A POSTPROCESSING TECHNIQUE FOR IMPROVED HARMONIC / PERCUSSION SEPARATION FOR POLYPHONIC MUSIC

Balaji Thoshkahna
Dept. of Electrical Engineering,
Indian Institute of Science,
Bangalore,India
balajitn@ee.iisc.ernet.in

K.R.Ramakrishnan
Dept. of Electrical Engineering,
Indian Institute of Science,
Bangalore,India
krr@ee.iisc.ernet.in

ABSTRACT

In this paper we propose a postprocessing technique for a spectrogram diffusion based harmonic/percussion decomposition algorithm. The proposed technique removes harmonic instrument leakages in the percussion enhanced outputs of the baseline algorithm. The technique uses median filtering and an adaptive detection of percussive segments in subbands followed by piecewise signal reconstruction using envelope properties to ensure that percussion is enhanced while harmonic leakages are suppressed. A new binary mask is created for the percussion signal which upon applying on the original signal improves harmonic versus percussion separation. We compare our algorithm with two recent techniques and show that on a database of polyphonic Indian music, the postprocessing algorithm improves the harmonic versus percussion decomposition significantly.

1. INTRODUCTION

Music source separation has been a very important topic of research with applications in transcription [1], audio coding [2], enhancement [3] and personalization [4]. Source separation involves separating a polyphonic mono or stereo music into its component instrument streams. As a preliminary step towards source separation, decomposition of the music signal into separate harmonic and percussive instrument streams has been a popular approach in recent years. The percussive instrument stream can be used for drums transcription [5], rhythm analysis [6], audio remixing [3] among the many applications. It has been shown that the percussion stream results in better drum transcription [5, 7] than the original music itself. Likewise, the harmonic instruments stream can be used for multipitch estimation [1], pitch modification [4], note transcription and lead vocals extraction [8] with greater ease.

McAulay et al. [9] first used sinusoidal modeling to decompose a signal into harmonic and noise components popularly known as the “sine+noise“ model. Verma et al. [10] introduced the idea of modeling transients in a signal leading to the development of ”sine+transients+noise” model. Various improvements to these models have been proposed in [11, 12]. Gillet et al. [7] used noise subspace projections to split polyphonic music into harmonic and noise components with the noise components predominantly having the percussive instruments. he noise signal was used for drum transcription and was found to be more effective than the original for the same task. Yoshii et al. [5] used a template based approach for harmonic instrument suppression to extract drums sounds from polyphonic music for transcription. Recently Ono et al. [3, 13] presented an iterative algorithm using spectrogram diffusion to split music signals into the component harmonic and percussion streams. The percussion streams were used for remixing and equalisation purposes. Fitzgerald [14] proposed a much simpler alternative to Ono’s algorithm using median filtering.

But most of the above discussed algorithms are aimed at Western music and specifically pop music which has strong percussion accompaniments. These algorithms do not perform well for Indian music which has somewhat muted percussion (often used just to give a basic beat to the lead instrument/vocalist) and an increased amount of vibratos in the instrumental sections. This leads to a lot of leakages of percussion into the harmonic stream and vice versa.

In this paper we develop a postprocessing technique that can applied to the output of Ono’s algorithm (called the baseline from here onwards) [3, 13] mentioned above. In Section 2 we briefly describe the baseline algorithm to establish the framework for our algorithm. Section 3 describes our post processing technique. The necessary framework to test the algorithm, the experiments and comparative results are described in Section 4. We conclude the paper in Section 5.

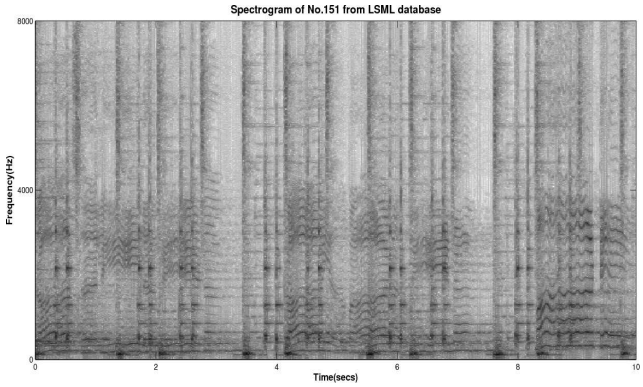

Figure 1. Spectrogram of song No.151 from LSML database. The strong vertical stripes are locations of percussion and the horizontal stripes are the harmonics of pitched instruments. The wavy horizontal lines between 8 secs and 10 secs are the vibratos in the lead male singing.

