International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 3 Issue: 9
ISSN: 2454-4248
52 – 58

Optimized Speaker Diarization System using Discrete Wavelet Transform and Pyknogram

Sukhvinder Kaur
Research Scholar, I.K. Gujral PTU, Jalandhar,
Kapurthala-144601, India
er1971sukhvinderkaur@rediffmail.com,

J. S. Sohal
Director, LCET, Ludhiana-141113, India
jssohal2001@yahoo.com

***Abstract*—**The aim of this paper is to present an optimized speaker diarization system that efficiently detects speaker change points in multispeaker speech data. Speaker diarization is the process to detect speaker turns and group together segments uttered by the same speaker. It can be used in speaker recognition, audio information retrieval, audio transcription, audio clustering, indexing and captioning of TV shows and movies. In this proposed technique, the daubechies 40-wavelet transform is used to compress the audio stream in the ratio of 1:4; their features are extracted by enhanced spectrogram called pyknogram based on Teaser Kaiser Energy Operator (TKEO). This method relies on resonances (formants) and harmonic structure of speech which are enhanced by decomposing the spectral sub-bands into amplitude and frequency components. The weighted average of the instantaneous frequency components are used to derive a short-time estimate value for the dominant frequency in each subband over a fixed period of time 0.12msec. Sudden changes in the dominant frequency correspond to the speaker change point and are detected by using traditional delta Bayesian Information Criteria (ΔBIC). This technique do not uses voice activity detection process (VAD). For re-segmentation, Information Change Rate (ICR) is used. Finally, hierarchical clustering algorithm make groups of homogeneous segments and are plotted by Dendrogram function in Matlab. The results are evaluated by F-measure and diarization error rate. It shows that the proposed method gives fast and better results as compared to traditional method with Mel frequency cepstral coefficient (MFCC) and Bayesian Information Criteria (BIC) algorithms.

Keywords*— Bayesian Information Criteria; Information Change Rate; MFCC; Pyknogram; Segmentation; Speaker Change Point; Teaser Kaiser Energy Operator; Wavelet Transform.*

*****

I. INTRODUCTION

Speaker diarization is the problem of determining "who spoke when" in a multiparty audio speech when the number and identities of the speakers are unknown[1]. Motivated by various applications in speaker indexing, speaker counting, automatic speech recognition, captioning of TV Shows and call routing, speaker diarization has been studied extensively over the past decade, and there are currently a wide variety of approaches including both top-down and bottom-up unsupervised clustering methods[2]. This research work is based on speaker indexing in which labels associating with speaker identities are assigned to different parts of an audio file. Feature extraction and segmentationfollowed by speaker clustering is also called speaker diarization as shown in Figure 1.The feature extraction converts the speech conversation into some parameterized representation. TheSegmentation and clustering of speakers in an audio recording are the two most challenging topics in speech processing. Speaker segmentation aims at finding speaker change points in an audio stream, whereas speaker clustering aims at grouping speech segments based on speaker characteristics. Model-based, metric-based, and hybrid speaker segmentation algorithms are reviewed in [3].A new speaker change detection method for two-speaker segmentation that requires neither a threshold nor existence of silence regions in the conversation is proposed in [4]. It assumes that one speaker initiates the conversation and he/she speaks for at least one second. Most commonly used criteria for speaker change detection like log likelihood ratio (LLR) and Bayesian information criterion (BIC) have an adjustable threshold/penalty parameter to make speaker change decisions. A criterion which can be used to identify speaker changes in an audio stream without such tuning of threshold or penalty is discussed in [5].

