An Overview of Automatic Speaker Diarization Systems

Sue E. Tranter, *Member, IEEE* and Douglas A. Reynolds, *Senior Member, IEEE*

***Abstract*—Audio diarization is the process of annotating an input audio channel with information that attributes (possibly overlapping) temporal regions of signal energy to their specific sources. These sources can include particular speakers, music, background noise sources, and other signal source/channel characteristics. Diarization can be used for helping speech recognition, facilitating the searching and indexing of audio archives, and increasing the richness of automatic transcriptions, making them more readable. In this paper, we provide an overview of the approaches currently used in a key area of audio diarization, namely speaker diarization, and discuss their relative merits and limitations. Performances using the different techniques are compared within the framework of the speaker diarization task in the DARPA EARS Rich Transcription evaluations. We also look at how the techniques are being introduced into real broadcast news systems and their portability to other domains and tasks such as meetings and speaker verification.**

***Index Terms*—Speaker diarization, speaker segmentation and clustering.**

I. Introduction

THE continually decreasing cost of and increasing access to processing power, storage capacity, and network bandwidth is facilitating the amassing of large volumes of audio, including broadcasts, voice mails, meetings and other "spoken documents." There is a growing need to apply automatic human language technologies to allow efficient and effective searching, indexing, and accessing of these information sources. Extracting the words being spoken in the audio using speech recognition technology provides a sound base for these tasks, but the transcripts are often hard to read and do not capture all the information contained within the audio. Other technologies are needed to extract meta-data which can make the transcripts more readable and provide context and information beyond a simple word sequence. Speaker turns and sentence boundaries are examples of such meta-data, both of which help provide a richer transcription of the audio, making transcripts more readable and potentially helping with other tasks such as summarization, parsing, or machine translation.

Manuscript received October 11, 2005; revised April 25, 2006. This work was supported by the Defense Advanced Research Projects Agency under Grant MDA972-02-1-0013 and in part by Air Force Contract FA8721-05-C-0002. Opinions, interpretations, conclusions, and recommendations are those of the authors and are not necessarily endorsed by the U.S. Government. This paper is based on the ICASSP 2005 HLT special session paper (Philadelphia, PA). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. John Makhoul.

S. Tranter is with the Engineering Department, Cambridge University, Cambridge CB2 1PZ, U.K. (e-mail: sej28@eng.cam.ac.uk).

D. Reynolds is with the Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, MA 02420-9185 USA (e-mail: dar@ll.mit.edu).

Digital Object Identifier 10.1109/TASL.2006.878256

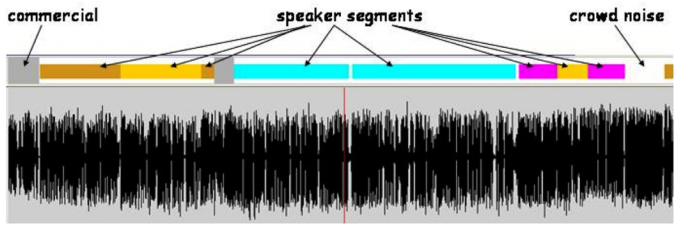


Fig. 1. Example of audio diarization on broadcast news. Annotated phenomena may include different structural regions such as commercials, different acoustic events such as music or noise, and different speakers. (Color version available online at http://ieeexplore.ieee.org.)

In general, a spoken document is a single-channel recording that consists of multiple audio sources. Audio sources may be different speakers, music segments, types of noise, etc. For example, a broadcast news program consists of speech from different speakers as well as music segments, commercials, and sounds used to segue into reports (see Fig. 1). Audio diarization is defined as the task of marking and categorising the audio sources within a spoken document. The types and details of the audio sources are application specific. At the simplest, diarization is speech versus nonspeech, where nonspeech is a general class consisting of music, silence, noise, etc., that need not be broken out by type. A more complicated diarization would further mark where speaker changes occur in the detected speech and associate segments of speech (a segment is a section of speech bounded by nonspeech or speaker change points) coming from the same speaker. This is usually referred to as speaker diarization (a.k.a. "who spoke when") or speaker segmentation and clustering and is the focus of most current research efforts in audio diarization. This paper discusses the techniques commonly used for speaker diarization, which allows searching audio by speaker, makes transcripts easier to read, and provides information which could be used within speaker adaptation in speech recognition systems. Other audio diarization tasks, such as explicitly detecting the presence of music (e.g., [2]), helping find the structure of a broadcast program (e.g., [3]), or locating commercials to eliminate unwanted audio (e.g., [4]), also have many potential benefits but fall outside the scope of this paper.

There are three primary domains which have been used for speaker diarization research and development: broadcast news audio, recorded meetings, and telephone conversations. The data from these domains differs in the quality of the recordings (bandwidth, microphones, noise), the amount and types of nonspeech sources, the number of speakers, the durations and sequencing of speaker turns, and the style/spontaneity of the speech. Each domain presents unique diarization challenges,

although often high-level system techniques tend to generalize well over several domains [5], [6]. The NIST Rich Transcription speaker evaluations [7] have primarily used both broadcast news and meeting data, whereas the NIST speaker recognition evaluations [8] have primarily used conversational telephone speech with summed sides (a.k.a two-wire).

The diarization task is also defined by the amount of specific prior knowledge allowed. There may be specific prior knowledge via example speech from the speakers in the audio, such as in a recording of a regular staff meeting. The task then becomes more like speaker detection or tracking tasks [9]. Specific prior knowledge could also be example speech from just a few of the speakers such as common anchors on particular news stations, or knowledge of the number of speakers in the audio, perhaps for a teleconference over a known number of lines, or maybe the structure of the audio recording (e.g., music followed by story). Most of this prior knowledge has been used to improve diarization performance although not all of it has proved beneficial within current systems. [10]. However, for a more portable speaker diarization system, it is desired to operate without any specific prior knowledge of the audio. This is the general task definition used in the Rich Transcription diarization evaluations, where only the broadcaster and date of broadcast are known in addition to having the audio data and we adopt this scenario when discussing speaker diarization systems.

The aim of this paper is to provide an overview of current speaker diarization approaches and to discuss performance and potential applications. In Section II, we outline the general framework of diarization systems and discuss different implementations of the key components within current systems. Performance is measured in terms of the diarization error rate (DER) using the DARPA EARS Rich Transcription Fall 2004 (RT-04F) speaker diarization evaluation data. Section IV looks at the use of these methods in real applications and the future directions for diarization research.

II. Diarization System Framework

In this section, we review the key subtasks used to build current speaker diarization systems. Most diarization systems perform these tasks separately, although it is possible to perform some of the stages jointly (for example speaker segmentation and clustering) and the ordering of the stages often varies from system to system. A prototypical combination of the key components of a diarization system is shown in Fig. 2. For each task, we provide a brief description of the common approaches employed and some of the issues in applying them.

