Speaker Diarization: A Review and Analysis

Aishwarya Balwani
Dept. of Electronics and Telecommunication
St. Francis Institute of Technology
Mumbai, India

Omkar Chogle
Dept. of Electronics and Telecommunication
St. Francis Institute of Technology
Mumbai, India

Shubhankar Kulkarni
Dept. of Electronics and Telecommunication
St. Francis Institute of Technology
Mumbai, India

Abstract-Speaker Diarization over the past few years has garnered tremendous attention, and a large amount of research has been carried out on the same by the audio and speech processing communities. The aim of speaker diarization is to answer the question of 'who spoke when?' Speaker diarization makes use of speaker recognition which is achieved by employing speaker segmentation and helps determine the change of speaker in the temporal dimension. Further, speaker clustering helps us group together speech segments on the basis of speaker characteristics. This paper is a general review of speaker diarization as a field of study, and then briefly studies some commonly used aspects of speaker diarization such as Bayesian models, Gaussian Mixture models and Hidden Markov models, before reviewing different types of clustering, the recent and increasing use of i-vectors for unsupervised calibration, and then also analyzes the same and compares their viability.

Keywords: ***Speaker diarization, speaker segmentation, speaker clustering, Bayesian models, Hidden Markov models, Gaussian mixture models, i-vectors.***

I. INTRODUCTION

Speaker diarization is the task of determining "who spoke when?" in an audio recording that contains an indefinite amount of speech and also an indefinite number of speakers. Earlier on, it was proposed as a research topic related to automatic speech recognition, where speaker diarization simply serves as an upstream processing step and nothing more. Over recent years, however, speaker diarization has developed into an important key technology for a variety of tasks, inclusive of navigation, retrieval, and also higher-level inference on audio data. Accordingly, many important improvements in precision and robustness have been reported since in the area, thanks to the efforts being put into the same. The application domains however also prove to be extremely challenging, as they range from broadcast news to college or school lectures and meetings, and therefore vary significantly and pose diverse problems, such as having access to multiple microphones and multimodal information or overlapping speech.

It is due to this vast potential that speaker diarization has emerged as an increasingly important and dedicated domain of speech research. In contrast to speech recognition which primarily revolves around transcription of speech; and also in contrast to speaker recognition which strictly identifies a particular speaker, speaker diarization is more temporal in nature and therefore many a times requires unsupervised identification of each speaker within an audio stream and the intervals during which each speaker is active.

Speaker diarization has utility in a majority of applications related to audio and/or video document processing, (information retrieval for example.) Indeed, it is often the case that audio and/or video recordings contain more than one active speaker. This is also the case for telephone conversations (especially those stemming from call centers and also offices which arrange for group calls regularly), broadcast news, debates, shows, movies, meetings, domain-specific videos (such as surgery operations for instance) or even lecture or conference recordings including multiple voices or questions/answers sessions. In all such cases, it can be beneficial to automatically determine the number of speakers involved in addition to the periods when each speaker is active. Clear examples of applications for speaker diarization algorithms include speech and speaker indexing, document content structuring, speaker recognition (in the presence of multiple or competing speakers) and speech translation.

This paper attempts to understand the basis of the subject, thus enabling us and other readers of the paper to work in the domain in the future.

II. BRIEF HISTORY

Over recent years the scientific community has developed research on speaker diarization in a number of different domains [1], with the focus usually being dictated by funded research projects. From early work with telephony data, broad-cast news (BN) became the main focus of research towards the late 1990's and early 2000's and the use of speaker diarization was aimed at automatically annotating TV and radio transmissions that are broadcast daily all over the world. Annotations included automatic speech transcription and metadata labeling, including speaker diarization. Interest in the meeting domain grew extensively from 2002, with the launch of several related research projects including the European Union (EU) Multimodal Meeting Manager (M4) project, the Swiss Interactive Multimodal Information Management (IM2) project, the EU Augmented Multi-party Interaction (AMI) project, subsequently continued through the EU Augmented Multi-party Interaction with Distant Access (AMIDA) project and, and finally, the EU Computers in the Human Interaction Loop (CHIL) project. All these projects addressed the research and development of multimodal technologies dedicated to the enhancement of human-to-human communications (notably in distant access) by automatically extracting meeting content, making the information available to meeting participants, or for archiving purposes.

