Speaker Change Detection

Ms Uzma Muzzaffar[1], Ms Sukhvinder Kaur[2], Mr. Muheet Ahmed Butt[3], Mr. Majid Zaman[4]
[1]*Dept. of Electronics and Communications, Swami Devi Dyal Institute Of Engineering & Technology, Kurukshetra University, Kurukshetra.*
[2]*Assistant Professor and Head of Department Of Electronics And Communications, Swami Devi Dyal Institute of Engineering & Technology, Kurukshetra University, Kurukshetra.*
[3]*Scientist, PG Department of Computer Sciences, University Of Kashmir, Srinagar.*
[4]*Scientist, Directorate of IT & SS, University of Kashmir, Srinagar.*

Abstract: Speaker Change Detection involves a multimedia indexing technology that makes use of audio information to answer the question "Who spoke when?" This thesis presents a step-by-step speaker diarization system implemented in MATLAB that is evaluated using the Diarization Error Rate (DER) metric. The proposed system, designed for segmenting audio recordings of broadcast news, provides implementations of state-of-the-art i-vectors as well as the traditional GMM speaker models. A graphical clustering algorithm introduced by Rouvier et al. in 2013 has also been implemented. This clustering algorithm offer lower DER as well as a computational advantage compared to the conventional GMM based hierarchical agglomerative clustering. An unsupervised speech activity detector (SAD) has also been developed that discards non speech in two stages - silence removal followed by music removal. The music removal subsystem has been adapted to classify speech segments with background music, e.g. news headlines sections, as speech. The proposed SAD achieves a favourable performance on the January 2013 subset of the REPERE corpus compared to the supervised SAD of the LIUM diarization toolkit.

Keywords **-** unsupervised, speech activity detection, MATLAB, ILP clustering, REPERE

I. INTRODUCTION

Initially in the late 1990's, when research in diarization was still in its nascent stages, few systems attempted to perform speech activity detection as a by-product of the segmentation and clustering [8]. Non speech was thought to be just another speaker. But owing to the acoustic variability of non speech, systems with explicit speech activity detectors performed much better. Often, the speaker segmentation and speaker clustering are performed iteratively. In this paper previously used methods in speaker diarization have been reviewed and the state-of-the-art algorithms implemented by various systems specialized in diarization of broad-cast news, meeting recordings and telephone conversations are compared. In the recent years, the National Institute of Science and Technology (NIST), USA have organised rich transcription tasks for broadcast news and telephone diarization (2003-`04) and for meeting diarization (2005, `07, `09). The Albayzin campaign of 2010, the ESTER (2008) [10] and REPERE (2012-14) broadcast audio and video diarization campaigns have fueled research in broadcast news diarization and attracted developers to participate with their diarization engines to set up benchmarks. Some of these competitor systems have also been reviewed in this chapter. diarization system has to answer the question \Who spoke when?" without any a-priori information about the speakers present in the audio recording. The output that is expected from the system is of the form shown in Figure 2.1. In particular, note that speakers segments are not expected to be labelled by their name, only by a unique speaker id, which is indicated by colour for the recording in the figure. Speaker diarization is thus different from speaker verification or speaker recognition where prior information for target speakers may be made available to the system beforehand in the form of speaker models or speaker biometrics.

A. Meeting Diarization

The evaluation of a diarization system is done using a metric called the Diarization Error Rate (DER) [4], which is the percentage of the time of the audio for which the speaker was wrongly labelled. The output of the system is compared with a segment level manually annotated temporal transcription indicating the speaker labels. In broadcast news, there is very little overlap. For the case where overlap is absent, the formula for DER can be simplified. This error calculation is used in the broadcast news diarization systems. The DER can be broken down systematically into 2 types. Consider an audio with S sec speech and NS sec non-speech as indicated by the annotation. Non-speech includes silences, speaker pauses, music, jingles, noise etc. Missed speech time is the time when the algorithm erroneously indicated a segment as non-speech. False alarm speech time, on the other hand, is the time when the algorithm erroneously indicated a segment as speech. These 2 errors occur during the speech activity detection, which is a pre-processing step in almost all diarization systems. They are numbered E1 = T1x100/S and

E2 = T2x100/S. E1 is called missed speech rate (MSR) and E2 is called false alarm speech rate (FASR)

