Active Learning Based Constrained Clustering For Speaker Diarization

Chengzhu Yu, *Student Member, IEEE*, and John H. L. Hansen, *Fellow, IEEE*

***Abstract*—Most speaker diarization research has focused on unsupervised scenarios, where no human supervision is available. However, in many real-world applications, a certain amount of human input could be expected, especially when minimal human supervision brings significant performance improvement. In this study, we propose an active learning based bottom-up speaker clustering algorithm to effectively improve speaker diarization performance with limited human input. Specifically, the proposed active learning based speaker clustering has two different stages: *explore* and *constrained clustering*. The *explore* stage is to quickly discover at least one sample for each speaker for boosting speaker clustering process with reliable initial speaker clusters. After discovering all, or a majority, of the involved speakers during *explore* stage, the *constrained clustering* is performed. *Constrained clustering* is similar to traditional bottom-up clustering process with an important difference that the clusters created during *explore* stage are restricted from merging with each other. *Constrained clustering* continues until only the clusters generated from the *explore* stage are left. Since the objective of active learning based speaker clustering algorithm is to provide good initial speaker models, performance saturates as soon as sufficient examples are ensured for each cluster. To further improve diarization performance with increasing human input, we propose a second method which actively select speech segments that account for the largest expected speaker error from existing cluster assignments for human evaluation and reassignment. The algorithms are evaluated on our recently created Apollo Mission Control Center dataset as well as augmented multiparty interaction meeting corpus. The results indicate that the proposed active learning algorithms are able to reduce diarization error rate significantly with a relatively small amount of human supervision.**

***Index Terms*—Active learning, bottom-up clustering, speaker diarization.**

I. Introduction

SPEAKER diarization is the process of automatically detecting *who spoke when* in an audio sequence. With an increasing amount of audio resources, speaker diarization becomes an important technology in many applications such as information retrieval [1], meeting annotations [2], [3], and conversation analysis [4]. Recently, speaker diarization has largely been applied for Rich Transcription (RT), where it plays the role of providing speaker indexes and other auxiliary information for improved speech-to-text transcriptions.

As a sequential process, speaker diarization normally involves several components such as voiced activity detection (VAD), speaker change detection (segmentation), clustering, and re-segmentation [5]–[7]. Among theses components, the core part of speaker diarization is clustering, where segments originated from the same audio sources such as speaker, music, and noise, are grouped together. Due to its significance, various speaker clustering solutions have been proposed. Theses include, but not limited to, bottom-up approach [8]–[10], also known as agglomerative hierarchical clustering (AHC), top-down approach [11]–[13], and recently proposed global optimization approaches [14], [15].

Bottom-up clustering is in general the most popular strategy among various clustering solutions. It starts by treating each individual segment, obtained in the segmentation stage, as separate clusters, and iteratively merging the closest two clusters until a specified stopping criteria is satisfied. While not as popular as its counterpart, the top-down approach has also been widely applied and some studies have reported that it could achieves comparable results with bottom-up clustering [16]. Different from bottom-up based approach, top-down clustering starts from modeling the entire audio as single model and iteratively splitting the model into sub clusters until a stopping criteria is met. Despite their differences, both bottom-up and top-down based approach are iterative processes and have the drawback of error propagation. A recently proposed clustering algorithm, Integer Linear Programming (ILP) [14], [15], attempts to overcome this drawback by finding the cluster assignments that minimize the within-cluster dispersion. While the ILP based solutions could avoid the drawbacks of error propagation, it must start with initial clusters containing a sufficient numbers of samples in order to model its attributes (e.g., i-Vector). Therefore, ILP is mostly performed after bottom-up clustering.

Along with the development in speaker clustering, the distance metrics used for measuring whether two segments belong to the same class, have also made significant improvement from the original Bayesian information criteria (BIC) [17], [18], generalized log-likelihood ratio (GLR) [19], Kullback-Leibler (KL) divergence [20], to the more recent i-Vector based distances such

Manuscript received February 23, 2017; revised June 20, 2017 and August 18, 2017; accepted August 20, 2017. Date of publication August 9, 2017; date of current version September 26, 2017. This work was supported by the National Science Foundation (NSF) under Grant 1219130, in part by AFRL under contract FA8750-15-1-0205, and in part by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. H. L. Hansen. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tan Lee. *(Corresponding author: John H. L. Hansen.)*

C. Yu is with the Jonsson School of Engineering and Computer, Center for Robust Speech Systems, Richardson, TX 75252 USA (e-mail: cxy110530@utdallas.edu).

J. H. L. Hansen is with the Center for Robust Speech, University of Texas at Dallas, Richardson, TX 75083-0688 USA (e-mail: john.hansen@utdallas.edu).

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/TASLP.2017.2747097

as cosine distance score (CDS) [21] and probabilistic linear discriminant analysis (PLDA) based distance [22], [23]. Other alternative distance metrics based on information theoretic frameworks, have also been proposed and showed competitive results [9]. Recently, deep neural network (DNN) based approaches have also been applied for speaker embedding feature extraction [24], i-Vector extraction [25], as well as speaker clustering [26].

Despite the success of recent improvement on speaker clustering algorithms, distance computations, as well as other non-trivial components, speaker diarization still remains a challenging task in many real-word applications. This is especially true when the audio quality is suboptimal or the speech communication comprises large proportions of fast speaker turns and short homogeneous speech segments. For example, diarization of telephone conversations are notably more challenging compared with broadcast news diarization or many meeting room diarization.

Due to the limitations within current speaker diarization systems using exclusively audio/speech information, a number of recent studies have proposed to exploit auxiliary information for improved speaker diarization performance. For example, the linguist information such as speaker name occurring patterns, are extracted from the speech transcripts to provide additional information during speaker clustering [27]. The speech transcript could be obtained from a manual transcriptions as well as automatic speech recognition (ASR) system. Another important supplementary information that is present in many speaker diarization applications is the visual information. Audio-visual speaker diarization has also been studied [28], [29]. However, these auxiliary information are obtainable only in certain scenarios and not applicable to the broad category of speaker diarization applications.

In this study, we propose an active learning based bottom-up speaker clustering algorithm that effectively utilize human input to improve speaker diarization performance. Our proposed algorithm is based on the assumption that human input could be engaged during the speaker diarization process in certain applications. This scenario is especially plausible if small amounts of human engagement could bring significant performance improvements. Another assumption we made in this study is that the human performs better than machine when answering the question of whether a given segment pair is from the same speaker or not. This assumption is based on the results from previous studies that while current automatic speaker recognition systems showed comparable performance with human in a clean speech condition, in adverse conditions human significantly outperform a machine [30], [31]. Besides, many audio streams contain higher level information such as video, spoken names, and contextual informations that could effectively be employed by a human in performing speaker recognition.

While humans can effectively determine whether two segments belong to the same speaker or not, tagging ground truth speaker labels of an audio stream containing a large number of participants is a significantly more difficult task. This is due to limitations in how human remembers voices from unfamiliar speakers. Therefore, the comprehensive labeling of speaker index for these tasks should be achieved by answering a series of queries: where a *yes or no* type question is suggested for given pair of segments if they belong to the same speaker or not. The total number of queries for obtaining perfect clustering results require a human to evaluate $\frac{N(N-1)}{2}$ queries in a worst case scenario, where N indicates the number of speech segments [32]. Therefore, an effective active query selection strategy is necessary in order to practically employ human input to boost the speaker diarization performance.