2. ONO'S ALGORITHM AND SHORTCOMINGS

The spectrogram diffusion based harmonic/percussion separation algorithm proposed by Ono et al. [3, 13] assumes that steady harmonic instruments show up as horizontal lines while percussive instruments show up as vertical lines in the signal spectrogram. This is because of the steady nature of harmonic instruments that play enduring discrete notes while percussive instruments have a short time burst of energy leading to a wideband spectral structure as shown in Figure 1. The diffusion algorithm uses a minimization of the spectrogram's vertical and horizontal derivatives using an auxiliary function approach.

Let $x[n]$ be a monaural polyphonic music signal sampled at 16kHz. Let $X(i,j)$ denote its STFT (Short Time Fourier Transform) at the i^{th} frequency bin and j^{th} frame. Let $W(i,j)$ be the range compressed version of the power spectrogram given by,

$$W(i,j) = |X(i,j)|^{2\gamma}, \tag{1}$$

where $\gamma = 0.3$.

Similarly let $H(i,j)$ and $P(i,j)$ represent the power spectrograms of the component harmonic and percussion signals.

A cost function $J(H,P)$ defined as below is used to minimize the gradients of the spectrograms.

$$J(H,P) = \frac{1}{\sigma_H^2}\sum_{i,j}\left(H(i,j) - H(i,j-1)\right)^2 + \frac{1}{\sigma_P^2}\sum_{i,j}\left(P(i,j) - P(i-1,j)\right)^2. \tag{2}$$

Then, we wish to find H and P that minimize the equation (2) under the constraint,

$$W = P + H. \tag{3}$$

An iterative update method using auxiliary function approach is used for the minimization of equation (2). This leads to the decomposition of the signal $x[n]$ into its component percussion and harmonic spectrograms P and H respectively for various values of the diffusion coefficient α ($0 < \alpha < 1$). P and H are "binarized" to P_{bin} and H_{bin} as in equations (4, 5) to attentuate the interference of harmonic instruments in the percussive stream and vice versa.

$$P_{bin}(i,j) = \begin{cases} X(i,j) & \text{if } P(i,j) > H(i,j), \\ 0 & \text{if } P(i,j) \le H(i,j). \end{cases} \tag{4}$$

$$H_{bin}(i,j) = X(i,j) - P_{bin}(i,j). \tag{5}$$

Depending on the value of α, either the percussive stream will be emphasized or the harmonic stream will be emphasized. The percussive and harmonic streams $p[n]$ and $h[n]$ are reconstructed by inverting the STFTs P_{bin} and H_{bin} respectively using the phase of the original signal $x[n]$ (at each frame during inversion) .

One of the shortcomings of this algorithm has been the leakage of harmonic instrument components into the P_{bin} component and the leakage of low strength percussion into the H_{bin} portion. As noted earlier there is a high presence of vibratos and muted percussion (tabla, mridangam[1]) in Indian music. This leads to a very bad decomposition scheme using baseline algorithm. A much faster algorithm using median filtering has been proposed in [14], but even that algorithm suffers from the same shortcomings.

3. THE PROPOSED ALGORITHM

We use only the percussion stream $p[n]$ from the baseline algorithm and the original signal $x[n]$ for the postprocessing technique we propose. Since percussion appears as a wideband signal in the spectrum and different harmonic instruments have different frequency characteristics, not all regions of the spectrum are equally affected by the harmonic leakage. Therefore we intend to remove the leakages using subband processing. The signal $p[n]$ is passed through an even stacked cosine modulated perfect reconstruction filterbank of 16 filters. The filterbank was designed using the LT-TOOLBOX[2] set of Matlab routines. The following operations are performed on each subband signal. Let $p_i[n]$ be the output of the i^{th} subband. The signal is split into frames of $40ms$ (Frame length $N_l = 640$ samples) with an overlap of $20ms$ (Frame shift $N_o = 320$ samples). Each frame is multiplied with a triangular window of length N_l samples