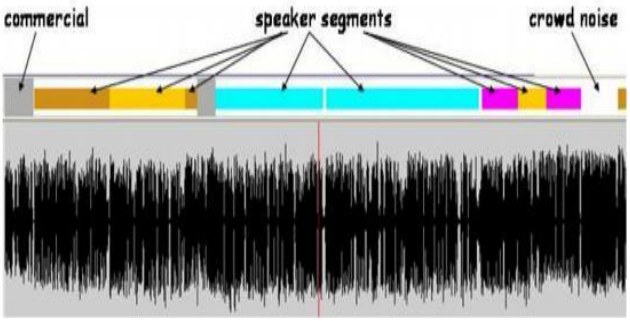

Fig.1: Rich transcription generated from a speaker diarization system [3]

The task of finding contiguous segments of speech in an audio and segregating them from other types of sounds is called speech activity detection (SAD). It is beneficial for speech processing systems since it is practical to process only speech segments rather than entire recordings. It makes a design more efficient by saving computation time and resources. Apart from the computational advantages, the absence of an SAD often causes insertion errors in ASR systems. Hence speech activity detection is a fundamental task in almost all fields of speech processing - coding, enhancement and recognition [8].

In speaker diarization, the error metric itself highlights the need for a speech activity detector since missed speech and false alarm speech are included in the diarization error rate metric. Moreover, with limited speaker data from small speech segments, presence of non-speech contaminates the estimated speaker models thereby affecting the performance of the diarization system. Initial approaches to diarization tried to let SAD be a by-product of the diarization system [8] by letting non speech be a single cluster which would be discarded at the end. However it was soon noticed that systems having an explicit SAD gave better results. SAD is often performed using frame-wise classification. Statistical models are trained and estimated on a feature space most suitable for discriminating the speech and non speech classes. In most cases, Gaussian mixture models are the statistical models used and the feature space is in most cases cepstral features.

II. RELATED WORK

For the task of speaker diarization, acoustic features that discriminate speaker information in the spectrogram but are invariant to the phone sequence being uttered are desired. Mel-frequency cepstral coefficients (MFCCs) or Perceptual Linear Prediction (PLP) coefficients, although not designed to distinguish between speakers, have been used widely in the areas of speaker verification and speaker recognition. Since a similar task of modelling speaker information is tackled in speaker diarization, MFCCs and other cepstral features are the most commonly used features. During speaker segmentation 12-19 MFCCs have been used along with the short time energy, while during clustering usage of higher order derivatives of these MFCCs has been reported LFCCs extracted using a linear filter bank instead of the Mel scale filter bank [12] and Linear Prediction Cepstral Coefficients (LPCCs) [13] have also been tested but no conclusion has been reached regarding the better performance of either. Typical sizes of analyses windows are 25-30ms with frame hops of 10ms.

For speech activity detection, acoustic features that discriminate between speech and non-speech are sought after. Features such as energy [13], zero-crossing rate, spectral centroid, spectral roll-o and spectral flux [14] have been used previously in speech activity detection. However the use of these feature vectors has always been seen in concatenation with cepstral features. Other than the above mentioned short time analysis features, 4Hz modulation frequency features that convey long term characteristics of the acoustic signal have also been investigated [15] and have been applied in the speaker overlap detection and speech activity detection. A major challenge faced in these features though is the high dimensionality of the features and the computational cost associated with it. Long term cumulative features drawn over texture windows of 500ms such as median of pitch, long time average spectrum, deviation of the 4th and 5th formants, harmonics to noise ratio, formant dispersion etc. have shown to be of use for fast cluster initialization [9], while features providing vocal source and vocal tract information [16] have shown better speaker discrimination when used along with MFCCs.

Recently Slaney et al. used features derived as activations of the bottleneck layer of a neural network. The artificial neural network was trained to discriminate 500ms segments as belonging to same or different speaker [17]. In another work [18], a 50% relative improvement was reported for speech activity detection on a large Youtube corpus when a two dimensional soft-max activation of a deep neural network was concatenated with 13 MFCC. Another interesting feature space explored in 2011 have sacrificed diarization error only slightly to obtain a 10x speed-up using binary valued features for performing clustering [19]. In this work, acoustic MFCC features of segments are transformed into a binary feature space using likelihoods obtained from GMMs.

III. PROPOSED METHODOLOGY

The speech activity detector in the proposed system is a model based classifier. It is independent of external training data for modelling the non speech and speech classes. The approach to such a model based speech activity detector is inspired by the SAD in the IIR-NTU submission to the NIST RT2009 evaluations [13]. In our system speech activity detection is done in two decoupled steps. First, silence is removed from the whole recording using an energy based bootstrapping followed by iterative classification. In the second step, music and other audible non speech are identified from the recording. For music removal the silence removed audio is fed to music vs. speech bootstrap discriminator. The frames of the audio which are music with a high confidence are used to train a music model which is iteratively refined. In both steps, only segments with duration 1s or longer have been labelled as non speech in order to avoid sporadic non speech to speech transitions. These constraints are incorporated using a GMM-HMM framework.