The use of active learning has been extensively studied for image clustering and other data mining tasks. However, only a limited number of studies in the area of speaker diarization, have investigated the use of active learning. For example, the study in [32] has proposed to use active learning to obtain background speaker labels from unlabeled data for training a PLDA system. While the study in [32] bears some similarity to our current study, the ultimate goal of [32] was to locate reliable samples sufficient enough to train a PLDA system to improve speaker recognition, rather than clustering the entire dataset as in speaker diairization. Another study, that employs active learning for speaker diarization is [33]. However, the criteria of active selection for human labeling in [33] is simply based on the length of speech segments, which is not applicable in many speaker diarization scenarios where the variance of segment length is small.

To effectively employ human input for speaker diarization, we first need to identify improving which part of speaker clustering components could bring the largest improvement in overall speaker diarization performance. A recent study [34] has evaluated several key components for speaker diarization and concluded that initial speaker models from pure and reliably labeled data could lead to a significant improvement in overall speaker diarization performance. Motivated by this study, we designed our active learning algorithm to quickly discover all or a majority of the speakers in an audio stream in the *explore* phase, and initiate speaker models using reliably labeled speech segments. We also propose to perform *constrained clustering* after the *explore* stage, where initial clusters from the *explore* stages, will constrained not to be merged with each other. The proposed algorithms could also be interpreted as a way of turning unsupervised speaker clustering problems into a slightly supervised or semi-supervised close-set speaker identification task [35], with speaker model updates at each iteration. Compared to other semi-supervised speaker diarization approaches such as transfer learning which requires a separate labeled dataset [24], [26], the proposed active learning based approaches does not have the problem of domain mismatch.

In addition to use active learning for improved bottom-up speaker clustering, we also investigate the use of active learning for cluster reassignment after the completion of the clustering process. The objective of the proposed active learning based cluster reassignment, is to actively select certain speech segments with clustered labels for human evaluation and reassignment. The essence of active learning based cluster reassignment is to effectively locate the most informative segments. In this study, we select speech segments with the largest expected speaker error as candidates for human evaluation and correction.

To summarize, in this study, we investigate the use of active learning for speaker diarization. We propose two alternative strategies where active learning is employed for bottom-up speaker clustering and post-clustering reassignment, respectively. The remainder of paper is organized as follows. In Section III, we introduce previous studies on the applications of active learning for bottom-up clustering. In Section II, we presents an overview of the bottom-up speaker clustering based on an i-Vector cosine distance score (CDS), which serves as our baseline system. In Sections III-A and III-B, we describe our proposed active learning algorithms for bottom-up speaker clustering and post-clustering cluster reassignment, respectively. We present the experiments and obtained results in Section V, and finally highlight algorithm advancements and impact, and draw conclusions in Section VI.

II. Baseline System

The baseline speaker diarization system used in this study, is a bottom-up speaker clustering algorithm with an i-Vector cosine distance score (CDS) as the distance metric. In this section, we briefly discuss i-Vector extraction, CDS measure and the overall bottom-up speaker clustering process.

A. i-Vector Extraction

In an i-Vector extraction framework, speaker and channel dependent GMM supervectors are modeled as follows:

$$M = m + Tw, \tag{1}$$

where m is the supervector obtained from the universal background model (UBM), T is the low rank total variability matrix representing the basis of the reduced total variability space, and w is the weights of the low rank factor loadings referred to as i-Vectors.

The estimation of the total variability matrix T employs expectation maximization (EM) method as described in [36]. After training the total variability matrix, the i-Vector of given speech utterance is extracted as the conditional expectation of i-Vector distribution given observation features.

$$w_s^* = E[P(w_s|X_s)], \tag{2}$$

where w_s^* is the i-Vector of the given speech utterance s, X_s is the clean observation features, $P(w_s|X_s)$ is the conditional distribution of the i-Vector given observation features, and $E[\cdot]$ indicates the expectation. Finally, the i-Vector of the given speech utterance can be represented using the Baum-Welch zeroth (N_s) and centralized first (F_s) order statistics,

$$w_s^* = (T' N_s \Sigma^{-1} T + I)^{-1} T \Sigma^{-1} F_s, \tag{3}$$

where Σ is the covariance matrix obtained from UBM model and I is the identity matrix.

B. Cosine Distance Score

A comparison of i-Vectors from two different segments or clusters could be successfully achieved with a simple cosine similarity measure. The i-Vector of each cluster is computed by concatenating all segments belong to that cluster. When it comes to computing the distance between a segment and a cluster, the segment is treated as a cluster with only one segment. The cosine distance score between two i-Vectors could be expressed as follows:

$$\text{score}\,(w_i, w_j) = \frac{w_i^T \cdot w_j}{||w_i|| \cdot ||w_j||}. \tag{4}$$

Note that, the score of the cosine distance ranges between $[-1, 1]$. The greater the number is towards 1, the greater the similarity exists between the two vectors. The cosine distance score has been a popular metric for speaker recognition in i-Vector space [37].

C. Bottom-Up Speaker Clustering

Bottom-up clustering, also known as hierarchical agglomerative clustering (HAC), has been the most popular speaker clustering approach used for speaker diarization. It typically starts by treating all homogeneous speech segments as separate clusters, and iteratively merging pairs of cluster that are close using a metric. In our study, for each iteration, we find two segments that have the highest cosine similarity, and merge them into a single cluster. After each iteration, i-Vectors are extracted from the updated clusters with newly merged segments and normalized to have zero mean followed by length normalization. The i-Vector of a cluster is extracted after concatenating all segments within that cluster. We continue the iterations until the CDS of the two closest cluster reach specified minimum stopping criteria.

III. Active Learning

Active learning is a semi-supervised machine learning algorithm where the purpose is to effectively and interactively engage human input for improved labelling performance. Active learning based constrained clustering has been extensively studied for image clustering tasks [38]–[40] and in the broader area of data mining [41], [42]. A number of algorithms have been proposed to use active learning for clustering unlabeled data with "human in the loop". The key idea behind these algorithms is to actively select pairs of appropriate data for human to provide answers in a form of binary: yes or no response. Most algorithms in these studies target flat clustering approaches such as k-means clustering, and normally composed of two stage: *explore* and *consolidate* [41], [42]. The purpose of the *explore* phase is to find the centroids of unique clusters, while the aim of the consolidate stage is to locate the most informative data pairs for human labelling.

While the fundamental problem of these studies is similar to the problem we have in speaker diarization, active learning for speaker diarization is a more challenging task due to the reasons listed below. First, different from flat clustering, hierarchical agglomerative clustering (HAC) in speaker diarization requires one to iteratively update the cluster statistics at each iteration. Therefore, the decisions in the current iteration are correlated to the decisions made in previous iterations, and therefore it is difficult to quantify the importance of each query pair during

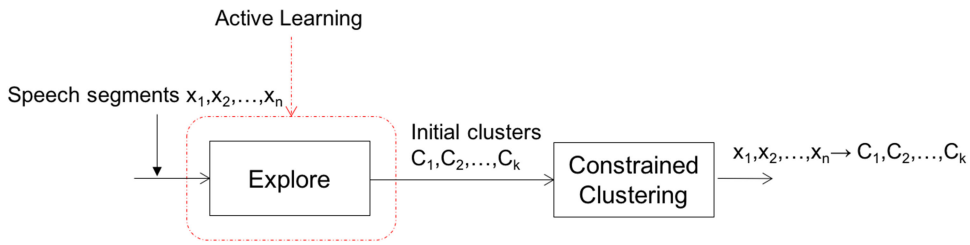


Fig. 1. Diagram of active learning based bottom-up speaker clustering. The red dotted block is active learning component where human involves.

clustering due to such dependencies. Another important difference is that, the total number of cluster centroids in speaker diarization are unknown most of time. Due to these differences, a direct replication of active learning algorithms developed in these studies is not viable for speaker diarization.

