiovisual diarization using only a single, low-resolution overview camera and also tests on meeting scenarios where the participants are able to move around freely in the room is [121]. The algorithm relies on very few assumptions and is able to cope with an arbitrary amount of cameras and subframes. Most importantly, as a result of training a combined audiovisual model, the authors found that speaker diarization algorithms can result in speaker localization as side information. This way joint audiovisual speaker diarization can answer the question "who spoken when and from where." This solution to the localization problem has properties that may not be observed either by audio-only diarization nor by video-only localization, such as increased robustness against various issues present in the channel. In addition, in contrast to audio-only speaker diarization, this solution provides a means for identifying speakers beyond clustering numbers by associating video regions with the clusters.

E. System Combination

System or component combination is often reported in the literature as an effective means for improving performance in many speech processing applications. However, very few studies related to speaker diarization have been reported in recent years. This could be due to the inherent difficulty of merging multiple output segmentations. Combination strategies have to accommodate differences in temporal synchronization, outputs with different number of speakers, and the matching of speaker labels. Moreover, systems involved in the combination have to exhibit segmentation outputs that are sufficiently orthogonal in order to ensure significant gains in performance when combined. Some of the combination strategies proposed consist of applying different algorithms/components sequentially, based on the segmentation outputs of the previous steps in order to refine boundaries (referred to as "hybridization" or "piped" systems in [122]). In [123] for instance, the authors combine two different algorithms based on the Information Bottleneck framework. In [124], the best components of two different speaker diarization systems implemented by two different French laboratories (LIUM and IRIT) are merged and/or used sequentially, which leads to a performance gain compared to results from individual systems. An original approach is proposed in [125], based on a "real" system combination. Here, a couple of systems uniquely differentiated by their input features (parameterizations based on Gaussianized against non-Gaussianized MFCCs) are combined for the speaker diarization of phone calls conversations. The combination approach relies on both systems identifying some common clusters which are then considered as the most relevant. All the segments not belonging to these common clusters are labeled as misclassified and are involved in a new re-classification step based on a GMM modeling of the common clusters and a maximum likelihood-based decision.

F. Alternative Models

Among the clustering structures recently developed some differ from the standard HMM insofar as they are fully nonparametric (that is, the number of parameters of the system depends on the observations). The Dirichlet process (DP) [126] allows for converting the systems into Bayesian and nonparametric systems. The DP mixture model produces infinite Gaussian mixtures and defines the number of components by a measure over distributions. The authors of [127] illustrate the use of the Dirichlet process mixtures, showing an improvement compared to other classical methods. Reference [128] proposes another nonparametric Bayesian approach, in which a stochastic hierarchical Dirichlet process (HDP) defines a prior distribution on transition matrices over countably infinite state spaces, that is, no fixed number of speakers is assumed, nor found through either split or merging approaches using classical model selection approaches (such as the BIC criterion). Instead, this prior measure is placed over distributions (called a random measure), which is integrated out using likelihood-prior conjugacy. The resulting HDP-HMM leads to a data-driven learning algorithm which infers posterior distributions over the number of states. This posterior uncertainty can be integrated out when making predictions effectively averaging over models of varying complexity. The HDP-HMM has shown promise in diarization [129], yielding similar performance to the standard agglomerative HMM with GMM emissions, while requiring very little hyperparameter tuning and providing a statistically sound model. Globally, these non parametric Bayesian approaches did not bring a major improvement compared to classical systems as presented in Section III. However, they may be promising insofar as they do not necessarily need to be optimized for certain data compared to methods cited in Section II. Furthermore, they provide a probabilistic interpretation on posterior distributions (e.g., number of speakers).

V. Performance Evaluation

In this section, we report an analysis of speaker diarization performance as reported during the four most recent NIST RT evaluations. The analysis focuses solely on conference meetings which are the core evaluation condition. We also present an analysis of the ground-truth references in order to underline the characteristics of the data with respect to meeting sources and the different evaluation campaigns. Finally we show state-of-the-art system results, collated from four NIST RT'07 and RT'09 evaluation participants, which aim at giving a baseline for future research.

A. Benchmarking Evaluations

Since 2004, NIST has organized a series of benchmark evaluations within the Rich Transcription (RT) campaigns.[3] One of the tasks involves speaker diarization of different sets of data. A common characteristic of these evaluations is that the only *a priori* knowledge available to the participants relates to the recording scenario/source (e.g., conference meetings, lectures, or coffee breaks for the meetings domain), the language (English), and the formats of the input and output files. Evaluation participants may use external or background data for building world models and/or for normalization purposes but no *a priori* information relating to speakers in the recordings is available. The number of speakers is also not known.

In recent years, the NIST RT evaluations have focussed on the conference meeting domain, where the spontaneous speaking style presents a considerable challenge for speaker diarization. Each meeting used in the evaluations was recorded using multiple microphones (of different types and quality) which are positioned on the participants or in different locations around the meeting room. By grouping these microphones into different classes, NIST created several contrastive evaluation conditions. These include: individual headphone microphones (IHM), single distant microphones (SDM), multiple distant microphones (MDM), multiple mark III arrays (MM3A), and all distant microphones (ADM). MM3A microphones are those exclusively found within the arrays built and provided by NIST. These are usually not included within the MDM condition, they are included within the ADM condition. In this section we show results for the MDM and SDM conditions since we consider them to be the most representative of standard meeting room recording equipment. These conditions have also proven to be the most popular among evaluation participants.

Participating teams are required to submit a hypothesis of speaker activity including start-stop times of speech segments with speaker labels, which are used solely to identify the multiple interventions of a given speaker, but do not need to reflect the speaker's real identity. These system outputs are compared to the ground-truth reference in order to obtain the overall DER. The DER metric is the sum of three sources of error: missed speech (percentage of speech in the ground-truth but not in the hypothesis), false alarm speech (percentage of speech in the hypothesis but not in the ground-truth) and speaker error (percentage of speech assigned to the wrong speaker). The speaker error can be further classified into incorrectly assigned speakers and speaker overlap error. In the first case, the hypothesized speaker does not correspond to the real (ground-truth) speaker. Speaker overlap error refers to the case when the wrong number of speakers is hypothesized when multiple speakers speak at the same time. The inclusion of overlapping speech error in the evaluation was restricted to a contrastive metric in the initial RT evaluations but has been the primary metric since 2006. Overlap errors can be classified as missed overlap (when fewer speakers than the real number are hypothesized) and false alarm overlap (when too many speakers are hypothesized). In the NIST evaluations up to four overlapping speakers are considered in the scoring.