A. Speech Detection

The aim of this step is to find the regions of speech in the audio stream. Depending on the domain data being used, nonspeech regions to be discarded can consist of many acoustic phenomena such as silence, music, room noise, background noise, or cross-talk.

The general approach used is maximum-likelihood classification with Gaussian mixture models (GMMs) trained on labeled training data, although different class models can be used, such as multistate HMMs. The simplest system uses just

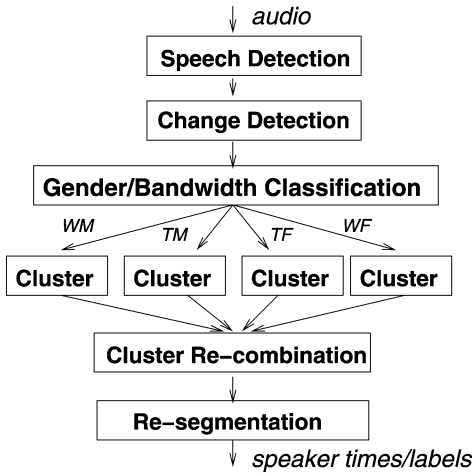


Fig. 2. Prototypical diarization system. Most diarization systems have components to perform speech detection, gender and/or bandwidth segmentation, speaker segmentation, speaker clustering, and final resegmentation or boundary refinement.

speech/nonspeech models such as in [11], while [12] is similar but four speech models are used for the possible gender/bandwidth combinations. Noise and music are explicitly modeled in [13]–[15] which have classes for speech, music, noise, speech + music, and speech + noise, while [16] and [17] use wideband speech, narrowband speech, music and speech + music. The extra speech + xx models are used to help minimize the false rejection of speech occurring in the presence of music or noise, and this data is subsequently reclassified as speech.The classes can also be broken down further, as in [18], which has eight models in total, five for nonspeech (music, laughter, breath, lip-smack, and silence) and three for speech (vowels and nasals, fricatives, and obstruents).

When operating on unsegmented audio, Viterbi segmentation, (single pass or iterative with optional adaptation) using the models is employed to identify speech regions. If an initial segmentation is already available (for example, the ordering of the key components may allow change point detection before nonspeech removal), each segment is individually classified. Minimum length constraints [11], [18] and heuristic smoothing rules [12], [15] may also be applied. An alternative approach which does not use Viterbi decoding, but instead a best model search with morphological rules is described in [19].

Silence can be removed in this early stage, using a phone recognizer (as in [17]) or energy constraint, or in a final stage processing using a word recognizer (as in [14]) or energy constraint (as in the MIT system for RT-03 [20]). Regions which contain commercials and thus are of no interest for the final output can also be automatically detected and removed at this early stage [4], [20]

For broadcast news audio, speech detection performance is typically less than 1% miss (speech in reference but not in the hypothesis) and 1%–2% false alarm (speech in the hypothesis but not in the reference), whereas for meeting audio, the figures are typically around 1% higher for both. When the speech detection phase is run early in a system, or the output is required for further processing such as for transcription, it is more important to minimize speech miss than false alarm rates, since the

former are unrecoverable errors in most systems. However, the DER, used to evaluate speaker diarization performance, treats both forms of error equally.

For telephone audio, typically some form of standard energy/spectrum-based speech activity detection is used since nonspeech tends to be silence or noise sources, although the GMM approach has also been successful in this domain with single-channel [21] or cross-channel [22] classes. For meeting audio, the nonspeech can be from a variety of noise sources, like paper shuffling, coughing, laughing, etc. and energy-based methods do not currently work well for distant microphones [23], [24], so using a simple pretrained speech/nonspeech GMM is generally preferred [6], [25], [23]. An interesting alternative uses a GMM, built on the normalized energy coefficients of the test data, to determine how much nonspeech to reject [24], while preliminary work in [6] shows potential for the future for a new energy-based method. When supported, multiple channel meeting audio can be used to help speech activity detection [26]. This problem is felt to be so important in the meetings domain that a separate evaluation for speech activity detection was introduced in the spring 2005 Rich Transcription meeting evaluation [27].

B. Change Detection

The aim of this step is to find points in the audio stream likely to be change points between audio sources. If the input to this stage is the unsegmented audio stream, then the change detection looks for both speaker and speech/nonspeech change points. If a speech detector or gender/bandwidth classifier has been run first, then the change detector looks for speaker change points within each speech segment.

Two main approaches have been used for change detection. They both involve looking at adjacent windows of data and calculating a distance metric between the two, then deciding whether the windows originate from the same or a different source. The differences between them lie in the choice of distance metric and thresholding decisions.

The first general approach used for change detection, used in [15], is a variation on the Bayesian information criterion (BIC) technique introduced in [28]. This technique searches for change points within a window using a penalized likelihood ratio test of whether the data in the window is better modeled by a single distribution (no change point) or two different distributions (change point). If a change is found, the window is reset to the change point and the search restarted. If no change point is found, the window is increased and the search is redone. Some of the issues in applying the BIC change detector are as follows. 1) It has high miss rates on detecting short turns ($<$ 2–5 s), so can be problematic to use on fast interchange speech like conversations. 2) The full search implementation is computationally expensive (order N^2), so most systems employ some form of computation reductions (e.g., [29]).

A second technique used first in [30] and later in [13], [17], and [31] uses fixed-length windows and represents each window by a Gaussian and the distance between them by the Gaussian Divergence (symmetric KL-2 distance). The step-by-step implementation in [19] and system for telephone audio in [32] are similar but use the generalized log likelihood ratio as the distance metric. The peaks in the distance function are then found and define the change points if their absolute value exceeds a predetermined threshold chosen on development data. Smoothing the distance distribution or eliminating the smaller of neighboring peaks within a certain minimum duration prevents the system overgenerating change points at true boundaries. Single Gaussians are generally preferred to GMMs due to the simplified distance calculations. Typical window sizes are 1–2 or 2–5 s when using a diagonal or full covariance Gaussian, respectively. As with BIC, the window length constrains the detection of short turns.

Since the change point detection often only provides an initial base segmentation for diarization systems, which will be clustered and often resegmented later, being able to run the change point detection very fast (typically less than $0.01\times\mathrm{RT}$ for a diagonal covariance system) is often more important than any performance degradation. In fact, [11] and [19] found no significant performance degradation when using a simple initial uniform segmentation within their systems.

Both change detection techniques require a detection threshold to be empirically tuned for changes in audio type and features. Tuning the change detector is a tradeoff between the desires to have long, pure segments to aid in initializing the clustering stage, and minimizing missed change points which produce contaminations in the clustering.

