Active Learning Based Constrained Clustering For Speaker Diarization

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

***Abstract*—Most speaker diarization researches focus on unsupervised scenarios, where no human supervision is available. However, in many real-world applications, certain amount of human input could be engaged, especially when minimal human supervision brings significant performance improvement. In this paper, we propose an active learning based bottom-up speaker clustering algorithm to effectively improve speaker diarization performance with limited human input. Specifically, proposed active learning based speaker clustering has two different stages: *explore* and *constrained clustering*. The purpose of *explore* stage is to quickly discover at least one sample for each speaker for boosting the speaker clustering process with reliable initial speaker clusters. After discovering all or majority of involved speakers during *explore* stage, the *constrained clustering* is performed. The *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. The *constrained clustering* continues until only the clusters generated from *explore* stage are left. As the objective of active learning based speaker clustering algorithm is to provide good initial speaker models, the performance saturates as soon as sufficient instances are ensured for each cluster. To further improve the diarization performance with increasing human input, we propose another active learning based algorithm, where the strategy is to actively select speech segments accounted for the largest expected speaker error (ESE) from existing cluster assignments for human evaluation and reassignment. We evaluate proposed active learning based algorithms on our recently created Apollo Mission Control Center (Apollo-MCC) dataset as well as AMI meeting corpus. The results indicate that proposed active learning algorithms could reduce diarization error rate (DER) significantly with relatively small amount of human supervison.**

***Index Terms*—Speaker diarization, active learning, bottom-up clustering**

I. Introduction

Speaker diarization is the process of automatically detecting *who spoke when* in an audio sequence. With 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, the improvements of speaker diarization algorithms have been largely driven by Rich Transcription (RT), where speaker diarization plays a role of providing speaker index and other auxilary information for improved speech-to-text transcriptions.

As a sequential process, speaker diarization normally involves several components such as voice activity detection, 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 segments, 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, top-down approach has also been widely applied and some studies have reported that it could achieve comparable result 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 process and has the drawback of error propagation. A recently proposed clustering algorithm, Interger Linear Programming (ILP) [14], [15], attempts to overcome this drawback by finding the cluster assignments that minimizing within-cluster dispersion. While the ILP could avoid the drawbacks of error propagation, it has to start with initial clusters containing sufficient samples for 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 same class, has also made significant improvement from original Bayesian informative criteria (BIC) [17], [18], generalized log-likelihood ration (GLR) [19], Kullback-Leibler (KL) divergence [20], to recent i-vector based distances such as cosine distance score (CDS) [21] as well as probablistic linear discriminant analysis (PLDA) based distance [22], [23]. Other alternative distance metrics, such as these based on information theoretic framework are also proposed and showed competitive results [9].

Despite the success of recent improvement on speaker clustering algorithms, distance computations, as well as other non-trivial components, speaker diarization still remains as a challenging tasks in many real-word applications, especially when the audio quality is suboptimal or the speech communications 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 auxilary informations 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 [24]. The speech transcript could be obtained from manual transcript as well as automatic speech recognition system. Another important supplementary information that presents in many speaker diarization applications is the visual information. The audio-visual speaker diarization has also been studied [25], [26]. However, these auxilary information are obtainable only in certain scenarios and not applicable to 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 speaker diarization process in certain applications. This scenario is especially plausible if small amount of human engagement could brings significant performance improvements. Another assumption we made in this paper is that, 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 as human in clean speech condition, in adverse conditions human significantly outperform machine [27], [28]. Besides, many audio streams contain higher level information such as video, spoken names, and contextual informations that could effectively be employed by human for performing speaker recognition.

While human can effectively determine whether two segments belong to the same speaker or not, tagging the ground truth speaker labels of an audio stream containing large number of participants is significantly more difficult task. This is due to the limitations of human in remembering voices from unfamiliar speakers. Therefore, the comprehensive labeling of speaker index for these tasks need to be achieved by answering a series of queries: a *yes or no* type of questions asking whether a given pair of segments belong to the same speaker or not. The total number of queries for obtaining perfect clustering results requires human to evaluate $\frac{N(N-1)}{2}$ queries in worst cases, where N indicates the number of speech segments [29]. Therefore, an effective active query selection strategy is necessary in order to practically employ human input to boosting the speaker diarization performance.

To effectively employ human input, we first need to identify improving which part of speaker clustering components could bring largest improvement on final speaker diarization performance. A recent study in [30] has evaluated several key components of speaker diarization and concluded that initial speaker models from pure and reliably labeled data could lead to significant improvement to the overall speaker diarization performance. Motivated by the above study, we designed our active learning algorithm to quickly discover all or majority of the speakers in audio stream in the *explore* phase, and initiate speaker clustering using reliable initial speaker models. We also propose to perform *constrained clustering* after *explore* stage, where initial clusters from *explore* stages, will not be merged together. The proposed algorithm could also be understood as a way of turning unsupervised speaker clustering problems into a slightly supervised close-set speaker identification tasks [31], with speaker models updates at each iteration .