[1]These technologies re being developed to potentially have to meet a range of challenging demands such as content indexing, linking and/or summarization of on-going or archived meetings, the inclusion of both verbal and non-verbal human communication (people movements, emotions, interactions with others, etc.). This is achieved by exploiting several synchronized data streams, such as audio, video and textual information (agenda, discussion papers, slides, etc.), that are able to capture different kinds of information that are useful for the structuring and analysis of meeting content. Speaker diarization plays an important role in the analysis of meeting data since it allows for such content to be structured in speaker turns, to which linguistic content and other metadata can be added (such as the dominant speakers, the level of interactions, or emotions).
While many other contrastive sub-domains such as lecture meetings or casual meetings such as breaks have also been considered, the conference meeting scenario has been the primary focus of many researchers. The meeting scenario is often referred to as “speech recognition complete”, i.e. a scenario in which all of the problems that can possibly arise in any speech recognition situation can be encountered in this domain. Conference meetings thus pose a number of new challenges to speaker diarization that typically were less relevant in earlier research.

III. Main Approaches

[1] Most speaker diarization systems fit into one of two categories: the bottom-up and the top-down approaches, as shown in Fig. 1. The top-down approach is initialized with very few clusters (usually one) whereas the bottom-up approach is initialized with many clusters (usually more clusters than expected speakers). In both cases the aim is to iteratively converge towards an optimum number of clusters. If the final number is higher than the optimum then the system is said to under-cluster. If it is lower it is said to over-cluster. Both bottom-up and top-down approaches are generally based on Hidden Markov Models (HMMs) where each state is a Gaussian Mixture Model (GMM) and corresponds to a speaker. Transitions between these states correspond to speaker turns, thereby signaling a change in speaker. In this section, we briefly discuss HMMs and GMMs before outlining the standard bottom - up and top-down approaches.

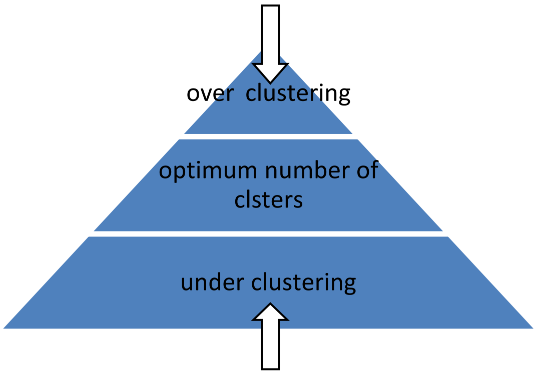

Figure 1: General diarization system

A. *Hidden Markov Models*

A Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. A HMM can be presented as the simplest dynamic Bayesian network.

In simpler Markov models (like a Markov chain), the state is directly visible to the observer, so the state transition probabilities are the only parameters. In a hidden Markov model, the state is not directly visible. The output however, which is dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. We see that the adjective 'hidden' in this case refers to the state sequence through which the model passes, not to the parameters of the model; the model is still referred to as a 'hidden' Markov model even if these parameters are known exactly.

Hidden Markov models are especially known for their application in temporal pattern recognition such as speech, handwriting, gesture recognition, part-of-speech tagging, musical score following, partial discharges and bioinformatics, thus making it an almost indispensible tool in speaker diarization.

A hidden Markov model can be considered a generalization of a mixture model where the hidden variables (which control the mixture component to be selected for each observation,) are related through a Markov process. Recently, hidden Markov models have been generalized to pairwise Markov models and triplet Markov models which allow consideration of more complex data structures and the modelling of non-stationary data. [2] Figure 2. gives an example of probabilistic parameters of an exemplary HMM.

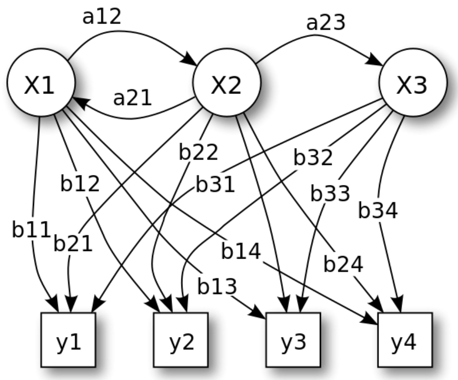

Figure 2: Probabilistic parameters of a hidden Markov model (example)
X — states
y — possible observations
a — state transition probabilities
b — output probabilities

B. Gaussian Mixture Models

In statistics, a mixture model is a probabilistic model for representing the presence of subpopulations within an overall population, without requiring that an observed data set should identify the sub-population to which an individual observation belongs. Formally a mixture model corresponds to the mixture distribution that represents the probability distribution of observations in the overall population. However, while problems associated with "mixture distributions" relate to deriving the properties of the overall population from those of the sub-populations, "mixture models" are used to make statistical inferences about the properties of the sub-populations when we are only given observations on the overall population, without sub-population identity information.