A. Active Learning Based Speaker Clustering

Due to error propagation characteristics of bottom-up speaker clustering, having a good initial cluster model has a significant impact on the final speaker diarization performance. In this section, we present a detailed description of our proposed active learning strategy for bottom-up speaker clustering. As noted in the introduction, the proposed algorithm has two clustering components: *explore* and *constrained clustering* as shown in the Fig. 1.

1) Explore: The purpose of the *explore* phase, is to quickly discover all speaker clusters contained within the audio streams, and finding at least one speech segment for each speaker cluster. To achieve this, we use the farthest first query search (FFQS) proposed by [41]. During FFQS, a speech segment is randomly selected from all speech segments to be used as a seed segment. The selected speech segment is then used to initialize the first cluster. After creating the first speaker cluster, the next segment is selected which is farthest from all existing clusters. The chosen segment is then made available for human comparison to provide expert opinion. If the chosen segment belongs to all existing clusters, the new segment is merged to corresponding cluster. Otherwise, a new cluster is created from the selected segment. Note that in order to decide whether a given speech segment belongs to a target cluster or not, we compose a query pair using the segment in question and the longest segment within the target cluster. If the answer from this query is "true", then the segment belongs to the target cluster, otherwise is separated. The FFQS process continues until the pairwise comparison operations reach the specified maximum number defined by the user. The details of the *explore* phase is detailed in Algorithm 1.

While the above algorithms can effectively identify and establish the speaker clusters in the audio streams, its performance varies a lot depending on which seed segment is selected. This problem is due to the randomness during initial seed selection. The previous studies in k-means clustering has revealed that if the initial seed data is closer to actual centroids, the procedure is more likely to achieve favorable clustering results [43]. Motivated by this, an initial unsupervised bottom-up clustering process is performed in our solution, where the centroid

Algorithm 1: FFQS with random seed during *explore* phase.

Data: Set of speech segments $X = \{x_i\}_{i=1}^{n}$, access to the answers of pairwise queries, maximum number queries Q user specified.
Result: $C_{k=1}^{k}$ initialized clusters
Start from null cluster $C = \{\}$;
Select a segment x at random, and create the first cluster as $C_1 = \{x_i\}$, $\lambda \leftarrow 1$;
while *maximum queries not reached,* $\lambda < Q$ **do**
  Find speech segment x_λ farthest from existing clusters in C, based on i-vector cosine similarity scores.;
  if x_λ *belongs to any clusters in* C **then**
    Add speech segment x_λ to matching cluster
  else
    Create new cluster C_k with speech segment x_λ;
  end
  Increase λ, after each query access.
end

segments of these clusters are used as initial seeds to start the FFQS algorithm for active learning based speaker clustering. Note that, the 'consolidation' step that normally performs after 'explore' stage in previous studies from image clustering i [41], [42] is not applied in our algorithm. This is mainly because the active learning based cluster reassignment algorithm that will be proposed in Section III-B could achieve the same thing, but more effectively as it only considers the segments that have higher chance of being incorrectly clustered.

2) Constrained Merging: After the initial *explore* phase, standard bottom-up clustering is performed with two important exceptions. First, the clusters $C_{k=1}^{K}$ created during the *explore* stage are restricted from merging with each other. Second, the stopping distance threshold during conventional bottom-up clustering is no longer necessary, as we have assumed all involved speakers are discovered during the *explore* phase. Here, bottom-up clustering will continues until only $C_{k=1}^{K}$ clusters remain.

B. Active Learning Based Cluster Reassignment

In the previous section, we proposed to use active learning during the speaker clustering process. Alternatively, human input could also be involved after clustering, to evaluate and also repair incorrectly labeled speech segments. This is quite similar to the use of active learning in automatic speech recognition

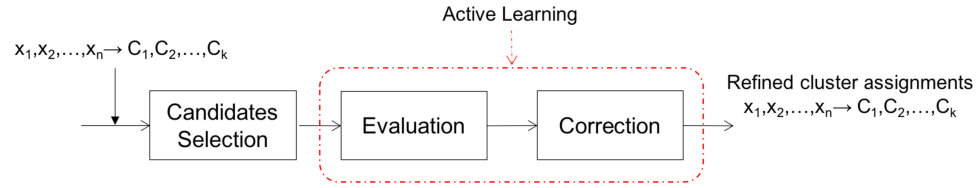


Fig. 2. Diagram of active learning based bottom-up speaker clustering. The red dotted block is active learning component where human involves.

(ASR), where the transcripts of sentences estimated with less confidence are selected for human to make corrections. However, the same algorithms used for ASR could not be directly applied in speaker diarization for several reasons. Firstly, the confidence measure used for ASR is not appropriate for speaker diarization. Moreover, the evaluation and reassignment process of potentially erroneous segments are also more diverse, and difficult in speaker diarization.

The proposed active learning based cluster reassignment procedure here has three major components as shown in Fig. 2.

1) Candidate Selection: The cluster reassignment solution starts with selecting the candidate speech segments for human experts to review. In order to effectively select the most informative segments, we rank order all speech segments in terms of expected speaker error (ESE). In other words, we select speech segments that will produce the largest expected speaker error reduction. In this study, we compute the expected speaker error for each speech segment as follows.

$$E(x_j, C_{x_j}) = P(x_j|C_{x_j}) \cdot J_{x_j \in C_{x_j}} + (1 - P(x_j|C_{x_j})) \cdot J_{x_j \notin C_{x_j}} \quad (5)$$

where x_j indicates jth speech segment, C_{x_j} is the cluster assigned to speech segment x_j, $P(x_j|C_{x_j})$ is the probability of segment x_j belonging to cluster C_{x_j}, $J_{x_j \in C_{x_j}}$ is the speaker error if x_j belongs to cluster C_{x_j}, and $J_{x_j \notin C_{x_j}}$ is the speaker error if x_j does not belong to cluster C_{x_j}. We could also write that

$$J_{x_j \in C_{x_j}} = 0$$
$$J_{x_j \notin C_{x_j}} = \frac{d_j}{\sum_{i=1}^{n} d_i} \quad (6)$$

where d_j is the length of speech segment x_j, and $\sum_{i=1}^{n} d_i$ is the total length sum of all speech segments in test audio stream.

We compute $P(x_j|C_{x_j})$ by modeling a multivariate Gaussian distribution using the i-Vectors of all the segments of a given cluster. Therefore,

$$P(x_j|C_{x_j}) = P(w_j|C_{w_j}), \quad (7)$$

where w_j is the i-Vector extracted from speech segment x_j. We normalize the probability so the sum is one.

$$\sum_{k=1}^{K} P(w_j|C_k) = 1. \quad (8)$$

2) Evaluation: After selecting candidate segments, a human expert will determine whether the given segment belongs to its assigned cluster. The process of employing human input to decide whether a segment belongs to a cluster, is more difficult than the strategy we used in *explore* phase in Section III-A. Here, we can not simply select a single longest segment as a representative of the cluster for comparison. This is due to the fact that the chosen longest segment of the cluster, could be an incorrect assignment. Therefore, we employ a majority voting based segment cluster evaluation strategy. Under this strategy, the segment in question is paired with each segment within a target cluster, resulting in multiple query pairs. If a majority of these answers for the pairs are true (two segments belong to the same cluster), we will make a decision that the given speech segment has the correct cluster assignment, and vice versa.

While the above strategy is robust for evaluating whether a given segment belongs to target cluster, it involves a significant number of query pairs for human evaluation. One heuristic used in our study is to set a maximum query number limit "V" for each segment evaluation. We rank speech segments assigned to target cluster by its confidence $P(x|c)$, and select the top V confident segments as representatives of that cluster. These selected representative segments will be paired with the test segment for a majority voting based evaluation.