Note that as the DER is time-weighted, it ascribes little importance to the diarization quality of speakers whose overall speaking time is small. Additionally, a nonscoring collar of 250 ms is generally applied either side of the ground-truth segment boundaries to account for inevitable inconsistencies in precise start and end point labeling. When comparing the system outputs with the ground-truth, and given that the labels identifying the speakers are just relative identifiers, the scoring algorithm

[3] See http://nist.gov/speech/tests/rt.

TABLE I

GROUND-TRUTH ANALYSIS FOR THE DATASETS OF THE LAST FOUR SPEAKER DIARIZATION EVALUATION CAMPAIGNS (RT'05 TO RT'09) AND MEETING SOURCE. COMPARISONS ARE BASED ON THE AVERAGE SPEAKER AND TURN DURATIONS (LEFT-HALF SIDE) AND THE PERCENTAGE OF SILENCE AND OVERLAPPING SPEECH (RIGHT-HALF SIDE)

		Av. speaker duration / Av. turn duration				silence / overlap			
Meeting Source	# meetings	RT'05	RT'06	RT'07	RT'09	RT'05	RT'06	RT'07	RT'09
AMI	2	2.7s/2.3s	-	-	-	10.7%/11.4%	-/-	-/-	-/-
CMU	6	2.0s/1.8s	1.7s/1.2s	1.8s/1.6s	-	6.8%/21.6%	20.1%/**13.6%**	29.6%/**8.8%**	-/-
ICSI	2	2.5s/2.2s	-	-	-	8.4%/20.7%	-/-	-/-	-/-
NIST	9	2.3s/1.9s	2.8s/1.6s	2.1s/1.5s	1.6s/1.3s	8.0%/21.0%	36.0%/5.8%	21.7%/6.6%	11.0%/**20.5%**
VT	6	3.0s/2.5s	2.8s/1.3s	2.3s/1.6s	-	17.6%/5.4%	44.8%/6.0%	23.9%/5.6%	-/-
EDI	6	-	2.1s/1.4s	2.0s/1.4s	1.8s/1.2s	-/-	27.4%/6.4%	24.2%/9.4%	27.3%/8.3%
TNO	1	-	2.1s/1.5s	-	-	-/-	26.5%/6.0%	-/-	-/-
IDI	2	-	-	-	2.2s/1.7s	-/-	-/-	-/-	17.4%/8.6%
Average	-	2.5s/**2.1s**	2.3s/1.4s	2.0s/1.5s	1.8s/1.4s	10.3%/**16.0%**	31.5%/7.7%	24.9%/7.6%	17.5%/**13.6%**

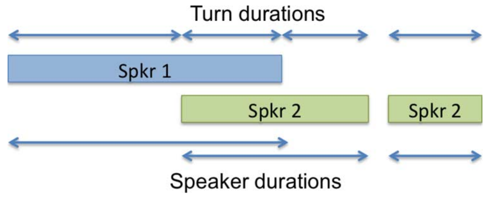

Fig. 2. Examples of turn and speaker durations in the presence of overlapped speech and silences.

first computes an optimum mapping between both sets of labels in order to obtain the DER. This is normally performed according to a standard dynamic programming algorithm defined by NIST.

B. Ground-Truth Analysis

Ground-truth references for evaluating speaker diarization were initially obtained via manual labeling of the acoustic data; however, high variations between different labelers proved to be problematic. Therefore, more recently, an automatically generated forced alignment has been used in order to extract more reliable speaker start and end points using an automatic speech recognition (ASR) system, human-created transcriptions, and the audio from individual head microphones (IHM).

As meeting data come from a variety of sources some differences between them are expected. Furthermore, large changes in the final DER scores from different evaluations would suggest that there are differences between the sets of meetings used each year. To gauge the differences we have analyzed over 20 different parameters computed on the ground-truth data. In Table I, we report four of these parameters, which we found most interesting, and group results by meeting source and by evaluation year.

In the left side of the table we report average speaker and turn durations. As exemplified in Fig. 2, the average speaker duration refers to the average time during which a speaker is active (i.e., a single line in the RTTM reference files). Conversely, the average turn duration refers to the average time during which there is no change in speaker activity and is thus always smaller than the average speaker duration. The difference between the two statistics reflects the degree of overlap and spontaneity. Without any overlap and a pause between each speaker exchange the average speaker and turn durations would be identical. Increases in overlap and spontaneity will result in a larger speaker/turn ratio. In the right side of Table I we report the percentage of silence and of overlapping speech.

For RT'05 the average speaker segment duration is 2.5 s. This value decreases continuously for subsequent datasets (2.3 s for RT'06, 2.0 s for RT'07, and 1.8 s for RT'09). This tendency leads to increasingly more frequent speaker turns and increases the chances of miss-classifying a speech segment. The average turn segment duration is 2.1 s for RT'05. This value falls to 1.4 s for RT'06 and remains stable for RT'07 and RT'09 (1.5 s and 1.4 s respectively). The consistent decrease in speaker/turn duration ratio highlights a general trend of increasing spontaneity and helps to explain the differences in results from one dataset to another. There are no distinct differences across different meeting sites.

There are also noticeable differences in silence and overlap statistics. The percentage of silence is lower for the RT'05 and RT'09 datasets than it is for the RT'06 and RT'09 datasets (10.3% and 17.5% cf. 31.5% and 24.9%). However, the RT'05 and RT'09 datasets have a higher overlap rate than the RT'06 and RT'07 datasets (16.0% and 13.6% cf. 7.7% and 7.6%). This is primarily due to three meetings (from CMU, ICSI, and NIST sites) which have overlap rates over 25% (note that values in Table I are averaged across sites, and do not reflect individual meeting scores). In the case of the RT'09 dataset, the slightly high average overlap of 13% is due to a single meeting (recorded by NIST) in which the overlap reaches 31%. Listening to this meeting we concluded that the reason of such overlap is that it is not a professional meeting but a social rendezvous. Conversely, RT'05 and RT'09 have in average a lower percentage of silence (10% and 17%) compared to RT'06 and RT'07 (31% and 25%). A lower silence rate and higher overlap might indicate that these meetings are more dynamic, with less idle time and more discussion, although this does not mean that they are more spontaneous, as their speech and speaker segment lengths are still high compared to the RT'09 dataset.

Overall, we see that, although all recordings belong to the same task, there are large differences between the datasets used for each evaluation campaign, as well as between recordings from the same source (recording site), but from different datasets. This emphasizes the need for robust systems which perform well regardless of particular dataset characteristics. It is important to note, however, that the NIST RT datasets discussed here typically contain around eight meetings per dataset, each of them contributing to a single DER score. Random variations on any meeting from these small datasets have a significant impact on average results. It is then difficult to reliably interpret results and hence also difficult to draw meaningful conclusions.