Alternatively, or in addition, a word or phone decoding step with heuristic rules may be used to help find putative speaker change points such as in [18] and the Cambridge 1998–2003 systems [16], [20]. However, this approach can over-segment the speech data and requires some additional merging or clustering to form viable speech segments, and can miss boundaries in fast speaker interchanges if relying on the presence of silence or gender changes between speakers.

C. Gender/Bandwidth Classification

The aim of this stage is to partition the segments into common groupings of gender (male or female) and bandwidth (low-bandwidth: narrow-band/telephone or high-bandwidth: studio). This is done to reduce the load on subsequent clustering, provide more flexibility in clustering settings (for example female speakers may have different optimal parameter settings to male speakers), and supply more side information about the speakers in the final output. If the partitioning can be done very accurately and assuming no speaker appears in the same broadcast in different classes (for example both in the studio and via a prerecorded field report) then performing this partitioning early on in the system can also help improve performance while reducing the computational load [33]. The potential drawback in this partitioning stage, however, is if a subset of a speaker's segments is misclassified the errors can be unrecoverable, although it is possible to allow these classifications to change in a subsequent resegmentation stage, such as in [19].

Classification for both gender and bandwidth is typically done using maximum-likelihood classification with GMMs trained on labeled training data. Either two classifiers are run (one for gender and one for bandwidth) or joint models for gender and bandwidth are used. This can be done either in

conjunction with the speech/nonspeech detection process or after the initial segmentation. Bandwidth classification can also be done using a test on the ratio of spectral energy above and below 4 kHz. An alternative method of gender classification, used in [17], aligns the word recognition output of a fast ASR system with gender dependent models and assigns the most likely gender to each segment. This has a high accuracy but is unnecessarily computationally expensive *if* a speech recognition output is not already available and segments ideally should be of a reasonable size (typically between 1 and 30 s). Gender classification error rates are around 1%–2% and bandwidth classification error rates are around 3%–5% for broadcast news audio.

D. Clustering

The purpose of this stage is to associate or cluster segments from the same speaker together. The clustering ideally produces one cluster for each speaker in the audio with all segments from a given speaker in a single cluster. The predominant approach used in diarization systems is hierarchical, agglomerative clustering with a BIC based stopping criterion [28] consisting of the following steps:

0) initialize leaf clusters of tree with speech segments;
1) compute pair-wise distances between each cluster;
2) merge closest clusters;
3) update distances of remaining clusters to new cluster;
4) iterate steps 1)–3) until stopping criterion is met.

The clusters are generally represented by a single full covariance Gaussian [5], [12], [15], [17], [31], [34], but GMMs have also been used [11], [19], [35], sometimes being built using mean-only MAP adaptation of a GMM of the entire test file to each cluster for increased robustness. The standard distance metric between clusters is the generalized likelihood ratio (GLR). It is possible to use other representations or distance metrics, but these have been found the most successful within the BIC clustering framework. The stopping criterion compares the BIC statistic from the two clusters being considered, x and y, with that of the parent cluster, z, should they be merged, the formulation being for the full covariance Gaussian case

$$\begin{aligned}\mathrm{BIC} &= \mathcal{L} - \frac{1}{2}\alpha M \log N \\ \Delta\mathrm{BIC} &= \frac{1}{2}[N_z \log(|S_z|) - N_x \log(|S_x|) \\ &\quad - N_y \log(|S_y|)] - \alpha P \\ P &= \left(\frac{d(d+3)}{4}\right)\log(N_z)\end{aligned}$$

where M is the number of free parameters, N the number of frames, S the covariance matrix, and d the dimension of the feature vector. (see, e.g., [20] for a more complete derivation.) If the pair of clusters are best described by a single full covariance Gaussian, the $\Delta\mathrm{BIC}$ will be low, whereas if there are two separate distributions, implying two speakers, the $\Delta\mathrm{BIC}$ will be high. For each step, the pair of clusters with the lowest $\Delta\mathrm{BIC}$ is merged and the statistics are recalculated. The process is generally stopped when the lowest $\Delta\mathrm{BIC}$ is greater than a specified threshold, usually 0. The use of the number of frames in the parent cluster N_z in the penalty factor P represents a "local" BIC decision, i.e., just considering the clusters being combined. This has been shown to perform better than the corresponding "global" BIC implementation which uses the number of frames in the whole show N_f instead [20], [31], [36].

Slight variations of this technique have also been used. For example, the system described in [18] uses essentially the local BIC score (with the number of parameters term incorporated within the penalty weight), but sets different thresholds for potential boundaries occurring during speech or nonspeech, motivated by an observation that most true speaker change points occurred during nonspeech regions. A further example, used in the system described in [11] and [37] removes the need for tuning the penalty weight α on development data, by ensuring that the number of parameters (M) in the merged and separate distributions are equal, although the base number of Gaussians and, hence, number of free parameters needs to be chosen carefully for optimal effect. Alternatives to the penalty term, such as using a constant [38], the weighted sum of the number of clusters and number of segments [13], or a penalized determinant of the within-cluster dispersion matrix [34], [39] have also had moderate success, but the BIC method has generally superseded these. Adding a Viterbi resegmentation between multiple iterations of clustering [31] or within a single iteration [11] has also been used to increase performance at the penalty of increased computational cost.

An alternative approach described in [40] uses a Euclidean distance between MAP-adapted GMMs and notes this is highly correlated with a Monte Carlo estimation of the Gaussian Divergence (symmetric KL-2) distance while also being an upper bound to it. The stopping criterion uses a fixed threshold, chosen on the development data, on the distance metric. The performance is comparable to the more conventional BIC method.

A further method described in [15] uses "proxy" speakers. A set of proxy models is applied to map segments into a vector space, then a Euclidean distance metric and an ad hoc occupancy stopping criterion are used, but the overall clustering framework remains the same. The proxy models can be built by adapting a universal background model (UBM) such as a 128 mixture GMM to the test data segments themselves, thus making the system portable to different shows and domains while still giving consistent performance gain over the BIC method.

Regardless of the clustering employed, the stopping criterion is critical to good performance and depends on how the output is to be used. Under-clustering fragments speaker data over several clusters, while over-clustering produces contaminated clusters containing speech from several speakers. For indexing information by speaker, both are suboptimal. However, when using cluster output to assist in speaker adaptation of speech recognition models, under-clustering may be suitable when a speaker occurs in multiple acoustic environments and over-clustering may be advantageous in aggregating speech from similar speakers or acoustic environments.