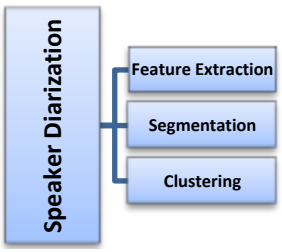

Fig 1 Basic Speaker Diarization System

For segmenting nonstationary signals of EEG, a new approach is discussed in [6] that uses nonlinear energy operator to accentuate the frequency and amplitude variations of the signal. A novel speaker segmentation and clustering algorithm is presented in [7]. This algorithm automatically performs both speaker segmentation and clustering without any prior knowledge of the identities or the number of speakers by using HMMs, agglomerative clustering, and the Bayesian Information Criterion. In this technique there is no need of threshold adjustment and training development data. A new online method [8] for speaker segmentation and clustering in real-world environments analyses and discusses the difficulties of online speaker diarization. It proposes a new segmentation and clustering method, in which the Bayesian information criterion (BIC) and the normalized cross-likelihood ratio

 ISSN: 2454-4248
52 – 58

(NCLR) are combined into an online speaker diarization system. There are three main application domains for speakerdiarization:Broadcast news, Meetings and Conversational telephone speech [2]. In this paper, the application domain for speaker diarization is TV show(i.e.Dr. Subhash Chandra show). It combines the featuresof Broadcast news and meeting domains as shown below:

- Numbers of speakers are4 to 10persons or more.
- Some parts of the file may containmusic or commercials.
- Variations in recording quality, including impulse noises,reverberation and variable speech levels may exist
- All the conversations take place in one place
- Average speaker change duration may be short
- Normally there are overlappingregions between speaker utterances and music.

The main application areas of this research are:

- Short speech utterances Diarization (rapid speaker change detection).
- High robustness to noise like music.
- Estimating number of speakers and its labelling.
- Overlapping speech detection.

This paper proposes an optimized method for speaker diarization by using discrete wavelet transform (DWT) for speech compression;Teaser-Kaiser Energy Operator (TKEO) based Enhanced spectrogram for feature extraction, Bayesian Information criteria (BIC) for speaker segmentation and Information change rate (ICR) with dendrogram for clustering. The results are evaluated by diarization error rate (DER). The remainder of this paper is organized as follows:Section 2 presents the speech compression technique. Section 3 describes feature extraction, section 4 speaker change detection method. Section 5 presents information change rate (ICR), hierarchical clustering. Section 6 introduces the proposed speaker diarization system and evaluation method. Section 7 discusses results and conclusion is given in section 8.

II. EFFICIENT COMPRESSION TECHNIQUE

Wavelet transforms have been applied to various research areas. Their applications include signal and image de-noising, compression, detection, and pattern recognition. The wavelet transform techniques and details of how to use them in speech compression is discussed in [9]. The more a wavelet concentrates energy in the approximation part of the coefficients, the better it is as a speech compressor. Voiced, unvoiced and mixed speech frames are determined in terms of how energy is concentrated into bands by the wavelet. For voiced frames, the energy is concentrated mostly in 2 bands, for unvoiced frames it spreads across the whole bandwidth, and for mixed frames it is confined in 3 to 5 bands. Optimum wavelets are selected based on energy conservation properties in the approximation part of the wavelet coefficients. Wavelet Transform (WT) is a modern parametrization method successfully used for some signal processing tasks[10].WT often out-performs parametrizations based on discrete time fouriertransform, due to its capability to represent the signal precisely, in both frequency and time domains. It is defined as the inner product of a signal $x(t)$ with the mother wavelet $\psi(t)$ as follows:

$$\psi_{a,b}(t)=\psi\left(\frac{t-b}{a}\right) \quad (1)$$

$$W_{\psi}x(a,b)=\frac{1}{\sqrt{a}}\int_{\infty}^{-\infty} x(t)\psi*\left(\frac{t-b}{a}\right)dt, \quad (2)$$

Where a and$b$ are scale and shift parameters respectively. Mother wavelet can be dilated or translated by changing a and$b$. The DWT functions at level m and time location t_m can beexpressed as :

$$d_m(t_m)= x(t)\ \psi_m\left(\frac{t-t_m}{2^m}\right) \quad (3)$$

Where, ψ_m is the decomposition filter at frequency level m. The effect of the decomposition filter is scaled by the factor 2^m at stage m, but otherwise the shape is the same at all stages. In this research work it is shown that the Daubechies 40 wavelet at level 2 concentrates more than 98% of the signal energy into the approximation part of the coefficients followed closely by the haar, daubechies 04, and daubichies12 wavelets as shown in Figure 2. So, it is used as an efficient method to compress the speech signal..

