Drum Source Separation using Percussive Feature Detection and Spectral Modulation

Dan Barry[ϕ], Derry Fitzgerald[^], Eugene Coyle[ϕ] and Bob Lawlor[*]

[ϕ] Digital Audio Research Group, Dublin Institute of Technology, Kevin St. Dublin, Ireland
barrydn@eircom.net & eugene.coyle@dit.ie

[^] Dept. Electrical Engineering, Cork Institute of Technology, Rossa Avenue, Bishopstown, Cork, Ireland
derry.fitzgerald@cit.ie

[*] Dept. of Electronic Engineering,
National University of Ireland, Maynooth, Ireland
rlawlor@eeng.may.ie

Abstract -- **We present a method for the separation and resynthesis of drum sources from single channel polyphonic mixtures. The frequency domain technique involves identifying the presence of a drum using a novel percussive feature detection function, after which the short-time magnitude spectrum is estimated and scaled according to an estimated time-amplitude function derived from the percussive measure. In addition to producing high quality separation results, the method we describe is also a useful pre-process for drum transcription techniques such as Prior Subspace Analysis in the presence of pitched instruments.**

Keywords-**Audio, Signal Processing, Source Separation, Drum.**

I INTRODUCTION

In recent years, some focus has shifted from pitched instrument transcription to drum transcription; and likewise in the field of sound source separation, some particular attention has been given to drum separation in the presence of pitched instruments [1]. Where metadata generation for music archive and retrieval systems is concerned, rhythm analysis is particularly important since broad genre categorization can be ascertained from simplistic aspects of rhythm such as tempo and meter. Automatic drum separation would facilitate more accurate transcription, thus giving access to the finer temporal aspects of rhythm such as polyrhythm and syncopation. Quite apart from this, drum separation and transcription is in itself a useful tool in such applications as computerised music education. Where the music consists of drums only, some existing algorithms give reasonably accurate results [2], however, in the presence of pitched instruments, the algorithms become less robust and less accurate by way of false beat detection and indeed missing beats altogether [3]. A drum separation algorithm in this case would be a viable pre-process in order to overcome some of the problems associated with drum transcription in the presence of pitched instruments. Algorithms such as ADRess [4] and those described in [5] are capable of drum separation in stereo signals if certain constraints are met. In particular, the drums must occupy a unique position within the stereo field. This condition of course is not always met and it is usually the case in popular music that elements of the drum kit share a stereo field position with other instruments. Other algorithms such as [6, 7] have attempted drum separation from single polyphonic mixture signals with varying results. The quality in these cases is usually described as tolerable for the purposes of rhythmic signature analysis. We present a fast and efficient way to decompose a spectrogram using a simple technique which involves percussive feature detection and spectral modulation which results in the extraction of the drum parts from a polyphonic mixture. The algorithm is applicable for the separation of almost any audio features which exhibit rapid broadband fluctuations such as drums in music or plosives, fricatives and transients in speech.

II METHOD OVERVIEW

Most of the drums used in popular music can be characterised by a rapid broadband rise in energy followed by a fast decay. This is particularly true of the kick and snare drum which could be considered as the most common drums found in modern music. Pitched instruments on the other hand will generally only exhibit energy at integer multiples of some fundamentals which correspond to the notes played in the music. There are of course exceptions in the case of mallet and hammer instruments which may exhibit drum like onsets prior to the stable harmonic regions of the note. With this in mind we develop an onset detector which is not concerned with measuring the rapid rises in energy; but rather an onset detector that measures the broadband nature or "percussivity" of the onset, independent of the actual energy present. In this way drum hits of varying velocity will be detected equally. A percussive temporal profile is derived by analysing each frame of a short-time Fourier transform (STFT) of the signal and assigning a percussive measure to it. The frame is then scaled according to this measure. It should be seen then that regions of the spectrogram with low percussive measures will be scaled down significantly. Upon resynthesis, only the percussive regions remain. Effectively the spectrogram is modulated by an envelope corresponding to the percussion detected within the signal.