In addition to using active learning for improved speaker clustering process, we also investigate the use of active learning for cluster reassignment after the completion of clustering process. The objective of 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 most informative segments. In this study, we select speech segments with largest expected speaker error as candidates for human evaluation and correction.

To summary, in this study, we investigate the use of active learning for speaker diarization. We propose two different strategies where active learning are employed for bottom-up speaker clustering and post-clustering reassignment, respectively. The remainder of paper is organized as follows. In Sec. II, we will introduce previous studies on the applications of active learning for bottom-up clustering. In Sec. III, we will give brief overview of the bottom-up speaker clustering based on i-vector cosine distance score (CDS), which serve as our baseline system. In Sec. IV and Sec. V, we describe our proposed active learning algorithms for bottom-up speaker clustering and post-clustering cluster reassignment, respectively. We put the experiments and obtained results in Sec. VII, and finally made conclusion in Sec. VIII.

II. Related Work

Active learning based constrained clustering has been extensively studied for image clustering tasks [32]–[34] and in broader area of data mining [35], [36]. 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 data for human to provide answers in a form of yes or no. Most algorithms in these studies, are targeting flat clustering algorithms such as k-means clustering, and normally composed of two stage: *explore* and *consolidate* [35], [36]. The purpose of *explore* phase is to find the centroids of different clusters, and the aim of consolidate stage is to locate most informative data pair for human labeling. While the fundamental problem of these studies is much similar to the problem we have in speaker diarization, active learning for speaker diarization is more challenging task due to the reasons below. Firstly, different from flat clustering, hierarchical agglomerative clustering (HAC) in speaker diarization requires to iteratively update the cluster statistics at each iteration. Therefore, the decisions in current iteration are correlated to the decisions made in previous iterations, and it is difficult to quantify the importance

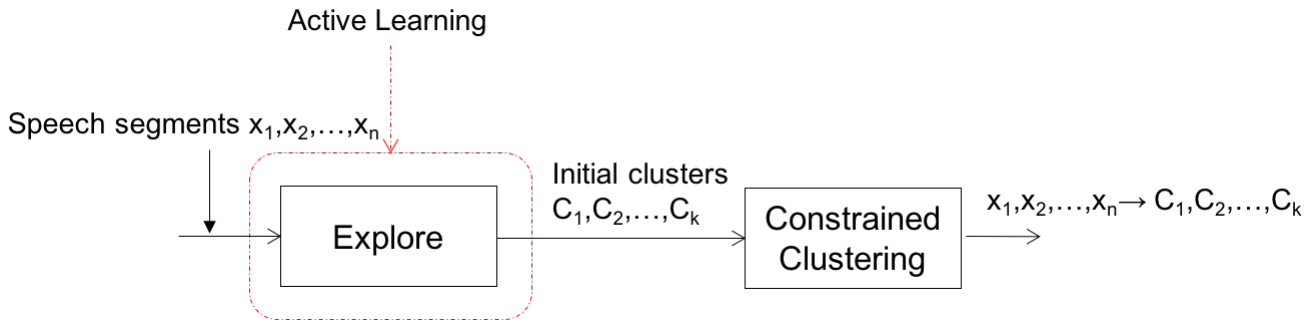

Fig. 1. Diagram of active learning based bottom-up speaker clustering. The red dotted block is active learning component where human involves.

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

On the other hand, only a limited number of studies in the area of speaker diarization, has investigated the use of active learning for speaker diarization tasks. For example, the study in [29] has proposed to use active learning to obtain background speaker labels from unlabeled data for training PLDA system. While the study in [29] bears similarity current study, the ultimate goal of [29] is to locate reliable samples sufficient enough to train PLDA system to improve speaker recognition task, rather than clustering entire dataset as in speaker diairization task. Another study, that employs active learning for speaker diarization is in [37]. However, the criteria of active selection for human labeling in [37] is simply based on the length of speech segment, which will not work in many speaker diarization task where the variance of segment length is small.

III. Baseline System

The baseline speaker diarization system used in this study, is a bottom-up speaker clustering algorithm with i-vector cosine distances score (CDS) as distance metric. In this section, we will give a brief discussion about i-vector extraction, CDS and bottom-up speaker clustering.

A. i-vector extraction

In i-vector extraction framework, speaker and channel dependent GMM supervector is 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 reduced total variability space, and w is 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 [38]. 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

Comparison of i-vectors from two different segments, could be successfully achieved with simple cosine similarity. The cosine distance score between two vectors could be expressed as follow:

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

Note that, the score of cosine distance ranges between [-1, 1]. The higher the number towards 1, the more similarity exist between two vectors. The cosine distance score has been a popular metric for speaker recognition in i-vector space.

