DRUM EXTRACTION FROM POLYPHONIC MUSIC BASED ON A SPECTRO-TEMPORAL MODEL OF PERCUSSIVE SOUNDS

François Rigaud, Mathieu Lagrange, Axel Röbel, Geoffroy Peeters

Analysis/Synthesis of Sound Team
IRCAM/CNRS-STMS, Paris, France
rigaud@telecom-paristech.fr, lagrange@ircam.fr, roebel@ircam.fr, peeters@ircam.fr

ABSTRACT

In this paper, we present a new algorithm for removing drums from a polyphonic audio signal. The aim of this algorithm is to discard time/frequency bins which present a percussive magnitude evolution, according to a pre-defined parametric model. Special care is taken to reduce the irrelevant removal of frequency modulated signal such as the ones produced by the singing voice. Performance evaluation is carried out using objective measures commonly used by the community. Compared with four state-of-the-art algorithms, the proposed algorithm shows competitive performances at a low computational cost.

***Index Terms*—** drum extraction, source separation, music information retrieval

1. INTRODUCTION

During the last decade, harmonic/drums separation has became a worthwhile research topic for Music Information Retrieval. For popular music, the study of drum track provides a lot of information for rhythm analysis. At the same time, removing drums from a polyphonic signal correspond to a denoising step before the harmonic part processing (multipitch, chord recognition, ...). Several algorithms have been proposed to deal with this problem according to three different approaches. **The first** approach uses Blind Source Separation methods such as the Independent Subspace Analysis [1] and the Non-Negative Matrix Factorization (NMF) [2]. The main limitations of those methods reside in assigning the separated signals into drums and harmonic ones. Also determining the correct number of basis vectors for modeling complex polyphonies remains an unsolved problem. **The second** approach is called the "match and adapt". The principle is to begin with a template of drum sound (temporal [3] or spectral [4]) defined *a priori*, search for similar patterns in the signal and refine the template according to those observations. These methods allow to compute every drum extraction in the same time as its transcription. **The last** approach is based on the determination of a discriminative model between harmonic and drums sounds. [5] considers drum sounds as noise and the harmonic part as a sum of sinusoids. [6] [7] consider that drum and harmonic parts are complementary in the spectrogram. The first part is mainly distributed according to vertical lines while the second according to horizontal lines.

The algorithm introduced in this paper belongs to the last approach. Indeed, we propose a method for discriminating in the spectrogram, at a given onset, the time/frequency bins that are more affected by drum sounds. To do so, this paper is concerned with three main contributions. In section 2 we propose a model to parametrize a drum event in the spectrogram. In section 3 we propose a way to discriminate drum and harmonic sounds using only one parameter of our model. In section 4 we detail the steps of the algorithm and introduce an efficient method that deals with the case of frequency modulations in order to reduce the irrelevant suppression of non-stationary musical instruments like the human singing voice. The performances of our algorithm are compared in Section 5 with the ones of several state-of-the-art algorithms.

2. A SPECTRO-TEMPORAL MODEL FOR DRUM EVENTS

2.1. Temporal envelope of a mode of vibration of a drum

In the temporal domain, the envelope of the sound corresponding to a mode of vibration of a drum is characterized by a fast attack (some ms) and a quick exponential decay (of about a hundred ms). We chose to model the attack by a linear growth. Therefore the envelope can be defined by :

$$p_{t_a,\alpha}(t) = \begin{cases} t/t_a & \text{for } t \leq t_a \\ e^{-\alpha(t-t_a)} & \text{for } t > t_a \end{cases}$$

where t_a is the time between the attack and the decay and α the damping factor.

2.2. Magnitude evolution in the spectrogram

Let us consider a complex signal $s(t) = p_{t_a,\alpha}(t) \cdot e^{j2\pi f_0 t}$ corresponding to a vibration mode of a drum, and $w(t)$ an analysis window. Its Short Time Fourier Transform (STFT) $S\big(\tau, f\big)$ calculated with a symmetric window, and taken at the frequency $f = f_0$ can be written :

$$S\big(\tau, f = f_0\big) = \int_{-\infty}^{\infty} p_{t_a,\alpha}(t) \cdot w(\tau - t) \cdot dt = (p_{t_a,\alpha} * w)(\tau)$$

To summarize, in order to generate a spectro-temporal pattern of a drum event according to this model, the parameters (t_a , α) are set and the corresponding temporal envelope is calculated. Then, a convolution with the analysis window, a normalization to a maximum magnitude of 1 and a subsampling according to the STFT hop size are applied.

