International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Robust Speaker Diarization Based on Daubechies Wavelet, Nonlinear Energy Operator and Pyknogram

Sukhvinder Kaur, J.S. Sohal, Amit Gupta

Abstract: ***Two common disciplines of speech processing are speaker recognition "identification and verification of speaker", and speaker diarization, "who spoke when". Motivated by various applications in automatic speaker recognition, speaker indexing, word counting, and audio transcription, speaker diarization (SD) becomes a significant area of signal processing. The basic designing steps of SD are feature extraction, voice activity detection (VAD), segmentation, and clustering. VAD process is accomplished by Daubechies 40, discrete wavelets transform (DWT). Initially, DWT was used for compression, scaling, and denoising of audio-stream and then partitioned into small frames of size 0.12 seconds. Next, features of each frame were extracted by applying nonlinear energy operator (NEO) based pyknogram. To measure the similarity between frames, a sliding window on delta-BIC distance metric was applied. A negative value of its output represents the same segments and vice-versa. To improve the output of the segmentation process, resegmentation was applied by information change rate method. At last, hierarchical clustering groups the homogeneous segments that correspond to a particular speaker and has been graphically represented by the dendrogram. The performance of SD was evaluated by F-measure and speaker diarization error rate (SER) and their results were compared with the traditional speaker diarization system that uses MFCC and BIC for segmentation and clustering. It reveals a significant reduction of 12.3% of SER in the proposed diarization system.***

Keywords: Bayesian information criteria, Dendrogram, Diarization Error Rate, Pyknogram

INTRODUCTION

Speaker diarization (SD) is the problem of determining "who spoke when" in a multiparty audio speech when the number and identities of the speakers are unknown. Motivated by various applications in automatic speech/speaker recognition (SR), speaker indexing, speaker counting, word counting and captioning of TV Shows, SD becomes an important discipline of speech processing over the past decade. An evaluation of SD is explored in [1].

According to this text, the scientific community has evolved research on speaker diarization with phone records, broadcast information in past 1990's and early 2000's. Interest within the assembly domain grew considerably from 2002 with the launch of several related studies projects on multimodal technology. SD become evaluated for Broadcast information statistics in English up to 2004, and the meeting domain became the principle consciousness of NIST critiques on account that 2005. The latest NIST RT assessment was held in 2009[2]. The speaker diarization machine became first developed at the International Computer Science Institute (ICSI) and used inside the NIST RT assessment. It encourages the studies in several automated speech technologies and affords not unusual datasets for assessment of overall performance. Commonly used steps for diarization described in [3]are feature extraction, voice activity detection (VAD), segmentation and clustering and depicted in Fig.1. This device can be used for the counting of speakers who collaborated in a conversation (the maximum possible without having a priori data on any of the speakers). It can be used for the detection of criminal activities; For example, in prisons, the call on three sides is illegal, detecting the presence of a third speaker in recorded conversations can be useful for identifying offenders.

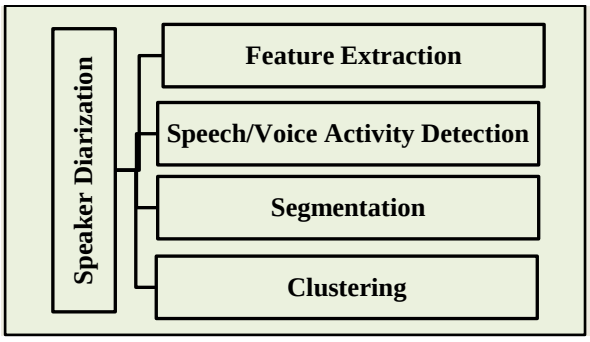

Fig.1. Basic steps of speaker diarization.

A. Feature Extraction

The audio processing starts with the pre-processing module for the filtering of noise and followed by extraction of features of acoustic signal like Mel frequency cepstral coefficient (MFCC), spectrogram, pitch, power, zero-crossing rate (ZCR), Discrete Wavelet Transform (DWT) and Teager Kaiser Energy Operator (TKEO). Mel-frequency cepstral coefficient [4]is broadly used as a feature in speech recognition.

