Speaker Change Detection using Teaser-Kaiser Energy Operator and Wavelet Transform

[1] Sukhvinder Kaur, [2] J. S. Sohal
[1] *Ph.D. Research Scholar, I.K. Gujral PTU, Jalandhar, Kapurthala-144601, India*
[2] *Director, LCET, Ludhiana-141113, India*

Abstract

The aim of this paper is to present an efficient, fast and optimized system that detects speaker change points in multispeaker speech data. It can be used in captioning of TV shows, movies and to split an audio stream into acoustically homogeneous segments, so that every segment ideally contains only one speaker. In this proposed technique, the daubechies 40-wavelet transform is used to compress the audio stream in the ratio of 1:4 with 99% of energy; their features are extracted by 5 level discrete wavelet transform and Teaser Kaiser Energy Operator (TKEO). This method relies on amplitude and frequency variation of the speech signal. Finally, sudden changes of energy in the output of TKEO that corresponds to the speaker change point, is detected by using sliding window. The results are evaluated by F-measure and shows that the proposed method gives fast and better results as compared to traditional method without using discrete wavelet transform.

Keywords

F-Measure, Segmentation, Sliding Window, Speaker Change Point, Teaser Kaiser Energy Operator, Wavelet Transform.

I. INTRODUCTION

The goal of audio segmentation and classification is to partition and label an input audio stream into speech, speaker, music, commercials, environmental background noise, or other acoustic conditions [1]. Segmentation of audio stream is used in many applications like speaker diarization, speaker recognition, audio information retrieval, audio transcription, audio clustering, indexing and captioning of TV shows and movies.

Speaker segmentation algorithms can be broadly classified into threecategories: model-based, metric-based, and hybrid (i.e. Combined metric- and model-based) ones [1]. Model based methods require training data to initialize the speakermodels. For example Universal Background Model (UBM), Hidden Markov Model (HMM), Gaussian Mixture Model (GMM). Metricbasedmethods do not require any prior knowledge of number of speakers, their identities, or the signal characteristics. A wide variety of distance metrics could be used. For example, a weighted squared Euclidean distance, Hotelling T2, the Kullback–Leibler divergence, Kullback–Leibler-2 divergence, generalizedlikelihood ratio (GLR) test, Bayesian Information Criteria (BIC), Cross likelihood ratio (CLR), Normalized cross likelihood ratio (NCLR). Hybrid algorithms combine metric and modelbasedtechniques. Usually, metric-based segmentationis used initially to pre-segment the input audio signal.The resulting segments are used then to create a set of speaker models. Next, model-based re-segmentationyields a more refined segmentation. Due to the rapid increase in the volume of computer-available recorded speech, reviewed methods for speaker segmentation and classification are very complex.

This paper proposes a method for segmentation by using speaker change detection algorithm in multiparty audio stream. In this study, a discrete wavelet transform (DWT) based compression and de-noising approach is presented to improve the performance of speaker change detection system. The frames of compressed signal are decomposed by 5 levels DWT and its energy is calculated by Teaser-Kaiser Energy Operator (TKEO)and act as features for segmentation [2]. The following section will introduce the principles of wavelet transform and TKEO, the experimental results shows an improvement in the speaker segmentation process.

II. SPEAKER CHANGE DETECTION ALGORITHMS

A Wavelet Transform Aspects

Wavelet transform (WT) has been widely used due to its most important property,which is to examine a signal simultaneously in the time–frequencydomain.[3]. WT emerged in the 1980s; however it only started being used tosolve engineering problems in the 1990s. Discrete wavelet transform (DWT) uses the fact that it is possible to resolve high frequency components within a small time window, and only low frequency components need large time windows. This is because a low frequency component completes a cycle in a large time interval whereas a

high frequency component completes a cycle in a much shorter interval. Therefore, slow varying components can only be identified over long time intervals but fast varying components can be identified over short time intervals. The wavelet transform is defined as the inner product of a signal *x(t)* with the mother wavelet $\psi(t)$ is as follows:

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

Where,

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

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 :

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

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.