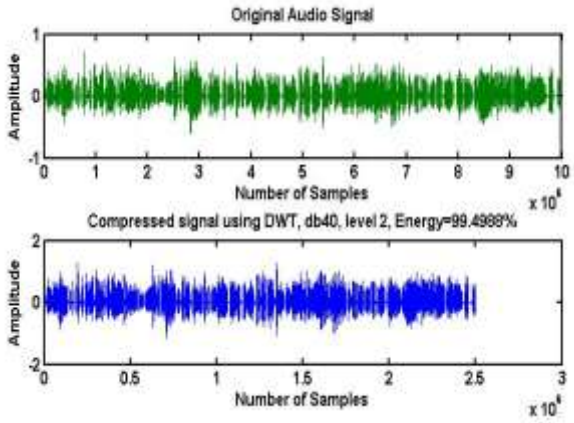

Fig.2 Waveform of Original Audio Signal and its Compressed Form

III. FEATURE EXTRACTION

In speaker change detection and overlapdetection process, Enhanced spectrogram based on TKEO is used as features of frames of compressed speech signal due to its high frequency and amplitude resolution.

A. Teaser Kaiser Energy Operator

The Teaser-Kaiser Energy Operator (TKEO) is a powerful nonlinear operator proposed by Kaiser, capable of extracting the signal energy based on mechanical and physical considerations[11]. It has been successfully used in various speech applications. This operator can be used to detect frequency and/or amplitude variations in a signal[12]. The output of TKEO can represent their spectral content of the

 ISSN: 2454-4248
52 – 58

signals having frequency less than sampling frequency. Since the frequency variation in the compressed signal is less than the one in the original signal, the problem of cross-terms is reduced by using TKEO. The TKEO measure is more effective than the traditional energy measure in detecting important parts of signal in a very noisy environment. For a bandlimited digital signal, this operator can be approximated by

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

B. Enhanced Spectrogram: Pyknogram

The enhanced spectrogram, called pyknogram, were first introduced in [13] to facilitate formant tracking and are calculated by applying multiband demodulation in the framework of the AM-FM modulaton model [14]. Overlaps in speech data can be detected by using pyknogram[15]. In pyknograms, the resonances (formants) and harmonic structure of speech are enhanced by decomposing the spectral sub-bands into amplitude and frequency components. The frequency and amplitude components of a given subband, x(n), are calculated using equation (5) of Teaser Kaiser Energy Operator (TKEO) 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 are used to derive a short-time estimate value for the dominant frequency in each subband over a fixed period of time, in this case the duration of a time-frame (typically 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 instantaneous frequency and amplitude functions calculated for each sample in the t^{th} frame over the frame length (T samples per frame). Resonances and harmonic peaks are located in each frame by comparing the average frequency estimates with filterbank center frequencies [13]. The motivation behind using an energy operator based approach [14]is to avoid assumptions on the number of speakers in the signal. The AM-FM decomposition method relies on signal resonances and does not restrict the signal to a specific structure.The final time-frequency representation is called a pyknogram and is denoted Spyk(t, f) as a function of time (t) and frequency (f) as shown in Figure 3.

IV. SPEAKER CHANGE DETECTION

Bayesian Information Criterion (BIC) is one of the most popular technique for detecting speaker change point in an audio recording presented in [16]. It's the statistical measure used in statistical hypothesis testing. Let's say the model trained on segment X_1 and X_2 is M_1 and M_2 respectively. Then BIC for each segments are,