3) Correction: After detecting segments with incorrect cluster assignments in the *evaluation* stage, we need to find the correct cluster designation for these segments. To accomplish this, we employ an N-best cluster evaluation. We find the N most possible cluster candidates of a given segment by ranking the i-Vector Gaussian posterior probabilities $P(x|C)$. Next, the human expert will evaluate whether the given segment belongs to any of these N clusters, using the majority voting scheme as in the *evaluation* stage.

IV. Test Data

We perform experiments on two different speech corpora: the CRSS-UTDallas Apollo Mission Control Center (MCC) audio corpus and AMI meeting corpus [44].

A. Apollo-MCC Audio Corpus

During the NASA Apollo mission, all communications between astronauts, flight controllers, and their backroom support teams inside NASA mission control center (MCC) are continuously recorded using a 30-track analog reel-to-reel recording machine. During each mission, a total of 60 audio channels are simultaneously recorded including the voices from more than hundreds of different participants of the mission. The University of Texas at Dallas (UTDallas), University of Maryland College Park (UMD), and Johnson Space Center (JSC) have combined the effort to digitize this data resource and have generated up to 19,000 hours of audio data from various missions of both Apollo and Gemini programs. The 19,000 hours are the

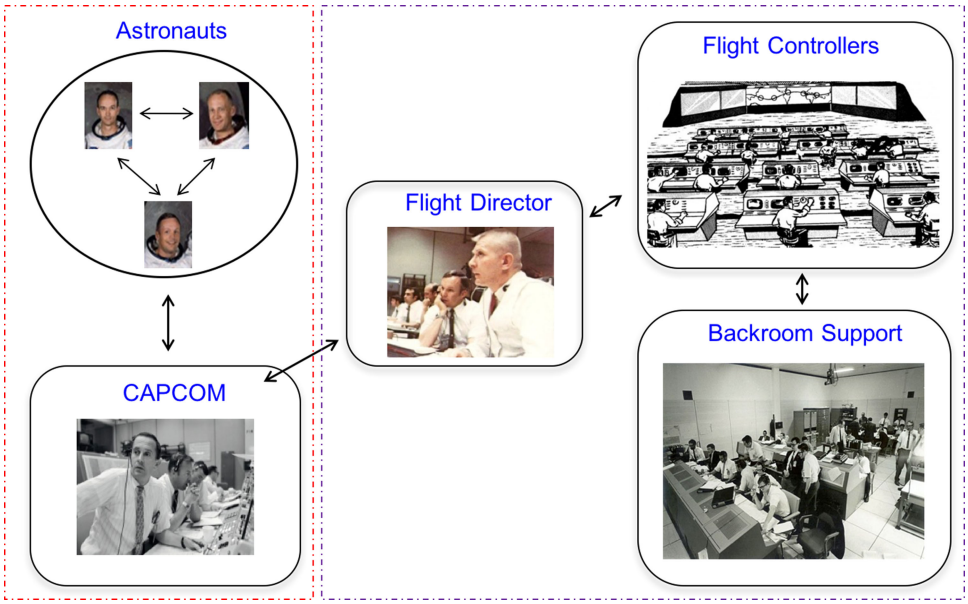


Fig. 3. Apollo Mission Control Center (MCC) communication overview. The red dotted parts are space-to-ground communications, including astronauts voice from space, and the black dotted parts are ground communications between hundreds of flight controllers and their 'backroom' support staffs.

combinations of audio from all 60 audio channels. These mission audio captures the full scope of each Apollo mission communication. Therefore, the corpus is extremely attractive for learning human-to-human communications, group interaction, as well as developing robust speech systems.

Moreover, as the speech community relies on labeled audio data to perform scientific research as well as algorithmic development, we have prepared a 'Task Specific' corpus based on a subset of Apollo-11 audio recordings. We performed our experiment here on this subset of Apollo-11 audio recordings which includes 3 synchronized channels: Flight director (FD) loop, Electrical, Environmental and Consumables Manager (EECOM) loop, and Guidance, Navigation, and Controls Systems Engineer(GNC) loop. Each of these audio recordings spanning approximately 10-hours before, and after the first lunar landing. This initial 28 hours task corpus has been transcribed to have speaker labels by well-trained speech science students from UTDallas[1].

The audios in Apollo-MCC datasets includes two types of communication: space-to-ground communications between astronauts and Capsule Communicator (CAPCOM), and ground communications between hundreds of flight controllers and "backroom" support staff, see Fig. 3. Most of the audio are recorded with close-talking microphones or traditional telephone handsets, and are in general good audio quality. Audio channels from astronauts are often transmitted through Earth's global dedicated telephone channels to Houston, TX from ground stations where the signal was received. Flight directors as well as their "backroom" support staff voice are recorded through intercom circuits called "loops". Each flight controller has their own loop, which records the entire communication within that channel. The Apollo-MCC audio corpus is separated into 28 individual audio streams, with each containing 60 minutes of audio. A summary of all information for this audio set, including the length of pure speech after removing silence, the number of homogeneous speech segments, and the total number of participants in each audio stream, are listed in Table I.

TABLE I
SYNOPSIS OF APOLLO-MCC AUDIO DATASET

Session Name	Speech (seconds)	Speech Segments	Participants
FD-01	252	161	9
FD-02	314	152	9
FD-03	123	63	10
FD-04	651	358	13
FD-05	457	226	14
FD-06	979	531	12
FD-07	394	267	13
FD-08	486	340	15
FD-09	217	126	13
FD-10	964	713	13
EECOM-01	1206	585	20
EECOM-02	563	252	20
EECOM-03	1014	471	31
EECOM-04	808	384	20
EECOM-05	812	357	26
EECOM-06	475	270	23
EECOM-07	553	337	21
EECOM-08	411	261	19
EECOM-09	744	430	31
GNC-01	859	270	17
GNC-02	735	346	21
GNC-03	653	291	25
GNC-04	1347	494	20
GNC-05	798	440	21
GNC-06	985	456	24
GNC-07	829	481	24
GNC-08	764	435	29
GNC-09	1728	995	29

[1]The task corpus will be released to the speech community for research and algorithmic development: http://crss.utdallas.edu

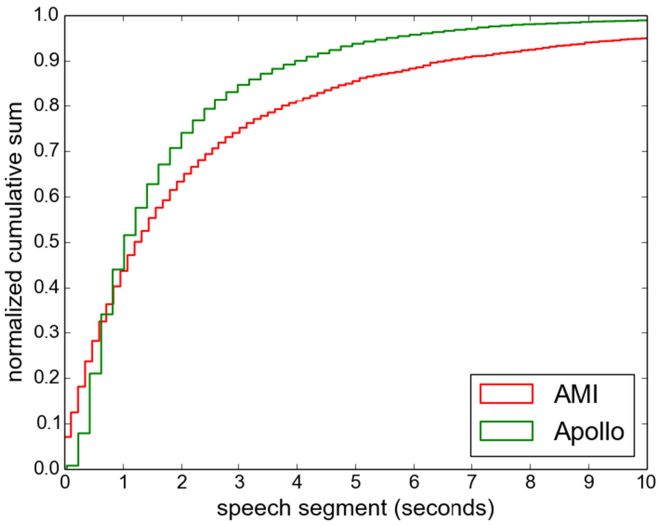


Fig. 4. Cumulative histogram as a function of speech segments length for Apollo-MCC audio corpus and AMI meeting corpus.