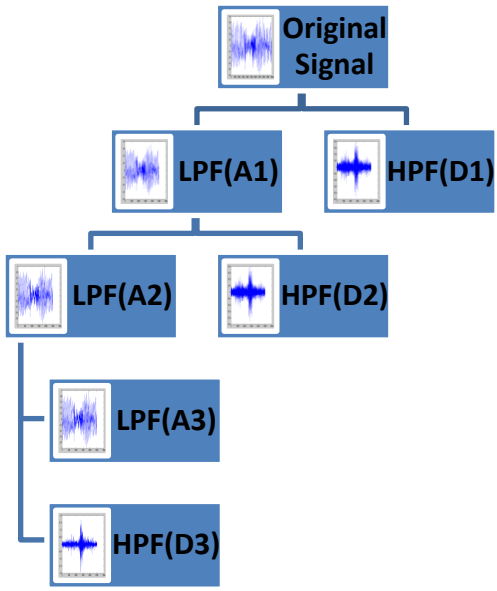

Fig. 1 Tree diagram of DWT

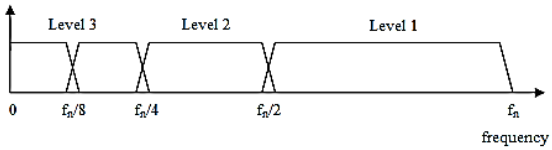

Fig.2 Frequency domain representation of the DWT

The signal decomposition represents the slowly changing features of the signal in the lower frequency bands and the rapidly changing features in the higher frequency bands. The DWT decomposes the original signal into approximation coefficients (A) and detail coefficients (D). The approximation coefficients store information regarding the low frequency components, while the detail coefficients store the high frequency information. This decomposition procedure can be repeated until the maximum level of decompositions is reached, i.e., there is only approximation and detail leaves containing just one coefficient. As shown in figure 1, the DWT decomposes the signal in dyadic form, and the recursive decomposition only act on the low frequency component, named "Approximation" (A). The high frequency component is preserved as "detail" (D). Its frequency response is given in figure2.The choice of wavelet transform depends on the application. Selecting a wavelet function which is very closely matches the signal to be processed is very important in wavelet applications. The most common family of wavelets, Daubechies, DWT-db has its low pass filter coefficients determined by solving the following system of equation (4) [4].

$$\sum_{k=0}^{n-m}(-1)^k h_{m-k}=0,\ \sum_{k=0}^{n-m}h_k=2,2\sum_{k=0}^{n-m}h_k h_{k-2m}=2\delta_{0,m}\,(4)$$

Characteristics of Daubichies and Haar wavelets used with 2 minutes audio stream at different levels are shown in Table 1 and found that db40 wavelet compressed the signal in the ratio of 1:4 which is closely matches with original speech signal at level 2 with higher energy as compared to db4. So, in this work, the daubechies wavelet db40 at level 2 is used to compress the speech signal.

Wavelet Type	DWT Level &Freq.(fs)	No of Samples in Approximation & Energy(%)	No. of Samples in Detail & Energy (%)	Time (Second)
Haar	0 44100	5000000 100		113.3787
	1 22050	2500000 99.4709	2500000 0.5291	226.7574
	2 11025	1250000 97.9855	1250000 2.0145	113.3787
db4	0 44100	5000000 100		113.3787
	1 22050	2500003 99.8583	2500003 0.1417	226.7576
	2 11025	125005 99.1539	125005 0.8461	113.3791
db40	0 44100	5000000 100		113.3787
	1 22050	2500039 99.9131	2500039 0.0869	226.7609
	2 11025	1250059 99.1999	250059 0.8001	113.3840

TABLE 1
CHARACTERISTICS OF THE DICRETE WAVELET TRANSFORM AT DIFFERENT LEVELS FOR 2 MIN. AUDIO STREAM

B) Teaser- Kaiser Energy Operators (TKEO)

In speaker change detection process, TKEO also named as nonlinear energy operator (NEO) is used as features of speech signal due to its high frequency and amplitude resolution. The TKEO measure is more effective than the traditional energy measure in detecting important parts of signal in a very noisy environment. It is defined as:

$$\Psi(x(t)) = [\dot{x}(t)^2] - [x(t)\ddot{x}(t)] \quad (5)$$

And the discrete version of the operator can be defined as

$$\Psi[x[n]] = x^2[n] - x[n-1]x[n+1] \quad (6)$$

This operator can be used to detect frequency and/or amplitude variations in a signal [5]. The output of TKEO can represent their spectral content of the signals having frequency less than sampling frequency. Since the frequency variation in the compressed/decomposed signal is less than the one in the original signal, the problem of cross-terms is ameliorated in using TKEO.

The sudden changes of energy in the output of TKEO correspond to the segmentation boundaries. To detect these changes, a sliding window can be used to produce the segmentation criteria. For a given time instant, n, the energy in the left half of the window is subtracted from the energy in the right half window resulting the signal L given in equation (7) [6].

$$L(n) = \left| \sum_{m=n-N+1}^{n} \Gamma(m) - \sum_{m=n+1}^{n+N} \Gamma(m) \right| \quad (7)$$

Where N is the window length of 25 samples and $\Gamma(m)$ is the m^{th}energy coefficient. If the two half windows have the same energy, the resulting signal L is zero and if the window is centered at a segment boundaries the L will be large. Consequently, the signal L contains local maxima indicating the segment boundaries. These segment boundaries are labeled and used for speaker change point.