A. Silence Removal

The silence removal in the proposed system is done using 19 MFCC features concatenated with short time energy (STE) and their first and second derivatives. A bootstrap segmentation assigns a confidence value to every frame for both silence and speech classes. The bootstrap silence model is trained using a Gaussian mixture of size 4 over the 60 dimensional feature spaces. A speech model is also trained with the same size from high confidence speech frames.

In an iterative classification step, each frame is classified into two classes' viz. speech and silence. The high confidence speech and silence frames from these are used to train the speech and silence models for the next iteration. As the number of iterations increase, the number of 60 dimensional Gaussians used to model the speech and silence GMMs are increased until a

IV. RESULTS

The proposed system has implementations of two speaker models which have been widely studied in speaker verification and speaker recognition tasks (i) Gaussian Mixture models and (ii) i-vector models. The GMM (equation 3.1) is a probabilistic model on the feature space. The features used here are short time energy concatenated with 19 MFCC features and their first and second derivatives in a 60 dimensional feature space. The similarity between GMMs is based on cross likelihoods of model of one segment fitting the data in the other. GMM for a segment is trained on the feature vectors of the segment using the Expectation-Maximization algorithm to obtain a diagonal covariance GMM of size 32. While evaluating the system using GMM speaker models, the CLR and NCLR distances have been tested along with HAC and ILP clustering algorithms. To obtain i-vectors, first a speech Universal Background Model (UBM) is trained on a training data. The UBM is a GMM with large number of Gaussians, so that it captures all possible variability's in speech in the feature space. In the proposed system the TIMIT and TIFR datasets have been used for the UBM training. The TIMIT set consists of 168 speakers uttering 10 English sentences each while the TIFR set consists of 100 speakers uttering 10 Hindi sentences each, both from native speakers of the respective languages. The UBM is a diagonal covariance GMM of size 512. UBM training is a onetime computation. The UBM is mean-adapted for the feature vectors of the concerned segment to obtain a GMM for the segment. The means of the UBM and the adapted segment GMM are concatenated together to get a 30720 sized super vectors (60x512). The Total Variability space is a subspace of the GMM superspace that captures all the speaker and channel related information. T is the low rank matrix whose columns span the Total variability subspace. For the proposed system, the matrix T is trained using the same speaker labelled dataset used for UBM training. The T matrix training is also a onetime computation. The i-vector of the segment is the projection of the GMM super vectors onto the Total Variability subspace. Thus for every segment, extraction of the i-vector x involves 2 steps {adapting the UBM to obtain its GMM super vector and extracting the factors of the total variability eigenvectors to get. The algorithm for training the T matrix from speaker labelled training data is detailed. The proposed system uses the MSR Identity toolbox for UBM training, training of the TV subspace and the i-vector extraction. Figure presents the experiments on the traditional hierarchical clustering using GMM speaker models and i-vector speaker models respectively. The next section presents the experiments using the ILP clustering algorithm on the GMM speaker models and i-vector speaker models in that order. The experiments presented below were performed on the NDTV dataset. The DER values presented are the overall diarization error rates, which are averages of the individual DER per episode weighted with the duration of the episode.

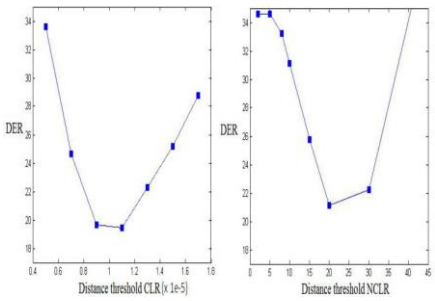

Fig.2: DER on NDTV dataset: HAC with distance threshold for GMM speaker models

Using HAC with i-vectors. New i-vectors were extracted for every segment obtained in the cluster merging step. The best result obtained was 16.69% DER for 75 dimensional TV space with the Mahalanobis distance.