$$BIC(X_1, M_1) = \log(p(X_1|M_1)) - \lambda d_1 \log N_1 \quad (9)$$

$$BIC(X_2, M_2) = \log(p(X_2|M_2)) - \lambda d_2 \log N_2 \quad (10)$$

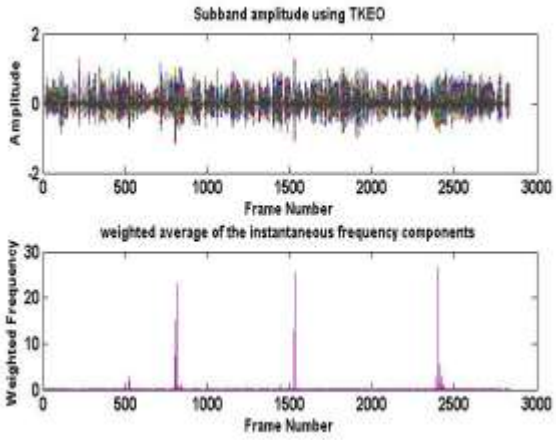

Fig. 3 Frames of Compressed Signal and weighted average of instantaneous frequencyComponent using pyknogram

The first term is likelihood term while second term checks for complexity and therefore controls over-fitting. Similarly BIC of segments concatenating X_1 and X_2, let's say X, with respect to modelM is calculated. Finally following BIC measure is calculated.

$$\Delta BIC = BIC(M) - BIC(M_1) - BIC(M_2) \quad (11)$$

For multivariate Gaussian distributions $M_1 = N(\mu_1, \sum_1)$, $M_2 = N(\mu_2, \sum_2)$ and $M = N(\mu, \sum)$ with model size N_1, N_2 and N_1+N_2 respectively, delta BIC is

$$\Delta BIC = (N_1 + N_2)\log(\sum) - N_1\log(\sum_1) - N_2\log(\sum_2)$$
$$- \lambda(0.5 *(d+0.5*(d+1)))\log N \quad (12)$$

Where λ is a penalty weight, d is a dimension of the feature space and $\sum_1$, $\sum_2$ and $\sum$ are determinants of covariance matrices for the segments X_1 , X_2 and X respectively. If $\Delta BIC > 0$, a local maximum of ΔBIC is found and time ti is considered to be a speaker change point. If $\Delta BIC < 0$, there is no speaker change point at time ti, as shown in Figure 4.

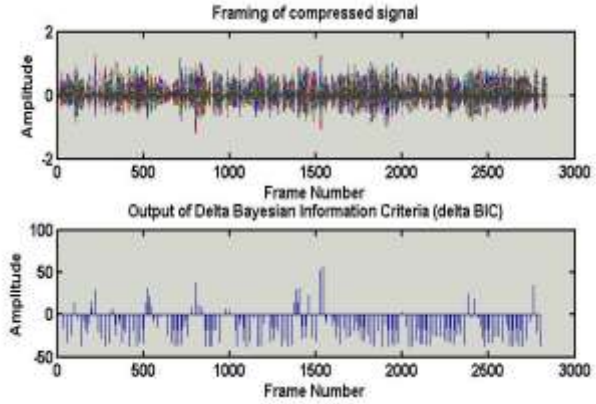

Fig. 4 Framing of Compressed Signal and output of Delta BIC

 ISSN: 2454-4248
52 – 58

V. Information Change Rate

The Information Change Rate (ICR) or Entropy can be used to characterize the similarity of two neighboring clusters[17]. The ICR determines the change in information that would be obtained by merging any two clusters under consideration. ICR of two clusters Cx and Cy is as:

$$\mathrm{ICR}(\mathrm{Cx},\mathrm{Cy}) \triangleq (1/(Ncx+Ncy)) * \ln BIC(Cx,Cy) \quad (13)$$

Where Ncx and Ncy are the number of features in clusters Cx and Cy respectively. This is 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 intraspeaker variance of the corresponding speaker identity.

VI. Hierarchical Clustering

Hierarchical clustering groups data over a variety of scales by creating a cluster tree or dendrogram [18]. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next level. This allows deciding the level or scale of clustering that is most appropriate for the application. The dendrogram function plots the cluster tree. A dendrogram consists of many U-shaped lines connecting objects in a hierarchical tree. The height of each U represents the distance between the two objects being connected. Each leaf in the dendrogram corresponds to one cluster.