Manuscript published on November 30, 2019.
* Correspondence Author
Sukhvinder Kaur*, Electronics and Communication Engineering, Research Scholar, IKG Punjab Technical University, Punjab, India.
J.S Sohal, LCET, Ludhiana, Punjab, India
Amit Gupta, Electronics and Communication Engineering IKGPTU, Punjab, India

Mel, an abbreviation of the word melody, is a unit of the pitch. Feature extraction can be stood as a step to lessen or compress the dimensionality of the input speech, a discount which inevitably ends in a few information losses. DWT has been extensively implemented to examine a signal concurrently in the time- frequency domain and widely used as compression and de-noising technique to improves device performance [5]. Teager Kaiser Energy Operator (TKEO) is a powerful nonlinear operator than traditional energy measure proposed by Kaiser which successfully used to detect the frequency and/or amplitude variations in a speech [6][7].

B. Voice Activity Detection (VAD)

VAD identifies speech and non-speech (silence, laughing, noises, and breathing, coughing

regions stated by [8]. It is performed using a Gaussian mixture model (GMM) trained on

speech and non-speech regions. Audio recording can also be segmented into speech and non-speech data by using Discrete wavelet transform and variance spectral flux (VSF)[9].

C. Speaker Segmentation and Clustering

Vocal regions are divided into speaker segments using segmentation algorithms. In general, it can be classified into three categories based on models, based on metrics and hybrid segmentation described in [10]. Model-based methods require training data to initialize speaker models. For example, the Universal Background Model (UBM), Hidden Markov Model (HMM), Gaussian Mixture Model (GMM). Methods based on metrics do not require any prior knowledge about the number of speakers, their identity or the characteristics of the signal. For example, the weighted square Euclidean distance, the divergence Kullback-Leibler divergence Kullback-Leibler-2 generalized likelihood ratio (GLR), Bayesian Information Criteria (BIC) [11], Cross likelihood ratio(CLR) and, Normalized cross likelihood ratio(NCLR)[13]. Hybrid algorithms are a combination of metrics and models based techniques. Execution of segmentation can be accurately evaluated by F-measure as described [14]. Due to the rapid increase in the recorded vocal volume available for the computer, the revised techniques for segmentation and speaker classification are very complicated. To address this discomfort, a new distance measure has been proposed that combines GLR with Information change rate (ICR)[2]. The latter compensates for the undesirable tendency of the previous one and, consequently, plays a vital role as a complementary criterion.

In this paper, a robust SD system model is proposed in which an improved spectrogram based on nonlinear energy has been used for feature extraction and the Bayesian information criteria and information change rate methods for segmentation. Furthermore, the homogeneous segments have been grouped by the application of hierarchical clustering and their performance was evaluated by the diarization error rate (DER).

The rest of this article is organized as follows: Section 2 illustrates the research methodology of speaker diarization system, including feature extraction, speaker segmentation, re-segmentation, and clustering. Their results are analyzed and evaluated in section 3. Finally, a conclusion is given in the last section.

II. PROPOSED WORK

To detect the speaker change point, music, clapping sound, and overlapping speakers in an audio recording of multiple speakers and grouping of homogeneous segments are very complex while designing a diarization model. Inadequate speaker detection and audio segmentation generate a significant amount of errors in most current systems. During clustering, four types of errors occur: false speech alarm (percentage of speech in the hypothesis, but not the fundamental truth), lost speech (percentage of word in fundamental truth, but not the hypothesis), speaker Error (percentage of voice assigned to 'wrong speaker) and overlapping speeches (when it is assumed that the wrong number of speakers speak when multiple speakers speak at the same time). The sum of all these errors constitutes the diarization error rate (DER), as shown in Table 2. So, for the demonstration of robust speaker diarization (SD) following modules are used:

- Data collection of the audio stream
- Ground truth analysis
- Design, development, and analysis of the model of SD System.
- Performance evaluation using DER.

A. Database Used

The recorded speech data was collected by using Smartphone and audacity software tool and its detail is revealed in Table 1. To use it in MATLAB software the recording must be in .wav form so, the recorded speeches are converted into .wav form by using a converter freely available on website www.zamzaar.com. It contains three recordings of multiple speakers of TV news and TV shows which were used for the demonstration and testing of the proposed speaker diarization system. The duration of recordings ranges from 2 to 8 minutes.