The voice communication style within the Apollo mission control center is quite different from traditional meeting corpora, including many focused short speech segments which was intended to improve communication efficiency for each lunar mission. Fig. 4 shows a cumulative histogram as a function of speech segments length. It can be seen that the Apollo-MCC audio dataset is composed of larger proportions of short speech segments (less than 3 sec) than the AMI meeting data. In addition, the Apollo-MCC audio dataset has relatively large number of participants (i.e., anywhere from 9 to 31 participants) as shown in Table I. Another important factor is that all data, speakers and communications are real and naturalistic, and reflect the collective effort of one of the most challenging engineering milestones for mankind. Overall, the diarization of Apollo-MCC audio dataset is clearly a realistic and challenging task.

B. AMI Meeting Dataset

We also evaluate the proposed algorithms on the popular 12-meeting subset of Augmented Multi-Party Interaction (AMI) corpus [44], [45]. This is approximately 5.4 hours of data with each session varying between 15–30 minutes. The AMI corpus contains both audio and visual data, while we only use the audio data recorded with headset microphones in our experiments. The corpus represents a natural meeting scenario. A total of three participants are involved in each of these 12-meetings, where the discussion focused on the task to design a new remote control device. The summary information this corpus is listed in Table II.

V. Experiments and Results

In this section, we perform experiments to evaluate proposed active learning based algorithms for speaker diarization. All experiments in our study use diarization error rate (DER) as evaluation metric.

TABLE II
Synopsis of 12 Meeting Subset of AMI Corpus

Session Name	Speech (seconds)	Speech Segments	Participants
IS1000a	809	309	3
IS1001a	165	102	3
IS1001b	912	315	3
IS1001c	591	212	3
IS1003b	868	342	3
IS1003d	661	441	3
IS1006b	1288	315	3
IS1006d	698	436	3
IS1008a	346	77	3
IS1008b	920	137	3
IS1008c	926	236	3
IS1008c	729	230	3

A. System Setup

1) Segmentation: Since the purpose of this study is to evaluate active learning based bottom-up clustering strategies, we use reference boundaries to define homogeneous segments. The use of such oracle segmentation information in our study is important, as we want to focus only on the bottom-up clustering strategy, and not introduce irrelevant errors caused by incorrect segmentation. Previous work has shown that the clustering step for speaker diarization could be developed independent of other modules [13], [34]. In addition, having a fixed segmentation with oracle pairwise query answers between segments, are needed for developing an active learning based solution in order to avoid expensive human labeling in the experimental stage.

2) i-Vector Extraction: The i-Vector is extracted using the Mel-Frequency Cepstral Coefficients (MFCCs). The 13 dimensional MFCC with delta and delta-delta (39-dim in total) are computed every 10 ms using a 25 ms analysis window. We use a 512 mixture universal background model (UBM) trained using the entire corpus data. The final i-Vector has 32 dimensions after factor analysis based dimension reduction.

B. Active Learning Based Speaker Clustering

In this experiment, we evaluate the performance of active learning based speaker clustering by varying the available amount of oracle query pairs. Note that the total number of queries to achieve a perfect clustering result is $\frac{N(N-1)}{2}$ in the worst case, where N is the total number of segments in the test audio. For all experiments in this study, it is assumed that the system has accessed to reference answers of selected query pairs, and therefore no errors from human experts. Future work could explore the impact of human errors on final system performance.

Here, we evaluate the proposed algorithm by varying the quantity of query pairs in proportion to the total number of segments N. For example, if a test audio stream has 1000 speech segments, an active learning of $0.1 \cdot N$ query pairs means we have access to 100 query pairs out of $1000 * (1000 - 1)/2$ total pairs. If we assume each speech segment has an average length of 2 seconds, the evaluation of 100 query pairs will require approximately 400 (100*2*2) seconds for human evaluation.

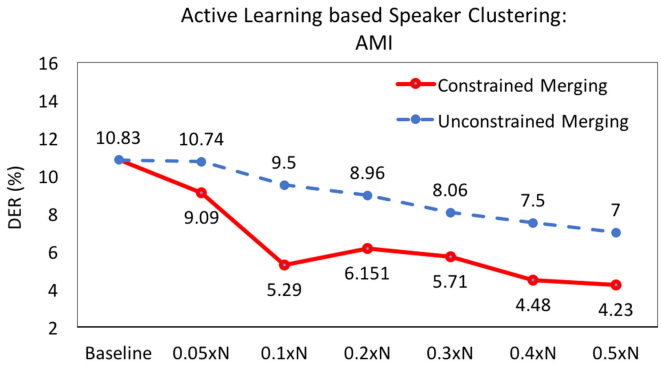


Fig. 5. Results of proposed active learning based speaker clustering algorithm on AMI meeting corpus. The soldi line is the diarization error rate (DER) with constrained clustering, while the dotted line is result obtained with unconstrained clustering as in baseline speaker clustering algorithms. Both constrained and unconstrained clustering performed after *explore* stage. The horizontal axis is the amount of query pairs proportional to total number of speech segments N.

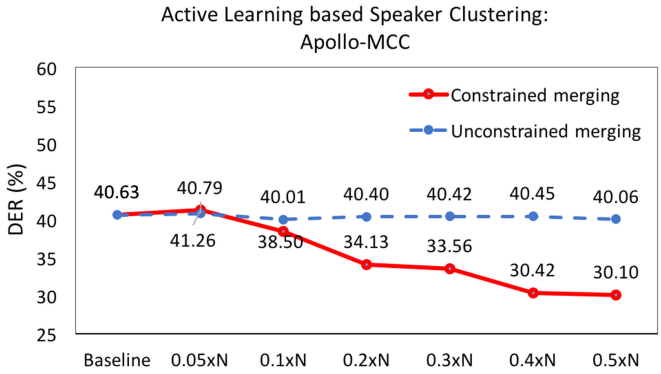


Fig. 6. Results of proposed active learning based speaker clustering algorithm on Apollo Mission Control Center (Apollo-MCC) audio dataset. The solid line is the diarization error rate (DER) with constrained clustering, while the dotted line is result obtained with unconstrained clustering as in baseline speaker clustering algorithms. Both constrained and unconstrained clustering performed after *explore* stage. The horizontal axis is the amount of query pairs proportional to total number of speech segments N.

This is a small amount of human evaluation time compared to the potential total time required for a human to obtain perfect speaker diarization: $(1000 * (1000 - 1)/2 * 2 * 2)$ sec for this example.

The solid line in Figs. 5 and 6 show the performance of the proposed active learning based speaker clustering algorithm using alternative amount of queries. The baseline result is obtained using a conventional bottom-up clustering with i-Vector cosine distance score (CDS). The first observation here is that the baseline DER in Apollo-MCC (40.63%) is much higher than that of AMI dataset (10.83%)[2]. Such difference in performance are expected due to the challenges within the Apollo-MCC dataset as noted in Section IV-A. The difficulty of speaker diarization is highly correlated with the number of speakers within the audio stream. For example, if an audio stream consisted of speech from only three speakers as in AMI meeting scenarios, we could achieve 66.6% DER by simply assigning single speaker label to all speech segments assuming every speakers speaks for the equal amount. On the other hand, if an audio stream consisted of speech from 20 speakers as in Apollo-MCC dataset, assigning single speaker label to all segments achieves 95% DER under the same assumption. Despite the difference in the baseline DER performance on AMI and Apollo-MCC dataset, the key outcome of this experiment is that the DER is measurably reduced in both dataset with a relatively small amount of query pairs. In

[2]Note that, for audio streams with a dominant speaker, a relatively low speaker DER could be achieved by simply assigning single speaker label to all speech segments without running any speaker diarization system. To better understand and compare the DER obtained in our experiments, we calculate the DER by blindly assigning a single speaker label to all speech segments in our datasets. We achieve 51.8% DER for AMI dataset and 67.8% DER for Apollo-MCC dataset.