In speaker diarization, we make use of Gaussian mixture models, which may be Bayesian or non-Bayesian models. Bayesian models are found to be more commonly used.

C. Bottom-Up Approach:

The bottom-up approach [1] is by far the most common in the literature. Also known as agglomerative hierarchical clustering (AHC or AGHC), the bottom-up approach trains a number of clusters or models and aims at successively merging and reducing the number of clusters until only one remains for each speaker. Various initializations have been studied and, whereas some have investigated k-means clustering, many systems use a uniform initialization, where the audio stream is divided into a number of equal length juxtaposed segments. This simple approach generally leads to good performance. In all cases the audio stream is initially over-segmented into a number of segments which exceeds the anticipated maximum number of speakers. The bottom-up approach then iteratively selects closely matching clusters to merge, hence reducing the number of clusters by one after the completion of each iteration. Clusters are generally modeled with a GMM, and upon merging, a single new GMM is trained on the data that was previously assigned to the two individual clusters. A reassignment of frames to clusters is usually performed after each cluster merging, (via Viterbi realignment for example,) and the whole process is repeated iteratively, until some stopping criterion is reached, upon which there should remain only one cluster for each detected speaker. Possible stopping criteria include thresholded approaches such as the Bayesian Information Criterion (BIC), Kullback-Leibler (KL)-based metrics, the Generalized Likelihood Ratio (GLR) or the recently proposed T_S metric. Bottom-up systems have been known to perform consistently well and therefore find application in a variety of instances.

D. Top-Down Approach:

In contrast with the previous approach [1], the top-down approach first models the entire audio stream with a single speaker model and successively adds new models to it until the full number of speakers are deemed to be accounted for. A single GMM model is trained on all the speech segments available, all of which are marked as unlabeled. Using some selection procedure to identify suitable training data from the non-labeled segments, new speaker models are iteratively added to the model one-by-one, with interleaved Viterbi realignment and adaptation. Segments attributed to any one of these new models are marked as labeled. Stopping criteria similar to those employed in bottom-up systems may be used to terminate the process or it can continue until no more relevant unlabeled segments with which to train new speaker models remain. Top-down approaches however are far less popular than their bottom-up counterparts. Whilst they are generally outperformed by the best bottom-up systems, top-down approaches have performed consistently and respectably well against the broader field of other bottom-up entries. Top-down approaches are also extremely computationally efficient and can be improved through cluster purification.

IV. Main Steps to Speaker Diarization

The procedure of speaker diarization can be outlined as follows and is further illustrated in Figure 3:

- Acoustic beam forming: It deals with the multiple microphones which are often used to record the same meeting from different locations in the room, handling the disturbances from the same and appropriating adequately formed voice segments.
- Speech activity Detection: Speech Activity Detection (SAD) involves the labeling of speech and non-speech segments. SAD can have a significant impact on speaker diarization performance for two reasons. The first stems directly from the standard speaker diarization

performance metric, namely the diarization error rate (DER), which takes into account both, the false alarm and missed speaker error rates. Poor SAD performance will therefore lead to an increased DER. The second follows from the fact that non-speech segments can disturb the speaker diarization process, and more specifically the acoustic models involved in the process. Indeed, the inclusion of non-speech segments in speaker modeling leads to less discriminant models and thus increased difficulties in segmentation. Consequently, a good compromise between missed and false alarm speech error rates has to be found to enhance the quality of the following speaker diarization process.

- Segmentation and Clustering: Speaker segmentation is core to the diarization process and aims at splitting the audio stream into speaker homogeneous segments or, alternatively, to detect changes in speakers, also known as speaker turns. The classical approach to segmentation performs a hypothesis testing using the acoustic segments in two sliding and possibly overlapping, consecutive windows. While the segmentation step operates on adjacent windows in order to determine whether or not they correspond to the same speaker, clustering aims at identifying and grouping together same-speaker segments which can be localized any-where in the audio stream. Ideally, there will be one cluster for each speaker. Clustering can further be classified into types such as hierarchical approach, spectral clustering, i-vector clustering, and more.

