Different Approaches for Speaker Diarization

Pooja M. Gaud
Electronics & Telecommunication Department
Dr.J.J.Magdum College of Engineering, Jaysingpur
Miraj, India
Email: gaudpooja19@gmail.com

Dr. V.V. Patil
H.O.D., Electronics Department
Dr. J. J. Magdum College of Engineering, Jaysingpur
Jysingpur, India
Email: vvpatil2429@gmail.com

***Abstract*—** Audio diarization is the process of annotating an input audio channel with information that attributes (possibly overlapping) temporal regions of signal energy to their specific sources. These sources can include particular speakers, music, background noise sources and other signal source/channel characteristics. Speaker diarization is the task of determining "who spoke when?" in an audio or video recording that contains an unknown amount of speech and also an unknown number of speakers. Diarization can be used for helping speech recognition, facilitating the searching and indexing of audio archives and increasing the richness of automatic transcriptions, making them more readable. Over recent years, however, speaker diarization has become an important key technology for many tasks, such as navigation, retrieval or higher-level inference on audio data. Accordingly, many important improvements in accuracy and robustness have been reported in the area of conferences. The application domains, from broadcast news, to lectures and meetings, vary greatly and pose different problems, such as access to multiple microphones and multimodal information or overlapping speech.

***Keywords*-** *speaker identification, segmentation, clustering*

*****

I. INTRODUCTION

Speaker diarization has emerged as an increasingly important and dedicated domain of speech research [1]. Whereas speaker and speech recognition involve, respectively, the recognition of a person's identity or the transcription of their speech, speaker diarization relates to the problem of determining 'who spoke when?' More formally this requires the unsupervised identification of each speaker within an audio stream and the intervals during which each speaker is active [2].

Although filterbank analysis was one of the earliest methods developed for speech processing, it remains one of the most effective techniques used in speaker recognition system. In this approach the short time magnitude spectrum of a speech signal is represented by the energy in the output signal of a set of band pass filter spaced evenly across the frequency range of interest. Approximately 20 filters are used in this process, producing a compact set of coefficients to represent the spectrum [3]. Also, in speaker diarization bottom-up approach & top-down approach is commonly used.

The section II gives information about speaker diarization process & speaker diarization architecture. Section III introduces different approaches used in speaker diarization system. Following section IV gives different applications of speaker diarization & section V is the general discussion on speaker diarization.

II. SPEAKER DIARIZATION

A. Speaker Diarization Process

The speaker diarization process consists of speech activity detection, speaker segmentation & speaker clustering [4]. Fig.1 shows the speaker diarization process**.**

1. Speech Activity Detection-

The purpose of this stage is to classify the audio into speech & non speech regions. It is important to identify & discard non speech regions such as music & noise early in the diarization process to avoid hindering subsequent speaker segmentation & clustering process [2].This stage is used to remove only prolonged periods of music or noise, rather than targeting short speaker pauses in the middle of speaker turns & thus breaking up homogeneous speaker segments.

2. Speaker Segmentation –

Speaker segmentation is the process of partitioning the audio data into homogeneous segments according to speaker identities. This stage is responsible for determining all boundary locations within each speech region that correspond to true speaker change points, providing clean, uncontaminated data for subsequent speaker clustering[1].

3. Speaker Clustering –

Speaker clustering is the process of associating segments of speech produced by the labelling all speech segments belonging to the same speaker with the same relative, show

 ISSN: 2321-8169
2350 – 2354

internal speaker label. The clustering stage is responsible for associating the segments belonging to the same speaker together. Clustering aims at identifying and grouping together same-speaker segments which can be localized anywhere in the audio stream. Ideally, at the end of this stage one single cluster is produced for each speaker in the audio, containing all speech segments belonging to that speaker [2].