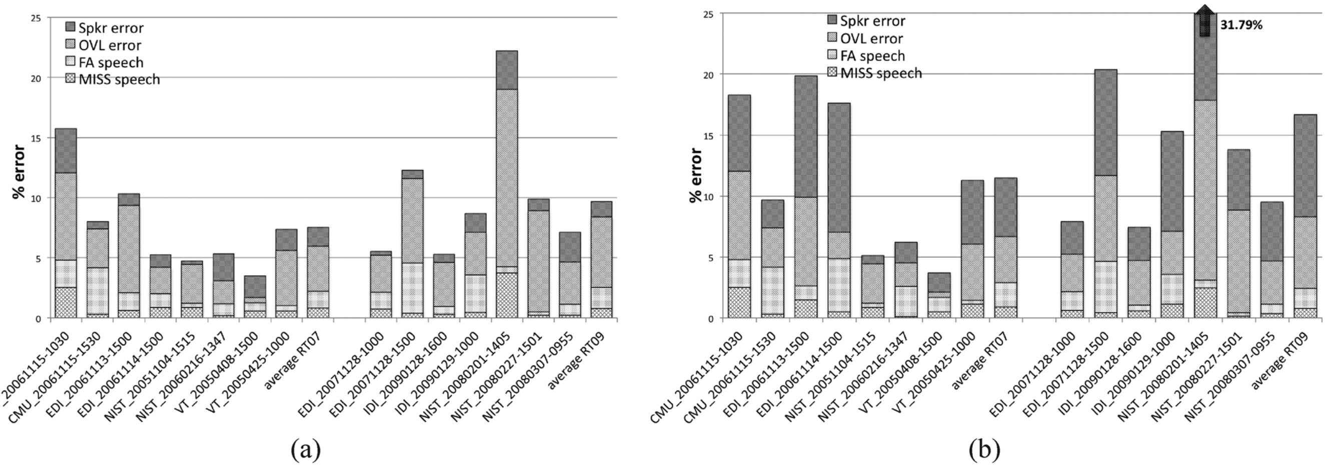

Fig. 3. DERs for the RT'07 and RT'09 (a) in multiple distant microphone (MDM) condition, and (b) single distant microphone (SDM) condition (note that spkr_error in meeting NIST_20080201–1405 has been trimmed to fit the screen, with a speaker error of 31.79% and a total DER of 49.65%).

Comparisons with the work of the speech and speaker recognition communities highlight the rapid acceleration in research effort and progress stemming from the availability of huge datasets. Advances in sophisticated modeling and normalization strategies have revolutionized research in these related fields over recent years. It becomes apparent that the fundamental lack of larger speaker diarization datasets, which makes it difficult to assess novel algorithms, is a critical barrier to further research in our field. Significantly larger datasets are needed in order to obtain more robust and meaningful performance estimates and comparisons. As a result of processing more data, faster algorithms will also need to be investigated for research in speaker diarization to be feasible with standard computing resources.

C. Evaluation Results

To assess the current state-of-the-art and provide a baseline for future research we present results for the RT'07 (Fig. 3 left half) and RT'09 (Fig. 3 right half) NIST evaluations for the MDM [Fig. 3(a)] and SDM [Fig. 3(b)] conditions. Both figures have been compiled from a comparison of results from four of the participating sites (LIA/Eurecom,[4] I2R/NTU, ICSI and UPC) and by selecting the result with lowest DER for each meeting recording. Given the volatility of the results described and studied in [3], by selecting the best result in each case we hypothesize that these results are a more meaningful estimation of the state-of-the-art performance in speaker diarization for conference meeting data than selecting all results from any single system output. To illustrate the variation in performance for different meetings we provide results for individual meetings. In both figures, errors are decomposed into the speaker error (Spkr error), overlap error (OVL error), false alarm speech error (FA speech), and missed speech error (MISS speech).

For the MDM condition [Fig. 3(a)] the average DER for the RT'07 and RT'09 datasets is 7.5% and 10.1%, respectively. Performance varies between 3.5% and 15.7% for the RT'07 dataset whereas for the RT'09 dataset performance varies between 5.3% and 22.2%. For the SDM condition the average DER is 11.6% and 17.7% for the RT'07 and RT'09 datasets, respectively. Performance is always poorer than that for the MDM condition and varies between 3.7% and 19.9% for the RT'07 dataset and between 7.4% and 49.7% for the RT'09 dataset. Thus, there is a large variation in performance across different meetings and in all cases we observe significant overlap errors and their often-dominant impact upon the final DER. Of particular note is the poor performance obtained on the single NIST_20080201–1405, which correlates with the particularly high percentage of overlapping speech for this meeting as illustrated in Table I. Hence, the detection and appropriate treatment of overlapping speech remains an unsolved problem. In fact, the overlap error shown in Fig. 3 is entirely due to missed overlap regions, as none of the speaker diarization systems considered in this analysis included an overlap detector. Also of note is the general stability of speech activity detection (SAD) algorithms which achieve impressive levels of performance in both MDM and SDM conditions (i.e., they are robust to the quality of the signal). Values of around 1% to 2% missed speech error rates and 2% to 3% false alarm error rates are currently typical. The main difference between MDM and SDM performance rests mainly in the speaker error. Here diarization systems are affected by the reduced signal quality which characterizes the SDM condition.

Overall, the large variations in DER observed among the different meetings and meeting sets originate from the large variance of many important factors for speaker diarization, which makes the conference meeting domain not as easily tractable as more formalized settings such as broadcast news, lectures, or court house trials. Previous work has highlighted the difficulty in assessing the performance of speaker diarization algorithms with the view of improving performance [130]. As reported in Section III, current approaches to speaker diarization involve a sequence of separate stages where each stage takes its input from the preceding stage(s). When combined in such a fashion, it is exceedingly difficult to assess the performance of each system component since every single one is affected by the performance of all previous processing stages. Furthermore, it is not guaranteed that improvements to one stage, for example that of segmentation, will lead unequivocally to improvements in later stages, for example that of clustering. This makes the optimization of different system components rather troublesome. Once again, by drawing comparisons to the speech and speaker recognition fields, it is reasonable to foresee more unified approaches, as is already in progress with the now commonplace

[4] Eurecom was associated with the LIA for the RT'09 campaign only.

combined approaches to segmentation and clustering. In particular, we believe that important decreases in DER will have to come in the near future from systems incorporating effective algorithms that can detect and correctly assign overlapping speech.