III. PROPOSED METHOD

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 99% (approx.) as shown in figure 3.The compressed signal is converted into overlapping frames by using hanning window to reduce the side-lobe artifacts at the boundary of the signal. The 5 level DWT and TKEO of the frames of compressed signal is taken as features for segmentation. On the output of TKEO per frame, sliding window is applied to find the boundaries in the frame. The flow chart of proposed method is shown in figure 4**.**

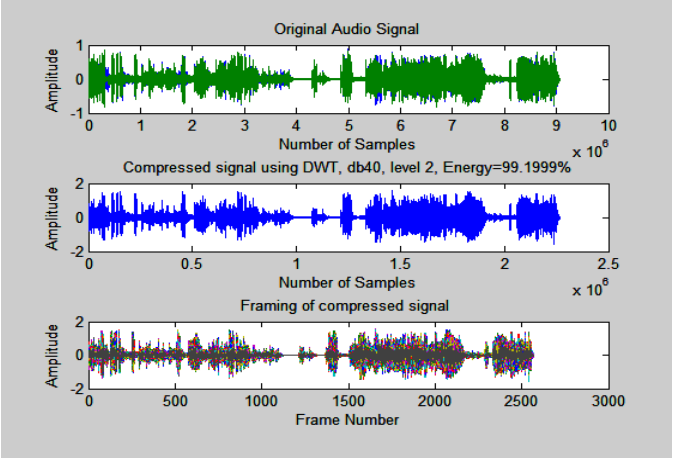

Fig. 3 Waveform of audio clip, its compressed form and frames.

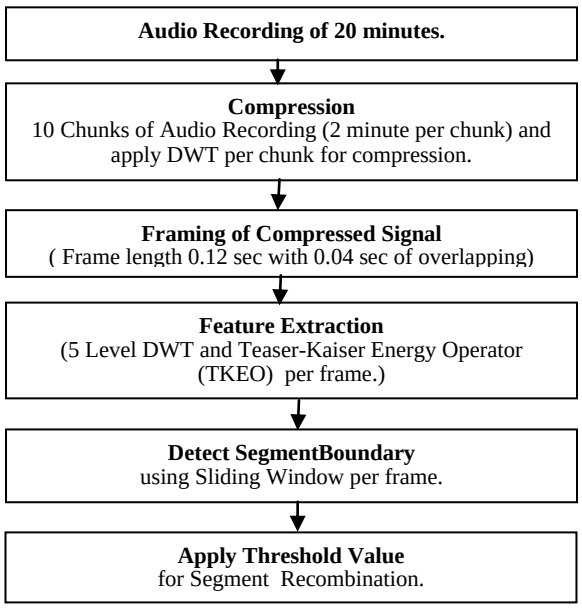

Fig. 4 Flow chart for Speaker Change Detection System

IV. EXPERIMENTAL FRAMEWORK AND RESULTS

A Database used

The data sources are four video clips free down loaded from youtube.com in MP4 format, further it is converted into .wav form and used in MATLAB. The parameters used in the system are shown in table 2.

B Performance Evaluation Measure

Speaker change detection is evaluated by F-measure defined as follows [7]:

$$\text{F-Measure} = \frac{2*Recall*Precision}{Recall+Precision} \quad (8)$$

Where,

$$\text{Recall} = \frac{\#\ of\ correct\ \ detected\ \ speaker\ \ changes}{\#\ of\ \ Speaker\ \ \ Changes} \quad (9)$$

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

TABLE II

SYSTEM PARAMETERS

Parameter	Value
Sampling frequency	44100Hz. 16 bits
Database	4 Video Clips (2 to 20 min)
No of speakers	3-4
Compression and signal enhancement technique	Discrete Wavelet Transform (DWT)
Wavelet type	Daubechies (db4 & db40)
Decomposition levels	2 and 5
Window type	Hanning
Analysis frame duration	0.12 sec.
Analysis frame shift	0.04 sec.

C) Results and Discussion

Speaker change detection evaluation was conducted on video clips free downloaded from youtube.com. The aim of this research is to correctly detect the speaker changes in order to help captioning of TV shows. This experiment is performed for two cases: without using DWT per frame and by using 5-level DWT per frame as shown in figure 5 and figure 6 respectively.