E. Joint Segmentation and Clustering

An alternative approach to running segmentation and clustering stages separately is to use an integrated scheme. This was

first done in [13] by employing a Viterbi decode between iterations of agglomerative clustering, but an initial segmentation stage was still required. A more recent completely integrated scheme, based on an evolutive-HMM (E-HMM) where detected speakers help influence both the detection of other speakers and the speaker boundaries, was introduced in [41] and developed in [19] and [42]. The recording is represented by an ergodic HMM in which each state represents a speaker and the transitions model the changes between speakers. The initial HMM contains only one state and represents all of the data. In each iteration, a short speech segment assumed to come from a nondetected speaker is selected and used to build a new speaker model by Bayesian adaptation of a UBM. A state is then added to the HMM to reflect this new speaker, and the transitions probabilities are modified accordingly. A new segmentation is then generated from a Viterbi decode of the data with the new HMM, and each model is adapted using the new segmentation. This resegmentation phase is repeated until the speaker labels no longer change. The process of adding new speakers is repeated until there is no gain in terms of comparable likelihood or there is no data left to form a new speaker. The main advantages of this integrated approach are to use all the information at each step and to allow the use of speaker recognition-based techniques, like Bayesian adaptation of the speaker models from a UBM.

F. Cluster Recombination

In this relatively recent approach [31], state-of-the-art speaker recognition modeling and matching techniques are used as a secondary stage for combining clusters. The signal processing and modeling used in the clustering stage of Section II-D are usually simple: no channel compensation, such as RASTA, since we wish to take advantage of common channel characteristics among a speaker's segments, and limited parameter distribution models, since the model needs to work with small amounts of data in the clusters at the start.

With cluster recombination, clustering is run to under-cluster the audio but still produce clusters with a reasonable amount of speech (> 30 s). A UBM is built on training data to represent general speakers. Both static and delta coefficients are used and feature normalization is applied to help reduce the effect of the acoustic environment. Maximum *a posteriori* (MAP) adaptation (usually mean-only) is then applied on each cluster from the UBM to form a single model per cluster. The cross likelihood ratio (CLR) between any two given clusters is defined [31], [43]

$$\mathrm{CLR}(c_i, c_j) = \log\left(\frac{L(x_i|\lambda_j)L(x_j|\lambda_i)}{L(x_i|\lambda_{\mathrm{ubm}})L(x_j|\lambda_{\mathrm{ubm}})}\right)$$

where $L(x_i|\lambda_j)$ is the average likelihood per frame of data x_i given the model λ_j. The pair of clusters with the highest CLR is merged and a new model is created. The process is repeated until the highest CLR is below a predefined threshold chosen from development data. Because of the computational load at this stage, each gender/bandwidth combination is usually processed separately, which also allows more appropriate UBMs to be used for each case.

Different types of feature normalization have been used with this process, namely RASTA-filtered cepstra with 10-s feature mean and variance normalization [15] and feature warping [44] using a sliding window of 3 s [14], [17]. The latter method had previously been found by one study to be more effective than other standard normalization techniques on a speaker verification task on cellular data [45]. In [17], it was found the feature normalization was necessary to get significant gain from the cluster recombination technique.

When the clusters are merged, a new speaker model can be trained with the combined data and distances updated (as in [14] and [17]) or standard clustering rules can be used with a static distance matrix (as in [15]). This recombination can be viewed as fusing intra- and inter- [43] audio file speaker clustering techniques. On the RT-04F evaluation it was found that this stage significantly improves performance, with further improvements being obtained subsequently by using a variable prior iterative MAP approach for adapting the UBMs, and building new UBMs including all of the test data [17].

G. Resegmentation

The last stage found in many diarization systems is a resegmentation of the audio via Viterbi decoding (with or without iterations) using the final cluster models and nonspeech models. The purpose of this stage is to refine the original segment boundaries and/or to fill in short segments that may have been removed for more robust processing in the clustering stage. Filtering the segment boundaries using a word or phone recognizer output can also help reduce the false alarm component of the error rate [31].

H. Finding Identities

Although current diarization systems are only evaluated using "relative" speaker labels (such as "spkr1"), it is often possible to find the true identities of the speakers (such as "Ted Koppel"). This can be achieved by a variety of methods, such as building speaker models for people who are likely to be in the news broadcasts (such as prominent politicians or main news anchors and reporters) and including these models in the speaker clustering stage or running speaker-tracking systems.

An alternative approach, introduced in [46], uses linguistic information contained within the transcriptions to predict the previous, current, or next speaker. Rules are defined based on category and word N-grams chosen from the training data, and are then applied sequentially on the test data until the speaker names have been found. Blocking rules are used to stop rules firing in certain contexts, for example, the sequence "*[name] reports* $*$" assigns the next speaker to be [name] unless $*$ is the word "that." An extension of this system described in [47], learns many rules and their associated probability of being correct automatically from the training data and then applies these simultaneously on the test data using probabilistic combination. Using automatic transcriptions and automatically found speaker turns naturally degrades performance but potentially 85% of the time can be correctly assigned to the true speaker identity using this method.

Although primarily used for identifying the speaker names given a set of speaker clusters, this technique can associate the same name for more than one input cluster and, therefore, could be thought of as a high-level cluster-combination stage.

I. Combining Different Diarization Methods

Combining methods used in different diarization systems could potentially improve performance over the best single diarization system. It has been shown that the word error rate (WER) of an automatic speech recognizer can be consistently reduced when combining multiple segmentations even if the individual segmentations themselves do not offer state-of-the-art performance in either DER or resulting WER [48]. Indeed, it seems that diversity between the segmentation methods is just as important as the segmentation quality when being combined. It is expected that gains in DER are also possible by combining different diarization modules or systems.

Several methods of combining aspects of different diarization systems have been tried, for example the "hybridization" or "piped" CLIPS/LIA systems of [35] and [49] and the "plug and play" CUED/MIT-LL system of [20] which both combine components of different systems together. A more integrated merging method is described in [49], while [35] describes a way of using the 2002 NIST speaker segmentation error metric to find regions in two inputs which agree and then uses these to train potentially more accurate speaker models. These systems generally produce performance gains, but tend to place some restriction on the systems being combined, such as the required architecture or equalizing the number of speakers. An alternative approach introduced in [50] uses a "cluster voting" technique to compare the output of arbitrary diarization systems, maintaining areas of agreement and voting using confidences or an external judging scheme in areas of conflict.

J. Sequential Speaker Clustering

For some applications, it can be important to produce speaker labels immediately without collecting all of the potential data from a particular scenario, for example real-time captioning of a broadcast news show. This constraint prevents the standard hierarchical clustering techniques being used, and instead requires the clustering to be performed sequentially or online. An elegant solution to this, described in [34], takes the segments in turn and decides if they match any of the existing speaker clusters using thresholds on distance metrics based on the generalized likelihood ratio and a penalized within-cluster dispersion. If a match is found, the statistics of the matched cluster are updated using the new segment information, whereas if no match is found, the segment starts a new speaker cluster. This process is much faster than the conventional hierarchical approach, particularly when there are a large number of initial segments, and has been used for both finding speaker turns [34] and for speaker adaptation within a real-time speech recognition framework [51].