Note that for percussive sounds t_a is generally small (less than 10ms) compared to the window length. In this range, a variation of t_a doesn't change significantly the curve after convolution. Therefore, t_a is fixed to 5ms in the reported experiments. Let us denote by $P_\alpha\big[n\big]$ the percussive spectro-temporal model generated for a given α, at time index n. Figure 1 illustrates the different steps of the generation of $P_\alpha\big[n\big]$.

978-1-4577-0539-7/11/$26.00 ©2011 IEEE

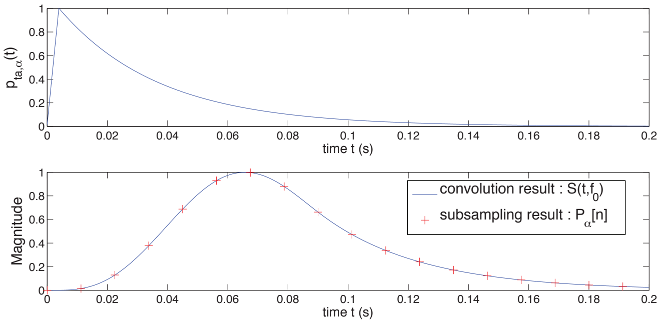

Fig. 1. Temporal model $p_{t_a,\alpha}(t)$ (top) and spectro-temporal model $P_\alpha\big[n\big]$ ($t_a = 5ms$ and $\alpha = 15$) (bottom) for a Hanning window of 90ms and a step size of 1/8 of the window length.

3. DISCRIMINATING HARMONIC AND DRUM EVENTS

3.1. Event detection

An onset detection algorithm [8] is run before the drum extraction. It allows the algorithm to search for drum events in the spectrogram only in intervals $\big[t_o(u), t_f(u)\big[$, where $t_o(u)$ is the frame index corresponding to the uth **o**nset time, and $t_f(u) = \min\big(t_o(u) + 400\text{ms}, t_o(u+1)\big)$, considered as being the **f**inal frame index of an hypothetical drum event.

We denote every event by $X_u\big[k,n\big]$ the magnitude of the kth bin extracted from the spectrogram on the time interval $\big[t_o(u), t_f(u)\big[$ with $n \in \big[1, t_f(u) - t_o(u)\big]$ and denote by $\tilde{X}_u\big[k,n\big]$ its normalization to a maximum value of 1.

3.2. Discriminating using α parameter

Because drum sounds present a faster decay than most harmonic instruments, the consideration of α parameter is an interesting approach to discriminate drum and harmonic signals. For every onset u and bin k, we estimate $\alpha_u(k)$ on the decay of $\tilde{X}_u\big[k,n\big]$ and calculate the associated energy defined by $E_u(k) = \sum_{n=1}^{t_f(u)-t_o(u)} \tilde{X}_u\big[k,n\big]^2$. To do so, we search for $\alpha_u(k)$ minimizing $|D_\alpha(k,u)|$, defined later in equation (1), which basically performs a divergence measure between the observed magnitude and the reference pattern. After summation according to α for every k and u, we obtain the distribution of the energy contained in all onset intervals as a function of α, denoted by $E(\alpha)$.

For every drum and harmonic signal of a multitrack database described in Section 5, $E(\alpha)$ is calculated. $E_{drums}(\alpha)$, $E_{harmo}(\alpha)$ and their cumulative distribution function are shown Figure 2. The two graphs point out that $E_{drums}(\alpha)$ is well localized for high values of α. $E_{harmo}(\alpha)$ is more uniformly distributed due to the diversity of instruments and their playing technique, and overlaps $E_{drums}(\alpha)$. Clearly, no ideal separation is possible but we can find a discriminative value of α maximizing the difference of harmonic and drum energies in the harmonic extracted signal : $\alpha_d = \max\limits_{\alpha_0} \sum_{\alpha=0}^{\alpha_0} E_{harmo}(\alpha) - E_{drums}(\alpha)$. On the considered data set, we find $\alpha_d \simeq 15$ so we choose to set $\alpha_d = 15$ for the algorithm presented Section 4. If we want to extract more drum sounds we can set $\alpha_d < 15$. If we want to preserve more the harmonic signal we can set $\alpha_d > 15$.

