IJMTST | Volume: 2 | Issue: 05 | May 2016 | ISSN: 2455-3778

A Novel Method for Silence Removal in Sounds Produced by Percussive Instruments

S. Srikanth[1] | B. Kishore Kumar[2] | Ch. Nagababu[3]

[1]Department of Electrical Engineering, (Ph.D. pursuing in KLU) ARMIET, Thane, Maharashtra, India
[2]Department Electronics and Telecommunication Engineering, ARMIET, Thane, Maharashtra, India
[3]Department Electronics and Telecommunication Engineering, ARMIET, Thane, Maharashtra, India

ABSTRACT

The steepness of an audio signal which is produced by the musical instruments, specifically percussive instruments is the perception of how high tone or low tone which can be considered as a frequency closely related to the fundamental frequency. This paper presents a novel method for silence removal and segmentation of music signals produced by the percussive instruments and the performance of proposed method is studied with the help of MATLAB simulations. This method is based on two simple features, namely the signal energy and the spectral centroid. As long as the feature sequences are extracted, a simple thresholding criterion is applied in order to remove the silence areas in the sound signal. The simulations were carried on various instruments like drum, flute and guitar and results of the proposed method were analyzed.

KEYWORDS: *Percussive instruments, Spectral energy, Spectral centroid, Silence removal.*

I. INTRODUCTION

Percussive instrument is a musical instrument [1] that is sounded by being creaked by someone (including attached or encircled beaters or clatters); creaked, scraped or wiped by hand; or creaked against another similar type of instrument. The percussive family is believed to include the ancient musical instruments, following the human voice. Percussive instruments are often separated into two types: Pitched percussive instruments, which produce tones with a detectable pitch, and unpitched percussive instruments, which generate tones without detectable pitch.

The sounds produced by the percussive instruments [1] range from simple clicks of wooden sticks to extremely complex cymbal sounds. Therefore the sound design of these instruments cannot be generalized into use of one synthesis technique. This is because of computational complexity and memory requirement. One way to reduce the memory utilization is by removing silence from the sounds and then processing the sounds using synthesizer.

Processing of audio signal is very critical in the applications where silence [2] or background noise is completely objectionable. So, we need to process the signal related to musical waveform consisting of music, silence, and other background noise. Basically, the speech or audio signals are segmented [2], [3] into well-interpreted regions of silence, unvoiced signal which is not exact sound; it is often hard to differentiate a feeble, unvoiced sound (like /q/ or /s/) from silence, or feeble voiced sound (like /p/ or /m/) from unvoiced sounds or even silence.

However, in general[4]-[6] it is not critical to divide the signal to an accuracy smaller than few seconds; hence, errors in peaks of the audio signal usually has no effect in most of the applications since these are very small. Since for most of the real time scenarios the unvoiced part has minimum energy content and thus silence (background noise) and unvoiced part is classified jointly as silence/unvoiced and is differentiated from voiced part.

In general, various applications need different

algorithms to meet their specific requirements in terms of computational accuracy, complexity, robustness, sensitivity, response time [4], [5] etc. The approaches are based on pitch detection, spectrum analysis, cepstral analysis, zero crossing rate, periodicity measure, hybrid detection, fusion and many other methods. Two extensively used methods for silence removal are namely Short Time Energy (STE) and Zeros Crossing Rate (ZCR) [7]. But the threshold used for the above features is fixed. Hence, it will be a problem for the practical cases such as in end point detection. The energy in silence/unvoiced sample is much less than voiced sample. This is the principle used in the short time energy. But, there is no specific variation of energy accompanied to case to case. On the other hand depending on demarcation rule of ZCR, if a portion of audio or speech exceeds 50 then this portion will be termed as unvoiced or silence and any portion which satisfies ZCR at about 12 is denoted as voiced one.

In applications of audio/voice activity detection [8], non-speech portion which includes silence, environmental conditions makes the endpoint detection [9] problem difficult substantially.

The method used in this paper is based on features of the sounds, spectral centroid [10] and signal energy. The reasons for considering these two features are: the spectral centroid will show spectral components of music which is mixed with noise and for simple cases, (where the level of background noise is not very high) and the energy of the voiced segments is larger than the energy of the silent segments.