III. Evaluation of Performance

In this section we briefly describe the NIST RT-04F speaker diarization evaluation and present the results when using the key techniques discussed in this paper on the RT-04F diarization evaluation data.

A. Speaker Diarization Error Measure

A system hypothesizes a set of speaker segments each of which consists of a (relative) speaker-id label such as "Mspkr1" or "Fspkr2" and the corresponding start and end times. This is then scored against reference "ground-truth" speaker segmentation which is generated using the rules given in [52]. Since the hypothesis speaker labels are relative, they must be matched appropriately to the true speaker names in the reference. To accomplish this, a one-to-one mapping of the reference speaker IDs to the hypothesis speaker IDs is performed so as to maximize the total overlap of the reference and (corresponding) mapped hypothesis speakers. Speaker diarization performance is then expressed in terms of the miss (speaker in reference but not in hypothesis), false alarm (speaker in hypothesis but not in reference), and speaker-error (mapped reference speaker is not the same as the hypothesized speaker) rates. The overall DER is the sum of these three components. A complete description of the evaluation measure and scoring software implementing it can be found at http://nist.gov/speech/tests/rt/rt2004/fall.

It should be noted that this measure is time-weighted, so the DER is primarily driven by (relatively few) loquacious speakers and it is, therefore, more important to get the main speakers complete and correct than to accurately find speakers who do not speak much. This scenario models some tasks, such as tracking anchor speakers in broadcast news for text summarization, but there may be other tasks (such as for speaker adaptation within automatic transcription, or ascertaining the opinions of several speakers in a quick debate) for which it is less appropriate. The same formulation can be modified to be speaker weighted instead of time weighted if necessary, but this is not discussed here. The utility of either weighting depends on the application of the diarization output.

B. Data

The RT-04F speaker diarization data consists of one 30-min extract from 12 different U.S. broadcast news shows. These were derived from TV shows: three from ABC, three from CNN, two from CNBC, two from PBS, one from CSPAN, and one from WBN. The style of show varied from a set of lectures from a few speakers (CSPAN) to rapid headline news reporting (CNN Headline News). Details of the exact composition of the data sets can be found in [52].

C. Results

The results from the main diarization techniques are shown in Fig. 3. Using a top-down clustering approach with full covariance models, arithmetic harmonic sphericity (AHS) distance metric and BIC stopping criterion gave a DER of between 20.5% and 22.5% [38]. The corresponding performance on the six-show RT diarization development data sets ranged from 15.9% to 26.9%, showing that the top-down method seems more unpredictable than the agglomerative method. This is thought to be because the initial clusters contain many speakers and segments may thus be assigned incorrectly early on, leading to an unrecoverable error. In contrast, the agglomerative scheme grows clusters from the original segments and should not contain impure multispeaker clusters until very late in the clustering process. The agglomerative clustering BIC-based scheme got around 17%–18% DER [11], [15], [17], [31], with Viterbi resegmentation between each step providing a slight benefit to 16.4% [11]. Further improvements to around 13% were made using CLR cluster recombination

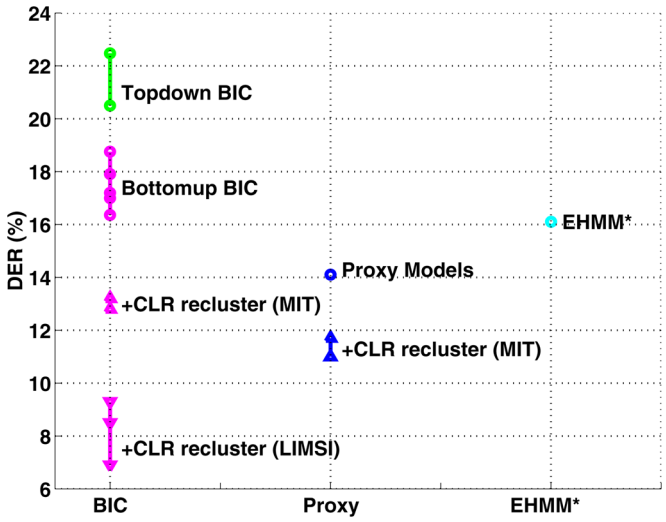


Fig. 3. DERs for different methods on the RT-04F evaluation data. Each dot represents a different version of a system built using the indicated core technique. *E-HMM not tuned on the U.S. broadcast news development sets. (Color version available online at http://ieeexplore.ieee.org.)

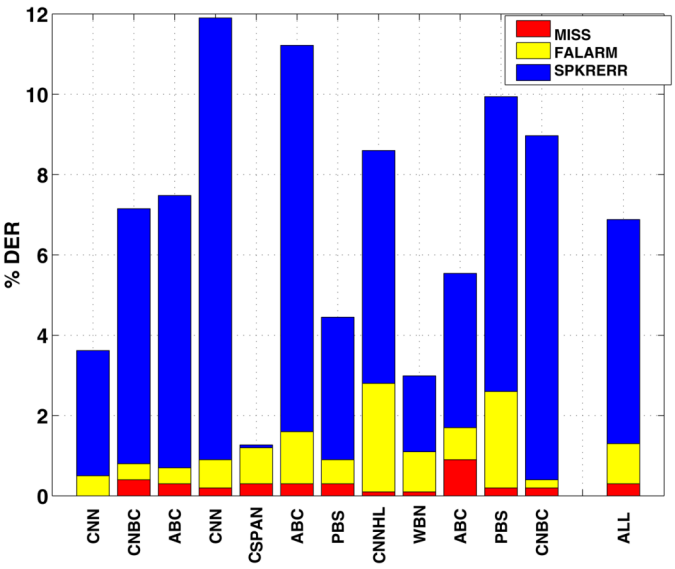


Fig. 4. DER per show for the system with lowest DER. There is a large variability between the different styles of show. (Color version available online at http://ieeexplore.ieee.org.)

and resegmentation [15]. The CLR cluster recombination stage which included feature warping produced a further reduction to around 8.5%–9.5% [14], [17], and using the whole of the RT-04F evaluation data in the UBM build of the CLR cluster recombination stage gave a final performance of 6.9% [17]. The proxy model technique performed better than the equivalent BIC stages, giving 14.1% initially and 11.0% after CLR cluster recombination and resegmentation [15]. The E-HMM system, despite not being tuned on the U.S. broadcast news development sets, gave 16.1% DER.

For the system with the lowest DER, the per-show results are given in Fig. 4. Typical of most systems, there is a large variability in performance over the shows, reflecting the variability in the number of speakers, the dominance of speakers, and the style and structure of the speech. Most of the variability is from the speaker error component due to over or under clustering.