[1] A south Indian classical instrument

[2] http://www.students.tut.fi/ jalhava/lt/intro.html

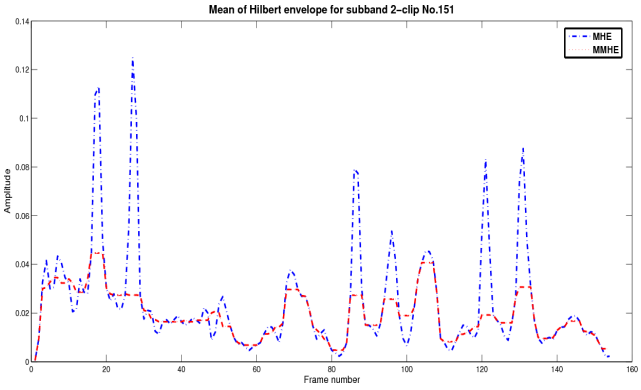

Figure 2. A plot of $\mathcal{H}_\mu$ (blue dash-dot) and $\mathcal{M}$ (red dotted) for subband 2 for No.151 from LSML database.

to facilitate the overlap and add at the reconstruction stage. The j^{th} frame is represented as $\mathcal{P}_i(j,:)$.

As noted by Scheirer [15], the amplitude envelope is more important than the frequency content for the perception of percussion. Therefore we intend to manipulate the envelope of the subband signals. The Hilbert envelope of a signal has been exploited for detection of transients in polyphonic music with great success [16]. We intend to use the same framework with a view of including temporal noise shaping (TNS) [2] for each frame in our future work.

Let the Hilbert transform for the j^{th} frame be $\hat{\mathcal{P}}_i(j,:)$. We find the Hilbert envelope of the signal [16] as:

$$\mathcal{H}_i(j,:) = \sqrt{\mathcal{P}_i(j,:)^2 + \hat{\mathcal{P}}_i(j,:)^2}. \quad (6)$$

We now use the sample mean of $\mathcal{H}_i(j,:)$ as a representative for the j^{th} frame (We also tried with the energy of each frame as a representative and the method works just as fine, but since we intend to use TNS in our future work, we choose to retain the Hilbert envelope within each frame).

$$\mathcal{H}_\mu(i,j) = \frac{1}{N}\sum_{k=1}^{N}\mathcal{H}_i(j,k). \quad (7)$$

$\mathcal{H}_\mu$ is used to detect the frames having percussion and harmonic instruments. In order to do this, $\mathcal{H}_\mu$ is median filtered with a l point median filter.

$$\mathcal{M}(i,j) = median\{\mathcal{H}_\mu(i, j-k : j+k), k = l-1/2\}, \quad (8)$$

where we used $l = 7$. We used a value of $l = 7$ since a median filter whose length is greater than the duration of the transient noise can suppress it [17] and most percussive transients are around 60-100ms long (3 to 5 frame shifts and hence we used the next odd numbered window length).

As shown in Figure 2, in $\mathcal{H}_i$ and $\mathcal{M}$ the presence of harmonic instruments creates a change of shape in the usual gamma function envelopes of percussion signals [18]. Therefore, the novelty function ρ, defined as the ratio between $\mathcal{H}_\mu$ and $\mathcal{M}$,

$$\rho(i,j) = \mathcal{H}_\mu(i,j)/\mathcal{M}(i,j), \quad (9)$$

is low at places of leakage while it retains a high value if percussion is present [17].

We now use two possible methods of finding a good threshold for detecting percussion in ρ. In the first method, we find the mean (μ) and variance (σ) of ρ for each subband. The threshold for the i^{th} subband, $T(i)$ is computed as,

$$T(i) = min(1.75, \mu(i) + 0.5 * \sigma(i)). \quad (10)$$

This threshold was decided empirically after testing on a small dataset of audio clips and is similar to the one used in [19].

In the second method, we assume that we have polyphonic audio with utmost 10% of the values of ρ are due to percussion. This is akin to the assumption that we have 2 percussion hits of $50ms$ duration per second of the signal. We find a threshold from the histogram of ρ such that 10% of the values of ρ lie to the right and the remaining 90% lie to the left of the threshold in the histogram.