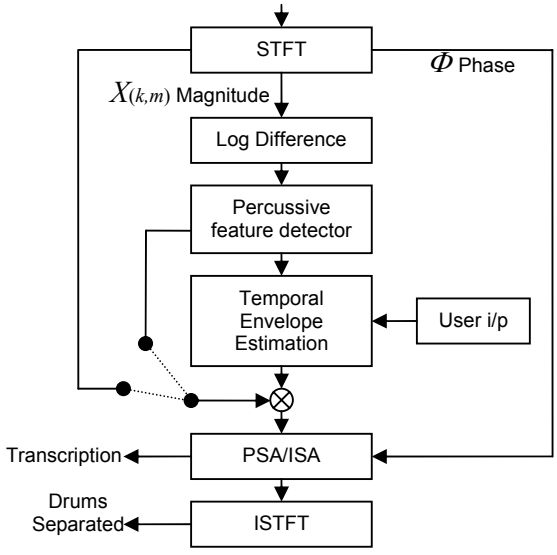

Figure 1: System Overview.

The figure above illustrates the general operation of the algorithm. The magnitude STFT of the signal is taken and the phase information is retained for resynthesis purposes later on. The log difference of each frequency component between consecutive frames is then calculated. This measure effectively tells us how rapidly the spectrogram is fluctuating. If the log difference exceeds a user specified threshold, it is deemed to belong to a percussive onset and a counter is incremented. The final value of this counter, once each frequency bin has been analysed, is then taken to be the measure of percussivity of the current frame. Once all frames have been processed, we have a temporal profile which describes the percussion characteristics of the signal. This profile is then used to modulate the spectrogram before resynthesis. Some specific options for resynthesis are discussed in section IV.

III TEMPORAL ESTIMATION

Firstly we take an STFT of the signal given by:

$$X(k,m) = abs\left[\sum_{n=0}^{N-1} w(n)x(n+mH)e^{-j2\pi mk/N}\right] \quad (1)$$

where *X(k,m)* is the absolute value of the complex STFT given in equation 1 and where *m* is the time frame index, *k* is the frequency bin index, *H* is the hopsize between frames and *N* is the FFT window size and where *w(n)* is a suitable window of length *N* also. Next we take the log difference of the spectrogram with respect to time as in equation 2.

$$X'(k,m) = 20\log_{10}\frac{X(k,m-1)}{X(k,m)} \quad (2)$$

for all *m* and $1 \le k \le K$

In order to detect the presence of a drum we define a percussive measure given in equation 3.

$$Pe(m) = \sum_{k=1}^{K}\begin{cases} P(k,m)=1 & if\ X'(k,m) > T \\ P(k,m)=0 & otherwise \end{cases} \quad (3)$$

Where, *T* is a threshold which signifies the rise in energy measured in dB which must be detected within a frequency channel before it is deemed to be a percussive onset. Effectively equation 3 acts like a counter; *Pe(m)* is simply a count of how many bins are positive going and exceed the threshold. *P(k,m)* contains a '1' if the threshold condition is met and a zero otherwise. Note that the actual energy present in the signal is not significant here; we simply want a measure of how "broadband" or percussive the onset is. The figure below shows the effectiveness of this approach. Standard energy based onset detectors such as [8] will not be able to distinguish between narrowband and broadband onsets. In these systems the level of detection will be intrinsically linked to the energy of the signal at any given time. The detection function we have described is independent of energy and so can deal with low energy onsets as long as they are broadband in nature.

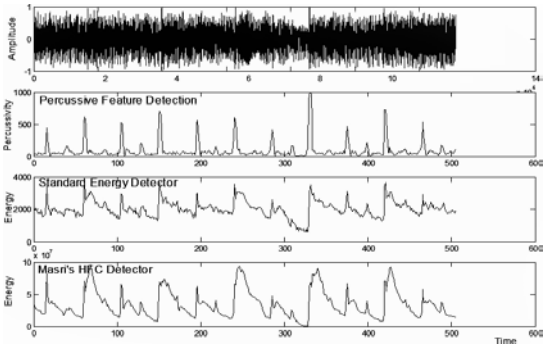

Figure2: The top plot shows the original audio clip. Plot 2 shows our percussive onset detector. The third plot shows the standard energy detector and the bottom plot shows Masri's high frequency weighted detection function [8]

Note that the percussive feature detection function we have described even manages to detect the low amplitude hi-hat strikes between the kick and snare events.