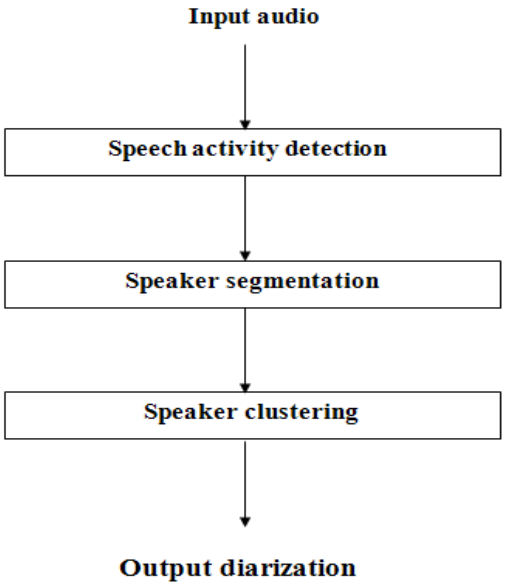

Fig.1 Speaker Diarization Process

B. Speaker Diarization Architecture

Speaker diarization architecture consists of data preprocessing, cluster initialization, merge/split operation, cluster distance & stopping criterion. Fig.2 shows the speaker diarization architecture.

Data preprocessing involves noise reduction, multichannel acoustic beam forming [5]-[6], the parameterization of speech data into acoustic feature and the detection of speech segments with a speech activity detection algorithm [2]. Cluster initialization depends on the approach to diarization that is the choice of an initial set of cluster in bottom up clustering [7],[8],[9] or a single segment in top down clustering[10],[11]. Split/merging mechanism is used to iteratively merge clusters [12] or to introduce new ones [13]. Cluster distance is the distance between clusters. Stopping criteria is used to determine when the optimum number of clusters has been reached [14], [15].

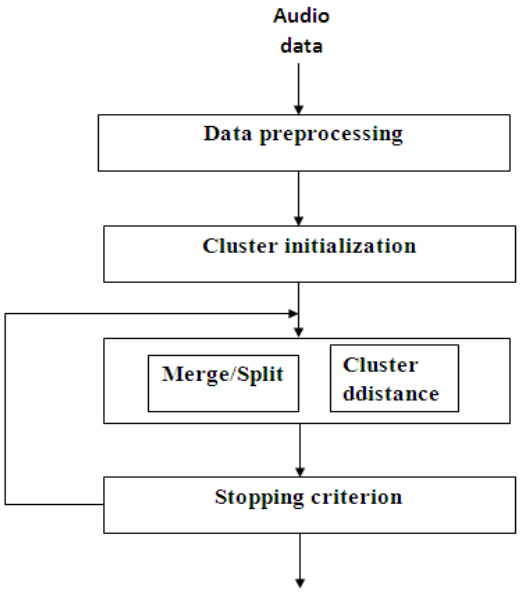

Fig.2 Speaker Diarization Architecture

III. DIFFERENT APPROACHES OF SPEAKER DIARIZATION

This section gives information about different approaches of speaker diarization system such as acoustic beamforming, speaker segmentation, speaker clustering & different modeling technique.

A. Acoustic Beamforming

The application of speaker diarization to the meeting domain triggered the need for dealing with multiple microphones which are often used to record the same meeting from different locations in the room. The microphones can have different characteristics: wall-mounted microphones (intended for speaker localization), lapel microphones, desktop microphones positioned on the meeting room table or microphone arrays. The use of different microphone combinations as well as differences in microphone quality called for new approaches to speaker diarization with multiple channels [5]. So the multiple distant microphone (MDM) condition was introduced in the NIST RT‘04 (spring) evaluation. A variety of algorithms have been proposed to extend mono-channel diarization systems to handle multiple channels. One option, proposed in, is to perform speaker diarization on each channel independently and then to merge the individual outputs. In order to do so, a two axis merging algorithm is used [6].

B. Speaker Segmentation

Speaker segmentation is the process of partitioning the audio data into homogeneous segments according to speaker identities. The classical approach to segmentation performs a hypothesis testing using the acoustic segments in two sliding and possibly overlapping, consecutive windows. For each

considered change point there are two possible hypotheses: first that both segments come from the same speaker (H0), and they can be well represented by a single model; and second that there are two different speakers (H1), and thus that two different models are more appropriate [2]. The segmentation stage is responsible for determining all boundary locations within a given audio that correspond to true speaker change points, providing clean, unconstrained data for subsequent speaker clustering. Speaker segmentation can be categorized into three classes- energy based, model selection based, metric based segmentation [1].