We use the threshold obtained from the first method since optimization process for the second approach is still under development at the time of writing this paper. We now use the threshold to determine the set of local maxima within each subband that belong to the percussion as:

$$\mathcal{F}(i,j) = \begin{cases} 1 & \text{if } \mathcal{H}_\mu(i,j) > T(i).\mathcal{M}(i,j) \\ 0 & \text{otherwise.} \end{cases}, \quad (11)$$

We locate local maxima in $\mathcal{H}_\mu$ for each subband and retain only frames corresponding to them in $\mathcal{F}$ while the rest of the frames are made 0. Since a percussive signal has a gamma function envelope, it has a minima to both sides of the local envelope maxima on the time axis. Upon finding the local maxima in the signal, we need to find the local minima on both its sides on the time axis in order to fully reconstruct the percussive signal as shown in Figure 3.

We rebuild the exact percussion signal by using the first local minima to the temporal left and right of each detected maxima as shown in Figure 4. This ensures that the entire percussion signal is preserved in the envelope. The set of non-zero frames in each subband are considered as the percussive frames.

The percussive frames from each subband are finally added using the overlap and add method to generate the subband signal that is percussion enhanced. The subband signals are then passed through the synthesis filterbank to generate the new percussion signal $p_{enh}[n]$.

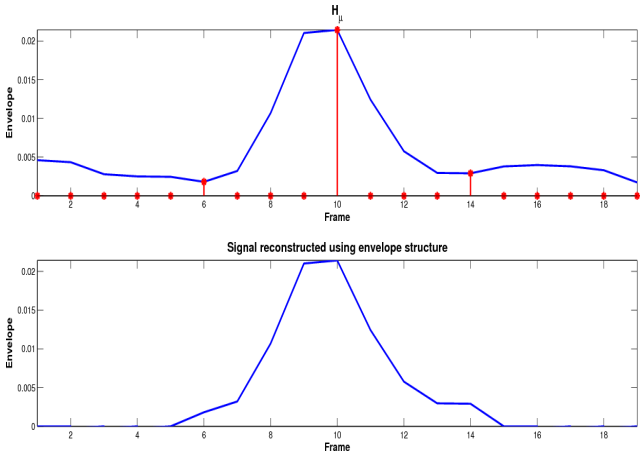

Figure 3. *Top*:A percussion envelope (solid line) and its local maximum and the minima (star). *Bottom*:Reconstructed percussion envelope using the piecewise reconstruction method described in this paper.

We use the newly generated percussion signal to enhance the H_{bin} signal given by the baseline algorithm. A STFT of the signal $p_{enh}[n]$ is computed as $P_{enh}(i,j)$. Now the STFT is averaged along the frequency axis as follows:

$$P_{avg}(i,j) = \frac{1}{m} \sum_{k=i-(m-1)/2}^{i+(m-1)/2} P_{enh}(k,j), \qquad (12)$$

where $m = 2$. P_{avg} changes from P_{enh} by a small amount if the frame is a percussive frame (since a percussion frame will have a wideband spectrum) while its value changes significantly if the frame has predominantly harmonic components. P_{avg} is compared with the spectrum of the original signal. If any component of P_{avg} is greater than a threshold ν times X, that component is assigned to percussion otherwise it is assigned to the harmonic stream of the signal.

$$\mathcal{P}_{fin}(i,j) = \begin{cases} X(i,j) & \text{if } P_{avg}(i,j) > \nu.X(i,j) \\ 0 & \text{otherwise.} \end{cases} \qquad (13)$$

$$\mathcal{H}_{fin}(i,j) = X(i,j) - \mathcal{P}_{fin}(i,j). \qquad (14)$$

We used a value of $\nu = 0.45$ in order to enhance even weak percussive segments.

The $\mathcal{P}_{fin}$ and $\mathcal{H}_{fin}$ are inverted to obtain the improved percussion $p_{fin}[n]$ and harmonic $h_{fin}[n]$ stream of the signal $x[n]$. As can be seen in Figure 5, the postprocessing reduces the harmonic leakages very well. In the next section we compare our output with both the baseline algorithm and Fitzgerald's algorithm.