VII. Proposed Speaker Diarization System

The goal of this research is to determine" who spoke when" in multiparty speech data of TV show. The main problem is to handle short speech utterances (2-5 seconds), high number of involved speakers, and high probability for missing speaker change point. We assume that the recorded signal is comprised of clean speech with commercial music and clapping. Our proposed speaker diarization system is depicted in Figure 5 is similar to the standard agglomerative clustering framework described in [1] except the two following main modifications.

- We don't use any training data and voice activity detection to presegment the audio stream into speech and non-speech regions. Instead of Mel frequency Cepstral Coefficient (MFCC)[1], we used enhanced spectrogram (pyknogram) as features of speech signal as discussed in section 2.
- After segmentation, for segment recombination Information change rate (ICR) is used. Clustering of homogeneous speakers is obtained by standard dendrogram used in MATLAB.

VIII. Experiments and Results

A. Experimental Framework

1) Database Used

The data source is recording of famous TV show "Dr. Subhash Chandra show" in MP4 format. It is converted into .wav form to use it in MATLAB, their parameters were shown in Table 1.

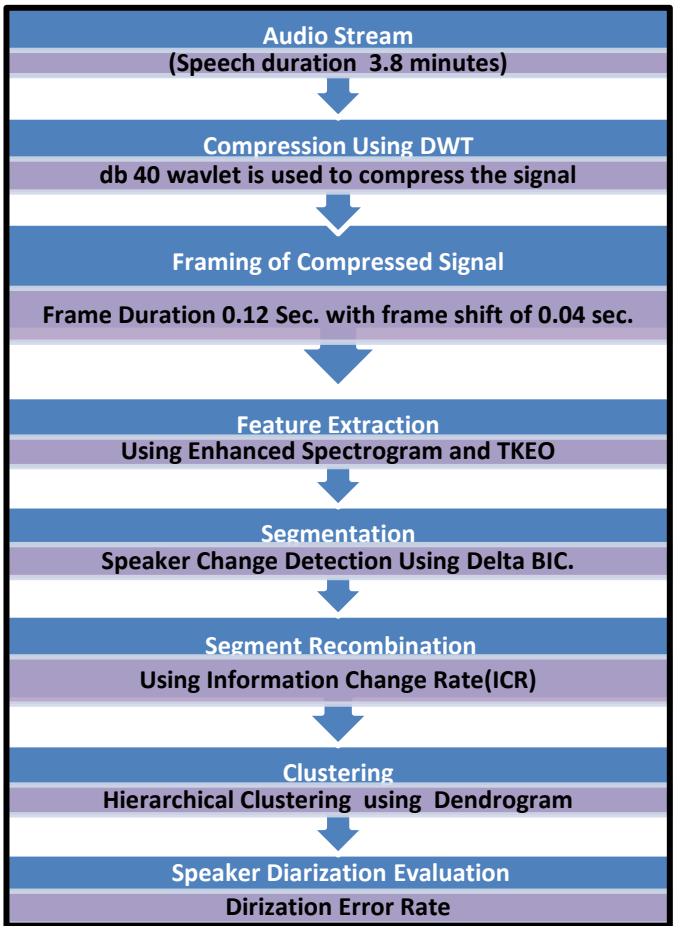

Fig.5 Proposed Speaker Diarization System

Table I System Parameters

Parameter	Value
Sampling frequency	44100Hz. 16 bits
Database	Recording of 3.8 minutes.
No of speakers	7 with clapping sound.
Compression and signal enhancement technique	Discrete Wavelet Transform (DWT)
Analysis frame duration	0.12 sec.
Analysis frame shift	0.04 sec.

In this research work, for the evaluation of speaker diarization, two files were extracted from an audio recording: reference file and hypothesized file. In first file speaker change points were manually detected using signal processing tool

 ISSN: 2454-4248
52 – 58

(SPTOOL) in MATLAB and consider as reference file, shown in Table 2. In second file, speaker change point is detected by applying different algorithms of audio compression, feature extraction,segmentation and clustering as discussed in previous section.