C. Bottom-up Speaker Clustering

Bottom-up clustering, also known as hierarchical agglomerative clustering (HAC), has been the most popular speaker clustering algorithm used for speaker diarization. It typically starts by treating all homogeneous speech segments as a separate cluster, and iteratively merge two closest clusters. In our study, for each iteration, we find two segments that has the highest cosine similarity, and merge them into a single cluster. After each iteration, i-vectors are extracted from updated clusters with newly merged segments and normalized to have zero mean. We continue the iterations until the CDS of two closest cluster reaches specified stopping criteria.

IV. Active Learning based Speaker Clustering

Due to the error propagation characteristics of bottom-up speaker clustering, having good initial cluster model, has significant impact on final speaker diarization performance. In this section, we will give a detailed description of our proposed active learning strategy for bottom-up speaker clustering. As

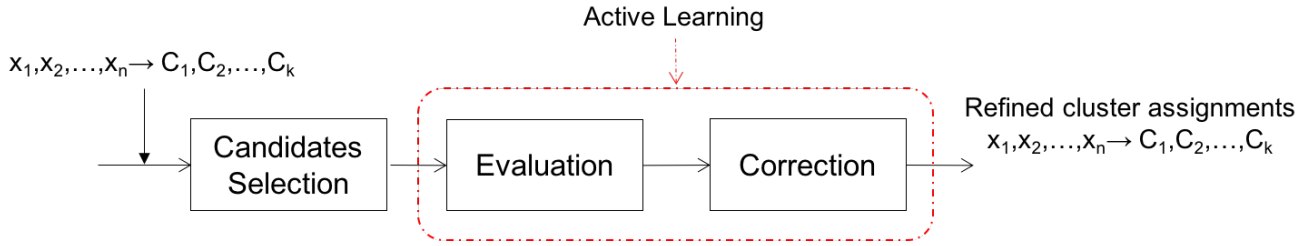

Fig. 2. Diagram of active learning based bottom-up speaker clustering. The red dotted block is active learning component where human involves.

mentioned in introduction, the proposed algorithm has two clustering components: *explore* and *constrained clustering* as in the Figure. 1.

A. Explore

The purpose of *explore* phase, is to quickly discover all speaker clusters within the audio streams, finding at least one speech segment for each speaker cluster. To achieve this, we use the farthest first query search (FFQS) proposed by [35]. During FFQS, a speech segment is randomly selected from all speech segments to be used as 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 existing clusters. The chosen segment is then provided for human to provide expert opinion. If the chosen segment belongs to 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 of this query returns by human expert is true, then the segment belongs to target cluster, and vice versa. The FFQS process continues until the pairwise comparison operations reached specified maximum specified number by user. The details of *explore* phase is detailed in Algorithm. 1.

Data: Set of speech segments $X = \{x_i\}_{i=1}^{n}$, acess 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

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

While the above algorithms could effectively explore 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, the initial seed data should be closer to actual centroids in order to achieve favorable clustering results [39]. Motivated by this, an initial unsupervised bottom-up clustering is performed, and centroid segments of these clusters are used as initial seeds to start FFQS algorithm for active learning based speaker clustering.

B. Constrained Merging

After initial *explore* phase, the standard bottom-up clustering is performed with two important exceptions. First, the clusters $C_{k=1}^{K}$ created during *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 *explore* phase. The bottom-up clustering will continues until only $C_{k=1}^{k}$ clusters remain.

V. Active Learning Based Cluster Reassignment

In previous section, we propose to use active learning during speaker clustering process. Alternatively, human input could be involved after clustering completes, to evaluate and fix incorrectly labeled speech segments. This is quite similar to the use of active learning in automatic speech recognition (ASR), where the transcripts of sentences estimated with less confidence is 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 much different, and difficult in speaker diarization.

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

A. Candidates selection

The cluster reassignment algorithm starts with selecting the candidate speech segments for human expert to review. In order to effectively select most informative segments, we rank all speech segments in terms of expected speaker error (ESE). In other words, we will select speech segments that