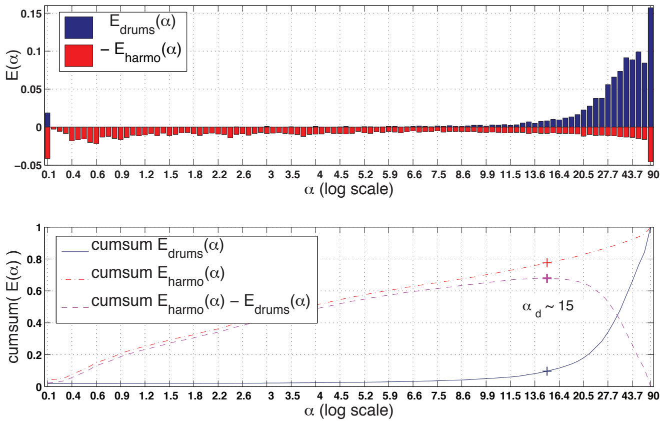

Fig. 2. Normalized distributions of energy of drum and harmonic (with minus sign for legibility) signals calculated individually from multitracks as a function of α (top) and their cumulated distribution functions (bottom).

Note that $E_{music}(\alpha)$ of the polyphonic music signals is quite similar [1] to $E_{drums}(\alpha) + E_{harmo}(\alpha)$. This is not an obvious result because the spectrograms of the sources are not additive and the sum of two decays in the mixture spectrogram gives an intermediate α. This justifies the consideration that one of the two sources (harmonic or drum) is dominant for most events of the mixture spectrogram.

4. PROPOSED ALGORITHM

Estimating $\alpha_u(k)$ for every bin k of every onset u is computationally demanding. In order to reduce computational effort and memory requirement of the algorithm, we propose to generate a single discriminative pattern $P_{\alpha_d}\big[n\big]$ and compare it to every $\tilde{X}_u\big[k,n\big]$.

A binary mask $M_P\big[k,t\big]$, of the same size as the spectrogram, having value 1 where $\tilde{X}_u\big[k,n\big]$ is detected as a drum event, is calculated. Finally, the mask is applied to the STFT of the original music signal. The drum signal is resynthesized by inverse STFT and subtracted from the music signal to obtain the harmonic signal.

4.1. Temporal adjustment

In polyphonic music the onsets corresponding to musical events are never perfectly synchronized. Thus, before comparing $\tilde{X}_u\big[k,n\big]$ to $P_{\alpha_d}\big[n\big]$, we need to adjust the relative position of the two curves. Two methods have been tested : adjustment by determination of the maximum of the intercorrelation function, and alignment of the maxima. As the results of the two algorithms are almost equivalent, the maxima alignment is considered because it is computationally more efficient. We denote by $P_{\alpha_d}\big[n-m\big]$ with $m \in \mathbb{N}$, the percussive reference curve translated by m samples.

4.2. Divergence evaluation

For a bin k at onset u, the divergence between the two curves is estimated using :

[1] For the subtraction of the 2 normalized distributions we find a mean value of about $1.1 \cdot 10^{-4}$ and a standard deviation value of $2.7 \cdot 10^{-3}$.