Table II Reference File

Start Time (sec.)	Stop Time (sec.)	Duration (sec.)	Start Frame No.	End Frame No.	Label
0	8	8	0	95	SPEAKER 1
8	13	6	96	165	S.CHANDRA2
13	19	6	166	240	SPEAKER 1
19	21	2	241	263	S.CHANDRA2
21	29	8	264	361	SPEAKER 3
29	33	4	362	412	S.CHANDRA2
33	35	2	413	440	SPEAKER 4
35	41	6	441	510	S.CHANDRA2
41	42	1	511	519	SPEAKER 4
42	45	4	520	566	S.CHANDRA2
45	49	4	567	617	SPEAKER 5
49	53	3	618	659	S.CHANDRA2
53	64	11	660	798	SPEAKER 5
64	66	2	799	826	CLAPPING
66	68	2	827	849	SC+CLAPPING
68	130	62	850	1627	S.CHANDRA2
130	157	26	1628	1957	SPEAKER 6
157	190	34	1958	2376	S.CHANDRA2
190	196	6	2377	2445	CLAPPING
196	221	25	2446	2758	SPEAKER 7
221	227	6	2759	2835	S.CHANDRA 2

2) Performance Evaluation Criteria

A change detection system has two possible types of error.Type-I errors occur if a true change is not spotted within a certain window. Type-II errors occur when a detected change does not correspond to a true change in the reference (false alarm)[5]. Type I and II errors are also referred to as and recall (RCL), precision (PRC) respectively, which are defined as

$$\text{Recall} = \frac{\#\ of\ correct\ \ detected\ \ speaker\ \ changes}{\#\ of\ Speaker\ \ Changes} \quad (10)$$

$$\text{Precision} = \frac{\#\ of\ correct\ \ detected\ \ speaker\ \ changes}{\#\ of\ detected\ \ Speaker\ \ Changes} \quad (11)$$

In order to compare the performance of different systems, the F-measure is often used and is defined as

$$\text{F-Measure} = \frac{2*Recall*Precision}{Recall + Precision} \quad (12)$$

The F-measure varies from 0 to 1, with a higher –measure indicating better performance. The main metric for Speaker Diarization System evaluation is the Diarization Error Rate (DER) described in [19] which is a sum of three contributing factors as shown in Eq.13, the miss detection error (speaker in reference but not in hypothesis), false alarm (speaker in hypothesis but not in reference), and speaker misclassification error (mapped reference speaker is not the same as the hypothesized speaker) rates.

$$DER = ESpkr + EFA + EMiss \quad (13)$$

A complete description of the evaluation measure and scoring software implementing it can be found [19]

B. Experimental Results

1) Speaker Change Detection Evaluation

In this section, we describe the experiments performed on different data sets. Based on the wavelet transform, the audio signals are compressed in the ratio of 1:4 at level 2 with energy of 98% (approx.).The compressed signal is converted into overlapping frames by using hanning window to reduce the side-lobe artefacts at the boundary of the signal. The Pyknogram based on TKEO of the frames of compressed signal is taken as features for speaker change detection. The sudden changes at the output of detected features correspond to speaker change points. These speaker change points were refined by traditional Bayesian Information Criteria algorithm as discussed in section 4, shown in figure 4.Comparison of hypothesized frames of final segmentation (shown in Figure 6) and reference frames from Table2are shown in Figure 7 for the performance evaluationof speaker change detection. It is observed that whenever speaker stops for more than 2 seconds while talking, a change is detected as in case of frame numbers 900 to 1500, six silent frames were detected as speaker change points. Also if speaker speaks for very small duration of 2-4 seconds, it can't bedetected by this system. Its response will be more than 95% if silent is removed and speaker speaks for more than 6 seconds. The performances of this system measured by F-measure for two cases are 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.