VI. Conclusion and Directions for Future Research

Research on speaker diarization has been developed in many domains, from phone calls conversations within the speaker recognition evaluations, to broadcast news and meeting recordings in the NIST Rich Transcription evaluations. Furthermore, it has been used in many applications such as a front-end for speaker and speech recognition, as a meta-data extraction tool to aid navigation in broadcast TV, lecture recordings, meetings, and video conferences and even for applications such as media similarity estimation for copyright detection. Also, speaker diarization research has led to various by-products. For example, with the availability of recordings using multiple microphones, a set of algorithms has been proposed in recent years both for signal enhancement and to take advantage of the extra information that these offer. In addition, the availability of other modalities, such as video, have started to inspire multimodal diarization systems, thus merging the visual and the acoustic domains.

This paper provides an overview of the current state-of-the-art in speaker diarization systems and underlines several challenges that need to be addressed in future years. For example, speaker diarization is not yet sufficiently mature so that methods can be easily ported across different domains, as shown in Section V, where small differences in meeting data (recorded at identical sites) lead to large variations in performance. In the meantime, larger datasets need to be compiled in order for results to become more meaningful and for systems to be more robust to unseen variations. Of course, with increasing dataset sizes, systems will have to become more efficient in order to process such data in reasonable time. Still, the biggest single challenge is probably the handling of overlapping speech, which needs to be attributed to multiple speakers. As a relatively embryonic community, at least compared to the more established fields of speech and speaker recognition, there are thus outstanding opportunities for significant advances and important changes to the somewhat ad hoc and heuristic approaches that currently dominate the field.

Overall, the future of the field seems even broader and brighter than the present, as more and more people acknowledge the usefulness of audio methods for many tasks that have traditionally been thought to be exclusively solvable in the visual domain. Speaker diarization is one of the fundamental problems underlying virtually any task that involves acoustics and the presence of more than one person.

Acknowledgment

The authors would like to thank those RT participating sites that lent them their results for this study, in particular J. Luque from UPC, Spain, and T.H. Nguyen and H. Li from I2R and NTU, Singapore.

References

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

[2] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 5, pp. 1557–1565, Sep. 2006.

[3] N. Mirghafori and C. Wooters, "Nuts and flakes: A study of data characteristics in speaker diarization," in *Proc. ICASSP*, 2006.

[4] X. Anguera, "Robust speaker diarization for meetings," Ph.D. dissertation, Univ. Politecnica de Catalunya, Barcelona, Spain, 2006.

[5] M. Kotti, E. Benetos, and C. Kotropoulos, "Computationally efficient and robust BIC-based speaker segmentation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 920–933, Jul. 2008.

[6] X. Zhu, C. Barras, L. Lamel, and J.-L. Gauvain, "Multi-stage speaker diarization for conference and lecture meetings," in *Proc. Multimodal Technol. Perception of Humans: Int. Eval. Workshops CLEAR 2007 and RT 2007, Baltimore, MD, May 8–11, 2007, Revised Selected Papers*, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 533–542.

[7] S. Jothilakshmi, V. Ramalingam, and S. Palanivel, "Speaker diarization using autoassociative neural networks," *Eng. Applicat. Artif. Intell.*, vol. 22, no. 4-5, pp. 667–675, 2009.

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

[9] 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, Baltimore, MD, USA, May 8–11, 2007, Revised Selected Papers*, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 509–519.

[10] J. Rougui, M. Rziza, D. Aboutajdine, M. Gelgon, and J. Martinez, "Fast incremental clustering of Gaussian mixture speaker models for scaling up retrieval in on-line broadcast," in *Proc. ICASSP*, May 2006, vol. 5, pp. 521–524.

[11] W. Tsai, S. Cheng, and H. Wang, in *Proc. ICSLP*, 2004.

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

[13] T. Nguyen *et al.*, "The IIR-NTU speaker diarization systems for RT 2009," in *Proc. RT'09, NIST Rich Transcription Workshop*, Melbourne, FL, 2009.

[14] 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 Lang. Recognition Workshop*, Chania, Creete, Jun. 2001, pp. 175–180.

[15] C. Fredouille and N. Evans, "The LIA RT'07 speaker diarization system," in *Proc. Multimodal Technol. for Perception of Humans: Int. Eval. Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8–11, 2007, Revised Selected Papers*, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 520–532.

[16] C. Fredouille, S. Bozonnet, and N. W. D. Evans, "The LIA-EURECOM RT'09 speaker diarization system," in *Proc. RT'09, NIST Rich Transcription Workshop*, Melbourne, FL, 2009.

[17] 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*, Dallas, TX, Mar. 14–19, 2010, pp. 4958–4961.

[18] D. Vijayasenan, F. Valente, and H. Bourlard, "Agglomerative information bottleneck for speaker diarization of meetings data," in *Proc. ASRU*, Dec. 2007, pp. 250–255.

[19] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 7, pp. 1382–1393, Sep. 2009.

[20] S. McEachern, "Estimating normal means with a conjugate style dirichlet process prior," in *Proc. Commun. Statist.: Simul. Comput.*, 1994, vol. 23, pp. 727–741.

[21] G. E. Hinton and D. van Camp, "Keeping the neural networks simple by minimizing the description length of the weights," in *Proc. 6th Annu. Conf. Comput. Learn. Theory*, New York, 1993, COLT '93, pp. 5–13.

[22] M. J. Wainwright and M. I. Jordan, "Variational inference in graphical models: The view from the marginal polytope," in *Proc. 41st Annu. Allerton Conf. Commun., Control, Comput.*, Urbana-Champaign, IL, 2003.

[23] F. Valente, "Variational Bayesian methods for audio indexing," Ph.D. dissertation, Eurecom Inst., Sophia-Antipolis, France, 2005.

[24] D. Reynolds, P. Kenny, and F. Castaldo, "A study of new approaches to speaker diarization," in *Proc. Interspeech*, 2009.

[25] P. Kenny, *"Bayesian Analysis of Speaker Diarization with Eigenvoice Priors," Technical Report*. Montreal, QC, Canada: CRIM, 2008.

[26] X. Anguera and J.-F. Bonastre, “A novel speaker binary key derived from anchor models,” in *Proc. Interspeech*, 2010.

[27] X. Anguera and J.-F. Bonastre, “Fast speaker diarization based on binary keys,” in *Proc. ICASSP*, 2011.