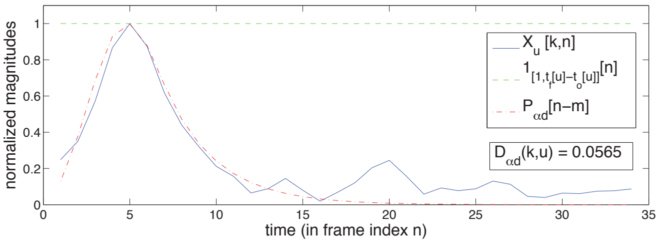

Fig. 3. Divergence measure for $P_{\alpha_d}\left[n\right]$ generated for ($t_a = 5ms$ and $\alpha_d = 15$). (same parameters of STFT as figure 1)

$$D_{\alpha_d}(k,u) = \frac{\sum_{n=1}^{t_f(u)-t_o(u)} \tilde{X}\left[k,n\right] - P_{\alpha_d}\left[n-m\right]}{\sum_{n=1}^{t_f(u)-t_o(u)} \mathbb{1}_{\left[1,t_f(u)-t_o(u)\right]}\left[n\right] - P_{\alpha_d}\left[n-m\right]} \tag{1}$$

If $\tilde{X}\left[k,n\right]$ presents beatings around $P_{\alpha_d}\left[n-m\right]$, the measure should be around zero (see Figure 3). Because only one onset is processed in $\left[t_o(u), t_f(u)\right[$, the sign of $D_{\alpha_d}(k,u)$ should allow us to know if $\tilde{X}\left[k,n\right]$ is mainly decreasing faster than $P_{\alpha_d}\left[n-m\right]$ $\left(D_{\alpha_d}(k,u) < 0\right)$ or slower $\left(D_{\alpha_d}(k,u) > 0\right)$. The denominator corresponds to a normalization by the least percussive bin magnitude evolution case (when the bin keeps a constant amplitude), it allows us to express the divergence in percentage of the area defined between dashed and dashdot curves on Figure 3.

4.3. Drum/harmonic events extraction

$D_{\alpha_d}(k,u) \leq 0$, is an equivalent condition to the $\alpha \geq \alpha_d$ condition presented in Subsection 3.2. In this case the bin should belong to the drum signal. On the contrary, if $D_{\alpha_d}(k,u) > 0$ $\big($equivalent to $\alpha < \alpha_d\big)$, the bin should belong to the harmonic signal.

Because small variations of $\tilde{X}\left[k,n\right]$ above $P_{\alpha_d}\left[n-m\right]$ can appear after the decay (see Figure 3) we introduce a threshold $D_t \geq 0$ on $D_{\alpha_d}(k,u)$. Thus, the criterion to assign a bin k at an onset u to the drum signal is :

$$D_{\alpha_d}(k,u) \leq D_t \tag{2}$$

If condition (2) is true we set the value 1 for bin index k and frame indexes $t \in \left[t_o(u), t_f(u)\right[$ in the mask $M_P\left[k,t\right]$. In the reported experiments we set $D_t = 0.1$.

4.4. Dealing with frequency modulations

Processing every bin independently raises a problem for signals containing frequency modulations (vibrato, glissando, ...). This phenomenon is illustrated by considering a synthesized chirp on Figure 4. For every single bin k, $\tilde{X}_u\left[k,n\right]$ presents a percussive shape, and verifies condition (2), but should not be considered as a drum event.

Indeed, we found by estimating $E(\alpha)$ for all the signals composing the multitracks (see Figure 2), that all the instruments entailing frequency modulations (for example the voice) present a lot of energy for $\alpha > 15$. Instead of working on the bins magnitude, one way to solve this issue would be to track partials over time, hence following the frequency modulations. However, this approach did not provide us with any improvement as performing partial tracking at onset location was found difficult.

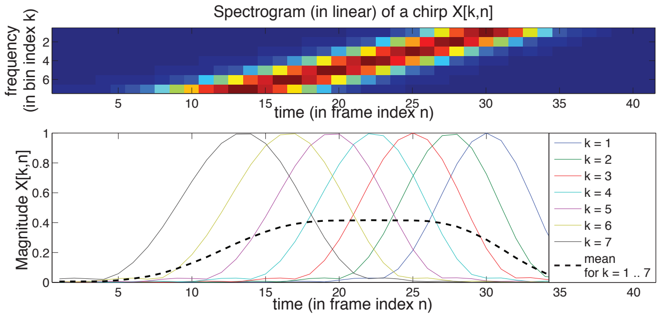

Fig. 4. Illustration of the frequency modulation consideration. On top, the linear spectrogram of a chirp. Down below, the magnitude evolution for every bin and the mean magnitude evolution for the block of bins.

