Approaches and Applications of Audio Diarization•

D. A. Reynolds and P. Torres-Carrasquillo
{dar,ptorres}@ll.mit.edu

MIT Lincoln Laboratory, Lexington, MA USA

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 has utility in making automatic transcripts more readable and in searching and indexing audio archives. In this paper we provide an overview of current audio diarization approaches and discuss performance and potential applications. We outline the general framework of diarization systems and present performance of current systems as measured in the DARPA EARS Rich Transcription Fall 2004 (RT-04F) speaker diarization evaluation. Lastly, we look at future challenges and directions for diarization research.

1. INTRODUCTION

With the continually decreasing cost of and increasing access to processing power, storage capacity and network bandwidth allowing for 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. In addition to the fundamental technology of speech recognition, to extract the words being spoken, other technologies are needed to extract meta-data that provides context and information beyond the words. *Audio diarization*, or the marking and categorizing of audio sources within a spoken document, is one such technology. Audio sources may be the speakers in an audio file, so diarization would allow searching for words spoken by a speaker or aiding speaker adaptation techniques for a speech recognition system. Sources may also be non-speech events like music, where diarization could help find the structure of a broadcast program or be used by speech recognition systems to skip sections for faster processing. As illustrated in the examples, the output audio annotations from diarization may be used directly for applications or as input to assist some downstream HLT system.

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 Figure 1). The types and details of the audio sources are application specific. At the simplest, diarization is speech versus non-speech, where non-speech 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 non-speech 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. For other applications it may be desired to have more or less detail in the annotation of speech and non-speech classes (e.g., explicitly locate music, detect the narrow-band speech, label speech only by sex of the speaker, etc.).

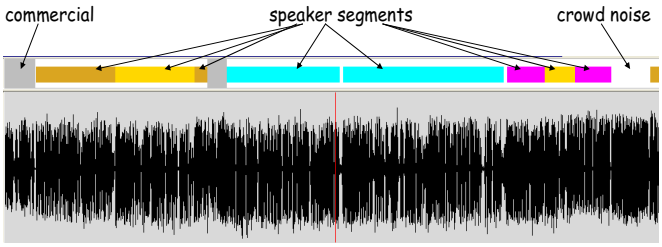

Figure 1: Broadcast news example of audio diarization.

There are three primary application domains for speaker diarization research and development: broadcast news audio, recorded meetings and telephone conversations. These domain data differ in the quality of the recordings (bandwidth, microphones, noise), the amount and types of non-speech sources, the number of speakers, and the style and structure of the speech (e.g., scripted, duration and sequencing of speaker turns). Each domain presents unique diarization challenges, although general system techniques tend to generalize well over all three. The NIST Rich Transcription evaluations have primarily used both broadcast news and meeting audio and the NIST speaker recognition evaluations have primarily used conversational telephone speech. See links at http://www.nist.gov/speech/tests for details on NIST speech evaluations.

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 [1]. Specific prior knowledge could also be example speech from just a few of the speakers, the number of speakers in the audio, or the structure of the audio recording (e.g., music followed by story). For a more portable speaker diarization system, it is desired to operate

• This work is sponsored by the Defense Advanced Research Agency under Air Force Contract F19628-00-C-0002. Opinions, interpretations, conclusions and recommendations are those of the author and are not necessarily endorsed by the United States Government

without specific prior knowledge of speakers or the number of speakers in the audio. This is the general task definition used in the Rich Transcription evaluations.

The aim of this paper is to provide an overview of current speaker diarization approaches and to discuss performance and potential applications. In the next section we outline the general framework of diarization systems. We then present performance of current systems as measured in the DARPA EARS Rich Transcription Fall 2004 (RT-04F) speaker diarization evaluation. Lastly, we look at challenges and future directions for diarization research.

2. DIARIZATION SYSTEM FRAMEWORK

In this section we review the key components found in current speaker diarization systems. A prototypical combination of these key components is shown in Figure 2. For each component, we provide a brief description of the common approaches employed and some of the issues in applying them. Readers are directed to the references for detailed technique descriptions and prior foundational research references on which they are built.