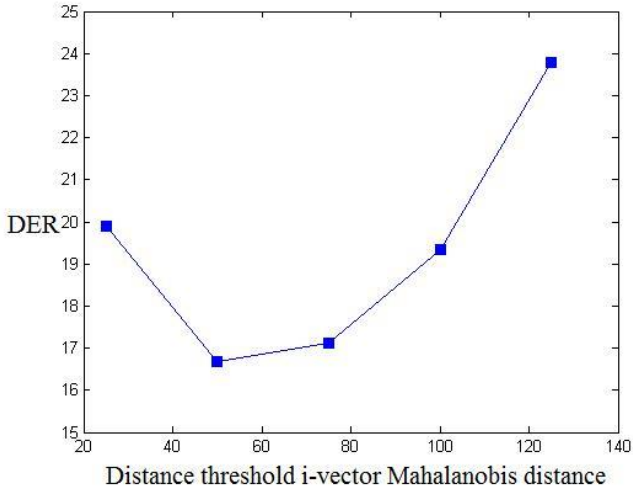

Fig.3: DER on NDTV dataset: HAC with distance threshold for GMM speaker models

The Integer Linear Programming formulation on the other hand offers a holistic trajectory to reach the optimum clustering. To verify this, the ILP formulation was implemented for the CLR and NCLR similarity matrix generated using the GMM speaker models and it gives an 11% relative improvement in the error compared to the best error from the GMM-HAC clustering algorithm. In literature the ILP has only been tried using i-vectors.

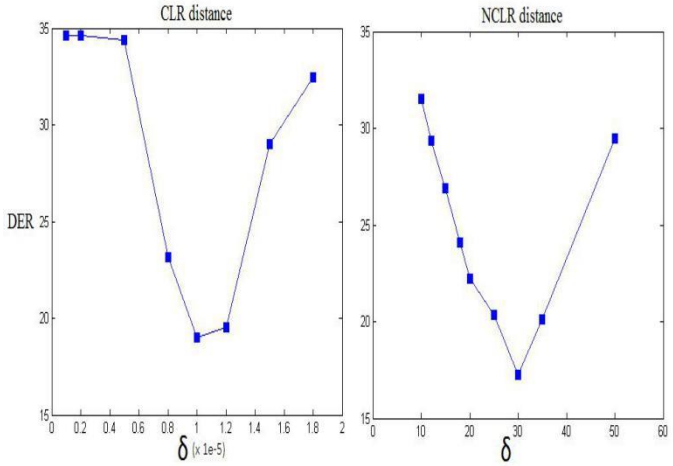

Fig.4: DER on NDTV dataset: ILP with distance threshold for GMM speaker models

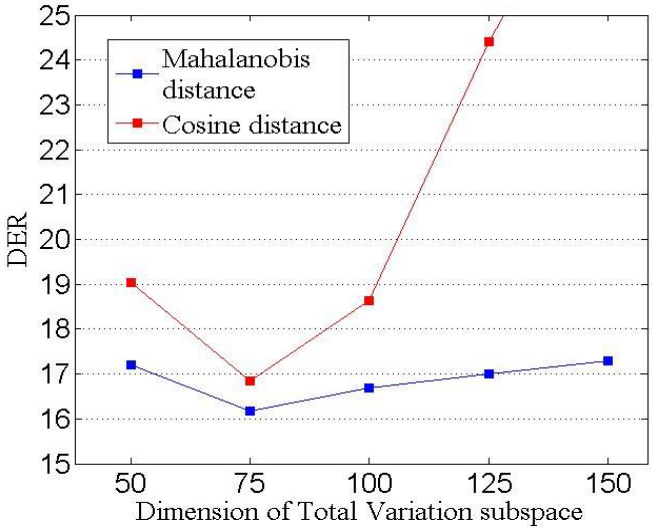

Fig.5: Performance of ILP clustering with i-vector speaker models with varying dimensions of the Total Variability subspace. Red plot is for the Mahalanobis similarity. Blue plot for the Cosine similarity.

Table 1: Best results from the 2 speaker models and 2 clustering algorithms

	HAC	ILP
GMM	19.45	17.27
i-vector	17.11	16.18

Previously indicated results for the dev0 subset of the REPERE show a 17.19% DER with GMM speaker models and 15.46% DER with i-vector speaker models. For the dev0 subset, we achieved a 23.19% DER with the HAC-GMM clustering and a 21.02% DER with ILP-i-vector clustering. The poorer performance compared to the previously attained results could be because of smaller sized UBM models (2048 as used by LIUM [25]). The overall DER for the 60 hour REPERE corpus is best for ILP-i vector clustering combination i.e. 24.4%

V. CONCLUSION

Experiments were performed on two broadcast news corpora {Indian news dataset from NDTV and the French REPERE corpus. The NDTV corpus is a 4h15m dataset from one news show. This dataset was manually annotated for the diarization experiments. The REPERE dataset of 60h04m was obtained from the French ELDA. The system is capable of performing speech activity detection without dependence

on external training data for non speech and speech models. Frame energy and zero crossing rate have been used as bootstrapping features to construct silence and music models from the audio recording being processed. Competitive speech activity detection has been achieved with two-stage SAD system silence detection, followed by music detection. The results are comparable to a state-of-the-art GMM-HMM based speech activity detector which uses external training data from a large dataset for creating non speech models. The i-vector speaker models, which are now state-of-the-art in speaker verification, provide a low dimensional representation of the speaker information compared to traditional GMM speaker models. They also offer a computational advantage since distance computation between i-vectors is much faster compared to cross-likelihood based similarity computation on GMM speaker models. Hence for real-time diarization systems, i-vectors seem more appealing.