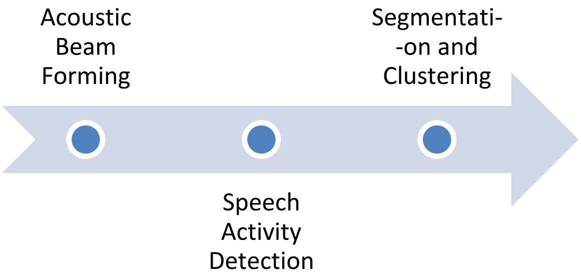

Figure 3: Sequential steps to Speaker Diarization

V. Comparison of Different Types of Clustering

Spectral clustering [3] can handle very complex and unknown cluster shapes, including those cases in which commonly used methods such as *K*-means or mixture models using EM may fail. Spectral clustering relies on analyzing the eigen-structure of an affinity matrix, rather than on estimating an explicit model of data distribution.

On qualitatively comparing the computational cost of the spectral clustering approach and of the hierarchical approach we see that as compared to the spectral approach, the computational cost of eigen-decomposition and *k*-mean clustering can be ignored, and most of the cost falls onto the calculation of affinity matrix which after unscented transformation approximation is drastically reduced. Hierarchical clustering requires a greater number of computations for Bayesian Information Criterion and estimating GMM and therefore it is much slower.

Speaker diarization [4] using unsupervised i-vector clustering has gained immense popularity in recent years. In this approach, i-vectors are extracted from short clips of speech which are segmented from larger multi-speaker conversations and are then organized into speaker clusters, typically according to their cosine score. This technique of using i-vectors to represent speakers has found great success in speaker recognition, and so the approach has broken into speaker diarization as well. A particular type of i-vectors clustering involves a temporal segmentation of the clip into clips that are 1-2 second long, and then the i-vectors are extracted for each of these segments. The i-vectors are then further broken down with the help of conversation-dependent Principal Component Analysis (PCA) and then scored with Probabilistic Linear Discriminant Analysis (PLDA) using parameters estimated on separate labeled data. The i-vectors are then clustered with these scores using Agglomerative Hierarchical Clustering (AHC), using a threshold which has been learned on unlabeled data, (hence proving the system to be unsupervised,) as the stopping criterion. Using PLDA scoring in diarization algorithms combined with conversation dependent PCA has yielded the results that there is a rich scoring metric which is unique to each conversation. This new scoring improves clustering performance over the same system with cosine scoring, which is the traditionally more popular technique. Performance is further improved with denser sampling of the i-vector subspace with overlapping segmentations.

On comparison, we establish that the earliest methods of using eigen vectors, also known as i-vectors result in the highest computation costs. The high costs associated wit basic eigen vector based techniques led to the development of spectral clustering; whose only computational cost is its affinity matrix. In most complex situations, spectral clustering is far more cost effective than its predecessor which is becoming increasingly obsolete except in the most simplest of requirements. Unsupervised i-vector scoring is now used in place of the traditional means of cosine scoring and is more efficient, thereby allowing us to have the option to choose between an affinity matrix, that is, spectral clustering and i-vectors as the situation demands.

VI. Conclusion.

This paper concludes that research on speaker diarization has been highly effective and has made great strides in an immensely short period of time. A variety of applications have been developed for diarization and all of speech processing on the whole, which hold a lot of promise for the future. There are opportunities for much growth, research and development, especially in the section of overlapping speakers, which is the biggest foreseeable problem for current systems.

A range of approaches are being applied, towards all aspects of the process of diarization, with the idea of making these systems more robust and effective. The future of speaker diarization indeed seems bright and can possibly accommodate many potential researchers in the near future.

REFERENCES

[1] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, Oriol Vinyals, Speaker Diarization: A Review of Recent Research, First draft submitted to the IEEE, 19th August, 2010.

[2] Tdunning, Hidden Markov Model with Output, Wikipedia.

[3] Huazhong Ning, Ming Liu, Hao Tang, Thomas Huang, A Spectral Clustering Approach to Speaker Diarization, 2006.

[4] Gregory Sell and Daniel Garcia-Romero, Speaker Diarization With PLDA I-Vector Scoring and Unsupervised Calibration, Spoken Language Technology Workshop (SLT), 2014 IEEE, 7-10 Dec. 2014.