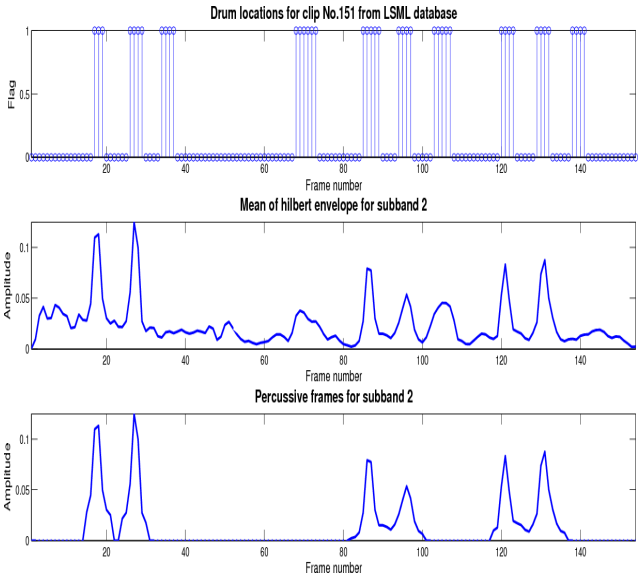

Figure 4. *Top*:Ground truth locations of percussion in No.151 from LSML database.*Middle*:Plot of $\mathcal{H}_{\mu}$ for subband 2.*Bottom*:Percussion located by signal rebuilding after local maxima detection.

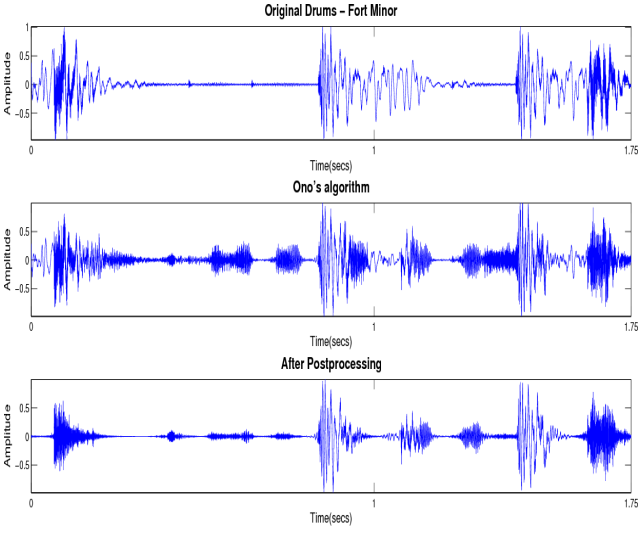

Figure 5. *Top*:Ground truth locations of percussion in the clip "Remember The Name-Fort Minor" from MTG-MASS database. *Middle*:Percussion stream from baseline Ono's algorithm.*Bottom*:Percussion stream after postprocessing.

4. EXPERIMENTS AND RESULTS

Since we did not have the individual instrument streams for Indian music as with the case in [13] for testing the efficacy of harmonic/percussion separation, we developed our own procedure as elaborated below.

To compare the working of our postprocessing technique, we prepared a database of 26 clips from various Indian film songs and also Western music songs. All songs have been sampled at 16kHz and are an average 10 seconds long. Each song was manually annotated using the gating technique [20] for percussive transients by two people independently. We annotated drums, mridangam, tabla, shakers and bass guitar slaps as percussive instruments. The percussive portions common to both the annotations were retained as the ground truth. We will call this the LSML database.

In order to compare the output of our postprocessing technique with the baseline Ono's algorithm, we derive the following measure.

Let $p[n]$ and $h[n]$ be the outputs of the baseline algorithm and $p_{fin}[n]$ and $h_{fin}[n]$ be the outputs of our postprocessing technique on the baseline algorithm. We now split each of these signals into frames of $40ms$ with an overlap of $20ms$. The energy in each frame of p and h are calculated as:

$$E_p(l) = \sum_{k=(l-1).N_o}^{(l-1).N_o+N_l} p^2[k], \tag{15}$$

$$E_h(l) = \sum_{k=(l-1).N_o}^{(l-1).N_o+N_l} h^2[k]. \tag{16}$$

$$\tag{17}$$

Similarly the energy for p_{fin} and h_{fin} are computed and stored in E_{pfin} and E_{hfin} respectively.