Table-1: Description of datasets used for the development and analysis of SD

Parameter	Value
Sampling frequency	44100Hz. 16 bits
Database-1	Recording of 3.5 minutes. Seven Speakers with clapping sound and music.
Database-2	Recording of 2.5 minutes. Two speakers with clapping sound.
Database-3	Recording of 8 minutes. Eight Speakers with music.

B. Ground-Truth Analysis

The references to the fundamental truth to evaluate the diarization of the audio speech was initially obtained through manual labeling of acoustic data, however, the high variations between the different labeling machines proved to be problematic [1]. Therefore, more recently, an automatically generated

Table-2: Performance evaluation measure of the speaker diarization system by mapping between reference speaker and hypothesis speakers.

Reference	X				Y		X & Y
Hypothesis	Speaker 1	Speaker 2				Speaker 3	Speaker 1
Error	Correct	Speaker Error	False Alarm	Correct	Missed Speech	Correct	Overlapping Speech
DER = Speaker Error + False Alarm + Missed Speech + Overlapping Speech Error							

forced alignment was used to extract more reliable starting and arrival points from the speakers using the automatic speech recognition system (ASR), transcriptions created by humans and audio from individual microphones heads (IHM). The standard performance metric of the SD system is the rich transcription metric of RTIST RT-09 known as the diarization error rate (DER) [11][12]. It is calculated by finding the optimal match between the true speaker and the hypothetical speaker and then calculating the percentage of time that is erroneously assigned based on the optimal match. Therefore, it is necessary to have databases of meetings that are accurately transcribed into segments of speakers to get their fundamental truth. In this research work, to find the start and end times of the vocal segments with the speaker labels, the audio files were transcribed using markers from the Signal Processing Tool (SPTOOL) in MATLAB, described in Fig. 2. The manual annotation of audio recording as described in Fig. 3 includes speaker change point, clapping, and overlapping speech and music.

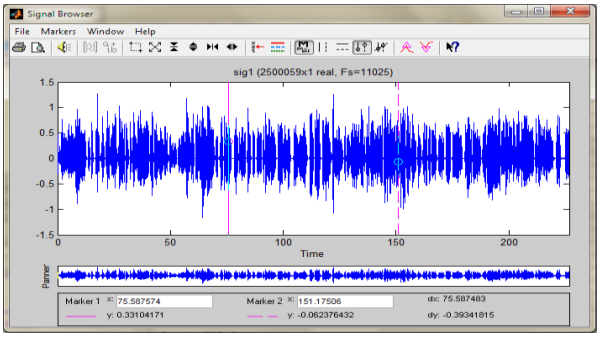

Fig.2. Transcription of the recording of Dr. Subhash Chandra show using SPTOOL.

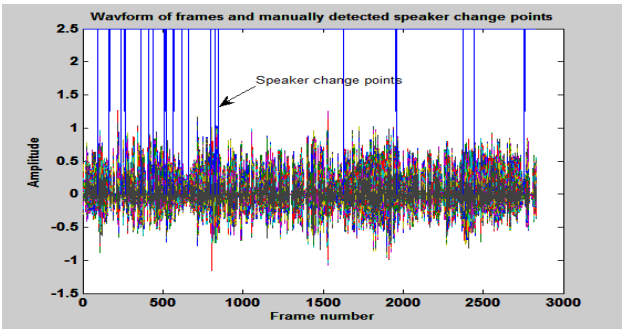

Fig.3. Detection of speaker change points and its corresponding frame number using SPTOOL.

C. Design and Development of Speaker Diarization Model

The goal of this research is to determine "who spoke when" in multiparty speech recordings and to reduce DER. Since, the recorded signals is comprised of clean speech with commercial music and clapping sound with no overlapping speech and hence have no overlapping speech error. Model of proposed speaker diarization system which is similar to the standard agglomerative clustering framework described in [15] is depicted in Fig. 4, except the following main modifications to remove false alarm and missed speech errors.