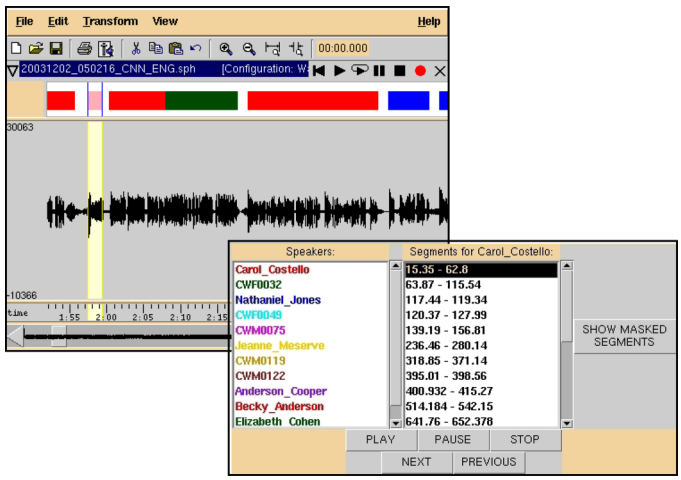


Fig. 5. Accessing broadcast news audio from automatically derived speaker names via a wavesurfer plug-in. (Color version available online at http://ieeexplore.ieee.org.)

Reducing this variability is a source of ongoing work. Certain techniques or parameter settings can perform better for different styles of show. Systems may potentially be improved by either automatically detecting the type of show and modifying the choice of techniques or parameters accordingly, or by combining different systems directly as discussed in Section II-A.

IV. Conclusion and Future Directions

There has been tremendous progress in task definition, data availability, scoring measures, and technical approaches for speaker diarization over recent years. The methods used in broadcast news diarization are now being deployed across other domains, for example, improving speaker recognition performance using multispeaker train and test data in conversational telephone speech [15] and finding speakers within meeting data [6], [23]–[25], [53]. The latter sometimes contains an additional stage when multiple microphones are present to either select the most prominent microphone [53] or to weight the audio signal from multiple microphones to form a single "superior" signal before further processing [6], [24]. Different methods for obtaining these channel weights have been tried including equal weighting [6], [24], using signal-to-noise ratio estimates [6], [24], using the correlation between different channels [6], and "delay-and-sum" beamforming [6]. However, the other components of the systems generally match those used in broadcast news with only the pretrained models and in some cases parameters being changed to reflect the new domain. Indeed, [6] is working toward the goal of complete portability between domains by trying to remove domain-specific models and parameters completely.

Other tasks, such as finding true speaker identities (see Section II-H) or speaker tracking (e.g., in [54]) are increasingly using diarization output as a starting point and performance is approaching a level where it can be considered "useful" for real human interaction.

Additions to applications which display audio and optionally transcriptions, such as wavesurfer[1] (see Fig. 5) or transcriber[2]

[1]Wavesurfer is available from www.speech.kth.se/wavesurfer

[2]Transcriber is available from http://trans.sourceforge.net/

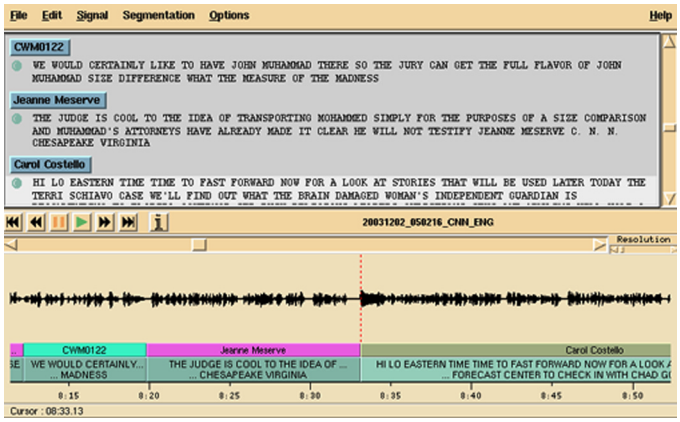


Fig. 6. Diarization information, including automatically found speaker identities, being used in the "Transcriber" tool to improve readability and facilitate searching by speaker. (Color version available online at http://ieeexplore.ieee.org.)

(see Fig. 6), and the inclusion in complete retrieval systems such as *Rough 'n' Ready* [55] and *SpeechFind* [56] allow users to see the current speaker information, understand the general flow of speakers throughout the broadcast, or search for a particular speaker within the audio. Experiments are also underway to ascertain if additional tasks, such as the process of annotating data, can be facilitated using diarization output.

The diarization tasks of the future will cover a wider scope than currently, both in terms of the amount of data (hundreds of hours) and information required (speaker identity, speaker characteristics, or potentially even emotion). Current techniques and toolkits (for example, ALIZE [57]) will provide a firm base to start from, but new methods, particularly combining information from many different approaches (as is currently done in the speaker recognition field [58]) will need to be developed to allow diarization to be maximally beneficial to real users and potential downstream processing such as machine translation and parsing. Additionally, further development of tools to allow user interactions with diarization output for specific jobs will help focus the research to contribute to high-utility human language technology.

ACKNOWLEDGMENT

The authors would like to thank C. Barras, J.-F. Bonastre, C. Fredouille, P. Nguyen, P. Torres-Carrasquillo, and C. Wooters for their help in the construction of this paper.

REFERENCES

[1] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. V, Philadelphia, PA, Mar. 2005, pp. 953–956.

[2] J. Saunders, "Real-time discrimination of broadcast speech/music," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. II, Atlanta, GA, May 1996, pp. 993–996.

[3] Z. Liu, Y. Wang, and T. Chen, "Audio feature extraction and analysis for scene segmentation and classification," *J. VLSI Signal Process. Syst.*, vol. 20, no. 1–2, pp. 61–79, Oct. 1998.

[4] S. E. Johnson and P. C. Woodland, "A method for direct audio search with applications to indexing and retrieval," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 3, Istanbul, Turkey, Jun. 2000, pp. 1427–1430.

[5] Y. Moh, P. Nguyen, and J.-C. Junqua, "Toward domain independent clustering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. II, China, Apr. 2003, pp. 85–88.

[6] X. Anguera, C. Wooters, B. Peskin, and M. Aguiló, "Robust speaker segmentation for meetings: The ICSI-SRI Spring 2005 Diarization System," in *Proc. Machine Learning for Multimodal Interaction Workshop (MLMI)*, Edinburgh, U.K., Jul. 2005, pp. 402–414.

[7] Benchmark Tests: Rich Transcription (RT). NIST. [Online]. Available: http://www.nist.gov/speech/tests/rt/

[8] Benchmark Tests: Speaker Recognition. NIST. [Online]. Available: http://www.nist.gov/speech/tests/spk/

[9] A. Martin and M. Przybocki, "Speaker recognition in a multi-speaker environment," in *Proc. Eur. Conf. Speech Commun. Technol.*, vol. 2, Aalborg, Denmark, Sep. 2001, pp. 787–790.