IV SPECTRAL MODULATION

By weighting each frame by the percussive measure *Pe(m)* , the spectrogram modulates in sympathy with the percussion. This results in the output of the algorithm only becoming active in the presence of a drum sound. There are some options when it comes to resynthesis; the simplest is to simply multiply the original frame by the percussive measure:

$$Y(k,m) = Pe(m)^{\Psi} X(k,m) \qquad (4)$$

for all m and $1 \le k \le K$

In order to control the decay characteristics of the percussive envelope we simply raise the percussive measure, *Pe(m),* to the power of Ψ . Larger values of Ψ will lead to faster decay. The parameter is set by the user such that satisfactory results are achieved upon audition. Equation 4 results in a time separation of the drum signals but not a frequency separation. Other sources which were present at the same time instant as the drums will also be present but will decay as the drum decays. This method is particularly useful for varying the level of the drums within a mixture signal. For this the separated drum signal is added back to the original signal in some ratio. This process allows for far greater control over the dynamic range of a signal than standard dynamic compression techniques.

The other option for resynthesis which does decouple the drums from the mixture in both the time and frequency domain is as follows:

$$Y(k,m) = Pe(m)^{\Psi} X(k,m)P(k,m) \qquad (5)$$

By multiplying the frame by the binary mask *P(k,m),* we are only resyntheising frequency components which were present during the percussive onset. This alters the timbre somewhat but it effectively suppresses non percussive sources in the mixture.

The separated drum signal is then resynthesised using the modulated magnitude spectrum with the original phase information, equation 6. It has been shown in [9] that using the original mixture phase information is more accurate than using a least squared error approximation such as that in [10].

$$y(n+mH) = w(n)\left(\frac{1}{K}\sum_{k=1}^{K} Y(k,m)\cdot e^{j\angle x_{\omega}(k,m)}\right)^{norm} \qquad (6)$$

The output must be normalised due to the fact that magnitude frames have been scaled according to the percussive measure. *w(n)* is a synthesis windowing function which is required to maintain smooth transitions at the frame boundaries since the process will alter the short-time magnitude spectrum. Since there is both an analysis and synthesis window, it is necessary to use a 75% overlap in order to have a constant sum reconstruction.

V RESULTS

The algorithm has been applied to many popular recordings and achieves high quality separations in most cases. The figure below shows the separation which has resulted from a typical piece of rock music. The drums are barely distinguishable by visual inspection in the time domain plot on top. However, the percussive feature detector has accurately discriminated between drum events and non drum events. The output of the feature detector is then used to modulate the spectrogram which is inverted to produce the bottom plot which is a time domain reconstruction of the drum events present in the signal.

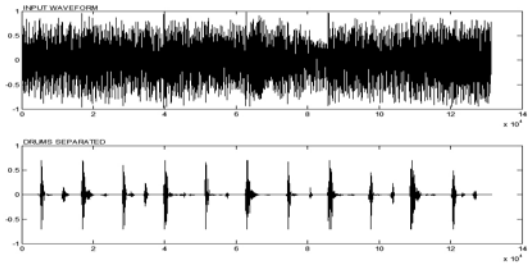

Figure 3: The plot shows the original input file and the drum separation which resulted.

To demonstrate the utility of the algorithm as a pre-processing stage before attempting drum transcription, an informal test was carried out on a highly compressed piece of audio which is a "worse case scenario" for drum transcription algorithms.

The compression we speak of is dynamic range compression as oppose to bit rate reduction compression. This sort of compression is used to increase the average level of the audio and is applied to many modern recordings in a stage known as 'mastering'. It effectively reduces peak levels and increases RMS levels dynamically, making it particularly difficult for variance based transcription techniques such as those in [2, 3] to distinguish the drums at all. The separation algorithm was applied to this recording.

Prior Subspace Analysis (PSA) [3], a technique for transcribing drums was then applied to both the unprocessed and separated spectrograms. The results obtained are shown in Tables 1 and 2. It can be seen that the use of the separation algorithm has substantially increased the performance of the PSA algorithm in transcribing drums in the presence of pitched instruments. The percentages are obtained using the following measure:

$$correct = \frac{total - undetected - incorrect}{total} \cdot 100$$

Type	Total	Missing	Incorrect	%
Snare	5	2	7	-80
Kick	6	1	2	50
Overall	11	3	9	-9

Table 1: Drum Transcription obtained using PSA on the unprocessed signal

Type	Total	Missing	Incorrect	%
Snare	5	0	0	100
Kick	6	0	1	83
Overall	11	0	1	91

Table 2: Drum Transcription obtained using PSA after the drum separation algorithm

In table 1, the percentage of detection overall is -9% (minus 9%). This was due to the fact that the PSA algorithm made several false positives, i.e. detected events which did not correspond to drum events. 2 out of 5 snares were missed and 1 out of 6 kicks were missed along with several false positives for both. The results in table 2 clearly show that the PSA algorithm has benefited greatly from the separation technique described in this paper. No events were missed and there was only one false positive in the case of the kick drum.