As an alternative to partial tracking, we propose the following criteria. If $\tilde{X}_u\left[k,n\right]$ verifies the condition (2) from bin index $K1$ to bin index $K2$ and the block of consecutive bins is really a drum event, it should have a percussive mean profile $\tilde{X}_{K_1..K_2}\left[n\right] = \underset{k\in[K_1,K_2]}{\text{mean}} \tilde{X}\left[k,n\right]$ which verifies the same condition. If the condition is not validated, the block $k \in \left[K_1, K_2\right]$ and $t \in \left[t_o(u), t_f(u)\right[$ is set to 0 in the mask $M_p\left[k,t\right]$.

5. EXPERIMENTS

The evaluation corpus comprises 6 multitrack of professionally produced music recordings covering a large range of musical genres. Most of them are extracted from the MASS evaluation database[2]. From those songs, 10 second clips are selected, and the tracks corresponding to a drum instrument are mixed-down (summed unchanged and converted to mono) to the Drum Track (DT). The remaining tracks are mixed-down to the Harmonic Track (HT) following the same procedure. The Mix Track (MT) is the sum of DT and HT.

In order to compare those approaches, the Signal to Degradation Ratio (SDR), the Signal to Artifact Ratio (SAR) and the Signal to Interference Ratio (SIR) are computed with HT as target and an estimate of HT computed from MT. All those metrics are expressed in dB. The SDR is related to the overall quality of the separation, but can be misleading. For example, computing the SDR with MT as estimate give us a performance of about 6dB which is better than any of the methods evaluated in this paper. This is because the drum events are only sparsely distributed. The SAR quantifies the degree to which the evaluated method did not removed any of the target signal, in our case HT. The SIR quantifies the degree to which the evaluated method was able to remove the unwanted signal, in our case DT, see [9] for more details.

The 6 mixtures are processed by 8 different methods[3]. P1 is the proposed method without frequency modulation treatment, P2 is the proposed method with frequency modulation treatment. For P1 and P2, $\alpha_d = 15$, $D_t = 0.1$ and the STFT is calculated with a 90ms Hanning window and a step size of 1/8 of the window length. Ono is a reimplementation of [6], Gi is the method proposed in [5], Oz is an instance of the framework proposed in [10]. V1, V2, V3 are 3 different versions of [2][4].

[2] `http://mtg.upf.edu/static/mass`

[3] The separated sounds are available at : `http://recherche.ircam.fr/equipes/analyse-synthese/lagrange/research/DrumSeparationIcassp2011`

[4] This method decomposes the signal into several NMF basis vectors

As the SAR and SIR interact in a similar way as precision and recall in estimation theory, we propose to plot the results over a 2 dimensional plot where the x-axis corresponds to the SAR and the y-axis corresponds to the SIR, see Figure 5. The Ono method performs very well in removing DT, but HT is largely impaired and the estimated signal is plagued with strong artifacts. The same comments apply to V1. V2 and V3 respectively improves the SAR by 7 and 12 dB compared to V1, but loose about 3 dB in SIR. Choosing between the different sets of features seems to depend on the task, and the last synthesis technique seems beneficial in any case. Gi shows balanced performances and very low variances. Compared to Gi, Oz achieves an SIR improvement of about 5 dB at the expense of a loss of about 3 dB in SAR. At the same SAR level, P1 achieves a better SIR than Oz. Compared to P1, P2 is better in SDR, SAR and SIR, so we can conclude that considering the frequency modulations as proposed in Section 4.4 is relevant.

The results discussed above corresponds to a given tuning of each algorithm. However, it would be more useful to measure the performance of the system given any parameter setting. The line with stars on Figure 5, depicts the performance of P2 for 30 values of α_d varying from 0.1 (left-most point) to 90 (right-most point). One should notice that the SIR is computed using the target and the interference signals extracted from the estimated HT. This explains the erratic evolution of the curve above 20 dB of SIR. Indeed, for low values of α_d, both the target and the interference signals are largely impaired. Below 20 dB of SIR, the curve decays smoothly to reach 6 dB of SIR which corresponds to the SIR that one would obtained by considering MT as an estimate of HT (dashed line). We see that the proposed algorithm performs well in most settings, except for high values of SAR.