[28] Y. Huang, O. Vinyals, G. Friedland, C. Muller, N. Mirghafori, and C. Wooters, “A fast-match approach for robust, faster than real-time speaker diarization,” in *Proc. IEEE Workshop Autom. Speech Recognition Understanding*, Kyoto, Japan, Dec. 2007, pp. 693–698.

[29] G. Friedland, J. Ching, and A. Janin, “Parallelizing speaker-attributed speech recognition for meeting browsing,” in *Proc. IEEE Int. Symp. Multimedia*, Taichung, Taiwan, Dec. 2010, pp. 121–128.

[30] X. Anguera, C. Wooters, and J. Hernando, “Friends and enemies: A novel initialization for speaker diarization,” in *Proc. ICSLP*, Pittsburgh, PA, Sep. 2006.

[31] J. Ajmera, “A robust speaker clustering algorithm,” in *Proc. ASRU*, 2003, pp. 411–416.

[32] X. Anguera, C. Wooters, and J. Hernando, “Purity algorithms for speaker diarization of meetings data,” in *Proc. ICASSP*, Toulouse, France, May 2006, pp. 1025–1028.

[33] S. S. Chen and P. S. Gopalakrishnan, “Speaker, environment and channel change detection and clustering via the bayesian information criterion,” in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, Feb. 1998, pp. 127–132.

[34] H. Gish and M. Schmidt, “Text independent speaker identification,” *IEEE Signal Process. Mag.*, vol. 11, no. 4, pp. 18–32, Oct. 1994.

[35] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede, “The ICSI meeting project: Resources and research,” in *Proc. ICASSP Meeting Recognition Workshop*, 2004.

[36] 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. Meas. Behavior*, 2005.

[37] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Cristoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet, “The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms,” *Lang. Resources Eval.*, vol. 41, Dec. 2007.

[38] C. Fredouille, D. Moraru, S. Meignier, L. Besacier, and J.-F. Bonastre, “The NIST 2004 spring rich transcription evaluation: Two-axis merging strategy in the context of multiple distant microphone based meeting speaker segmentation,” in *Proc. NIST 2004 Spring Rich Transcript. Eval. Workshop*, Montreal, QC, Canada, 2004.

[39] Q. Jin, K. Laskowski, T. Schultz, and A. Waibel, “Speaker segmentation and clustering in meetings,” in *Proc. ICSLP*, Jeju, Korea, Sep. 2004.

[40] D. Istrate, C. Fredouille, S. Meignier, L. Besacier, and J.-F. Bonastre, “NIST RT05S evaluation: Pre-processing techniques and speaker diarization on multiple microphone meetings,” in *Proc. NIST 2005 Spring Rich Transcript. Eval. Workshop*, Edinburgh, U.K., Jul. 2005.

[41] X. Anguera, C. Wooters, B. Peskin, and M. Aguilo, “Robust speaker segmentation for meetings: The ICSI-SRI spring 2005 diarization system,” in *Proc. NIST MLMI Meeting Recognition Workshop*, Edinburgh, U.K., 2005.

[42] X. Anguera, C. Wooters, and J. Hernando, “Acoustic beamforming for speaker diarization of meetings,” *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 7, pp. 2011–2023, Sep. 2007.

[43] X. Anguera, BeamformIt (The Fast and Robust Acoustic Beamformer) [Online]. Available: http://www.xavieranguera.com/beamformit/

[44] N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. New York: Wiley, 1949.

[45] 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, vol. 1, pp. 4–7.

[46] M. L. Seltzer, B. Raj, and R. M. Stern, “Likelihood maximizing beamforming for robust hands-free speech recognition,” *IEEE Trans. Speech Audio Process.*, vol. 12, no. 5, pp. 489–498, Sep. 2004.

[47] L. J. Griffiths and C. W. Jim, “An alternative approach to linearly constrained adaptive beamforming,” *IEEE Trans. Antennas Propagat.*, vol. AP-30, no. 1, pp. 27–34, Jan. 1982.

[48] M. Woelfel and J. McDonough*, Distant Speech Recognition*. New York: Wiley, 2009.

[49] C. Wooters, J. Fung, B. Peskin, and X. Anguera, “Towards robust speaker segmentation: The ICSI-SRI fall 2004 diarization system,” in *Proc. Fall 2004 Rich Transcript. Workshop (RT04)*, Palisades, NY, Nov. 2004.

[50] J. Ramirez, J. M. Girriz, and J. C. Segura, M. Grimm and K. Kroschel, Eds., “Voice activity detection. Fundamentals and speech recognition system robustness,” in *Proc. Robust Speech Recognit. Understand.*, Vienna, Austria, Jun. 2007, p. 460.

[51] C. Fredouille and G. Senay, “Technical improvements of the E-HMM based speaker diarization system for meeting records,” in *Proc. MLMI Third Int. Workshop, Bethesda, MD, USA, Revised Selected Paper*, Berlin, Heidelberg: Springer-Verlag, 2006, pp. 359–370.

[52] D. A. V. Leeuwen and M. Konečný, “Progress in the AMIDA speaker diarization system for meeting data,” in *Proc. Multimodal Technol. for Percept. of Humans: Int. Eval. Workshops CLEAR 2007 and RT 2007, Baltimore, MD, May 8–11, 2007, Revised Selected Papers*, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 475–483.

[53] A. Rentzeperis, A. Stergious, C. Boukis, A. Pnevmatikakis, and L. Polymenakos, “The 2006 Athens information technology speech activity detection and speaker diarization systems,” in *Proc. Mach. Learn. Multimodal Interaction: 3rd Int. Workshop, MLMI 2006, Bethesda, MD, Revised Selected Paper*, Berlin, Heidelberg: Springer-Verlag, 2006, pp. 385–395.

[54] A. Temko, D. Macho, and C. Nadeu, “Enhanced SVM training for robust speech activity detection,” in *Proc. ICASSP*, Honolulu, HI, 2007, pp. 1025–1028.

[55] X. Anguera, C. Wooters, M. Anguilo, and C. Nadeu, “Hybrid speech/non-speech detector applied to speaker diarization of meetings,” in *Proc. Speaker Odyssey Workshop*, Puerto Rico, Jun. 2006.

[56] H. Sun, T. L. Nwe, B. Ma, and H. Li, “Speaker diarization for meeting room audio,” in *Proc. Interspeech’09*, Sep. 2009.

[57] T. L. Nwe, H. Sun, H. Li, and S. Rahardja, “Speaker diarization in meeting audio,” in *Proc. ICASSP*, Taipei, Taiwan, 2009, pp. 4073–4076.