- Voice activity detection (SAD) is used to pre-segment the audio stream into speech and non-speech region which is accomplished by discrete wavelet transform (DWT) to remove false alarm and missed speech errors [6]. It also successfully resolves high-frequency components of speech signal within a small time window and low-frequency components in large time windows. The wavelet transform is defined as the inner product of a signal x(t) with the mother wavelet ψ(t) is as follows:

$$W_{\psi}x(a,b)= \frac{1}{\sqrt{a}}\int_{\infty}^{-\infty} x(t)\psi^{*}_{a,b}(t)\,dt, \quad (1)$$

$$\text{Where, } \psi_{a,b}(t)=\psi\left(\frac{t-b}{a}\right) \quad (2)$$

Where *a* and *b* are scale and shift parameters respectively. Varying these parameters mother wavelet can be dilated or translated.

In this research, metric-based segmentation is used, which do not require any training data and it yields a high recall rate at a moderate precision rate [11]. Since there is no overlapping speech in the audio recordings so, there is no overlapping error. The false alarm and missed speech errors will be removed by applying daubechies 40 DWT. So, in this proposed work the only error of DER which is to be computed is speaker diarization error rate (SER).

1) The Designing of Steps of Proposed Speaker Diarization

i. A multiparty audio recording of the TV shows and TV News is taken in MP3 format. It is converted into .wav form by using freely available software at www.zamzaar.com.

ii. The .wav file of audio recording is read in the MATLAB software for processing. To increase the speed of processing of long .wav files, first it is compressed in the ratio of 1:4 and denoised using the wavelet transform technique with help of Daubechies wavelet (db 40) at level-2. It decomposes the audio signal into two halves: approximations which contain low-frequency component and second is details which contain high-frequency component.

iii. About 99% of speech information is present in approximation coefficients.

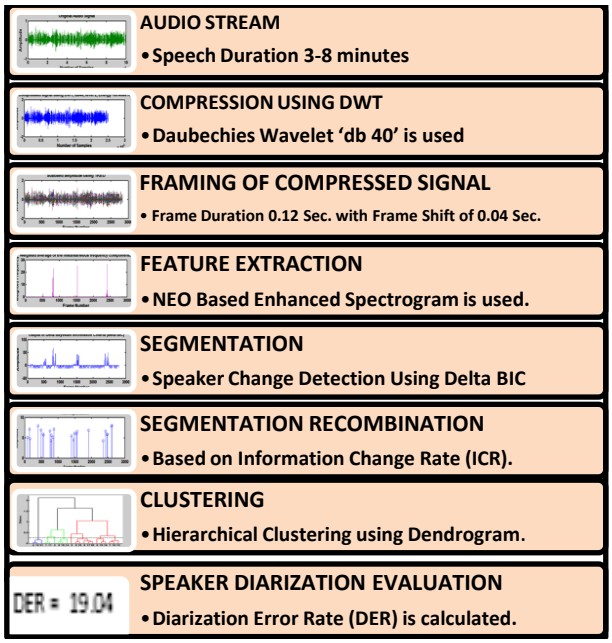

Fig.4. Proposed model of speaker diarization system.

iv. Framing of compressed signal (approximation coefficients): Small frames of duration 0.12 seconds with a frame shift of 0.04 sec are produced. (Number of frames =1700 to 6500).

v. Features of each frame are extracted using NEO based enhanced spectrogram and MFCC (traditional method).

vi. To detect the speaker change point or segmentation boundaries in the audio recording, the distance between extracted features is calculated by using Bayesian information criteria and sliding window. The value of its output may be positive or negative. The negative output corresponds to a similar speaker. The Peak value shows segmentation boundaries. After applying threshold value to these peak values, segments are produced and decoded to its original state. In this research, window length is of 50 samples with sliding of 15 samples.

vii. For Resegmentation, again features of segments are extracted by using DWT at level 1 with Daubechies wavelet (db4). The distance between to segments is determined by Bayesian information criteria based information change rate (ICR). The distance between various segments is sorted and the threshold value is applied for its purification. Its performance is evaluated by F_measure.

$$\text{F-Measure} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall}+\text{Precision}} \quad (3)$$

viii. For making clusters of homogeneous segments hierarchical clustering using dendrogram is used.