TABLE III
MEAN ABSOLUTE PERCENTAGE DEVIATION (MAPD) OF PREDICTED SPEAKER NUMBERS WHEN USING ACTIVE LEARNING BASED SPEAKER CLUSTERING ALGORITHM WITH DIFFERENT AMOUNT OF QUERY ACCESS

	Baseline	0.2 × N	0.4 × N	0.6 × N
MAPD	38%	36%	31%	24 %

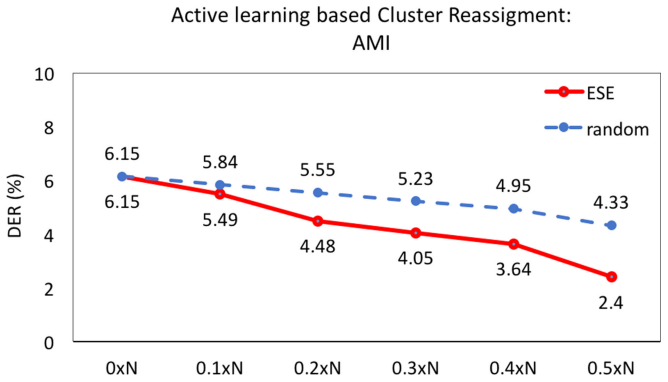


Fig. 7. Results of proposed active learning based cluster reassignment algorithm on AMI meeting corpus. The solid line is the diarization error rate (DER) using the expected speaker error (ESE) as criteria for candidates selection, while the dotted line is DER obtained by randomly selecting segments as candidates for evaluation and reassignment. The horizontal axis is the amount of candidate segments proportional to total number of speech segments N.

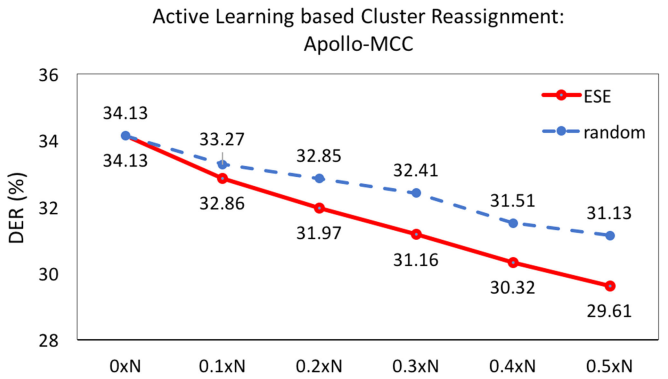


Fig. 8. Results of proposed active learning based cluster reassignment algorithm on Apollo Mission Control Center (Apollo-MCC) corpus. The solid line is the diarization error rate (DER) using the expected speaker error (ESE) as criteria for candidates selection, while the dotted line is DER obtained by randomly selecting segments as candidates for evaluation and reassignment. The horizontal axis is the amount of candidate segments proportional to total number of speech segments N.

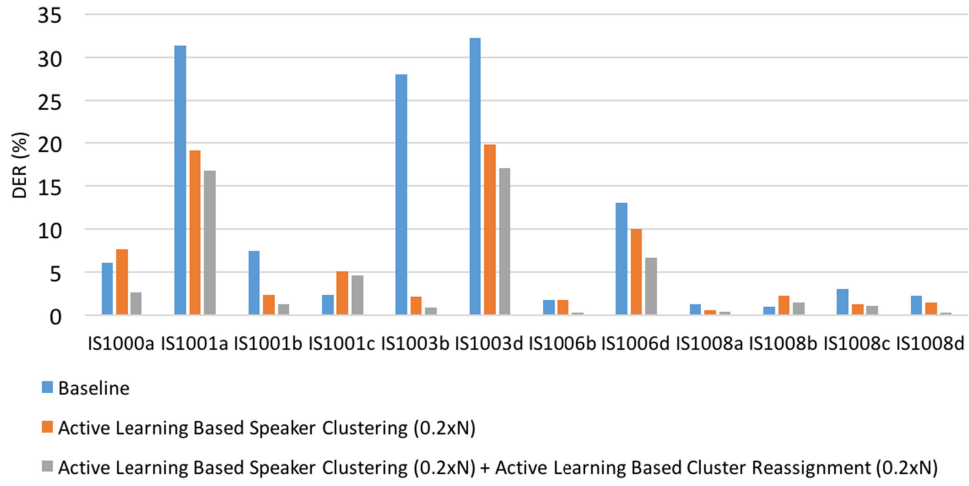


Fig. 9. The changes of DER on each session of 12-meeting subset of AMI corpus after applying proposed active learning algorithms.

the case of the AMI dataset, the DER is reduced from 10.83% to 9.09%, a relative of 16% reduction with only 0.05xN query pairs, while the DER is further reduced to 5.29% with 0.1xN queries. Fig. 9 illustrates how the DER of each AMI meeting session improves (or decreases) using the proposed algorithm. It can be seen that most (8 out of 12) sessions showed some degree of improvement, with only a small increase in DER for other sessions.

In the case of the Apollo-MCC dataset, the proposed algorithm is also capable of effectively reducing the DER, although it requires access to many more queries compared with the AMI dataset. This points to larger number of participants within Apollo-MCC dataset, which requires more human input during the *explore* stage, to discover all involved speakers. In addition, the Table III indicates the mean absolute percentage deviation (MAPD) of predicted speaker numbers reduces as more queries are being accessed when using active learning based speaker clustering algorithm. And it is consistently lower than using the baseline bottom-up clustering algorithm.

1) Importance of Constrained Clustering: We also evaluate the importance of *constrained clustering* in active learning based speaker clustering. The dotted lines in Figs. 5 and 6 indicates performance using traditional bottom-up clustering without any constraint in merging, after the *explore* stage. And solid lines in Figs. 5 and 6 show the performance with *constrained clustering*. In comparing these two results, we see that it is extremely important to perform *constrained clustering*. Only a small improvement in the AMI dataset and nearly no improvement in speaker diarization is observed if *constrained clustering* is not applied. This indicates that our proposed active learning based speaker diarization essentially transforms an unsupervised speaker diarization task into something similar to a supervised close-set speaker identification task, and therefore achieves significant improvement.

2) Limitations: While the experimental results in Figs. 5 and 6 have shown that active learning based speaker clustering algorithm reduces DER with a relatively small amount of human input, performance saturates as we use more queries. This is expected, as the objective of proposed algorithm is to discover all involved speakers, and initialize reliable speaker models for each. As soon as the majority of speakers are discovered with a relative sufficient number of queries, the benefit of using more queries significantly decreases. This poses a limitation in certain scenarios, where human input is needed to further drop a given DER value to some desired level of performance.

C. Active Learning Based Cluster Reassignment

In this experiment, we continue to explore gains in speaker diarization performance using our second active learning algorithm: active learning based cluster reassignment. For both experiments using the Apollo-MCC and AMI meeting corpora, we use at most 10 instances per cluster, to compose query pairs for majority voting based evaluation of whether a particular segment belongs to a target cluster. We also fix our n-best search to the rank of 3, during the search for correct cluster assignment. We evaluate our active learning algorithm by varying the amount of segments we will select for evaluation and reassignment. We also define this amount to be proportional to the total number of segments N. For example, if the total number of speech segments in an audio stream is $N = 1000$, an evaluation of 0.1xN segments means the human expert will review 100 speech segments, which requires access to 100*3*10 queries with correct answers if we use an n-best rank of 3, and 10 instances per cluster. Note that if some of these 100 speech segments correctly labelled initially, less queries are required.