[58] E. El-Khoury, C. Senac, and J. Pinquier, “Improved speaker diarization system for meetings,” in *Proc. ICASSP*, Taipei, Taiwan, 2009, pp. 4097–4100.

[59] L. Lu, H.-J. Zhang, and H. Jiang, “Content analysis for audio classification and segmentation,” *IEEE Trans. Speech Audio Process.*, vol. 10, no. 7, pp. 504–516, Oct. 2002.

[60] R. Li, Q. Jin, and T. Schultz, “Improving speaker segmentation via speaker identification and text segmentation,” in *Proc. Interspeech*, Sep. 2009, pp. 3073–3076.

[61] 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.

[62] D. Van Leeuwen and M. Huijbregts, “The AMI speaker diarization system for NIST RT06s meeting data,” in *Machine Learning for Multimodal Interaction*. Berlin, Germany: Springer-Verlag, 2007, vol. 4299, Lecture Notes in Computer Science, pp. 371–384.

[63] A. Vandecatseye, J.-P. Martens, J. Neto, H. Meinedo, C. Garcia-Mateo, J. Dieguez, F. Mihelic, J. Zibert, J. Nouza, P. David, M. Pleva, A. Cizmar, H. Papageorgiou, and C. Alexandris, “The cost278 pan-European broadcast news database,” in *Proc. LREC*, Lisbon, Portugal, 5, 2004, vol. 4, pp. 873–876.

[64] K. Mori and S. Nakagawa, “Speaker change detection and speaker clustering using VQ distortion for broadcast news speech recognition,” in *Proc. ICASSP*, 2001, pp. 413–416.

[65] J. Ajmera and I. McCowan, “Robust speaker change detection,” *IEEE Signal Process. Lett.*, vol. 11, pp. 649–651, 2004.

[66] L. Lu and H.-J. Zhang, “Real-time unsupervised speaker change detection,” in *16th Int. Conf. Pattern Recognit.*, 2002, vol. 2, pp. 358–361.

[67] X. Anguera and J. Hernando, “Evolutive speaker segmentation using a repository system,” in *Proc. Interspeech*, 2004.

[68] X. Anguera, C. Wooters, and J. Hernando, “Speaker diarization for multi-party meetings using acoustic fusion,” in *Proc. ASRU*, Nov. 2005, pp. 426–431.

[69] A. Malegaonkar, A. Ariyaeeinia, P. Sivakumaran, and J. Fortuna, “Unsupervised speaker change detection using probabilistic pattern matching,” *IEEE Signal Process. Lett.*, vol. 13, no. 8, pp. 509–512, Aug. 2006.

[70] M.-H. Siu, G. Yu, and H. Gish, “Segregation of speakers for speech recognition and speaker identification,” in *Proc. ICASSP’91*, 1991, pp. 873–876.

[71] P. Delacourt and C. Wellekens, “DISTBIC : A speaker-based segmentation for audio data indexing,” *Speech Commun.*, pp. 111–126, 2000.

[72] S. S. Han and K. J. Narayanan, “Agglomerative hierarchical speaker clustering using incremental Gaussian mixture cluster modeling,” in *Proc. Interspeech’08*, Brisbane, Australia, 2008, pp. 20–23.

[73] R. Gangadharaiah, B. Narayanaswamy, and N. Balakrishnan, “A novel method for two speaker segmentation,” in *Proc. ICSLP*, Jeju, Korea, Sep. 2004.

[74] D. Liu and F. Kubala, "Fast speaker change detection for broadcast news transcription and indexing," in *Proc. Eurospeech'99*, Sep. 1999, pp. 1031–1034.

[75] M. A. Siegler, U. Jain, B. Raj, and R. M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA Speech Recognit. Workshop*, 1997, pp. 97–99.

[76] P. Zochová and V. Radová, "Modified DISTBIC algorithm for speaker change detection," in *Proc. 9th Eur. Conf. Speech Commun. Technol.*, Bonn, Germany, 2005, pp. 3073–3076.

[77] X. Zhu, C. Barras, L. Lamel, and J.-L. Gauvain, "Speaker diarization: From broadcast news to lectures," in *Proc. MLMI*, 2006, pp. 396–406.

[78] K. Han and S. Narayanan, "Novel inter-cluster distance measure combining GLR and ICR for improved agglomerative hierarchical speaker clustering," in *Proc. ICASSP*, Apr. 2008, pp. 4373–4376.

[79] D. Moraru, M. Ben, and G. Gravier, "Experiments on speaker tracking and segmentation in radio broadcast news," in *Proc. ICSLP*, 2005.

[80] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Improving speaker diarization," in *Proc. DARPA RT04*, 2004.

[81] H. Aronowitz, "Trainable speaker diarization," in *Proc. Interspeech*, Aug. 2007, pp. 1861–1864.

[82] H. Hung and G. Friedland, "Towards audio-visual on-line diarization of participants in group meetings," in *Proc. Workshop Multi-Camera and Multi-Modal Sensor Fusion Algorithms Applicat. –M2SFA2*, Marseille, France, 2008.

[83] G. Friedland and O. Vinyals, "Live speaker identification in conversations," in *Proc. MM'08: Proc. 16th ACM Int. Conf. Multimedia*, New York, 2008, pp. 1017–1018.

[84] G. Friedland, O. Vinyals, Y. Huang, and C. Muller, "Prosodic and other long-term features for speaker diarization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 5, pp. 985–993, Jul. 2009.

[85] J. Luque, X. Anguera, A. Temko, and J. Hernando, "Speaker diarization for conference room: The UPC RT07s evaluation system," in *Proc. Multimodal Technol. Perception of Humans: Int. Eval. Workshops CLEAR 2007 and RT 2007, Baltimore, MD, May 8–11, 2007, Revised Selected Papers*, Berlin, Heidelberg: Springer-Verlag, 2008, pp. 543–553.

[86] J. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences," in *Proc. Interspeech*, 2006.

[87] G. Lathoud and I. M. Cowan, "Location based speaker segmentation," in *Proc. ICASSP*, 2003, vol. 1, pp. 176–179.

[88] D. Ellis and J. C. Liu, "Speaker turn detection based on between-channels differences," in *Proc. ICASSP*, 2004.

[89] J. Ajmera, G. Lathoud, and L. McCowan, "Clustering and segmenting speakers and their locations in meetings," in *Proc. ICASSP*, 2004, vol. 1, pp. 605–608.

[90] J. M. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple distant microphone meetings: Mixing acoustic features and inter-channel time differences," in *Proc. Interspeech*, 2006.