ix. Finally, the performance results were evaluated by speaker diarization error rate (SER) instead of overall diarization error (DER) which is expressed in terms of missspeech, false alarm and SER. Its computation requires two files: hypothesized file (designed outputs) and reference file (Ground truth) as discussed in the previous section. The speaker diarization error is defined as [13]:

$$SER = \frac{\text{Number of frames incorreclty speaker labeled}}{\text{Number of speech frames}} \quad (4)$$

2) Description of various algorithms used in various stages of proposed speaker diarization of the multiparty audio stream is as follows.

a) Proposed Feature Extraction Algorithm: Enhanced spectrogram based on NEO is used as features of frames of compressed speech signal due to its high frequency and amplitude resolution. The enhanced spectrogram, also called Pyknogram, was first introduced in [16] to facilitate formant tracking and are calculated by applying multiband demodulation in the framework of the AM-FM modulation model described in [7]. Overlaps in speech data can be detected by using Pyknogram and is used in [17]. In Pyknogram, the resonances (formants) and harmonic structure of speech are enhanced by decomposing the spectral sub-bands into amplitude and frequency components. It is based on a powerful nonlinear energy operator (NEO),which is successfully used to detect the frequency and/or amplitude variations in a speech and computed as:

$$\psi[x(n)] = x^2(n) - x(n-1)x(n+1) \quad (5)$$

The frequency (f) and amplitude (|a|) components of a given subband, x(n), are calculated using equation (5) as follows:

$$f = \frac{1}{2\pi}\arccos\left(1 - \left(\frac{\psi|x(n)-x(n-1|}{2\psi|x(n)|}\right)\right) \quad (6)$$

$$|a| = \sqrt{\frac{\psi|x(n)|}{\sin^2(2\pi f)}} \quad (7)$$

The weighted average of the instantaneous frequency components is used to derive a brief-time estimate rate forthe dominant frequency in each subband, in this case the length of a time-period is 12 msec.

$$F_w(t) = \frac{\sum_t^{n+T} f(n)a^2(n)}{\sum_t^{n+T} a^2(n)} \quad (8)$$

Where f(n) and a(n) are the immediate frequency and amplitude functions calculated for each sample in the t_{th} frame over the frame length (T samples consistent with frame). Logarithmic value of $F_w(t)$ improves the results. Fig.5 shows the frames of compressed signal and weighted average of instantaneous frequency.

b) Speaker Segmentation: The sudden change at the output of detected features in the previous section corresponds to speaker change points. These speaker change points were refined by traditional Bayesian Information Criteria and are calculated as:

$$\Delta BIC = (N1+N2)*\log(\sum) - N1*\log(\sum 1) - N2\log(\sum 2) - \lambda*(0.5(d+0.5(d+1)))*\log(N) \quad (9)$$

Where λ is a penalty weight, d is a dimension of the feature space, N_1, N_2 and N_1+N_2 are model sizes and $\sum_1$, $\sum_2$ and $\sum$ are determinants of covariance matrices for the segments X_1 , X_2, and X respectively.

.

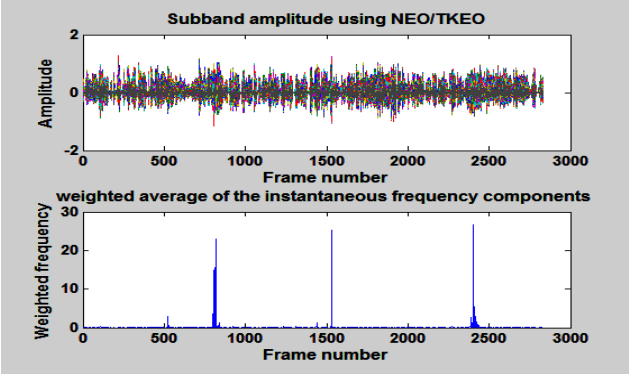

Fig.5. Frames of compressed signal and weighted average of the instantaneous frequency component.