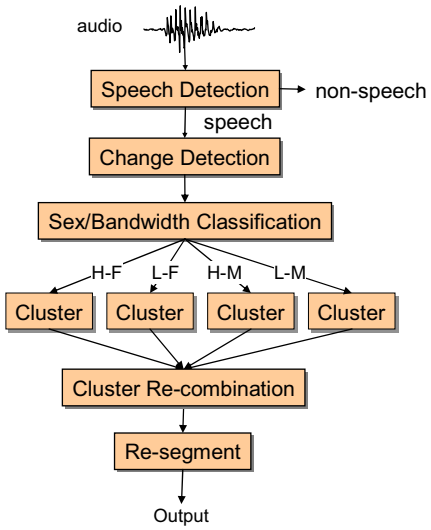

Figure 2: Example combination of speaker diarization key components.

2.1 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, the non-speech classes can consist of silence, music, room noise, street noise, etc. For broadcast news audio, the general approach used is maximum likelihood classification with Gaussian Mixture Models (GMMs) trained on labeled training data. Usually five class models are used: speech, music, noise, speech+music, and speech+noise. The extra speech models are used to help minimize false rejects of speech occurring in the presence of music or noise. Due to the high variability in realization, noise is the most difficult non-speech class to characterize and classify.

When operating on un-segmented audio, Viterbi segmentation, (single pass or iterative) using the models is employed to identify speech regions. With some initial segmentation (see next section) each segment is individually classified. A word or phone decoding step may also be used for finer grain speech boundary detection. For broadcast news audio, speech detection performance is about 1% miss and 1-2% false alarm. It is more important to minimize speech miss rates since these are unrecoverable errors in most systems.

For telephone audio, typically some form of standard energy/spectrum based speech activity detection is used since non-speech tends to be silence or noise sources. For meeting audio, the non-speech can be from a variety of noise sources, like paper shuffling, coughing, etc. When supported, multiple channel meeting audio can be used to help speech activity detection [2].

2.2 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 un-segmented audio stream, then the change detection is looking for both speaker and speech/non-speech change points. If a speech detector has been run first, then the change detector is looking for speaker change points in each speech segment.

The general approach used for change detection is some variation on the Bayesian Information Criterion (BIC) technique introduced in [3]. 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: (a) it has high miss rates on detecting short turns (< 2-5 seconds), so can be problematic to use on fast interchange speech like conversations, (b) the full search implementation is computationally expensive (order N^2), so most systems employ some form of computation reductions (e.g., [4]), and (c) the detection threshold needs 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 may be used to find putative speaker change points at pauses longer than some specified duration. This approach can over-segment the speech and may miss boundaries in fast speaker interchanges.

2.3 Sex/Bandwidth Classification

The aim of this stage is to partition the segments into common groupings of sex (male or female) and bandwidth (low-bandwidth: narrowband/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 than male speakers), and supply more side information about the speakers in the final output. The potential drawback in this partitioning stage prior to clustering is if a subset of a speaker's segments is misclassified the errors are unrecoverable.

Classification for both sex and bandwidth is typically done using maximum likelihood classification with GMMs trained on labeled training data. Either two classifiers are run (one for sex

and one for bandwidth) or joint models for sex and bandwidth are used. Bandwidth classification can also be done using a test on the ratio of spectral energy above and below 4 kHz. Sex classification error rates are around 1-2% and bandwidth classification error rates are around 3-5% for broadcast news audio

2.4 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 [3] 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.

In the BIC based clustering, the distance between clusters is a generalized likelihood ratio testing whether the pair of clusters is best described by two individual or one single full covariance Gaussian distribution. If merged, the data from both clusters are combined to estimate the single distribution. The process is stopped when the penalized minimum distance is greater than a specified threshold (typically 0).