The solid lines in Figs. 7 and 8 indicates the performance of proposed active learning based speaker clustering algorithm on AMI and Apollo-MCC datasets, respectively. We use clustering output from active learning based speaker clustering (0.2xN condition) as a base for performing cluster reassignment. We can see that the DER drops consistently as more segments are selected for reassignments. In the case of AMI dataset, the DER is reduce from 6.15% to 5.49%, a relative of 10% reduction with reviewing only 0.1xN segments, and this number continues to reduce as more segments are selected for reassignment.

We also evaluate the effectiveness of using the expected speaker error (ESE) as a criteria for selecting segment

candidates for human reassignment. We compare the largest ESE based candidate selection with baseline random segment selection scheme. The solid lines in Figs. 7 and 8 indicates performance of using ESE as the criterion, while the dotted line indicates performance with random segment selection. The results clearly show that the DER drops at a much faster rate using ESE based candidate selections, in both AMI and Apollo datasets.

Fig. 9 illustrates the DER of each AMI meeting session improved (or decreased) using proposed algorithm combined with active learning based clustering. We also notice that all sessions of AMI dataset showed different degrees of improvement. A similar trend is also observed in results from the Apollo-MCC dataset, although the relative improvement in Apollo-MCC is relatively smaller than that in AMI dataset. This is mostly because of the larger number of participants, causing the cluster reassignment process to be more difficult. Also, the majority voting based approach we used to determine whether a segment belongs to target cluster is sensitive to the initial clustering results. Overall, the DER is consistently dropping as more queries are allowed.

VI. Conclusion

In this study, we have proposed two active learning based algorithms for speaker diarization. The first algorithm employs active learning in order to obtain reliable initial speaker models and to perform constrained acoustic clustering. This essentially converts a fully unsupervised speaker clustering tasks into a semi-supervised task similar to speaker identification, where cluster models are updated after each iteration. By incorporating such information, the proposed algorithm reduces the DER significantly, with only access to a relatively small amount of queries.

Since performance of proposed active learning algorithm for speaker clustering saturates when a sufficient amount of tokens are collected for each cluster, we proposed another active learning algorithm to perform cluster reassignment after the completion of the initial speaker clustering. Active learning based cluster reassignment was shown to select the clustered segments with the largest expected speaker error for human evaluation and reassignment. Experiments on both the AMI meeting dataset and Apollo-MCC dataset indicate a clear reduction in the DER, with greater improvement as more queries are allowed. It was also shown that the expected speaker error based segment selection strategy was significantly more effective than random segment selection.

In this study, we have assumed that the human assistance provides perfect answers to any query pair on whether the two segments belong to the same speaker. However, in reality, human errors are always expected and future studies could explore how the proposed algorithms would performs or could be improved with human errors.

References

[1] M. Huijbregts, "Segmentation, diarization and speech transcription: Surprise data unraveled," Ph.D. dissertation, Dept. Electr. Eng., Mathematics Comput. Sci., Univ. Twente, Enschede, The Netherlands, 2008.

[2] 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. 2011–2022, Sep. 2007.

[3] A. Janin *et al.*, "The ICSI meeting project: Resources and research," in *Proc. 2004 Int. Conf. Acoust., Speach, Signal Process. NIST Meeting Recognit. Workshop*, 2004, pp. 201–204.

[4] O. Vinyals and G. Friedland, "Towards semantic analysis of conversations: A system for the live identification of speakers in meetings," in *Proc. 2008 IEEE Int. Conf., Semantic Comput.*, 2008, pp. 426–431.

[5] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2005, vol. 5, pp. 953–956.

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

[7] X. Anguera, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker diarization: A review of recent research," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 356–370, Feb. 2012.

[8] C. Wooters and M. Huijbregts, "The ICSI RT07s speaker diarization system," in *Multimodal Technologies for Perception of Humans*. Berlin, Germany: Springer, 2008, pp. 509–519.

[9] 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.

[10] H. Sun, B. Ma, S. Z. K. Khine, and H. Li, "Speaker diarization system for RT07 and RT09 meeting room audio," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, 2010, pp. 4982–4985.

[11] 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.*, vol. 20, no. 2, pp. 303–330, 2006.

[12] S. Bozonnet, N. W. Evans, and C. Fredouille, "The lia-eurecom RT'09 speaker diarization system: Enhancements in speaker modelling and cluster purification," in *Proc. 2010 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2010, pp. 4958–4961.

[13] S. H. Shum, N. Dehak, R. Dehak, and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 10, pp. 2015–2028, Oct. 2013.

[14] M. Rouvier and S. Meignier, "A global optimization framework for speaker diarization," in *Proc. Odyssey*, 2012, pp. 146–150.

[15] G. Dupuy, S. Meignier, P. Deléglise, and Y. Esteve, "Recent improvements on ILP-based clustering for broadcast news speaker diarization," in *Proc. Odyssey*, 2014, pp. 187–193.

[16] N. Evans, S. Bozonnet, D. Wang, C. Fredouille, and R. Troncy, "A comparative study of bottom-up and top-down approaches to speaker diarization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 2, pp. 382–392, Feb. 2012.

[17] S. Chen and P. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the Bayesian information criterion," in *Proc. DARPA Broadcast News Transcription Understanding Workshop*, 1998, vol. 8, pp. 127–132.

[18] B. Zhou and J. H. L. Hansen, "Efficient audio stream segmentation via the combined T/sup 2/statistic and Bayesian information criterion," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 4, pp. 467–474, Jul. 2005.

[19] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," in *Proc. IEEE Int. Conf., Acoust., Speech Signal Process.*, 1998, vol. 2, pp. 757–760.

[20] 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.

[21] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 788–798, May 2011.

[22] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

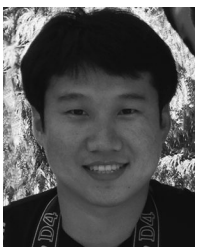
[23] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proc. Speaker Lang. Recognit. Workshop*, 2010, pp. 14–18.

[24] M. Rouvier, P. Bousquet, and B. Favre, "Speaker diarization through speaker embeddings," in *Proc. 23rd Eur. IEEE Signal Process. Conf*, 2015, pp. 2082–2086.

[25] G. Sell, D. Garcia-Romero, and A. McCree, "Speaker diarization with i-vectors from DNN senone posteriors," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3096–3099.

[26] R. Milner and T. Hain, "DNN-based speaker clustering for speaker diarisation," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 2185–2189.