1. Energy Based Segmentation

Energy based segmentation algorithm rely on silence location within the input audio stream to partition the audio, based on the assumption that speaker segment boundaries occur at these locations. Approaches for detecting silence locations include the use of a viterbi decoder & by measuring & thresholding the signal energy [16].

2. Model Selection Based Segmentation

Model selection based segmentation relies on the use of competing models to determine the locations of speaker change points. In order to locate the most appropriate speaker change point within the given speech segment, every frame within the segment is examined using this approach [16].

3. Metric Based Segmentation

In metric based segmentation a distance metric is used to determine the statistical dissimilarity between two segments in order to make decision as to whether a segment boundary is appropriate at a given locations. This is most commonly achieved via the sliding window approach using constant sized window [16]. Metric based segmentation is improved by using heuristic rules; these rules were developed, to determine which peaks on the distance curve correspond to true speaker change points. The proposed heuristic rules govern the smoothing of the distance curve, detection of peaks on the smoothed curve & the selection of significant peaks on segment boundary hypothesis [3].

C .Speaker Clustering

The clustering stage is responsible for associating the segments belonging to the same speaker together. The aim of the clustering is to identify and grouping together same-speaker segments which can be localized anywhere in the audio stream. Ideally, there will be one cluster for each speaker [2]. Speaker clustering consist of two commonly used approach such as bottom-up approach & top-down approach.

1. Bottom-Up

Bottom-up approach 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. In all cases the audio stream is initially over-segmented into a number of segments which exceeds the anticipated maximum number of speakers [2].

2. *Top-Down*

The top-down approach is initialized with very few clusters (usually one). In this case the aim is to iteratively converge towards an optimum number of clusters [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. Top-down approaches are far less popular than their bottom-up counterparts. Top-down approaches are also extremely computationally efficient and can be improved through cluster purification [2].

D. Modeling Techniques for Speaker Diarization

Different modeling techniques are used in speaker diarization to improve the diarization output. These are modeling uncertainty in speaker model estimates, modeling uncertainty in Eigen voice speaker modeling, Gaussian mixture speaker modeling.

1. Modeling Uncertainty in Speaker Model Estimates

The uncertainty associated with the direct estimation of model parameters, the Bayes factor is proposed as a decision criterion for speaker segmentation & clustering, replacing the popular maximum likelihood criteria widely. The Bayes factor approach is able to incorporate the information regarding the whole audio as prior information, to aid segmentation & clustering decisions. Considering the speaker clustering task in a hypothesis testing framework, a solution to the Bayes predictive density for single, full covariance multivariate Gaussian speaker model is derived, which forms the basic building block for constructing the Bayes factor criterion. The concept of using Bayes factor for speaker clustering is then extended to the segmentation task. The solution to the Bayes predictive density integral is used to construct a Bayes factors as a distance metric that is more suitable for this application, replacing the popular generalized likelihood ratio. The Bayes factor, derived specifically for multivariate Gaussian speaker modeling, was introduced to account for the uncertainty of the speaker model estimates. The use of the Bayes factor also enabled the incorporation of prior information regarding the audio to aid segmentation and clustering decisions [18].

2. Modeling Uncertainty in Eigen voice Speaker Modeling

An alternative approach to speaker modeling, based on joint factor analysis techniques is the use of the Eigen voice model to represent the speakers. Under this approach GMM that best represents the observations of a particular speaker is given by the combination of a speaker independent universal background model with an additional speaker dependent offset constrained to lie in a low dimensional speaker variability subspace [3].