Systems differ mainly in the selection of the distance function, how clusters are merged and the stopping criterion. For example, the system described in [5], applies a set of anchor models to map segments into a vector space, then uses Euclidean distances and an ad hoc occupancy stopping criterion. Other clustering schemes, like divisive [6] and integrated segmentation and clustering [7,8] have also been used successfully.

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.

2.5 Cluster Re-combination

In this relatively recent approach [8], state-of-the-art speaker recognition modeling and matching techniques are used as a secondary test for combining clusters. The speech processing and modeling used in the tree clustering stage 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 cluster data at the start. With cluster recombination, clustering is run to under-cluster the audio and produce clusters with a reasonable amount of speech (> 30s). Each cluster's data is then used to train an adapted GMM with channel compensated features and a cross-cluster likelihood ratio distance is computed between clusters by scoring each cluster's data against all cluster models. These distances are then used to drive an agglomerative clustering with an empirically derived stopping threshold. For each merge a new speaker model can be trained with the combined data and distances updated or standard clustering rules can be used with a static distance matrix.

This recombination can be viewed as fusing intra and inter [9] audio file speaker clustering techniques. On the RT-04F evaluation it was found that this stage significantly improves performance.

2.6 Re-segmentation

The last stage found in diarization systems is a re-segmentation of the audio via Viterbi decoding (with or without iterations) using the final cluster models and non-speech 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.

3. RT-04F EVALUATION

In this section we briefly describe the NIST RT-04F speaker diarization evaluation and present some representative system results.

3.1 Speaker Diarization Error Measure

A system hypothesizes a set of speaker segments each of which consists of a speaker-id label and the corresponding start and end times. This is then scored against reference 'ground-truth' speaker segmentation. 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 diarization error (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. Note that this measure is time-weighted, so the DER is primarily driven by loquacious speakers. The same formulation can be modified to be speaker weighted. The utility of either weighting depends on the end use (is finding all speakers important or finding the most talkative ones?).

3.2 Data

The RT-04F speaker diarization data consists of one 30 minute extract from 12 different US 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 [10].

3.3 Results

There were four participants in the RT-04F diarization evaluation: Cambridge University [6], ICSI-SRI [11], LIMSI [8],

and MIT Lincoln Laboratory [5]. Most of the systems used were built upon the basic system as outlined in the previous section. For basic system configurations, the error rates were quite similar (17-18%). Some differences in final systems consisted of applying cluster recombination (LIMSI), using recognition words to refine segment boundaries (LIMSI), using anchor model vectors for clustering (MITLL), using a threshold free cluster stopping criterion (ICSI-SRI) and using a top-down clustering scheme and a AHS distance (CUED). Important system details can be found in their respective references. The evaluation results of the participants' primary systems are shown in Figure 3.

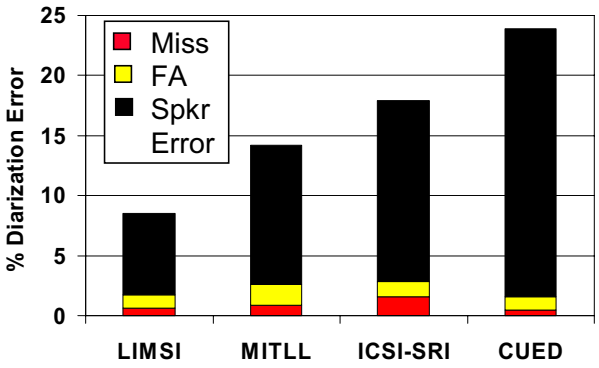

Figure 3: DERs for RT04F evaluation. Results are primary system submissions from participants.

For the MITLL primary system, the per-show results are given in Figure 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. When optimal per-show a-posteriori clustering is used total DER decreases from 14.1% to 9.8%.