[91] J. Pardo, X. Anguera, and C. Wooters, "Speaker diarization for multiple-distant-microphone meetings using several sources of information," *IEEE Trans. Comput.*, vol. 56, no. 9, pp. 1212–1224, Sep. 2007.

[92] N. W. D. Evans, C. Fredouille, and J.-F. Bonastre, "Speaker diarization using unsupervised discriminant analysis of inter-channel delay features," in *Proc. ICASSP*, Apr. 2009, pp. 4061–4064.

[93] M. Wölfel, Q. Yang, Q. Jin, and T. Schultz, "Speaker identification using warped MVDR cepstral features," in *Proc. Interspeech*, 2009.

[94] E. Shriberg, "Higher-level features in speaker recognition," in *Speaker Classification I*, C. Müller, Ed. Berlin, Heidelberg, Germany: Springer, 2007, vol. 4343, Lecture Notes in Artificial Intelligence.

[95] D. Imseng and G. Friedland, "Tuning-robust initialization methods for speaker diarization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 8, pp. 2028–2037, Nov. 2010.

[96] D. Imseng and G. Friedland, "Robust speaker diarization for short speech recordings," in *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, Dec. 2009, pp. 432–437.

[97] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findings and implications for automatic processing of multi-party conversations," in *Proc. Eurospeech'01*, Aalborg, Denmark, 2001, pp. 1359–1362.

[98] O. Çetin and E. Shriberg, "Speaker overlaps and ASR errors in meetings: Effects before, during, and after the overlap," in *Proc. ICASSP*, Toulouse, France, 2006, pp. 357–360.

[99] K. Boakye, B. Trueba-Hornero, O. Vinyals, and G. Friedland, "Overlapped speech detection for improved speaker diarization in multiparty meetings," in *Proc. ICASSP*, 2008, pp. 4353–4356.

[100] B. Trueba-Hornero, "Handling overlapped speech in speaker diarization," M.S. thesis, Univ. Politecnica de Catalunya, Barcelona, Spain, 2008.

[101] K. Boakye, "Audio segmentation for meetings speech processing," Ph.D. dissertation, Univ. of California, Berkeley, 2008.

[102] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *Proc. ASRU*, Kyoto, Japan, 2007, pp. 686–6.

[103] B. E. D. Kingsbury, N. Morgan, and S. Greenberg, "Robust speech recognition using the modulation spectrogram," *Speech Commun.*, vol. 25, no. 1-3, pp. 117–132, 1998.

[104] H. J. Nock, G. Iyengar, and C. Neti, "Speaker localization using audio-visual synchrony: An empirical study," *Lecture Notes in Comput. Sci.*, vol. 2728, pp. 565–570, 2003.

[105] C. Zhang, P. Yin, Y. Rui, R. Cutler, and P. Viola, "Boosting-based multimodal speaker detection for distributed meetings," in *Proc. IEEE Int. Workshop Multimedia Signal Process. (MMSP)*, 2006, pp. 86–91.

[106] A. Noulas and B. J. A. Krose, "On-line multi-modal speaker diarization," in *Proc. 9th Int. Conf. Multimodal Interfaces ICMI '07*, New York, 2007, pp. 350–357.

[107] Z. Ghahramani and M. I. Jordan, "Factorial hidden Markov models," *Mach. Learn.*, vol. 29, pp. 245–273, Nov. 1997.

[108] A. K. Noulas, G. Englebienne, and B. J. A. Krose, "Mutimodal speaker diarization," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2011, preprint, to be published.

[109] S. Tamura, K. Iwano, and S. Furui, "Multi-modal speech recognition using optical-flow analysis for lip images," *Real World Speech Process.*, vol. 36, no. 2–3, pp. 117–124, 2004.

[110] T. Chen and R. Rao, "Cross-modal prediction in audio-visual communication," in *Proc. ICASSP*, 1996, vol. 4, pp. 2056–2059.

[111] J. W. Fisher, T. Darrell, W. T. Freeman, and P. A. Viola, "Learning joint statistical models for audio-visual fusion and segregation," in *Proc. NIPS*, 2000, pp. 772–778.

[112] J. W. Fisher and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Trans. Multimedia*, vol. 6, no. 3, pp. 406–413, Jun. 2004.

[113] R. Rao and T. Chen, "Exploiting audio-visual correlation in coding of talking head sequences," in *Proc. Int. Picture Coding Symp.*, Mar. 1996.

[114] M. Siracusa and J. Fisher, "Dynamic dependency tests for audio-visual speaker association," in *Proc. ICASSP*, Apr. 2007, pp. 457–460.

[115] E. K. Patterson, S. Gurbuz, Z. Tufekci, and J. N. Gowdy, "CUAVE: A new audio-visual database for multimodal human–computer interface research," in *Proc. ICASSP*, 2002, pp. 2017–2020.

[116] D. McNeill, *Language and Gesture*. New York: Cambridge Univ. Press, 2000.

[117] H. Vajaria, T. Islam, S. Sarkar, R. Sankar, and R. Kasturi, "Audio segmentation and speaker localization in meeting videos," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR'06)*, 2006, vol. 2, pp. 1150–1153.

[118] H. Hung, Y. Huang, C. Yeo, and D. Gatica-Perez, "Associating audio-visual activity cues in a dominance estimation framework," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognition (CVPR) Workshop Human Communicative Behavior*, Anchorage, AK, 2008, pp. 1–6.

[119] N. Campbell and N. Suzuki, "Working with very sparse data to detect speaker and listener participation in a meetings corpus," in *Proc. Workshop Programme*, May 2006, vol. 10.

[120] G. Friedland, H. Hung, and C. Yeo, "Multimodal speaker diarization of real-world meetings using compressed-domain video features," in *Proc. ICASSP*, Apr. 2009, pp. 4069–4072.

[121] G. Friedland, C. Yeo, and H. Hung, "Visual speaker localization aided by acoustic models," in *Proc. 17th ACM Int. Conf. Multimedia MM'09:* , New York, 2009, pp. 195–202.

[122] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," in *Proc. CSL, Sel. Papers from Speaker Lang. Recognit. Workshop (Odyssey'04)*, 2006, pp. 303–330.

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

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

[125] V. Gupta, P. Kenny, P. Ouellet, G. Boulianne, and P. Dumouchel, "Combining Gaussianized/non-Gaussianized features to improve speaker diarization of telephone conversations," in *IEEE Signal Process. Lett.*, Dec. 2007, vol. 14, no. 12, pp. 1040–1043.

[126] T. S. Ferguson, "A Bayesian analysis of some nonparametric problems," *Ann. Statist.*, vol. 1, no. 2, pp. 209–230, 1973.