Fig. 6 represents the output of delta BIC of frames of compressed signal and it is concluded that negative values of ΔBIC represent the same speaker and highest positive value detected at time *t* is the speaker change point

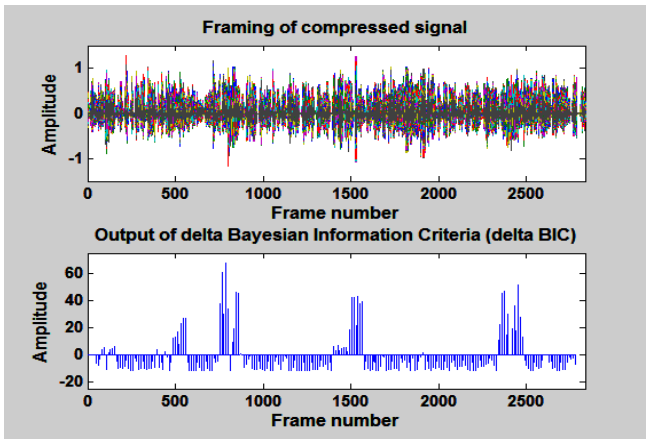

Fig.6. Frames of compressed signal and output of delta BIC

c) Speaker Resegmentation and Clustering: The Information Change Rate (ICR) or Entropy is used to characterize the similarity of two neighboring clusters. It determines the change in information that would be obtained by merging any two clusters Cx and Cy under consideration and is given as follows:

$$\mathrm{ICR(Cx, Cy)} \triangleq \left(\frac{1}{(Ncx + Ncy)}\right) * \ln BIC(Cx, Cy) \qquad (10)$$

Where Ncx and Ncy are the number of features in clusters Cx and Cy respectively. This is a statistical measure between cluster represents how much entropy would be increased by merging the clusters considered. ICR should be small when the clusters considered are homogeneous in terms of speaker characteristics and each cluster is large to fully cover the intra speaker variance of the corresponding speaker identity. In this research, after segmentation, their features were again detected by DWT and then go for final segmentation by using ICR. Its results are depicted in Fig. 7. Comparison of hypothesized frames of final segmentation and reference frames (ground truth) are shown in Fig. 8 This figure was used to calculate precision, recall, F_measure, and SER. Finally, Hierarchical clustering groups the homogeneous segments over a variety of scales by creating a cluster tree or dendrogram.

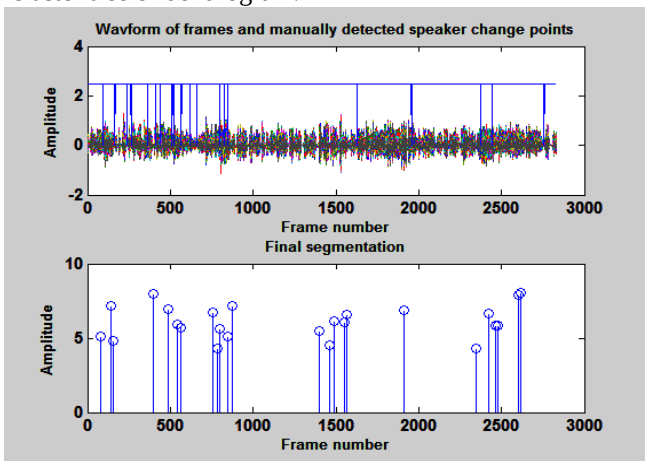

Fig.7. Frames with manual segmentation and hypothesized segments.

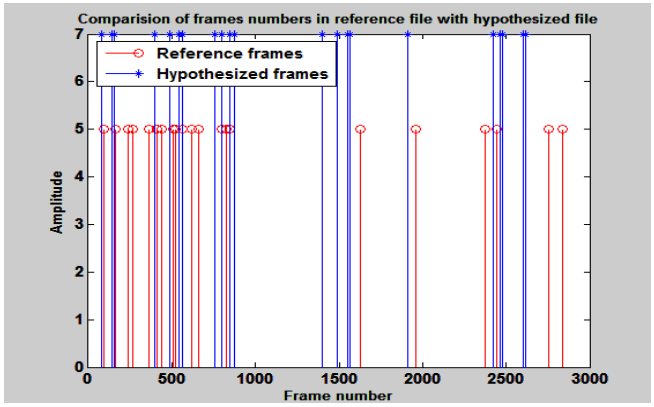

Fig.8. Comparisons of frames of the reference file and hypothesize file.