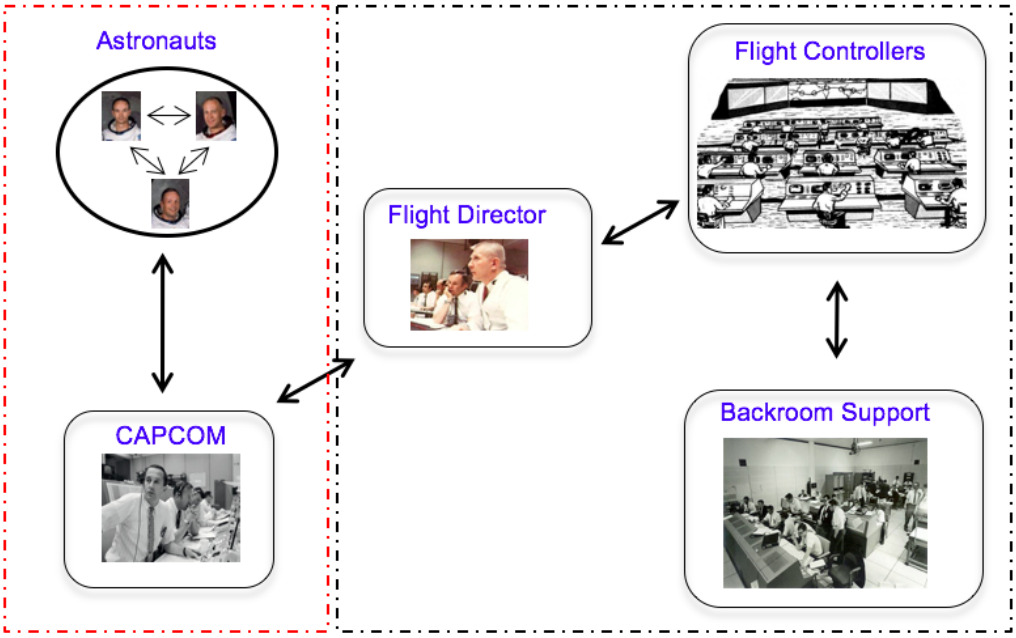

Fig. 3. Apollo Mission Control Center (MCC) communication overview. The red dotted parts are space-to-ground communications, including astronauts voice from space, and the black dotted parts are ground communications between hundreds of flight controllers and their 'backroom' supports.

will produce largest expected speaker error reduction. In this study, we compute the expected speaker error for each speech segment as below.

$$\begin{aligned} E(x_j) =& P(x_j|C_j) \cdot J_{x_j \in C_j} \\ &+ (1 - P(x_j|C_j)) \cdot J_{x_j \notin C_j} \end{aligned} \tag{5}$$

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

$$\begin{aligned} J_{x_j \in C_j} &= 0 \\ J_{x_j \notin C_j} &= \frac{d_j}{\sum_{i=1}^{n} d_i} \end{aligned} \tag{6}$$

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

We compute $P(x_j|C_j)$ by modeling a multivariate Gaussian distribution for each $C_{k=1}^{k}$ using i-vectors of all the segments of given cluster. We normalize the probability to sum to one.

$$\sum_{i=1}^{k} P(x_j|C_i) = 1; \tag{7}$$

B. Evaluation

After selecting the candidate segments, 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 Sec. IV. We could not simply select a single longest segment as a representative of the cluster to be compared with. 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 target cluster, resulting in multiple query pairs. If majority answers of these pairs are true (two segments belong to the same cluster), we will make a decision that the given speech segment has correct cluster assignment, and vice versa.

While the above strategy is robust for evaluating whether a given segment belongs to target cluster, it involves significant number of query pairs for human evaluation. One heuristic that we 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 top V confident segments as representatives of that cluster. These selected representative segments will be paired with test segment for majority voting based evaluation.

C. Correction

After detecting segments with incorrect cluster assignment in *evaluation* stage, we need to find the correct cluster designation for these segments. To do this, we employ N-best cluster evaluation. We find the N most possible cluster candidates of given segment by ranking the i-vector Gaussian posterior probabilities $P(x|C)$. 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.

VI. Test Data

We perform experiments on two different speech database: Apollo Mission Control Center (MCC) audio corpus and AMI meeting corpus [40].

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

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. Apollo-MCC Audio Corpus

During the NASA Apollo mission, all communications between astronauts, flight controllers, and their backroom support teams inside 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 (UTD), University of Maryland College Park (UMD), and Johnson Space Center(JSC) has 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 missions. Those mission audio records the full detail of the Apollo communication, therefore extremely attractive for learning human-to-human communications, group interaction, as well as developing robust speech system.

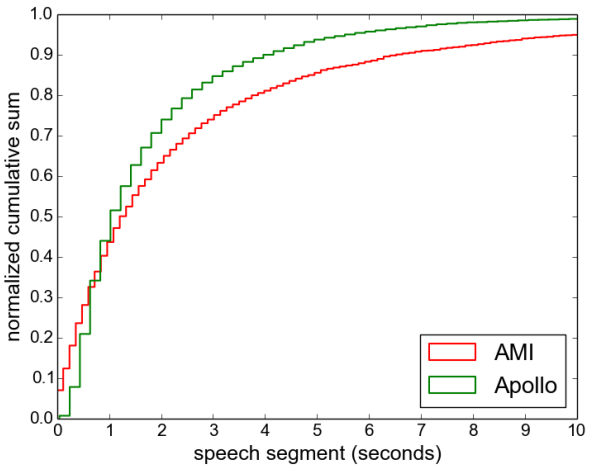

Fig. 4. Cumulative histogram as a function of speech segments length for Apollo-MCC audio corpus and AMI meeting corpus.