This paper has been organized as follows: Section II, details the background of the features of the segments and features of music signals. Section III explains about the proposed method which involves thresholding and segmentation. In Section IV the experimental results of the proposed method are analyzed. Finally, the conclusion is summarized in Section V.

II. Segment Classification and Features

A. Audio-Segment Classification

Submit your manuscript electronically for review.

B. Processing flow of audio segmentation

The basic processing flow of the audio segmentation [9] and voice activity detection is shown in the Figure 1. The audio stream is classified into speech and non-speech signals depending on the features extracted from the audio stream. The speaker segmentation is carried out by speaker identity. The non-speech is signal further classified as music, environmental effects and background noise classification.

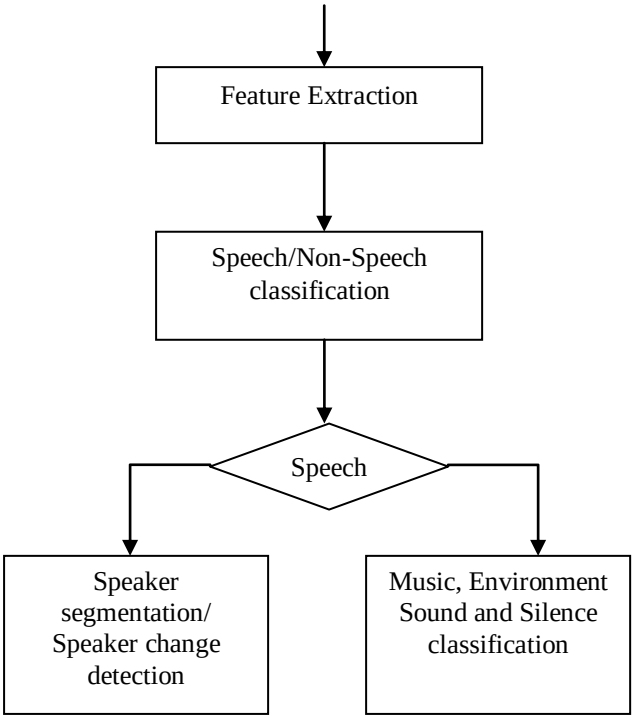

Fig.1: Basic processing flow diagram of segmentation

C. Segment features

The features of the segment [9] are segment length, boundary length, boundary confidence and signal type. Segment length is the duration of the segment. Boundary length is the duration of the silence or non-speech region recognized at the end of segment. Boundary confidence is the value which represents the part of silence or non-speech recognized. Signal type specifies whether the signal is speech or the non-speech. The degree of signal type can be stated as the amount of maximum length of segments represented the non-speech or silence.

D. Features of Audio/Music Signals

1.**Spectrum flux**: The average spectrum variation value between two adjacent frames is known as the spectrum flux.

$$SF = \frac{1}{(N-1)(K-1)} \sum_{n-1}^{N-1} \sum_{k-1}^{K-1} [\log(F(n,k) + \partial) - \log(F(n-1),k) + \partial]^2 \quad (1)$$

Where $F(n,k)$ is the DFT of the input signal.

$$F(n,k) = \sum_{n=-\infty}^{\infty} x(n) w(mL - n) e^{-j(2\pi/L)kn} \quad (2)$$

Where $x(n)$ is the original input signal, L is the length of the window, k is the order of transform and N is the number of frames.

2.**Zero crossing rate** [5]: The zero-crossing rate is the estimate of sign-changes along a signal, i.e., the rate at which the signal changes from negative to positive or vice versa. This feature has been used heavily in both speaker identification and to retrieve music information from music signal which is further useful to distinguish percussive sounds.

ZCR is described as

$$ZCR = \frac{1}{T-1}\sum_{t=1}^{T-1} II\{X(t)X(t-1) < 0\} \quad (3)$$

Where X is a signal of duration T and the indicator function II is 1 if its argument X(t)X(t-1) is true and 0 otherwise.

In some cases, instead of all positive and negative zero crossings either only the "positive-going" or only "negative-going" crossings are counted. Since, between a pair of adjacent negative zero-crossings there must be one and only one positive zero-crossing and vice versa.

Noise frame ratio [5]: The ratio of noise frames in a given audio file is called as the noise frame ratio (NFR). If the peak value of normalized correlation function of music file is less than the fixed threshold then it is called as noise frame. NFR value for a noise environment is higher than that of music.