Independent Subspace Analysis (ISA) techniques [2] also benefit greatly when the separation algorithm presented here is used as a pre-process. The plot below shows the differences between applying ISA directly to the unprocessed audio, figure 4, and applying ISA to the separated spectrogram, figure 5.

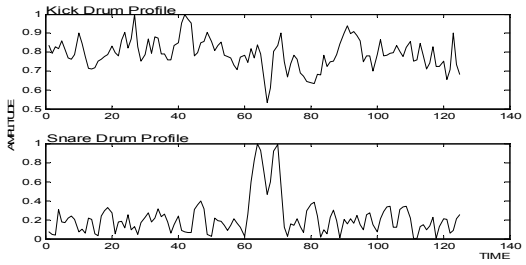

Figure 4: ISA was applied directly to the same audio clip shown in figure 3.

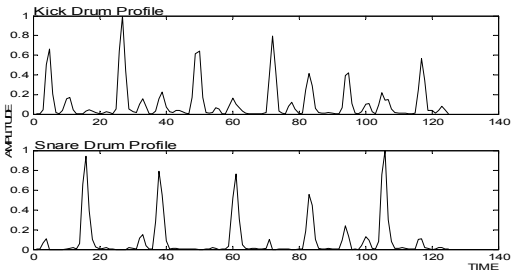

Figure 5: ISA after the separation algorithm has been applied

V CONCLUSIONS

A system capable of separating drum sources from a single polyphonic mixture has been presented. The algorithm is useful in the context of audio processing for music production and education. It has also been illustrated that the use of this algorithm as a pre-processing step for drum transcription algorithms greatly improves the transcription results.

REFERENCES

[1] M. Helen and T. Virtanen, "Separation of Drums From Polyphonic Music Using Non-Negative Matrix Factorization and Support Vector Machine," EUSIPCO 2005, Antalya, Turkey, Sept. 4-8. 2005.

[2] FitzGerald, D., Coyle E, Lawlor B., "Sub-band Independent Subspace Analysis for Drum Transcription", *Proceedings of the. Digital Audio Effects Conference (DAFX02)*, Hamburg, pp. 65-69, 2002.

[3] FitzGerald, D., Coyle E, Lawlor B., "Drum Transcription in the presence of pitched instruments using Prior Subspace Analysis" *Proc. Irish Signals and Systems Conference 2003*, Limerick. July 1-2 2003

[4] Barry, D., Lawlor, R. and Coyle E., "Real-time Sound Source Separation using Azimuth Discrimination and Resynthesis", *Proc. 117th Audio Engineering Society Convention*, October 28-31, San Francisco, CA, USA, 2004

[5] C. Avendano, " Frequency Domain Source Identification and Manipulation In Stereo Mixes for Enhancement, Suppression and Re-Panning Applications," *In Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics,* pp. 55-58 New Paltz, NY, October 19-22 2003

[6] Zils A., Pachet F., Delerue O., Gouyon F., "Automatic Extraction of Drum Tracks from Polyphonic Music Signals" *Proc. of the 2nd International Conference on Web Delivering of Music(WedelMusic2002),* Darmstadt, Germany, Dec. 9-11, 2002

[7] Uhle C., Dittmar C., Sporer T., "Extraction of drum tracks from polyphonic music using independent subspace analysis" *Proc*. of *4th International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003),* April 2003, Nara, Japan

[8] Masri P. Bateman A. 1996. "Improved modelling of attack transients in music analysis resynthesis" *in Proc. International Computer Music Conference* (ICMC). pp. 100-103

[9] Barry, D., Lawlor, R. and Coyle E., "Comparison of Signal Reconstruction Methods for the Azimuth Discrimination and Resynthesis Algorithm", *Proc. 118th Audio Engineering Society Convention*, May 28-31, Barcelona, Spain, 2005

[10] Griffin D. W., Lim J.S., "Signal Estimation from Modified Short-Time Fourier Transform", *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. ASSP-32, no. 2, April 1984