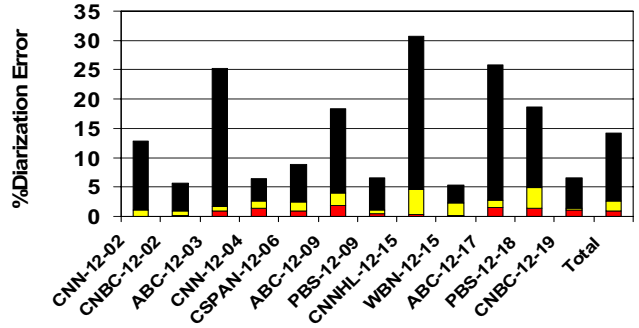

Figure 4: DER per show for the primary MITLL system.

4. FUTURE DIRECTIONS

There has been tremendous progress in task definition, data availability, scoring measures and technical approaches for audio diarization. Time-weighted error rates on broadcast news audio are less then 10%. Going forward the following challenges and directions should be addressed:

Tighter integration with speech recognition systems: Current diarization systems only use speech recognition information in a superficial way to adjust segment times. Some very interesting work being done at LIMSI [12], shows how spoken cues ("Back to you, Bob") can be exploited to improve and add more information to diarization output.

Utilizing prosodic information: Systems currently do not exploit such features as pitch trajectories and other super-segmental features to aid in segmentation or clustering.

Relating diarization error measure to utility: While the current diarization numbers are good, it is difficult to relate them to a system's utility to some downstream process. Is 10% error good enough for a browsing task in broadcast news audio? Is it good enough to help improve speaker recognition performance using "contaminated" test data from more than one speaker? This latter question is addressed in [5].

Making systems robust and portable: Currently systems are developed and evaluated with particular domain data. While many of the techniques found to work in one domain transfer to new domains, there is often considerable tuning and domain-specific parameters training. Ideally a system should work adequately over multiple domains. Even within a single domain, as seen in Figure 4, there is still considerable need for robustness over varying data sources.

Performing intra- and inter-audio clustering and indexing: An interesting future task would be to combine intra-audio speaker diarization with inter-audio clustering so one could automatically find linkages and connections of speakers or groups of speakers in audio archives.

REFERENCES

[1] A. Martin and M. Pryzbocki, "Speaker Recognition in a Multi-Speaker Environment," Eurospeech, 2001

[2] T. Pfau, D.P.W. Ellis, and A. Stolcke, "Multispeaker Speech Activity Detection for the ICSI Meeting Recorder," ASRU 2001, Madonna di Campiglio, Italy

[3] S. Chen and P. Gopalakrishnam, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," 1998 DARPA Broadcast News Workshop

[4] B. Zhou, J.H.L. Hansen, "Unsupervised Audio Stream Segmentation and Clustering via the Bayesian Information Criterion," ICSLP 2000

[5] D. A. Reynolds and P. Torres-Carrasquillo, "The MIT Lincoln Laboratory RT-04F Diarization Systems: Applications to Broadcast Audio and Telephone Conversations," RT-04F Workshop, Nov. 2004

[6] S. E. Tranter, M. J. F. Gales, R. Sinha, S. Umesh, P. C. Woodland, "The Development Of The Cambridge University RT-04 Diarisation System," RT-04F Workshop, Nov. 2004

[7] D. Moraru, L. Besacier, S. Meignier, C. Fredouille and J.-F. Bonastre, "Speaker Diarization In The Elisa Consortium Over The Last 4 Years,": RT-04F Workshop, Nov. 2004

[8] C. Barras, X. Zhu, S. Meignier and J.-L. Gauvain, "Improving Speaker Diarization," RT-04F Workshop, Nov 2004

[9] 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," ICSLP 1998

[10] J. Fiscus, et. al, "Results of the Fall 2004 STT and MDE Evaluation," RT-04F Workshop, Nov. 2004

[11] C. Wooters, J. Fung, B. Peskin, and X. Anguera, "Towards Robust Speaker Segmentation: The ICSI-SRI Fall 2004 Diarization System," RT-04F Workshop, Nov. 2004

[12] L. Canseco-Rodriguez, L. Lamel, and J-L. Gauvain, "Speaker Diarization from Speech Transcripts", ICSLP 2004, pp. 1272-1275, Oct. 2004.