A novel criterion is proposed in this work by incorporating eigenvoice modeling technique into the popular cross likelihood ratio (CLR) criterion, for speaker clustering, allowing the system to capitalize on the advantages of eigenvoice modeling technique in a CLR framework. This novel criterion uses Bayesian methods to estimate the conditional probabilities in computing the CLR, effectively combining the eigenvoice CLR framework with the advantage of a Bayesian approach to the diarization problem [18].

3. Gaussian Mixture Speaker Modeling

A GMM is a parametric model that can be used to estimate continuous probability density function for multidimensional signal features [2].The use of GMM for speaker modeling have been reported in speaker processing literature for a wide range of application including speaker diarization, speaker identification & speaker verification tasks. The approaches for estimating the parameters of a GMM based on Maximum likelihood, Maximum posteriori criterion.

i. Maximum Likelihood Estimation

The aim of maximum likelihood estimation is to find the set of model parameters which maximum the likelihood of the training data. For single multivariate Gaussian modeling, the estimation of model parameter using the maximum likelihood criterion is a relatively trivial task. However, for GMM, this task is not so straightforward due to the assumption made for mixture models [17].

ii. Maximum Posteriori Estimation

A maximum posteriori estimation is based on Bayesian estimation theory, this approach is widely reported in speaker verification such as more recently MAP estimation of GMM have also proven useful in training speaker cluster models for the clustering task in speaker diarization system [3].The MAP approach is able to produce more robust models given limited training data, by incorporating prior knowledge about the speaker model parameters into the training procedure. This is achieved through the use of a universal background model (UBM) .The UBM is a GMM trained on a large selection of representative speech, often using the ML approach [17].

IV. APPLICATIONS OF SPEAKER DIARIZATION

Speaker diarization has utility in a majority of applications related to audio and/or video document processing, such as information retrieval. It is also used in telephone conversation, broadcast news, debates, shows, movies, meetings, domain-specific videos, even lecture or conference recordings including multiple speakers or questions/answers sessions. In all such cases, it can be advantageous to automatically determine the number of speakers involved in addition to the periods when each speaker is active[2].

In broadcast news, speech data is usually acquired using boom or lapel microphones with some recordings being made in the studio and others in the field. Here signal to noise ratio is better. Broadcast news speech is often read or at least prepared in advance. The no of speaker is larger in broadcast news

Meetings are usually recorded using desktop or far-field microphones which are more convenient for users than head-mounted or lapel microphones. In meetings Signal to noise ratio is less than broadcast news. Meeting speech is spontaneous in nature & contains more overlapping speech. The no of speaker used is less in meetings application [18].

V. DISCUSSION

The speaker diarization is developed to improve the performance and practicality of speaker diarization technology in the broadcast news audio domain through the reduction of diarization error rates. Study on speaker diarization has been used in many domains, from phone calls conversations within the speaker recognition evaluations, to broadcast news and meeting recordings in the NIST Rich Transcription evaluations. Furthermore, it has been used in many applications such as a front-end for speaker and speech recognition, as a metadata extraction tool to aid navigation in broadcast TV, lecture recordings, meetings, and video conferences and even for applications such as media similarity estimation for copyright detection. Also, speaker diarization research has led to various by-products.

This study gives the information about the different approaches used in speaker diarization. In the speaker diarization, larger datasets need to be compiled in order for results to become more meaningful and for systems to be more robust to unseen variations, with increasing dataset sizes, systems will have to become more efficient in order to process such data in reasonable time. Still, the biggest single challenge is probably the handling of overlapping speech, which needs to be attributed to multiple speakers.

This article provides an overview of the current state-of the art in speaker diarization systems and underlines several challenges that need to be addressed in future years.

REFERENCES

[1] S. Tranter and D. Reynolds, "An overview of automatic speaker diarization systems," IEEE TASLP, vol. 14, no. 5, pp. 1557–1565, 2006.

[2] Xavier Anguera, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, Oriol Vinyals, "Speaker Diarization - A Review of Recent Research," IEEE TASLP, 2010.

[3] R.Mammone, X. Zhang and R. Ramchandra, "Robust speaker recognition: a feature – based approach," IEEE signal processing magazine, vol.13, no.5, pp.58-71, 1996.