We now compare the energies between E_p and E_{pfin}. Since both p and p_{fin} are percussive components, we use the ground truth to find the total energy in the non-percussive frames of both these signals. Let $\mathcal{F}_\mathcal{P}$ represent the set of frames marked as percussive and $\mathcal{F}_\mathcal{H}$ represent the non-percussive frames. Then we find the energy in the percussive and non-percussive frames of $p[n]$ as:

$$E_p^P = \sum_{l \in \mathcal{F}_\mathcal{P}} E_p(l), \tag{18}$$

$$E_p^H = \sum_{l \in \mathcal{F}_\mathcal{H}} E_p(l). \tag{19}$$

Similarly we compute the same for the p_{fin} as E_{pfin}^P and E_{pfin}^H.

We now compare the energies E_p^H and E_{pfin}^H after normalizing the energies E_p^P and E_{pfin}^P. We compute β_P, where,

$$\beta_P = \frac{E_p^P}{E_{pfin}^P}. \tag{20}$$

Now ,

$$\Gamma_P = \frac{E_p^H}{\beta_P . E_{pfin}^H}, \tag{21}$$

computes the ratio between energies in the non-percussive frames of p and p_{fin} when the energies in the percussive frames are equal. A value of $\Gamma_P > 1$ indicates that the signal p_{fin} has lesser energy than p in the non-percussive segments.

Likewise, we compute the ratio Γ_H by normalizing the energies of h and h_{fin} in the non-percussive sections and finding the ratio of the energies in the percussive sections as,

$$\Gamma_H = \frac{E_h^P}{\beta_H . E_{hfin}^P}, \tag{22}$$

where β_H is ,

$$\beta_H = \frac{E_h^H}{E_{hfin}^H}. \tag{23}$$

We form Γ_{Tot} as,

$$\Gamma_{Tot} = \Gamma_P + \Gamma_H, \tag{24}$$

to give us an overall measure of how well p_{fin} and h_{fin} compare with p and h respectively. Γ_{Tot} attains a value of 2 when the baseline algorithm is compared with itself.

We show the performance of our postprocessing algorithm (PP1) and Fitzgerald's method against the baseline Ono's technique in Figure 6. Both the postprocessing technique and Fitzgerald's technique are compared against the baseline algorithm. As can be seen, our method performs better than both Fitzgerald's technique and the baseline algorithm for any value of diffusion coefficient α. Also, the postprocessing technique performs better for a lower diffusion coefficient α. With increasing α, energy in the percussion stream p_{bin} decreases and hence the leakage too decreases. Therefore our postprocessing algorithm performs better for lower α.

5. FUTURE WORK AND CONCLUSIONS

In this paper we have proposed a simple postprocessing technique for Ono's harmonic/percussion decomposition algorithm using no prior information about the sources except their production mechanism and the envelope structure. We also are currently working on a technique that uses the harmonic stream along with the percussive stream for improved separation.

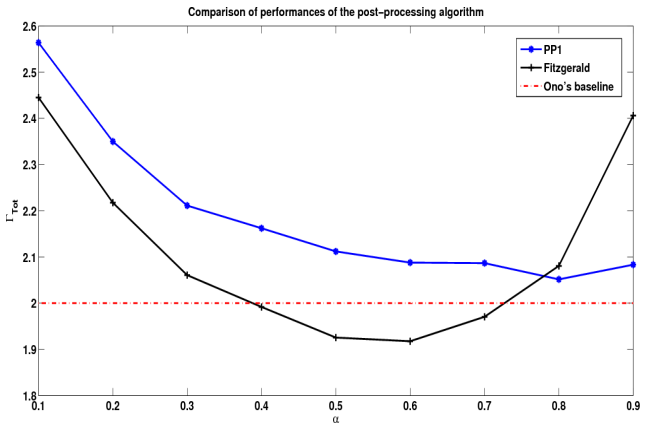

Figure 6. Performance of the postprocessing technique (PP1) against the Fitzgerald's method and Ono's baseline algorithms for varying α.