E. Silence in a Signal

Silence [11], [12] is the absence of perceptible sound or existence of minimum intensity sound signal. Music basically depends on silence in different forms to differentiate remaining portions of sound and allow characteristics, melodies and rhythms to have higher influence. For example, most music scores feature *rests* represents durations of silence as shown in Figure 2.

Fig.2: Representation of silence part in a signal

Apart from that silence in music can be seen as a time for examine to reflect on the piece. The listeners feel the effects of the music tones and can reflect on that instant deliberately. Silence doesn't obstruct musical brilliance but can enlarge the sounds of instruments and vocals within the piece. Some composers consider the use of silence in music to an utmost. *4'33"* is an experimental musical work by avant-garde composer John Cage. Though first did on the piano, the piece was formulated for any instrument or instruments and is arranged in three movements. The composer can able to set the length of the combination of three movements but it is not possible to set individual length of the each movement.

III. PROPOSED METHOD

A. Block Diagram

In general, the following steps are executed: First, two feature sequences are extracted frame by frame for the whole audio signal. Second, for each sequence two thresholds are dynamically estimated. Third, a simple thresholding criterion is applied on the sequences. Fourth, Speech segments are detected based on the above criterion and finally a simple post-processing stage is applied.

The block diagram showing the sequence of steps for the proposed method is shown in Figure 3.

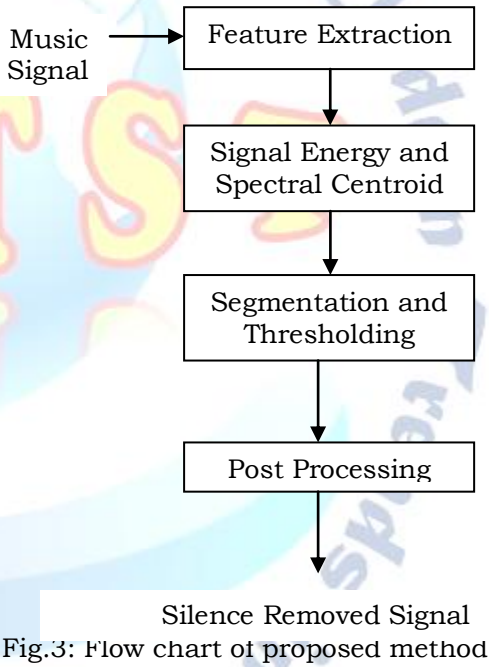

Fig.3: Flow chart of proposed method

B. Feature extraction

In order to improve the accuracy of segmentation for audio sequence, it is important to select good features that can extract temporal and spectral characteristics of musical sounds. To extract the features, first we have to divide the signal into frames of 50ms each. For each frame, we have to extract the features i.e., signal energy and spectral centroid. The reason to consider the above two features is their simplicity in calculation.

The frames are divided using hamming window of 50ms length. Figure 4 is an example which shows the hamming windowed signal.

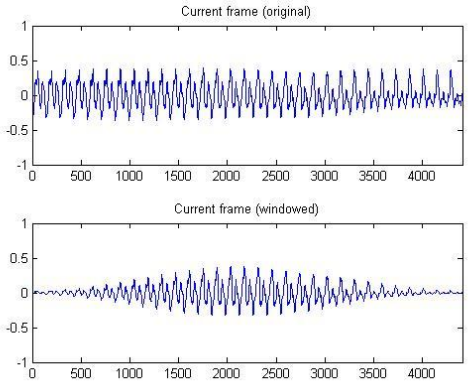

Fig.4: Segmented waveform of a signal

1. Signal Energy: The amplitude of signal varies over time. The energy of the speech signal provides a representation that reflects these amplitude variations. Let $x_i(n)$; n = 1… N is the number of music samples existed in the i^{th} frame which is having length N. Then, for each frame i the energy is computed based on equation (4):

$$E(i) = \sum_{n=1}^{N} | X_i(n) |^2 \quad (4)$$

2. Spectral Centroid: The spectral centroid is generally correlated with the measure of the intensity of the sound. This measure is acquired by calculating the spectral centroid or centre of gravity by considering frequency and magnitude information which are calculated by using the Fourier Transform. The separate centroid of each spectral frame is defined as the average frequency weighted by amplitudes, divided by the sum of the amplitudes:

$$C_i = \frac{\sum_{k=1}^{N} (k+1) X_i(K)}{\sum_{k=1}^{N} X_i(k)} \quad (5)$$

C. Segment Detection

For the classification purpose some threshold value needs to be set. If the threshold value was set to be static for example in the particular case it was observed for many words that the energy in the voiced region was greater than 30dB but it could not be taken into consideration if the words were spoken in different intensity or loudness. So threshold value was calculated dynamically according to the music data. The procedure to calculate the threshold value is as follows:

1. Compute the histogram of the feature sequence values.
2. Detect the histogram local maxima.
3. Let M1 and M2 be the positions of the first and second local maxima respectively.

The threshold value is computed using the following equation:

$$T = \frac{W.M1+M2}{W+1} \quad (6)$$

W is a user-defined parameter. Large values of W obviously lead to threshold values closer to M1.

The above thresholding is applied to both features which will give two threshold values T1 and T2 corresponding to signal energy and spectral centroid. Based on the values of T1 and T2 musical signal is separated into segments.

For the segment whose values are lesser than T1 and T2, corresponds to silence part and remaining segments are considered as music. Thus the silence removal task from musical signal is accomplished.

D. Post Processing

As a post-processing step, the detected speech segments are lengthened by 5 short term windows (i.e., 250msec), on both sides. Finally, successive segments are merged.

IV. Experimental Results

We focus our attention on musical instruments which are pitched percussive in nature. The results are analyzed using Matlab simulations. The performance of various instruments like drum, flute cajon, kick and Djembe is shown.

In Figure 5, subplot 5(a) shows the energy of the drum in time domain with respect to frames, 5(b) shows the spectral centroid of the drum and 5(c) shows the music after removal of the silence. Similar analysis is done for the other instruments and some are shown in Figures 6 and 7.

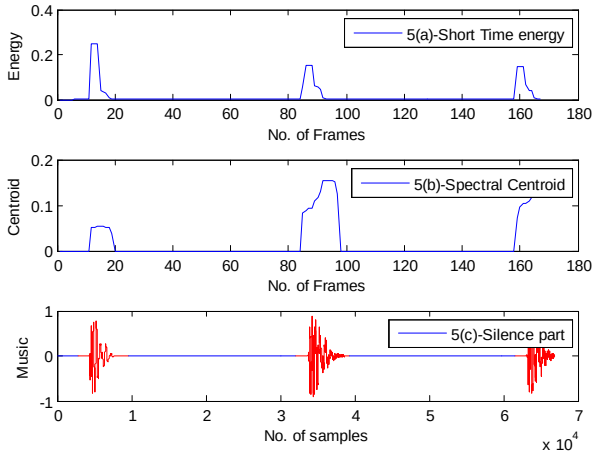

Fig.5: Result of silence removal in drum sound.

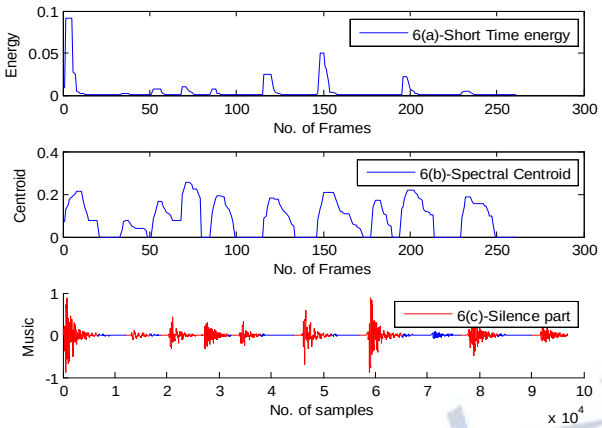

Fig. 6: Results of silence removal in Cajon sound

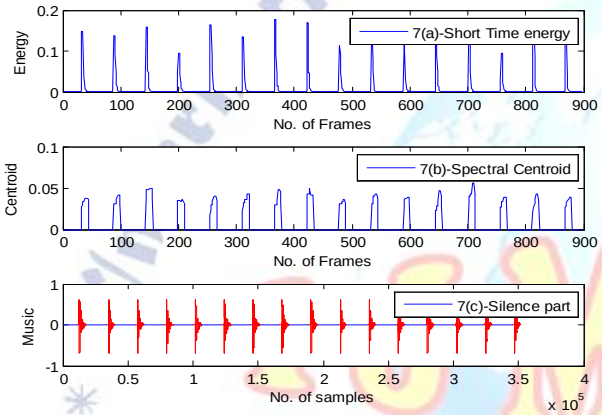

Fig.7: Results of silence removal in Kick1 sound

Table – 1: Simulation Results

Name of the Instrument	No. of Effective Voiced Segments	No. of Detected Segments
Drum	3	3
Flute	5	5
Cajon	10	9
Kick1	16	16
Kick2	1	1
Djembe	9	8
Total	44	42