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 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 recording spanning approximately 10-hours before, and after lunar landing. This initial 28 hours task corpus has been transcribed to have speaker labels by well-trained speech science students from UTD [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" supports, see Figure. 3. Most of these audios are recorded with close-talking microphones and are in general good audio quality. The audios from astronauts transmitted through Earth's dedicated telephone channels to Houston from ground stations where the signal was received. The flight directors as well as their "backroom" supports voice are recorded through intercom circuit 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 of them containing 60 minutes of audios. The summary information about these audio files, 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.

The voice communication style within Apollo mission con-

[1]The task corpus will be released to the speech community for research and algorithmic development.

trol center is much different from both traditional meeting corpus, including many short speech segments which was intended to improve communication efficiency. The Fig. 4 shows cumulative histogram as a function of speech segments length. It can be observed that the Apollo-MCC audio dataset is composed of larger proportions of short speech segments (less than 3sec) than AMI meeting data. Another major challenges in Apollo-MCC audio dataset is the relatively large number of participants as shown in Table. I. Overall, the diarization of Apollo-MCC audio dataset is a realistic and challenging task.

B. AMI Meeting Dataset

We also evaluate proposed algorithms in the popular 12-meeting subset of Augmented MultiParty Interaction (AMI) corpus [40], [41]. 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, and we uses only the audio data recorded with headset microphones in our experiments. The corpus represents a natural meeting scenarios. A total of three participants are involved in each of these 12-meeting, discussing about the task to design a new remote control device. The summary information about different sessions are listed in Table. II.

VII. Experiments and Results

In this section, we will 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.

A. System Setup

1) Segmentation: As the purpose of this paper 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 part, and not confused by the errors caused by incorrect segmentation. The previous study has shown that the clustering part of speaker diarization could be developed independent of other modules [13], [30]. Besides, having fixed segmentation and oracle pairwise query answers between segments, are necessary for developing active learning based algorithm in order to avoid expensive human labeling in the experiment stage.

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

B. Active Learning Based Speaker Clustering

In this experiments, we evaluate the performance of active learning based speaker clustering, by varying the amount of oracle query pairs we have access to. Note that, the total

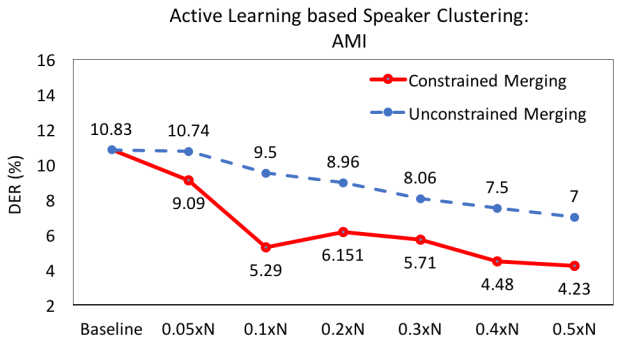

Fig. 5. Results of proposed active learning based speaker clustering algorithm on AMI meeting corpus. The red line is the diarization error rate (DER) with constrained clustering, while the blue dotted line is result obtained with unconstrained clustering as in baseline speaker clustering algorithms. 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.

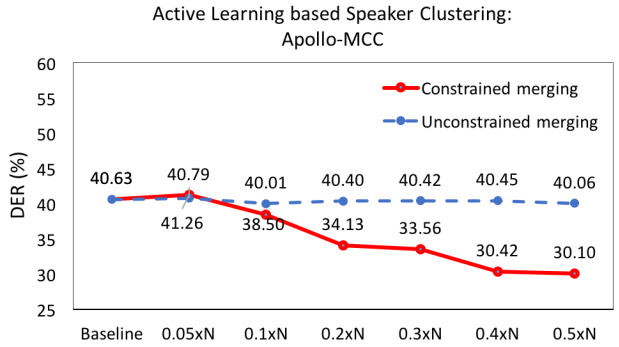

Fig. 6. Results of proposed active learning based speaker clustering algorithm on Apollo Mission Control Center (Apollo-MCC) audio dataset. The red line is the diarization error rate (DER) with constrained clustering, while the blue dotted line is result obtained with unconstrained clustering as in baseline 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.

number of queries to achieve perfect clustering result is $\frac{N(N-1)}{2}$ in the worst case, where N is total number of segment in test audio. For all experiments in this study, we will access to the reference answers of to selected query pairs, assuming no errors from human experts. However, in future study, the human errors should be taken into account.

We will evaluate the proposed algorithm by varying the quantity of query pairs proportional to the total segment number of N. For example, if a test audio stream has 1000 speech segments, active learning $0.1 \cdot N$ query pairs means we have access to 100 query pairs out of $1000 * (1000-1)/2$ total pairs. If we assume each speech segments has an average length of 2 seconds, the evaluation of 100 query pairs will spend about 400 (100*2*2) seconds for human evaluation. This is very small amount of human evaluation time compared to the total time requires human to obtain perfect speaker diarization which is $(1000*(1000-1)/2*2*2)$ in this example.