[127] F. Valente, "Infinite models for speaker clustering," in *Proc. Int. Conf. Spoken Lang. Process.*, 2006, iDIAP-RR 06–19.

[128] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1566–1581, 2006.

[129] E. B. Fox, E. B. Sudderth, M. I. Jordan, and A. S. Willsky, "An HDP-HMM for systems with state persistence," in *Proc. ICML*, Jul. 2008.

[130] M. Huijbregts and C. Wooters, "The blame game: Performance analysis of speaker diarization system components," in *Proc. Interspeech*, Aug. 2007, pp. 1857–60.

Xavier Anguera Miro (M'06) received the Telecommunications Engineering and European Masters in Language and Speech (M.S.) degrees from the Universitat Politècnica de Catalunya (UPC), Barcelona, Spain, in 2001 and the Ph.D. degree from UPC, with a thesis on "Robust Speaker Diarization for Meetings."

From 2001 to 2003, he was with Panasonic Speech Technology Lab, Santa Barbara, CA. From 2004 to 2006, he was a Visiting Researcher at the International Computer Science Institute (ICSI), Berkeley, CA, where he pursued research on speaker diarization for meetings, contributing to ICSI's participation in the NIST RT evaluations in 2004 (broadcast news) and 2005–2007 (meetings), obtaining state-of-the-art results. He briefly joined LIMSI, Paris, France, in 2006. He has been with Telefonica Research, Barcelona, Spain, since 2007, pursuing research in multimedia. His current research interests include speaker characterization (including diarization, recognition, etc.), language identification (including a participation in NIST LRE'07 evaluation) and several topics in multimodal multimedia analysis (e.g., video copy detection, involving the participation in NIST TRECVID 2009 and 2010 evaluations). He has authored or coauthored over 50 peer-reviewed research articles. He is the main developer of the BeamformIt toolkit, extensively used by the RT community for processing multiple microphone recordings.

Dr. Anguera Miro is a member of ISCA, ACM, and IEEE Signal Processing Society and has been involved in the organization of several ACM and IEEE conferences. He has been a reviewer for many conferences, as well as for several journals in the multimedia domain.

Simon Bozonnet (S'08) received the diploma in electrical engineering from INSA de Lyon, France, in 2008 with specialization in signal processing and the Master of Research in Images and Systems from INSA. He undertook his M.S. thesis at the Nuclear Energy Center (CEA), Bruyères-le-Châtel, France, where he worked on signal fusion and intelligent systems for source localization. He is currently pursuing the Ph.D. degree from Telecom ParisTech, Paris, France, and joined the Multimedia Communications Department as a Ph.D. candidate with LIA-EURECOM, Sophia-Antipolis, France.

As part of his studies, he spent one year at KTH (Royal Institute of Technology), Stockholm, Sweden. His research interests include multimedia indexing, and specifically speaker diarization. He participated in LIA-EURECOM recent submission to the NIST RT'09 evaluation and contributes his expertise in speaker diarization to the national "ACAV" project which aims to improve web accessibility for the visually and hearing impaired.

Nicholas Evans (M'06) received the M.Eng. and Ph.D. degrees from the University of Wales Swansea (UWS), Swansea, U.K., in 1999 and 2003, respectively.

From 2002 and 2006, he was a Lecturer at UWS and was an Honorary Lecturer until 2009. He briefly joined the Laboratoire Informatique d'Avignon (LIA), at the Université d'Avignon et des Pays de Vaucluse (UAPV), Avignon, France, in 2006 before moving to EURECOM, Sophia Antipolis, France, in 2007 where he is now an Assistant Professor. At EURECOM, he heads research in speech and audio processing and is currently active in the fields of speaker diarization, speaker recognition, multimodal biometrics, speech enhancement, and acoustic echo cancellation. His team led LIA-EURECOM's joint entry to the NIST Rich Transcription evaluations in 2009. He has authored or coauthored in excess of 50 peer-reviewed research articles and participates in several national and European projects, all involving speech processing.

Dr. Evans is a member of the IEEE Signal Processing Society, ISCA, and EURASIP and he serves as an Associate Editor of the *EURASIP Journal on Audio, Speech, and Music Processing*.

Corinne Fredouille received the Ph.D. degree from the Laboratoire Informatique d'Avignon (LIA), University of Avignon, Avignon, France, in 2000

She was appointed as an Assistant Professor at LIA 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 and acoustic-based characterization. She has participated in several national and international speaker diarization system evaluation campaigns and has published over 15 research papers in this field.

Prof. Fredouille is a member of the International Speech Communication Association (ISCA) and secretary of the French speaking communication association (AFCP), Special Interest Group (SIG) of ISCA.

Gerald Friedland (M'08) received the diplom and doctorate (*summa cum laude*) degrees in computer science from Freie Universität Berlin, Berlin, Germany, in 2002 and 2006, respectively.

He is a Senior Research Scientist at the International Computer Science Institute (ICSI), Berkeley, CA, an independent nonprofit research lab associated with the University of California (UC) at Berkeley where he, among other functions, is currently leading the speaker diarization research. Apart from speech, his interests also include image and video processing and multimodal machine learning. He is a Principal Investigator on an IARPA project on video concept detection and a Co-Principal Investigator on an NGA NURI grant on multimodal location estimation. Until 2009 he had been a site Coordinator for the EU-funded AMIDA and the Swiss-funded IM2 projects which sponsored the research on multimodal meeting analysis algorithms.

Dr. Friedland is a member of the IEEE Computer Society and the IEEE Communication Society, and he is involved in the organization of various ACM and IEEE conferences, including the IEEE International Conference on Semantic Computing (ICSC), where he served as cochair and the IEEE International Symposium on Multimedia (ISM2009), where he served as program cochair. He is also cofounder and Program Director of the IEEE International Summer School for Semantic Computing at UC Berkeley. He is the recipient of several research and industry recognitions, among them the Multimedia Entrepreneur Award by the German government and the European Academic Software Award. Most recently, he won the first prize in the ACM Multimedia Grand Challenge 2009.

Oriol Vinyals received a double degree in mathematics and telecommunication engineering from the Polytechnic University of Catalonia, Barcelona, Spain, and the M.S. degree in computer science from the University of California, San Diego, in 2009. He is currently pursuing the Ph.D. degree at the University of California, Berkeley.

His interests include artificial intelligence, with particular emphasis on machine learning, speech, and vision. He was a Visiting Scholar at the Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, in 2006, where he worked in computer vision and robotics.

Dr. Vinyals received a Microsoft Research Ph.D. Fellowship in 2011.