[27] L. Canseco-Rodriguez, L. Lamel, and J. Gauvain, "Speaker diarization from speech transcripts," in *Proc. ICSLP*, 2004, pp. 1272–1275.
[28] G. Friedland, H. Hung, and C. Yeo, "Multi-modal speaker diarization of real-world meetings using compressed-domain video features," in *Proc. 2009 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2009, pp. 4069–4072.
[29] A. Noulas, G. Englebienne, and B. J. A. Krose, "Multimodal speaker diarization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 79–93, Jan. 2012.
[30] S. J. Wenndt and R. L. Mitchell, "Machine recognition vs. human recognition of voices," in *Proc. 2012 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2012, pp. 4245–4248.
[31] A. Schmidt-Nielsen and T. H. Crystal, "Human vs. machine speaker identification with telephone speech," in *Proc. ICSLP*, 1998, pp. 121–124.
[32] S. H. Shum, N. Dehak, and J. R. Glass, "Limited labels for unlimited data: Active learning for speaker recognition," in *Proc. Interspeech*, 2014, pp. 383–387.
[33] B. Mateusz, J. Poignant, L. Besacier, and G. Quénot, "Active selection with label propagation for minimizing human effort in speaker annotation of TV shows," in *Proc. Workshop Speech, Lang. Audio Multimedia*, 2014, pp. 5–8.
[34] M. Sinclair and S. King, "Where are the challenges in speaker diarization?" in *Proc. 2013 IEEE Int. Conf. Acoust., Speech Signal Process.*, 2013, pp. 7741–7745.
[35] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, "Combining speaker identification and BIC for speaker diarization," in *Interspeech, 2005*, vol. 5, pp. 2441–2444.
[36] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech Audio Process.*, vol. 13, no. 3, pp. 345–354, May 2005.
[37] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal Process. Mag.*, vol. 32, no. 6, pp. 74–99, Nov. 2015.
[38] Y. Guo and R. Greiner, "Optimistic active-learning using mutual information," in *Proc. 20th Int. Joint conf. Artif. Intell.*, 2007, vol. 7, pp. 823–829.
[39] S. Vijayanarasimhan, P. Jain, and K. Grauman, "Far-sighted active learning on a budget for image and video recognition," in *Proc. 2010 IEEE Conf., Comput. Vis. Pattern Recognit.*, 2010, pp. 3035–3042.
[40] A. Biswas and D. Jacobs, "Active image clustering with pairwise constraints from humans," *Int. J. Comput. Vis.*, vol. 108, no. 1-2, pp. 133–147, 2014.
[41] S. Basu, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering," in *Proc. 2004 SIAM Int. Conf. Data Mining*, 2004, vol. 4, pp. 333–344.
[42] P. K. Mallapragada, R. Jin, and A. K. Jain, "Active query selection for semi-supervised clustering," in *Proc. IEEE 19th Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.
[43] S. S. Khan and A. Ahmad, "Cluster center initialization algorithm for k-means clustering," *Pattern Recognit. Lett.*, vol. 25, no. 11, pp. 1293–1302, 2004.
[44] I. McCowan *et al.*, "The ami meeting corpus," in *Proc. 5th Int. Conf. Methods Techn. Behavioral Res.*, vol. 88, 2005, pp. 28–39.
[45] E. Gonina, G. Friedland, H. Cook, and K. Keutzer, "Fast speaker diarization using a high-level scripting language," in *Proc. 2011 IEEE Workshop, Autom. Speech Recognit. Understanding*, 2011, pp. 553–558.



Chengzhu Yu (S'13) was born in HunChun, China. He received the B.S. degree from China University of Petroleum, Beijing, China, in 2008, and the Ph.D. degree from the University of Texas at Dallas (UTDallas), TX, USA in 2017. During the Ph.D. studies, he was a Research Assistant in the Center for Robust Speech Systems (CRSS), UTDallas. He is currently a Research Scientist at Tencent AI Lab, Seattle, WA, USA. Before joining Tencent AI lab, he had internship with Nippon Telegraph and Telephone Corporation Communications Science Laboratories, speech and dialog research group in Microsoft Research, and Apple Siri Team. His research interests include speech recognition, speaker recognition, speaker diarization, as well as other interesting machine learning applications.

John H. L. Hansen (S'81–M'82–SM'93–F'07) received the Ph.D. and M.S. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, GA, USA, in 1988 and 1983, respectively, and the B.S.E.E. degree from Rutgers University, College of Engineering, New Brunswick, NJ, USA, in 1982. He received the honorary degree Doctor Technices Honoris Causa from Aalborg University, Aalborg, Denmark, in April 2016 in recognition of his contributions to speech signal processing and speech/language/hearing sciences. He joined the University of Texas at Dallas (UTDallas), Erik Jonsson School of Engineering and Computer Science, Richardson, TX, USA in 2005, where he currently serves as Jonsson School Associate Dean for Research, as well as a Professor of electrical and computer engineering, the Distinguished University Chair in Telecommunications Engineering, and a joint appointment as a Professor in the School of Behavioral and Brain Sciences (Speech & Hearing). He previously served as the Department Head of Electrical Engineering from August 2005–December 2012, overseeing a +4x increase in research expenditures (4.5 M–22.3 M) with a 20% increase in enrollment along with hiring 18 additional T/TT faculty, growing UTDallas to the 8th largest EE program from ASEE rankings in terms of degrees awarded. At UTDallas, he established the Center for Robust Speech Systems (CRSS). Previously, he served as the Department Chairman and a Professor of the Department of Speech, Language and Hearing Sciences (SLHS), and a Professor in electrical & computer engineering, University of Colorado - Boulder (1998–2005), where he co-founded and served as the Associate Director of the Center for Spoken Language Research. In 1988, he established the Robust Speech Processing Laboratory and continues to direct research activities in CRSS, UTDallas. He is author/coauthor of 661 journal and conference papers including 12 textbooks in the field of speech processing and language technology, signal processing for vehicle systems, coauthor of textbook *Discrete-Time Processing of Speech Signals*, (IEEE Press, 2000), co-editor of *DSP for In-Vehicle and Mobile Systems* (Springer, 2004), *Advances for In-Vehicle and Mobile Systems: Challenges for International Standards* (Springer, 2006), *In-Vehicle Corpus and Signal Processing for Driver Behavior* (Springer, 2008), and lead author of the report The Impact of Speech Under Stress on Military Speech Technology, (NATO RTO-TR-10, 2000). His research interests include the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech enhancement, feature estimation in noise, robust speech recognition with emphasis on spoken document retrieval, and in-vehicle interactive systems for hands-free human-computer interaction. He has been named IEEE Fellow (2007) for contributions in "Robust Speech Recognition in Stress and Noise," International Speech Communication Association (ISCA) Fellow (2010) for contributions on research for speech processing of signals under adverse conditions, and received The Acoustical Society of Americas 25 Year Award (2010) in recognition of his service, contributions, and membership to the Acoustical Society of America. He is currently serving as the ISCA President (2017–19) and a member of the ISCA Board, having previously served as the Vice-President (2015–17). He also was selected and is serving as the Vice-Chair on U.S. Office of Scientific Advisory Committees (OSAC) for OSAC-Speaker in the voice forensics domain (2015–2017). Previously, he served as the IEEE Technical Committee (TC) Chair and Member of the IEEE Signal Processing Society: Speech-Language Processing Technical Committee (SLTC) (2005–08; 2010–14; elected IEEE SLTC Chairman for 2011–13, Past-Chair for 2014), and elected ISCA Distinguished Lecturer (2011–12). He has served as member of IEEE Signal Processing Society Educational Technical Committee (2005–08; 2008–10); Technical Advisor to the U.S. Delegate for NATO (IST/TG-01); IEEE Signal Processing Society Distinguished Lecturer (2005/06), an Associate Editor of the IEEE TRANSACTIONS ON SPEECH & AUDIO PROCESSING (1992–99), an Associate Editor of the IEEE SIGNAL PROCESSING LETTERS (1998–2000), Editorial Board Member of the *IEEE Signal Processing Magazine* (2001–03); and a Guest Editor (October 1994) for special issue on *Robust Speech Recognition* of IEEE TRANSACTIONS ON SPEECH & AUDIO PROCESSING. He has served on Speech Communications Technical Committee for Acoustical Society of America (2000–03), and previously on ISCA Advisory Council. He has supervised 82 Ph.D./M.S. thesis candidates (45 Ph.D., 37 M.S./M.A.), received The 2005 University of Colorado Teacher Recognition Award as voted on by the student body.He also organized and served as the General Chair for ISCA Interspeech-2002, September 16–20, 2002, Co-Organizer and Technical Program Chair for IEEE ICASSP-2010, Dallas, TX, March 15–19, 2010, and Cochair and Organizer for IEEE SLT-2014, December 7–10, 2014 in Lake Tahoe, NV, USA.