VI. REFERENCE

[1]. Inside the secret technology that makes `the daily show' and `last week tonight' work, http://splitsider.com/2015/03/inside-the-secret-technology-that-makes-the-daily-show-and-last-night-tonight-work/.

[2]. Gerald Friedland, Luke Gottlieb, and Adam Janin. Joke-o-mat: browsing sitcoms punchline by punchline. In Proceedings of the 17th ACM international conference on Multimedia, pages 1115{1116. ACM, 2009.

[3]. Sue E Tranter, Douglas Reynolds, et al. An overview of automatic speaker diarization systems. Audio, Speech, and Language Processing, IEEE Transactions on, 14(5):1557{1565, 2006.

[4]. Xavier Anguera Miro. Robust speaker diarization for meetings. Universitat Politecnica de Catalunya, 2007.

[5]. Pyannote - collaborative annotation of audio-visual documents, http://pyannote.github.io/.

[6]. Juliette Kahn, Olivier Galibert, Ludovic Quintard, Matthieu Carre, Aude Giraudel, and Philippe Joly. A presentation of the repere challenge. In Content-Based Multimedia Indexing (CBMI), 2012 10th International Workshop on, pages 1{6. IEEE, 2012.

[7]. Nist: The nist rich transcription 2009 (rt'09) evaluation.

[8]. Xavier Anguera Miro, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Fried-land, and Oriol Vinyals. Speaker diarization: A review of recent research. Audio, Speech, and Language Processing, IEEE Transactions on, 20(2):356{370, 2012.

[9]. Gerald Friedland, Adam Janin, David Imseng, Xavier Anguera Miro, Luke Gottlieb, Marijn Huijbregts, Mary Tai Knox, and Oriol Vinyals. The icsi rt-09 speaker diarization system. Audio, Speech, and Language Processing, IEEE Transactions on, 20(2):371{381, 2012.

[10]. Martin Zelenak, Henrik Schulz, Francisco Javier Hernando Pericas, et al. Albayzin 2010 evaluation campaign: speaker diarization. 2010.

[11]. Sylvain Meignier and Teva Merlin. Lium spkdiarization: an open source toolkit for diariza-tion. In CMU SPUD Workshop, volume 2010, 2010.

[12]. Simon Bozonnet, Nicholas WD Evans, and Corinne Fredouille. The lia-eurecom rt'09 speaker diarization system: enhancements in speaker modelling and cluster puri cation. In Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, pages 4958{4961. IEEE, 2010.

[13]. T Nguyen, H Sun, S Zhao, SZK Khine, HD Tran, TLN Ma, B Ma, ES Chng, and H Li. The iir-ntu speaker diarization systems for rt 2009. In RT'09, NIST Rich Transcription Workshop, May 28-29, 2009, Melbourne, Florida, USA, volume 14, pages 17{40, 2009.

[14]. Arlindo Veiga, Carla Lopes, and Fernando Perdig~ao. Speaker diarization using gaussian mixture turns and segment matching. Proc. FALA, 2010.

[15]. Hari Krishna Maganti, Petr Motlicek, and Daniel Gatica-Perez. Unsupervised speech/non-speech detection for automatic speech recognition in meeting rooms. In Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, volume 4, pages IV{1037. IEEE, 2007.

[16]. Wai Nang Chan, Tan Lee, Nengheng Zheng, and Hua Ouyang. Use of vocal source features in speaker segmentation. In Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on, volume 1, pages I{I. IEEE, 2006.

[17]. Sree Harsha Yella, Andreas Stolcke, and Malcolm Slaney. Arti cial neural network features for speaker diarization. In Spoken Language Technology Workshop (SLT), 2014 IEEE, pages 402{406. IEEE, 2014.

[18]. Neville Ryant, Mark Liberman, and Jiahong Yuan. Speech activity detection on youtube using deep neural networks. In INTERSPEECH, pages 728{731, 2013.

[19]. Xavier Anguera and Jean-Francois Bonastre. Fast speaker diarization based on binary keys. In Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on, pages 4428{4431. IEEE, 2011.