[10] D. Moraru, L. Besacier, and E. Castelli, "Using a-priori information for speaker diarization," in *Proc. Odyssey Speaker and Language Recognition Workshop*, Toledo, Spain, May 2004, pp. 355–362.

[11] C. Wooters, J. Fung, B. Peskin, and X. Anguera, "Toward Robust speaker segmentation: The ICSI-SRI Fall 2004 Diarization System," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, Nov. 2004, [Online]. Available: http://www.icsi.berkeley.edu/cgi-bin/pubs/publication.pl?ID=000100.

[12] P. Nguyen, L. Rigazio, Y. Moh, and J. C. Junqua. Rich transcription 2002 site report. Panasonic speech technology laboratory (PSTL). presented at *Proc. Rich Transcription Workshop (RT-02)*. [Online]. Available: http://www.nist.gov/speech/tests/rt/rt2002/presentations/rt02.pdf

[13] J.-L. Gauvain, L. Lamel, and G. Adda, "Partitioning and transcription of broadcast news data," in *Proc. Int. Conf. Spoken Lang. Process.*, vol. 4, Sydney, Australia, Dec. 1998, pp. 1335–1338.

[14] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, "Combining speaker identification and BIC for speaker diarization," in *Proc. Eur. Conf. Speech Commun. Technol.*, Lisbon, Portugal, Sep. 2005, pp. 2441–2444.

[15] D. A. Reynolds and P. Torres-Carrasquillo, "The MIT Lincoln Laboratory RT-04F diarization systems: Applications to broadcast audio and telephone conversations," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, Nov. 2004.

[16] T. Hain, S. E. Johnson, A. Tuerk, P. C. Woodland, and S. J. Young. Segment generation and clustering in the HTK broadcast news transcription system. presented at *Proc. 1998 DARPA Broadcast News Transcription and Understanding Workshop*. [Online]. Available: http://mi.eng.cam.ac.uk/reports/abstracts/hain_darpa98.html

[17] R. Sinha, S. E. Tranter, M. J. F. Gales, and P. C. Woodland, "The Cambridge University March 2005 speaker diarization system," in *Proc. Eur. Conf. Speech Commun. Technol.*, Lisbon, Portugal, Sep. 2005, pp. 2437–2440.

[18] D. Liu and F. Kubala, "Fast speaker change detection for broadcast news transcription and indexing," in *Proc. Eur. Conf. Speech Commun. Technol.*, vol. III, Budapest, Hungary, Sep. 1999, pp. 1031–1034.

[19] S. Meignier, D. Moraru, C. Fredouille, J.-F. Bonastre, and L. Besacier, "Step-by-Step and integrated approaches in broadcast news speaker diarization," Comput. Speech Lang., no. 20, pp. 303–330, Sep. 2005, to be published.

[20] S. E. Tranter and D. A. Reynolds, "Speaker diarization for broadcast news," in *Proc. Odyssey Speaker and Language Recognition Workshop*, Toledo, Spain, Jun. 2004, pp. 337–344.

[21] S. E. Tranter, K. Yu, G. Evermann, and P. C. Woodland, "Generating and evaluating segmentations for automatic speech recognition of conversational telephone speech," in *Proc. ICASSP*, vol. I, Montreal, QC, Canada, May 2004, pp. 753–756.

[22] D. Liu and F. Kubala, "A cross-channel modeling approach for automatic segmentation of conversational telephone speech," in *Proc. IEEE ASRU Workshop*, St. Thomas, U.S. Virgin Islands, Dec. 2003, pp. 333–338.

[23] D. A. van Leeuwan, "The TNO speaker diarization system for NIST RT05s meeting data," in *Proc. Machine Learning for Multimodal Interaction Workshop (MLMI)*, Edinburgh, UK, Jul. 2005, pp. 440–449.

[24] D. Istrate, C. Fredouille, S. Meignier, L. Besacier, and J.-F. Bonastre, "NIST RT'05 evaluation: Preprocessing techniques and speaker diarization on multiple microphone meetings," in *Proc. Machine Learning for Multimodal Interaction Workshop (MLMI)*, Edinburgh, U.K., Jul. 2005, pp. 428–439.

[25] S. Cassidy, "The macquarie speaker diarization system for RT05s," in *Proc. NIST Spring Rich Transcription Evaluation Workshop (RT-05s)*, Edinburgh, UK, Jul. 2005.

[26] T. Pfau, D. Ellis, and A. Stolcke, "Multispeaker speech activity detection for the ICSI meeting recorder," in *Proc. IEEE ASRU Workshop*, Trento, Italy, Dec. 2001, pp. 107–110.

[27] J. G. Fiscus, N. Radde, J. S. Garofolo, A. Le, J. Ajot, and C. Laprun, "The rich transcription 2005 spring meeting recogntion evaluation," in *Proc. Machine Learning for Multimodal Interaction Workshop (MLMI)*, Edinburgh, UK, Jul. 2005, pp. 369–389.

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

[29] B. Zhou and J. Hansen, "Unsupervised audio stream segmentation and clustering via the Bayesian information criterion," in *Proc. Int. Conf. Spoken Language Process.*, vol. 3, Beijing, China, Oct. 2000, pp. 714–717.

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

[31] C. Barras, X. Zhu, S. Meignier, and J.-L. Gauvain, "Improving speaker diarization," in *Proc. Fall Rich Transcription Workshop (RT-04)*, Palisades, NY, Nov. 2004, [Online]. Available: http://www.limsi.fr/Individu/barras/publis/rt04f_diarization.pdf.

[32] A. E. Rosenberg, A. Gorin, Z. Liu, and S. Parthasarathy, "Unsupervised speaker segmentation of telephone conversations," in *Proc. Int. Conf. Spoken Language Process.*, Denver, CO, Sep. 2002, pp. 565–568.

[33] S. Meignier, D. Moraru, C. Fredouille, L. Besacier, and J.-F. Bonastre, "Benefits of prior acoustic segmentation for automatic speaker segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. I, Montreal, QC, Canada, May 2004, pp. 397–400.

[34] D. Liu and F. Kubala, "Online speaker clustering," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. I, Hong Kong, China, Apr. 2003, pp. 572–575.

[35] D. Moraru, S. Meignier, L. Besacier, J.-F. Bonastre, and I. Magrin-Chagnolleau. The ELISA consortium approaches in speaker segmentation during the NIST 2002 speaker recognition evaluation. presented at *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*. [Online]. Available: http://www.lia.univ-avignon.fr/fich_art/339-moricassp2003.pdf

[36] M. Cettolo, "Segmentation, classification and clustering of an Italian corpus," in *Proc. Recherche d'Information Assisté par Ordinateur (RIAO)*, Paris, France, Apr. 2000, [Online]. Available: http://munst.itc.it/people/cettolo/papers/riao00a.ps.gz.