III. RESULT AND DISCUSSION

For the performance evaluation of speaker change detection (SCD), it is observed that whenever speaker stops for more than 2 seconds while talking, six change points were detected as in case of frame numbers 850 to 1600 shown in Fig.8. Also if the speaker speaks for a very small duration of 2-4 seconds, it can't be detected by this system. Its response is more than 95% if silent is removed and the speaker speaks for more than 6 seconds. The performance of SCD is measured by F-measure and shown in Table 3. It is clear from the results that F-measure is improved by using Pyknogram with BIC and ICR as compared to MFCC with BIC. Also, the proposed speaker diarization system used hierarchical clustering to group the homogeneous speakers into clusters by using Dendrogram function in MATLAB. It plots the cluster tree which consists of many U-shaped lines connecting objects in a hierarchical form as depicted in Fig.9. The height of each U represents the distance between the two objects being connected. Each leaf in the dendrogram corresponds to one cluster. The figure shows that at distance 2.5 numbers of clusters/speakers are 8.

Table-3: Computation of F-measure and SER for various databases in SD System.

Speaker Change Detection Method		Recall (%age)	Precision (%age)	F-measure (%age)	SER (%age)
Pyknogram with BIC and ICR *(Proposed Method)*	Database 1	76.19	80	78.05	19.04
	Database 2	80	80	80.00	16.66
	Database 3	62.5	100	76.92	12.3
MFCC and BIC *(Traditional Method)*	Database 1	55.39	51.78	53.52	22.46
	Database 2	61	64	62.46	19.74
	Database 3	49.1	75	59.35	14.2

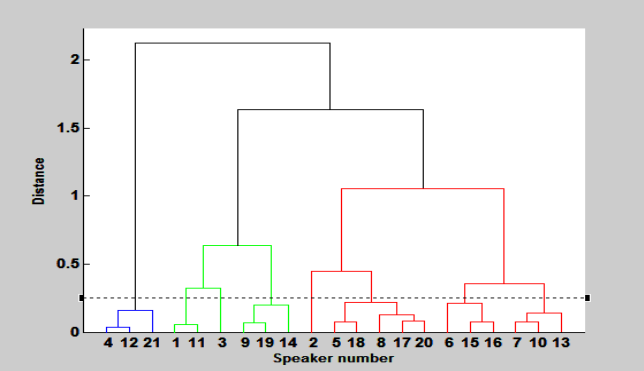

Fig.9. Hierarchical clustering of homogeneous speakers.

Overall performance of SD is evaluated by speaker misclassification error or speaker diarization error rate (SER) by applying eq. (4). It is obtained by matching hypothesized speaker to the true speaker names in the reference file. To accomplish this, a one-to-one mapping of the reference speaker IDs to the hypothesis speaker IDs is performed to maximize the total overlap of the reference and (corresponding) mapped hypothesis speakers. When the SER calculation is performed, a time window of 0.5s around each changing point is excluded (i.e. Errors that take place within the 0.25s on either side of a given changing point are not taken into account). The SER of the proposed system at distance 2.5 marked in Fig.9 is 19.04 which is comparable with the standard diarization system using MFCC and BIC shown in Table 3. Least SER is obtained by analyzing database 3 with the proposed method. Two more test databases were applied to the same proposed SD system, which also shows better results than by using the traditional method.

IV. CONCLUSION AND FUTURE SCOPE

In this research work, an optimized model of speaker diarization (SD) system is proposed to reduce the diarization error rate (DER) in the multiparty audio stream by enhancing the existing algorithms. The main phase of SD is feature extraction which is accomplished by applying DWT and enhanced spectrogram based on nonlinear energy operator (NEO) on speech signal. The use of DWT removes the non-speech signals so; there is no need of computing false alarm and missed speech errors. For speaker change point detection delta BIC is applied successfully and is evaluated by precision, recall and F-measure. Moreover, Information change rate (ICR) algorithm is used for re-segmentation and at last hierarchical clustering algorithm make clusters of homogenous speakers. These algorithms were applied on three different databases and its performance was evaluated by speaker diarization error rate (SER). To reveal the effectiveness of this research, we have also implemented the speaker diarization system based on traditional features and clustering criteria and it was tested using the same dataset. This implementation is equivalent to the current-state-of-the art speech diarization approaches and serves as the baseline for performance comparisons. Performance scores of two systems show that the proposed system gives better results as compared to the traditional system. Lowest SER value is 12.3 which are obtained with database 3 due to the long duration of speeches in TV news.