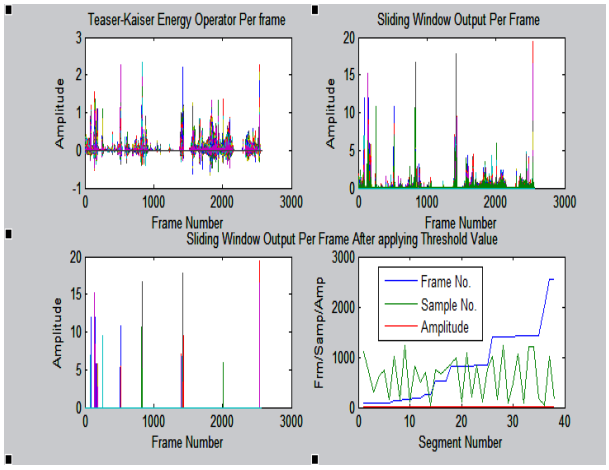

Fig. 5 TKEO per frame, sliding window output and sliding window outputs after applying threshold value 5.

Each figure has four sub figures. Top two figures represent TKEO per frame and output of sliding window. Bottom two figures show the output after applying threshold value. Fourth sub-figure represents the speaker change point and its amplitude in its respective frame and sample number. By using equation (7), sliding window detects the sudden changes of energy that corresponds to speaker change. Finally threshold value5 and 30 is applied for segment recombination. It is clear that by using DWT, it improves the capability of TKEO by scaling and decomposing signals with different frequency bands. It detects all those segments which were not detected in the first case. The performance is measured by F-measure for two cases as shown in Table 3. It is clear from the results that by using DWT, F-measure is improved.

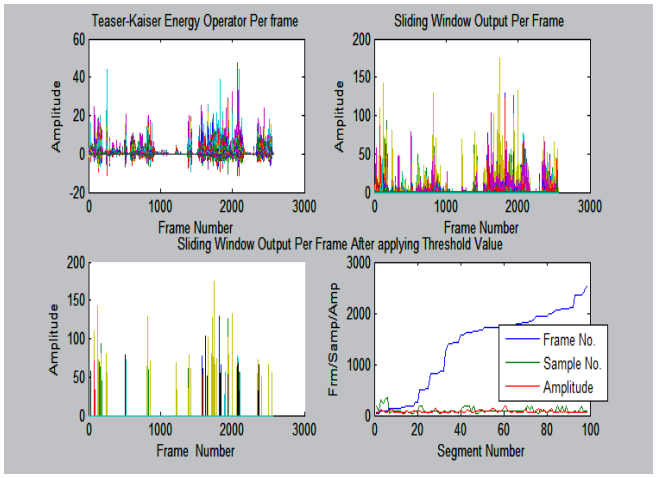

Fig. 6 TKEO after applying DWT at 5 levels per frame, sliding window output and sliding window outputs after applying threshold value 30.

TABLE III

SPEAKER CHANGE DETECTION EVALUATION

Speaker Change Detection Method and Threshold Value	Recall (%age)	Precision (%age)	F-measure (%age)
Without DWT, Threshold Value 5	55.39	51.78	51.55
With DWT, Threshold value 30	61.13	77.46	67.46

V. CONCLUSION

This paper proposed a new technique for speaker segmentation in an audio stream. In this approach, the multi-speaker speech data is compressed using DWT and their frames are processed.TKEO is used to accentuate the frequency and amplitude variations of the frames of speech signal. To reduce the effect ofcross-terms introduced by TKEO, the signal is

initially scaled and decomposed intosignals with different frequency bands by using 5 levels DWT. The Result of applying the proposed technique on audio stream indicate that using wavelet transform improvesthe capability of TKEO indetecting speaker change point in multiparty speech data.Further work will be separated into two research areas: Overlapping speech detection and clustering.

REFERENCES

[1] Margarita Kotti, Vassiliki Moschou, Constantine Kotropoulos, " Review Speaker segmentation and clustering" Signal Processing 88 pp 1091–1124, 2008.

[2] Z. Tufekci and J.N. Gowdy, "Feature extraction using discrete wavelet transform for speech recognition," in Proc. IEEE Southeastcon, USA, pp. 116-123, 2000.

[3] Da Wu, J and B Fu Lin, "Speaker Identification using discrete wavelet packet transform technique with irregular decomposition", Expert System with Applications 36, pp 3136-3143, 2009, DOI:10.1016/j.eswa.2008.01.038

[4] I. Daubechies, "Orthonormal bases of compactly supported wavelets", Commun. on Pure and Appl. Math., Vol. 41, pp. 909-996, Nov. 1988.

[5] J.F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal", Proceedings of the IEEE ICASSP-90, Albuquerque, NM, pp-381-384, April 1990.

[6] R. Agarwal and J. Gotman, "Adaptive Segmentation of Electroencephalographic Data Using a Nonlinear Energy Operator" Proc. IEEE International Symposium on Circuits and Systems (ISCAS'99), vol. 4, pp. 199-202, 1999.

[7] Jitendra Ajmera, Iain Mccowan, And Hervé Bourlard, "Robust Speaker Change Detection", IEEE Signal Processing Letters,Vol. 11, No. 8, pp 649-651, August 2004