[37] J. Ajmera and C. Wooters, "A Robust Speaker Clustering Algorithm," in *Proc. IEEE ASRU Workshop*, St Thomas, U.S. Virgin Islands, Nov. 2003, pp. 411–416.

[38] S. E. Tranter, M. J. F. Gales, R. Sinha, S. Umesh, and P. C. Woodland, "The development of the Cambridge University RT-04 diarization system," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, Nov. 2004, [Online]. Available: http://mi.eng.cam.ac.uk/reports/abstracts/tranter_rt04.html.

[39] H. Jin, F. Kubala, and R. Schwartz, "Automatic speaker clustering," in *Proc. DARPA Speech Recognition Workshop*, Chantilly, VA, Feb. 1997, pp. 108–111.

[40] 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. Int. Conf. Spoken Language Processing*, Jeju Island, Korea, Oct. 2004, pp. 2329–2332.

[41] S. Meignier, J.-F. Bonastre, C. Fredouille, and T. Merlin, "Evolutive HMM for multispeaker tracking system," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. II, Istanbul, Turkey, Jun. 2000, pp. 1201–1204.

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

[43] D. Reynolds, E. Singer, B. Carlson, J. O'Leary, J. McLaughlin, and M. Zissman, "Blind clustering of speech utterances based on speaker and language characteristics," in *Proc. Int. Conf. Spoken Language Process.*, vol. 7, Sydney, Australia, Dec. 1998, pp. 3193–3196.

[44] J. Pelecanos and S. Sridharan, "Feature warping for Robust speaker verification," in *Proc. Odyssey Speaker and Language Recognition Workshop*, Crete, Greece, Jun. 2001, pp. 213–218.

[45] C. Barras and J.-L. Gauvain, "Feature and score normalization for speaker verification of cellular data," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. II, Hong Kong, China, Apr. 2003, pp. 49–52.

[46] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain, "Speaker Diarization from Speech Transcripts," in *Proc. Int. Conf. Spoken Language Process.*, Jeju Island, Korea, Oct. 2004, pp. 1272–1275.

[47] S. E. Tranter, "Who really spoke when?—Finding speaker turns and identities in audio," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. I, Toulouse, France, May 2006, pp. 1013–1016.

[48] M. J. F. Gales, D. Y. Kim, P. C. Woodland, H. Y. Chan, D. Mrva, R. Sinha, and S. E. Tranter, "Progress in the CU-HTK transcription system," *IEEE Trans. Audio, Speech, Lang, Process.*, vol. 14, no. 5, pp. 1511–1523, Sep. 2006.

[49] D. Moraru, S. Meignier, C. Fredouille, L. Besacier, and J.-F. Bonastre, "The ELISA consortium approaches in speaker segmentation during the NIST 2003 Rich Transcription evaluation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. 1, Montreal, QC, Canada, May 2004, pp. 373–376.

[50] S. E. Tranter, "Two-way cluster voting to improve speaker diarization performance," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. I, Philadelphia, PA, Mar. 2005, pp. 753–756.

[51] D. Liu, D. Kiecza, A. Srivastava, and F. Kubala, "Online speaker adaptation and tracking for real-time speech recognition," in *Proc. Eur. Conf. Speech Commun. Technol.*, Lisbon, Portugal, Sep. 2005, pp. 281–284.

[52] J. G. Fiscus, J. S. Garofolo, A. Le, A. F. Martin, D. S. Pallett, M. A. Przybocki, and G. Sanders, "Results of the fall 2004 STT and MDE evaluation," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, Palisades, NY, Nov. 2004.

[53] Q. Jin, K. Laskowski, T. Schultz, and A. Waibel, "Speaker segmentation and clustering in meetings," in *Proc. ICASSP Meeting Recognition Workshop*, Montreal, QC, Canada, May 2004, [Online]. Available: http://isl.ira.uka.de/publications/SchultzJin_NIST04.pdf.

[54] D. Istrate, N. Scheffler, C. Fredouille, and J.-F. Bonastre, "Broadcast news speaker tracking for ESTER 2005 campaign," in *Proc. Eur. Conf. Speech Commun. Technol.*, Lisbon, Portugal, Sep. 2005, pp. 2445–2448.

[55] F. Kubala, S. Colbath, D. Liu, A. Srivastava, and J. Makhoul, "Integrated technologies for indexing spoken language," *Commun. ACM*, vol. 43, no. 2, pp. 48–56, Feb. 2000.

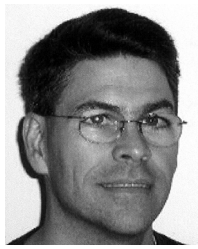
[56] J. H. L. Hansen, R. Huang, B. Z. M. Seadle, J. J. R. Deller, A. R. Gurijala, M. Kurimo, and P. Angkititrakul, "Speechfind: Advances in spoken document retrieval for a national gallery of the spoken word," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 5, pp. 712–730, Sep. 2005.

[57] J. F. Bonastre, F. Wils, and S. Meignier, "Alize: A free toolkit for speaker recogntion," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. I, Philadelphia, PA, Mar. 2005, pp. 737–740.

[58] D. Reynolds, W. Andrews, J. Campbell, J. Navratil, B. Peskin, A. Adami, Q. Jin, D. Klusacek, J. Abramson, R. Mihaescu, J. Godfrey, D. Jones, and B. Xiang, "The superSID project: Exploiting high-level information for high-accuracy speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, vol. IV, Hong Kong, China, Apr. 2003, pp. 784–787.

Sue E. Tranter (M'04) received the M.Eng. degree in engineering science, specializing in information engineering, from the University of Oxford, Oxford, U.K., in 1996 and the M.Phil. degree in computer speech and language processing from the University of Cambridge, Cambridge, U.K., in 1997.

Following this, she worked as a Research Assistant on MultiMedia Document Retrieval at the University of Cambridge until 2000, and then on nonlinear control theory at the University of Oxford. Since 2002 she has been a Research Associate on the Effective Affordable Reusable Speech-To-Text (EARS) project at the University of Cambridge, specializing in speaker segmentation and clustering.



Douglas Reynolds (SM'98) received the B.E.E. degree (with highest honors) and the Ph.D. degree in electrical engineering, both from the Georgia Institute of Technology, Atlanta.

He joined the Speech Systems Technology Group (now the Information Systems Technology Group), Lincoln Laboratory, Massachusetts Institute of Technology, Cambridge, in 1992. Currently, he is a Senior Member of Technical Staff and his research interests include robust speaker and language identification and verification, speech recognition, and general problems in signal classification and clustering.

Douglas is a Senior Member of IEEE Signal Processing Society and a cofounder and member of the steering committee of the Odyssey Speaker Recognition workshop.