To detect and handle speeches of duration less than 5 seconds, overlapping speech in the multiparty audio stream and effective clustering algorithm are the challenges which directly affect the performance of SD System. These challenges are the future scope of this research.

REFERENCES

1. X. A. Miro and S. Bozonnet, “Speaker Diarization: A Review of Recent Research,” IEEE Trans. Audio, Speech Language Processing, 20, 2012, 1–15.
2. M. H. Moattar and M. M. Homayounpour, “A Review on Speaker Diarization Systems and Approaches,” Speech Communication, 54(10), 2012, 1065–1103.
3. N. Evans et. al., “A Comparative Study Of Bottom-Up and Top-Down Approaches to Speaker Diarization,” IEEE Trans. Audio, Speech Language Processing, 20(2), 2012, 382–392.
4. A. Klautau, “The MFCC,” 2005, 1–14.
5. J.-D. Wu and B.-F. Lin, “Speaker Identification using Discrete Wavelet Packet Transform Technique with Irregular Decomposition,” Expert Systems Applications, 36(2), 2009, 3136–3143.
6. M. Bahoura and J. Rouat, “Wavelet Speech Enhancement based on The Teager Energy Operator,” IEEE Signal Process. Letters, 8(1), 2001, 10–12.
7. P. Maragos, J. F. Kaiser, T. F. Quatieri, “Energy Separation in Signal Modulations with Application to Speech Analysis,” IEEE Transaction on Signal Processing, 41(10), 1993, 3024–3051.
8. M. W. Mak and H. B. Yu, “A Study of Voice Activity Detection Techniques for Nist Speaker Recognition Evaluations,” Computer Speech Language, 28(1), 2014, 295–313.

9. R. Huang, J. H. L. Hansen, "Advances in Unsupervised Audio Classification and Segmentation for the Broadcast News and NGSW Corpora,"IEEE Transactions on Audio, Speech, and Language Processing, 14(3), 2006, 907–919.
10. M. Kotti, V. Moschou, and C. Kotropoulos, "Speaker Segmentation and Clustering," Signal Processing, 88(5), 2008, 1091–1124.
11. J. Ajmera, I. Mccowan, and H. Bourlard, 2004, "Robust Speaker Change Detection,"IEEE Signal Processing Letter, 11(8): 649–651.
12. M. Sinclair and S. Kingt, "Where Are The Challenges In Speaker Diarization? Mark Sinclair*, Simon Kingt The Centre for Speech Technology Research, The University of Edinburgh, UK,", 2013, 7741–7745.
13. Le, V.B., Mella, O. & Fohr, D., "Speaker Diarization using Normalized Cross Likelihood Ratio", Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2, 2007, 873–876.
14. Trec, "Tutorial: Common Evaluation Measures," NIST, 2006.
15. X. A. Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker Diarization: A Review of Recent Research," IEEE Trans. Audio, Speech Lang. Process., 20(2), 2012, 356–370.
16. Potamianos and P. Maragos, "Speech Formant Frequency and Bandwidth Tracking using Multiband Energy Demodulation," Acoustical Society of America, 99(6), 1996, 3795–3806.
17. N. Shokouhi, A. Ziaei, A. Sangwan, and J. H. L. Hansen, "Robust Overlapped Speech Detection and its Application in Word-Count Estimation For Prof-Life-Log Data, IEEE International Conference on Acoustics, Speech And Signal Processing, No. 978, 2015, 4724–4728.

AUTHOR PROFILE

Sukhvinder Kaur, received the BE degree in Electronics and Electrical Communication Engineering from Punjab University, Chandigarh, India in 1995 and M.Tech degree from Punjab Technical University in 2010. She is currently defended the Ph.D. degree in the department of Electronics and Communication Engineering , I K Gujral PTU, Kapurthala, Punjab, India Her research interests include multimedia indexing and automatic speaker diarization.