[4] X. Zhu, C.Barras, L.Lamel and J. –L. Gauvain, " Multi- stage speaker diarization for conference & lecture meetings," in multimodal technologies for perception of humans, pp.533-542, 2008.

[5] A. Janin, J. Ang, S. Bhagat, R. Dhillon, J. Edwards, J. Macias-Guarasa, N. Morgan, B. Peskin, E. Shriberg, A. Stolcke, C. Wooters, and B. Wrede, "The ICSI meeting project: Resources and research," in Proc. ICASSP Meeting Recognition Workshop, 2004.

[6] D. Mostefa, N. Moreau, K. Choukri, G. Potamianos, S. M. Chu, A. Tyagi, J. R. Casas, J. Turmo, L. Cristoforetti, F. Tobia, A. Pnevmatikakis, V. Mylonakis, F. Talantzis, S. Burger, R. Stiefelhagen, K. Bernardin, and C. Rochet, The CHIL audiovisual corpus for lecture and meeting analysis inside smart rooms. Language Resources and Evaluation, December 2007, vol. 41.

[7] X. Anguera, C. Wooters, and J. Hernando, "Robust speaker diarization for meetings: ICSI RT06s evaluation system," in Proc. ICSLP, Pittsburgh, USA, September 2006.

[8] T. Nguyen et al., "The IIR-NTU Speaker Diarization Systems for RT 2009," in RT'09, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, Florida, USA, 2009.

[9] X. Anguera, C. Wooters, and J. Hernando, "Friends and enemies: A novel initialization for speaker diarization," in Proc. ICSLP, Pittsburgh, USA, September 2006.

[10] C. Fredouille and N. Evans, "The LIA RT'07 speaker diarization system," in Multimodal Technologies for Perception of Humans: International Evaluation Workshops CLEAR 2007 and RT 2007, Baltimore, MD, USA, May 8-11, 2007, Revised Selected Papers. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 520–532.

[11] C. Fredouille, S. Bozonnet, and N. W. D. Evans, "The LIA-EURECOM RT'09 Speaker Diarization System," in RT'09, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, Florida, USA, 2009.

[12] T. Nguyen et al., "The IIR-NTU Speaker Diarization Systems for RT 2009," in RT'09, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, Florida, USA, 2009.

[13] C. Fredouille, S. Bozonnet, and N. W. D. Evans, "The LIA-EURECOM RT'09 Speaker Diarization System," in RT'09, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, Florida, USA, 2009.

[14] S. S. Chen and P. S. Gopalakrishnan, "Speaker, environment and channel change detection and clustering via the bayesian information criterion," in Proc. of DARPA Broadcast News Transcription and Understanding Workshop, Lansdowne, Virginia, USA, February 1998, pp. 127–132.

[15] H. Gish and M. Schmidt, "Text independent speaker identification," in IEEE Signal Processing Magazine, Octomber 1994, pp. 18–32.

[16] C.Barras, Z. Xuan, S.Meignier and L.Lamel and J. –L. Gauvain, "Multi- stage speaker diarization for broadcast news," IEEE Trans, Audio, speech and language processing, vol.14, pp.1505-1512, 2006.

[17] 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 and language, no.20, pp. 303-330, 2006.

[18] David I-Chung Wang, "Speaker Diarization- Who Spoke When", Octomber 2012.

BIOGRAPHY

Miss.P.M.Gaud received the BE degree in 2013 in Electronics & Telecommunication Dept. from Govt. College Of Engg. Karad & persuing ME Electronics & Telecommunication in Dr. J.J.Magdum College of Engg. Jaysingpur.

Dr.Mrs.V.V.Patil received the BE degree in1994 & ME degree in 2004 in Electronics Engg. Dept. of Walchand College of Engg. sangli & PHD degree from Electrical Engg. Dept. of I.I.T Bombay in 2014.
She is currently professor & Head in Electronics Engg. Dept. of Dr. J.J.Magdum College of Engg. Jaysingpur.
Her research interests are in the area of speech & signal processing applications.