6. REFERENCES

[1] A. Klapuri: "Automatic Transcription of Music", *MSc Thesis, Tampere University*,1998.

[2] J. Herre: "Temporal Noise Shaping, Quantization and Coding methods in Perceptual Audio Coding: A Tutorial Introduction", *AES 17th Intl Conference:High-Quality Audio Coding*,1999.

[3] N. Ono and K. Miyamoto and H. Kameoka and S. Sagayama: "A Real-Time Equalizer of Harmonic and Percussive Components in Music Signals", *Intl Society of Music Information Retrieval Conference)*,2008.

[4] M. Shashanka, P.Smaragdis, B. Raj and R.Singh: "Separating a Foreground Singer from Background Music.", *International Symposium on Frontiers of Research on Speech and Music (FRSM)*,2007.

[5] K. Yoshii, M. Goto and G. Okuno: "Drum Sound Recognition for Polyphonic Audio Signals by Adaptation and Matching of Spectrogram Templates with Harmonic Structure Suppression", *IEEE Transactions on Audio, Speech and Language Processing*,Vol-15,No.1,2007.

[6] J. Paulus and A. Klapuri: "Measuring the Similarity of Rhythm Patterns", *Intl Society for Music Information Retrieval Conference*,2002.

[7] O. Gillet and G. Richard: "Transcription and Separation of Drums Signals from Polyphonic Music", *IEEE Transactions on Audio, Speech and Language Processing*,Vol-16,No.3,2008.

[8] Y. Li and D. Wang : "Separation of Singing Voice from Music Accompaniment for Monaural Recordings", *IEEE Transactions on Audio, Speech and Language Processing*,Vol-15,No.4,2007.

[9] R. McAulay and T. F. Quatieri: "Speech Analysis/Synthesis Based on a Sinusoidal Representation", *IEEE Transactions on Acoustics,Speech and Signal Processing*,Vol 34,pp-744-754,1990.

[10] T. S. Verma and S. N. Levine and T. H. Y. Meng: "Transient Modeling Synthesis:A Flexible Analysis/Synthesis Tool for Transient Signals", *Intl Computer Music Conference (ICMC)*,1997.

[11] X. Serra and J. O. Smith: "Spectral Modeling Synthesis:A Sound Analysis/Synthesis System Based on a Deterministic plus Stochastic Decomposition", *Computer Music Journal*,Vol 14(4),pp-14-24,1990.

[12] R.Badeau, R. Boyer and B. David: "EDS Parametric Modeling and Tracking of Audio Signals", *Intl Conference on Digital Audio Effects (DAFx)*,2002.

[13] N. Ono and K. Miyamoto and J. Le Roux and H. Kameoka and S. Sagayama: "Separation of a Monaural Audio Signal into Harmonic/Percussive Components by Complementary Diffusion on Spectrogram", *European signal Processing Conference(EUSIPCO))*,2008.

[14] D. Fitzgerald: "Harmonic/Percussive Separation using Median Filtering", *Intl Conference on Digital Audio Effects (DAFx)*,2010.

[15] E. Scheirer: "Music Listening Systems", *PhD Thesis, Massachusetts Institute of Technology*,2000.

[16] F. X. Nsabimana and U. Zolzer: "Transient Encoding of Audio Signals using Dyadic Approximations", *Intl Conference on Digital Audio Effects (DAFx)*,2007.

[17] I. Kauppinen: "Methods for Detecting Impulsive Noise in Speech and Audio Signals", *Intl Conference on Digital Signal Processing*,2002.

[18] M. G. Christensen and S. van de Par: "Efficient Parametric Coding of Transients", *IEEE Transactions on Audio, Speech and Language Processing*,Vol-14,No.4,2006.

[19] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies and M. B. Sandler : "A Tutorial on Onset Detection in Music Signals", *IEEE Transactions on Audio, Speech and Language Processing*,Vol-13,No.5,2005.

[20] D. J. Hermes: "Vowel Onset Detection", *Journal of Acoustical. Society. of America*,Vol-87(2), 866-873,1990.