SPEAKER SEGMENTATION USING TEASER-KAISER ENERGY OPERATOR AND WAVELET TRANSFORM

Sukhvinder Kaur  
Ph.D. Research Scholar, I.K. Gujral PTU,  
Jalandhar, Kapurthala-144601, India  

J. S. Sohal  
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 three categories: model-based, metric-based, and hybrid (i.e. Combined metric- and model-based) ones [1]. Model based methods require training data to initialize the speaker models. For example Universal Background Model (UBM), Hidden Markov Model (HMM), Gaussian Mixture Model (GMM). Metric based methods 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, generalized likelihood ratio (GLR) test, Bayesian Information Criteria (BIC), Cross likelihood ratio (CLR), Normalized cross likelihood ratio (NCLR). Hybrid algorithms combine metric and model based techniques. Usually, metric-based segmentation is 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-segmentation yields 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–frequency domain.[3]. WT emerged in the 1980s; however it only started being used to solve 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 $\psi(t)$ as follows:

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

(1)

Where,

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

(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 be expressed as:

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

(3)

Where, $\psi_m$ 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 figure 2. 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].

\[ \sum_{k=0}^{n-m} (-1)^k h_{m-k} = 0 \sum_{k=0}^{n-m} h_k = 2, \sum_{k=0}^{n-m} h_{k-2m} = 25, m \] (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.

<table>
<thead>
<tr>
<th>Wavelet Type</th>
<th>DWT Level</th>
<th>No. of Samples in Approximation (%)</th>
<th>No. of Samples in Detail (%)</th>
<th>Time (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Haar</td>
<td>0</td>
<td>4410</td>
<td>100</td>
<td>113.278</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>260000</td>
<td>20</td>
<td>128.764</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>120000</td>
<td>10</td>
<td>113.578</td>
</tr>
<tr>
<td>db4</td>
<td>0</td>
<td>4410</td>
<td>100</td>
<td>113.278</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>250000</td>
<td>147</td>
<td>113.578</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>120000</td>
<td>94</td>
<td>113.278</td>
</tr>
<tr>
<td>db40</td>
<td>0</td>
<td>4410</td>
<td>100</td>
<td>113.278</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>250000</td>
<td>147</td>
<td>113.578</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>120000</td>
<td>94</td>
<td>113.278</td>
</tr>
</tbody>
</table>

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:

\[ \Psi(x(t)) = [x(t)^2] - [x(t)x(t)] \] (5)

And the discrete version of the operator can be defined as

\[ \Psi[x[n]] = x^2[n] - x[n-1]x[n+1] \] (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].

\[
L(n) = \left| \sum_{m = n - N + 1}^{n} \Gamma(m) - \sum_{m = n + 1}^{N} \Gamma(m) \right|
\]  

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 boundary 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.

![Fig. 3 Waveform of audio clip, its compressed form and frames.](image)

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. 4 Flow chart for speaker Change Detection system](image)

###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.

#### TABLE II

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling frequency</td>
<td>44100 Hz, 16 bits</td>
</tr>
<tr>
<td>Database</td>
<td>4 Video Clips (2 to 20 min)</td>
</tr>
<tr>
<td>No of speakers</td>
<td>3-4</td>
</tr>
<tr>
<td>Compression and signal enhancement technique</td>
<td>DWT</td>
</tr>
<tr>
<td>Wavelet type</td>
<td>Daubechies (db4 &amp; db40)</td>
</tr>
<tr>
<td>Decomposition levels</td>
<td>2 and 5</td>
</tr>
<tr>
<td>Window type</td>
<td>Hanning</td>
</tr>
<tr>
<td>Analysis frame duration</td>
<td>0.12 sec</td>
</tr>
<tr>
<td>Analysis frame shift</td>
<td>0.04 sec</td>
</tr>
</tbody>
</table>

#### B Performance Evaluation Measure

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

\[
F\text{-Measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}
\]  

Where, \( \text{Recall} = \frac{\# \text{ of correct detected speaker changes}}{\# \text{ of Speaker Changes}} \)  

\[
\text{Recall} = \frac{\# \text{ of correct detected speaker changes}}{\# \text{ of Speaker Changes}}
\]
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. 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 value 5 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>
<thead>
<tr>
<th>Speaker Change Detection Method</th>
<th>Recall (Wag)</th>
<th>Precision (Wag)</th>
<th>F-measure (Wag)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without DWT, Threshold Value 5</td>
<td>51.39</td>
<td>51.78</td>
<td>51.55</td>
</tr>
<tr>
<td>With DWT, Threshold value 30</td>
<td>61.13</td>
<td>77.46</td>
<td>67.46</td>
</tr>
</tbody>
</table>

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 of cross-terms introduced by TKEO, the signal is initially scaled and decomposed into signals with different frequency bands by using 5 levels DWT. The results of applying the proposed technique on audio stream indicate that using wavelet transform improves the capability of TKEO in detecting speaker change point in multiparty speech data. Further work will be separated into two research areas: Overlapping speech detection and clustering.

REFERENCES