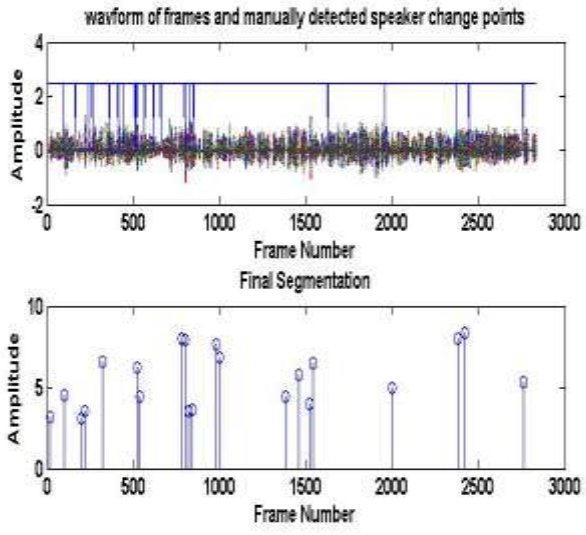

Fig. 6 Frames with manual segmentation and hypothesized segments.

 ISSN: 2454-4248
52 – 58

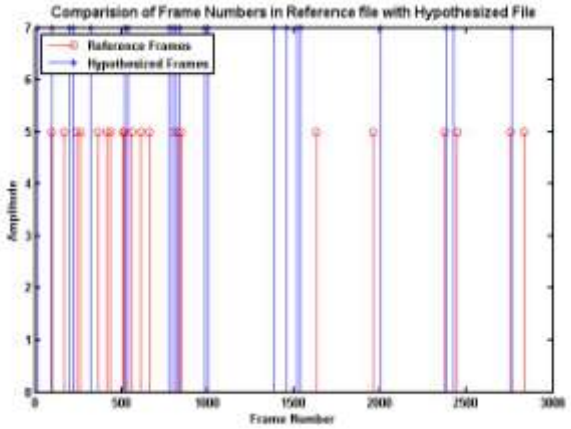

Fig. 7 Comparison of frames of reference and hypothesized files.

2) Speaker Diarization Error

The proposed speaker diarization system uses hierarchical clustering to group the homogeneous speakers into clusters. Dendrogram function in MATLAB plots the cluster tree as shown in Figure 8. This system do not uses speech activity detection algorithm so, there is no need of evaluating miss detection error (Miss) and false alarm (FA). Only speaker misclassification error is evaluated by matching hypothesized speakerto 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 so as to maximize the total overlap of the reference and (corresponding) mapped hypothesis speakers. The DER of proposed system at distance 3.7 is 22.34 whichis comparable with standard diarization system shown in Table 3.

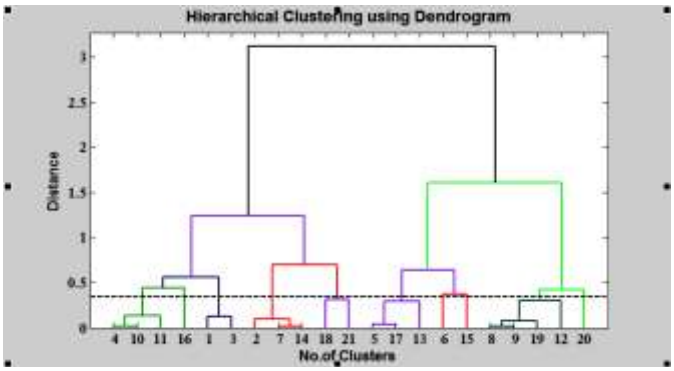

Fig. 8 Hierarchical Clustering using Dendrogram.

Table III Speaker Change Detection Evaluation

Speaker Change Detection Method without Voice Activity Detection (VAD)	Recall (%age)	Precision (%age)	F-measure (%age)	Diarization Error Rate (%age)
Pyknogram with BIC and ICR	66.67	70	68.29	20.34
MFCC and BIC	55.39	51.78	53.52	22.46

IX. CONCLUSION AND FUTURE SCOPE

This paper proposed a new technique for speaker diarization of an audio stream. In this approach the signal is initially compressed using 2 levels DWT and partition into overlapping frames of duration 0.12 seconds with frame shift of 0.04 seconds. The enhanced spectrogram known as Pyknogram based on TKEO is used as features of compressed speech signal. The sudden changes at the output of feature extraction correspond to speaker change point and were detected by Delta BIC. Resegmentation is used to refine the segments by Information change rate (ICR). For clustering dendrogram function in MATLAB is used.Theresults of applying the proposed and traditional technique (using MFCC and BIC) on audio stream indicates that using wavelet transform and pyknogram improves the speaker change point detection evaluation parameter F-measure and reduces the diarization error rate (DER) also improves the capability of diarization process .Further work will be separated into two research areas: Handling of speeches of short duration of 2-4 seconds and overlapping speech detection to improve the performance of speaker diarization system.