The red line in Figure. 5 and Figure. 6 shows the perfor-

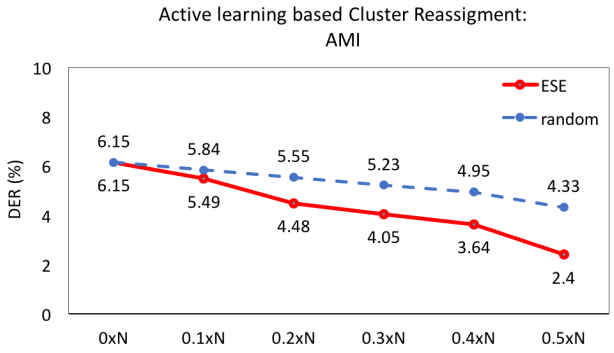

Fig. 7. Results of proposed active learning based cluster reassignment algorithm on AMI meeting corpus. The red line is the diarization error rate (DER) using the expected speaker error (ESE) as criteria for candidates selection, while the blue dotted line is DER obtained by randomly selecting segments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

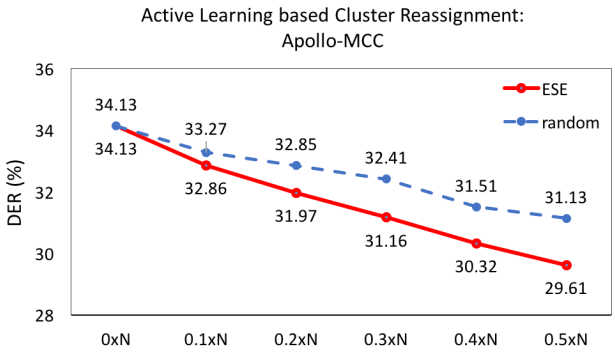

Fig. 8. Results of proposed active learning based cluster reassignment algorithm on Apollo Mission Control Center (Apollo-MCC) corpus. The red line is the diarization error rate (DER) using the expected speaker error (ESE) as criteria for candidates selection, while the blue dotted line is DER obtained 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.

mance of proposed active learning based speaker clustering algorithm using different amount queries. The baseline result is obtained using conventional bottom-up clustering with i-vector cosine distance score (CDS). The first thing we notice is that the baseline DER in Apollo-MCC (40.63%) is much higher than that of AMI dataset (10.83%). Such difference in performances are expected due to the challenges within Apollo-MCC dataset we illustrated in Section. VI-A. We could also see from the results, DER reduce rapidly with relatively small amount of human input. In case of AMI dataset, the DER reduce from 10.83% to 9.09%, a relative of 16% reduction with only 0.05xN query pairs, and the DER further reduce to 5.29% with 0.1xN queries. Figirue. 9 illustrates how DER of each AMI meeting session improved (or decreased) using proposed algorithm. It could be noticed that most (8 out of 12) sessions showed different degree of improvement, and only small increase in DER on other sessions.

In case of Apollo-MCC dataset, the proposed algorithm is also capable of effectively reducing the DER, although it requires access to much more queries compared with AMI dataset. This is attributes to the significantly larger participants within Apollo-MCC dataset, which requires more human input during *explore* stage, to discover all involved speakers.

1) Importance of constrained clustering: We also evaluate the importance of *constrained clustering* in active learning based speaker clustering. The blue dotted lines in Figure. 5 and Figure. 6 indicates the performance using traditional bottom-up clustering without any constraint in merging, after *explore* stage. And red dotted lines in Figure. 5 and Figure. 6 shows the performance with *constrained clustering*. Comparing these two results, we could see that it is extremely important to perform *constrained clustering*. Only small improvement in AMI dataset and nearly no improvements in speaker diarization is observed if *constrained clustering* is not applied. This indicates that our proposed active learning based speaker diarization is essentially transforms unsupervised speaker diarization task into something similar to supervised close-set speaker identification tasks, and therefore be able to achieve significant improvements.

2) limitations: While the experiment results in Figure. 5 and Figure. 6 has shown that proposed active learning based speaker clustering algorithm drops the DER quickly with relatively small amount of human input, the performance saturates as we uses more queries. This is expected behavior, as the objective of proposed algorithm is to discover all involved speakers, and initialize reliable speaker models for each of them. As soon as the majority of speakers are discovered with relative sufficient number of instances, the benefit of using more queries will be inconsequential. This poses a limitation in certain scenarios, where human input are needed to further drops the DER.