Simulation results are shown in the Table – 1. The proposed method is applied to six different instruments. This method detects 42 segments out of 44 effective voiced segments. The segmentation accuracy is 95.45%.

V. Conclusion

Segmentation and silence removal are applied to sounds produced by percussive musical instruments. This proposed method is completely based on two spectral features, signal energy and spectral centroid. So, it is shown to be computationally efficient for real time applications. Silence part is removed from the sound. And we observed that the number of segments for each sound is different. The threshold is exclusively stated and there is no requirement for trial and error method.

References

[1] Saima Anwar Lashari, Rosziati Ibrahim and Norhalina Senan, "Soft Set Theory for Automatic Classification of Traditional Pakistani Musical Instruments Sounds", 2012 International Conference on Computer & Information Science (ICCIS), 978-1-4673-1938-6/12/$31.00 ©2012 IEEE.

[2] A. Pikrakis, T. Giannakopoulos, and S. Theodoridis, "An overview of speech/music discrimination techniques in the context of audio recordings", vol. 120, pp. 81–102, 2008.

[3] J. Saunders, "Real-time discrimination of broadcast speech/music", Proceedings of the Acoustics, Speech, and Signal Processing (ICASSP96), pp. 993–996.

[4] Norhalina Senan , Rosziati Ibrahim, Nazri Mohd Nawi , Musa Mohd Mokji, "Feature Extraction for Traditional Malay Musical Instruments Classification System", 978-0-7695-3879-2/09 $26.00 © 2009 IEEE

[5] Bojana GajiL, Kuldip K. Paliwa, "Robust Feature Extraction using Subband Spectral Centroid Histograms", l0-7803-7041-4/01/$10.00 02001 IEEE

[6] S.E Tanter, K. Yu, G. Evermann, P.C.Woogland , "Generating and evaluating segmentations for automatic recognition of conversational telephone speech", In proc., ICASSP, Montreal, Canada, May 2004, Vol-1, pg 753-756.

[7] Qi Li, , Jinsong Zheng, Augustine Tsai, and Qiru Zhou, "Robust Endpoint Detection and Energy Normalization for Real-Time Speech and Speaker Recognition", IEEE Transactions On Speech And Audio Processing, Vol. 10, No. 3, March 2002

[8] Naoki Nitanda, Miki Haseyama, and Hideo Kitajima, "Audio Signal Segmentation and Classification For Scene-Cut Detection", IEEE transactions on speech and audio, 2005, pg: 4030-4033.

[9] L. Lu, H. Zhang, and S.Z. Li, "Content-Based Audio Classification and Segmentation by using Support Vector Machines", Multimedia Systems, vol. 8, no. 6, pp. 482–492, 2003.

[10], Jia Min Karen Kua, Tharmarajah Thiruvaran, Mohaddeseh Nosratighods Eliathamby Ambikairajah, Julien Epps, "Investigation of Spectral Centroid Magnitude and Frequency for Speaker Recognition", Odyssey 2010 The Speaker and Language Recognition Workshop 28 June – 1 July 2010, Brno, Czech Republic

[11] Dabrowski, A. ; Marciniak, T. ; Krzykowska, A. ; Weychan,R. "Influence of silence removal on speaker recognition based on short Polish sequences", Signal Processing Algorithms, Architectures, Arrangements, and Applications Conference Proceedings (SPA),2011 Publication Year: 2011 , Page(s): 1 – 5.

[12] Yingyong Qi, Bobby R. Hunt, "Voiced-Unvoiced-Silence Classifications of Speech Using Hybrid Features and a Network Classifier", IEEE Transactions On Speech And Audio Processing, Vol. 1, No. 2, April 1993.