REFERENCES

[1] X. Anguera Miró, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland, and O. Vinyals, "Speaker Diarization: A Review of Recent Research," *IEEE Trans. Audio Speech Lang. Process.*, vol. 20, no. 2, pp. 356–370, 2012.

[2] M. H. Moattar and M. M. Homayounpour, "A review on speaker diarization systems and approaches," *Speech Commun.*, vol. 54, no. 10, pp. 1065–1103, 2012.

[3] M. Kotti, V. Moschou, and C. Kotropoulos, "Speaker segmentation and clustering," *Signal Processing*, vol. 88, no. 5, pp. 1091–1124, 2008.

[4] A. G. Adami, S. S. Kajarekar, and H. Hermansky, "A New Speaker Change Detection Method for Two-speaker Segmentation," *IEEE Int. Conf. Acoust. Speech, Signal Process.*, vol. 4, pp. 3908–3911, 2002.

[5] J. Ajmera, I. McCowan, and H. Bourlard, "Robust speaker change detection," *Signal Process. Lett. IEEE*, vol. 11, no. 8, pp. 649–651, 2004.

[6] H. Hassanpour and M. Shahiri, "Adaptive segmentation using wavelet transform," 2007.

[7] J. Ajmera and C. Wooters, "A robust speaker clustering algorithm," *2003 IEEE Work. Autom. Speech Recognit. Underst. (IEEE Cat. No.03EX721)*, pp. 413–416, 2003.

[8] M. Grašič, M. Kos, and Z. Kačič, "Online speaker segmentation and clustering using cross-likelihood ratio calculation with reference criterion selection," *IET signal Process.*, vol. 4, no. 6, p. 673, 2010.

[9] J. I. Agbinya and N. S. Wales, "Processing," pp. 1–6, 1996.

[10] M. Vetierli, "~ velets and Signal," no. October, pp. 14–38, 1991.

[11] M. Bahoura and J. Rouat, "Wavelet speech enhancement based on the Teager energy operator," *IEEE Signal Process. Lett.*, vol. 8, no. 1, pp. 10–12, 2001.

[12] F. Kaiser and S. Street, "+ 4) (5)," vol. 2, no. 10, pp. 381–384, 1990.

[13] A. Potamianos and P. Maragos, "multiband energy

 ISSN: 2454-4248
52 – 58

demodulation," vol. 99, no. 6, pp. 3795–3806, 1996.

[14] P. Maragos, S. Member, J. F. Kaiser, T. F. Quatieri, and S. Member, "Application to Speech Analysis S :, s :," vol. 41, no. 10, pp. 3024–3051, 1993.

[15] 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 Navid Shokouhi , Ali Ziaei , Abhijeet Sangwan , John H . L . Hansen Center for Robust Speech Systems (CRSS) The University of Texas at Dallas , Richar," no. 978, pp. 4724–4728, 2015.

[16] P. S. Gopalakrishnan, "Clustering Via The Bayesian Information Criterion With," pp. 645–648, 1998.

[17] S. California and L. Angeles, "A Novel Inter-Cluster Distance Measure Combining Glr And Icr For Improved Agglomerative Hierarchical Speaker Clustering Kyu J . Han and Shrikanth S . Narayanan Speech Analysis and Interpretation Laboratory (SAIL) Ming Hsieh Department of Electrical Engineering , Viterbi School of Engineering," pp. 4373–4376, 2008.

[18] A. Solomonoff, A. Mielke, M. Schmidt, and H. Gish, "Clustering speakers by their voices," *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.*, vol. 2, pp. 757–760, 1998.

[19] 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," pp. 7741–7745, 2013.