C. Active Learning Based Cluster Reassignment

In this experiment, we continue to improve speaker diarization performance using our second active learning algorithm: active learning based cluster reassignment. For both experiments on Apollo-MCC audio dataset and AMI meeting corpus, we use at most 10 instances per cluster, to compose query pairs for majority voting based evaluation of whether a particular segment belongs to target cluster. We also fix our n-best search to the rank of 3, during the search of 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 proportional to total number of segments N. For example, if the total number of speech segments in an audio stream is $N = 1000$, evaluation of 0.1xN segments means human expert will review 100 speech segments, which requires access to 100*3*10 queries with correct answer if we uses n-best rank of 3, and 10 instances per cluster.

The red lines in Figure. 7 and Figure. 8 indicates the performance of proposed active learning based speaker clustering algorithm on AMI and Apollo-MCC dataset, respectively. We uses clustering output from active learning based speaker clustering algorithm (0.2xN condition) as a base for performing cluster reassignment. We could see that the DER drops

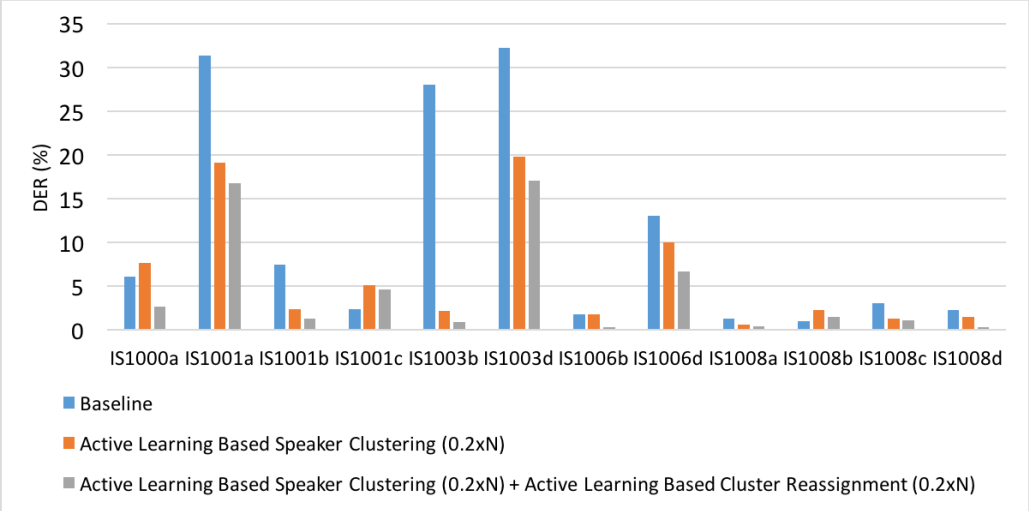

Fig. 9. The changes of DER on each session of 12-meeting subset of AMI corpus after applying proposed active learning algorithms.

consistently as more segments are selected for reassignments. In case of AMI dataset, the DER reduce from 6.15% to 5.49%, a relative of 10% reduction with only $0.1\text{x}N$ query pairs, and this number continue to reduce as more segments are selected for reassignment. Figirue. 9 illustrates how DER of each AMI meeting session improved (or decreased) using proposed algorithm combined with active learning based clustering. We could notice that all sessions of AMI dataset showed different degree of improvements.

The similar trend is also observed in the results from 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 more difficult. Also, the majority voting based approach we use for determine whether a segment belongs to target cluster, is sensitive to initial clustering results. But overall, the DER is consistently dropping as more queries are allowed.

1) Expected speaker error based candidate selection: In this experiment, we evaluate the effectiveness of using expected speaker error (ESE) as criteria for selecting segment candidates for human reassignment. We compare largest ESE based candidate selection with random segment selection. The red lines in Figure. 7 and Figure. 8 indicates the performance of using ESE as criteria, while the blue dotted line in Figure. 7 and Figure. 8 indicates the performance of random segment selection. The results clearly shows that the DER drops much faster with ESE based candidate selections, in both AMI and Apollo dataset.

VIII. Conclusion

In this study, we propose two active learning based algorithms for speaker diarization. The first algorithm is to use active learning for obtaining reliable initial speaker models and to perform constrained clustering. This is essentially turning an fully unsupervised speaker clustering tasks into a supervised tasks similar to speaker identification, with cluster models updated after each iteration. Due to such supervised information, the proposed algorithm reduce the DER significantly with access to relatively small amount of queries with true answer.

As the performance of proposed active learning algorithm for speaker clustering saturates when sufficient amount of instances are collected for each cluster, we propose another active learning algorithm to perform cluster reassignment after the completion of speaker clustering. The active learning based cluster reassignment, selects the clustered segments with largest expected speaker error for human evaluation and reassignment. The experiments on both AMI meeting dataset and Apollo-MCC dataset indicates that the DER drops continuously as more queries are allowed. And we also show that expected speaker error based segment selection strategy is significantly more effective than random segments selections.

In this study, we assume that human provides perfect answer to any query pair on whether the two segments belong to the same speaker. However, in reality, human errors are always expected and future study will evaluate how proposed algorithms performs with human errors.

Acknowledgments

This research was supported by National Science Foundation (NSF) under Grant 1219130.

REFERENCES

[1] M. A. H. Huijbregts, "Segmentation, diarization and speech transcription: surprise data unraveled," 2008.

[2] X. Anguera, C. Wooters, and J. Hernando, "Acoustic beamforming for speaker diarization of meetings," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 7, pp. 2011–2022, 2007.

[3] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke *et al.*, "The icsi meeting project: Resources and research," in *Proceedings of the 2004 ICASSP NIST Meeting Recognition Workshop*, 2004.

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

[5] D. A. Reynolds and P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proceedings.(ICASSP'05). IEEE International Conference on Acoustics, Speech, and Signal Processing, 2005.*, vol. 5. IEEE, 2005, pp. v–953.

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

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

[8] C. Wooters and M. Huijbregts, "The icsi rt07s speaker diarization system," in *Multimodal Technologies for Perception of Humans*. Springer, 2008, pp. 509–519.

[9] D. Vijayasenan, F. Valente, and H. Bourlard, "An information theoretic approach to speaker diarization of meeting data," *IEEE transactions on audio, speech, and language processing*, vol. 17, no. 7, pp. 1382–1393, 2009.

[10] H. Sun, B. Ma, S. Z. K. Khine, and H. Li, "Speaker diarization system for rt07 and rt09 meeting room audio." in *ICASSP*, 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," *Computer Speech & Language*, 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 *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 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 Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2015–2028, 2013.

[14] M. Rouvier and S. Meignier, "A global optimization framework for speaker diarization." in *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 *Proceedings of Odyssey*. Citeseer, 2014.

[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 Transactions on Audio, speech, and language processing*, vol. 20, no. 2, pp. 382–392, 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 and Understanding Workshop*, vol. 8. Virginia, USA, 1998, pp. 127–132.

[18] B. Zhou and J. Hansen, "Unsupervised audio stream segmentation and clustering via the bayesian information criterion," in *in Proc. ISCLP 2000*. Citeseer, 2000.

[19] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol. 2. IEEE, 1998, 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 recognition workshop*, vol. 1997, 1997.

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

[22] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, 2007, pp. 1–8.

[23] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Speaker and Language Recognition Workshop (IEEE Odyssey)*, 2010.

[24] L. Canseco-Rodriguez, L. Lamel, and J.-L. Gauvain, "Speaker diarization from speech transcripts." ICSLP, 2004.

[25] G. Friedland, H. Hung, and C. Yeo, "Multi-modal speaker diarization of real-world meetings using compressed-domain video features," in *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2009, pp. 4069–4072.

[26] A. Noulas, G. Englebienne, and B. J. Krose, "Multimodal speaker diarization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 1, pp. 79–93, 2012.

[27] S. J. Wenndt and R. L. Mitchell, "Machine recognition vs human recognition of voices," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 4245–4248.

[28] A. Schmidt-Nielsen, "Human vs. machine speaker identification with telephone speech."

[29] S. H. Shum, N. Dehak, and J. R. Glass, "Limited labels for unlimited data: active learning for speaker recognition." in *INTERSPEECH*, 2014, pp. 383–387.

[30] M. Sinclair and S. King, "Where are the challenges in speaker diarization?" in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7741–7745.

[31] X. Zhu, C. Barras, S. Meignier, and J.-L. Gauvain, "Combining speaker identification and bic for speaker diarization." in *INTERSPEECH*, vol. 5, 2005, pp. 2441–2444.

[32] Y. Guo and R. Greiner, "Optimistic active-learning using mutual information." in *IJCAI*, vol. 7, 2007, pp. 823–829.

[33] S. Vijayanarasimhan, P. Jain, and K. Grauman, "Far-sighted active learning on a budget for image and video recognition," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3035–3042.

[34] A. Biswas and D. Jacobs, "Active image clustering with pairwise constraints from humans," *International Journal of Computer Vision*, vol. 108, no. 1-2, pp. 133–147, 2014.

[35] S. Basu, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering." in *SDM*, vol. 4. SIAM, 2004, pp. 333–344.

[36] P. K. Mallapragada, R. Jin, and A. K. Jain, "Active query selection for semi-supervised clustering," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[37] 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 *Workshop on Speech, Language and Audio in Multimedia (SLAM 2014)*, 2014, pp. 5–p.

[38] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *Speech and Audio Processing, IEEE Transactions on*, vol. 13, no. 3, pp. 345–354, 2005.

[39] S. S. Khan and A. Ahmad, "Cluster center initialization algorithm for k-means clustering," *Pattern recognition letters*, vol. 25, no. 11, pp. 1293–1302, 2004.

[40] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos *et al.*, "The ami meeting corpus," in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, vol. 88, 2005.

[41] E. Gonina, G. Friedland, H. Cook, and K. Keutzer, "Fast speaker diarization using a high-level scripting language," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 553–558.