6. ACKNOWLEDGMENTS

This work has been partially funded under the OSEO Quaero program.

7. CONCLUSION

We proposed an efficient method for removing the drums from a polyphonic music signal. This method discriminate between drums and other instruments using a parametric model of the evolution of the STFT magnitude. A special care is taken in order to discriminate between magnitude modulations due to drums and to frequency modulated signals such as voice. Those two contributions leads to a causal algorithm, of low complexity and that can be parameterized easily. Furthermore, it compares favorably to several state of the art algorithms.

Future work will focus on optimizing further the algorithm, evaluate it using perceptive tests, and determine its usefulness as a preprocessing method for Music Information Retrieval (MIR) tasks. For the latter, studying the optimal parameter settings for a given task is of great interest.

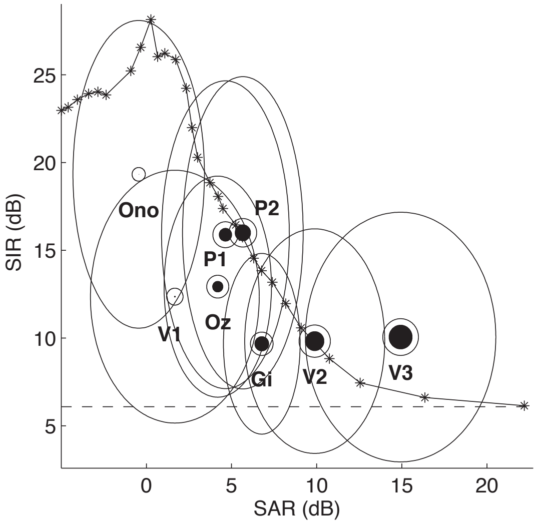

Fig. 5. Performance evaluation. The size of the filled dots and the circle around the dot are one tenth of respectively the average SDR and the average SDR plus its variance. The vertical and horizontal diameters of the ellipses are respectively the SAR and SIR variances.

8. REFERENCES

[1] C. Uhle, C. Dittmar, and T. Sporer, “Extraction of drum tracks from polyphonic music using independent subspace analysis,” in *Proc. of the 4th Int. Symposium on Independent Component Analysis and Blind Signal Separation*, 2003, pp. 843–848.

[2] M. Helén and T. Virtanen, “Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine,” in *Proc. EUSIPCO*, 2005.

[3] A. Zils, F. Pachet, O. Delerue, and F. Gouyon, “Automatic extraction of drum tracks from polyphonic music signals,” in *Proc. of the 2nd Int. Conf. on Web Delivering Of Music*, 2002.

[4] K. Yoshii, M. Goto, and H. G. Okuno, “Automatic drum sound description for real-world music using template adaptation and matching methods,” in *Proc. of the Int. Conf. on Music Information Retrieval*, 2004, pp. 184–191.

[5] O. Gillet and G. Richard, “Extraction and remixing of drum tracks from polyphonic music signals,” in *IEEE Workshop on App. of Signal Processing to Audio and Acoustics*, 2005.

[6] N. Ono, K. Miyamoto, J. L. Roux, H. Kameoka, and S. Sagayama, “Separation of a monaural audio signal into harmonic/percussive components by complementary diffusion on spectrogram,” in *Proc. of the EUSICO European Signal Processing Conf.*, aug 2008.

[7] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, “A real-time equalizer of harmonic and percussive components in music signals,” in *ISMIR 2008 - Session 1c - Timbre A*, 2008.

[8] A. Röbel, “A new approach to transient processing in the phase vocoder,” in *Pro. of the 6th Int. Conf. on Digital Audio Effects (DAFx’03)*, 2003, pp. 344–349.

[9] E. Vincent, R. Gribonval, and C. Févotte, “Performance Measurement in Blind Audio Source Separation,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 1, pp. 1462–1471, Jan. 2006.

[10] A. Ozerov, E. Vincent, and F. Bimbot, “A General Modular Framework for Audio Source Separation,” in *9th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA’10)*, 2010, pp. 27–30.

which are then classified as drums or non-drums according to a given set of features. V1 considers the best feature set used in [2] for classification, whereas V2 and V3 consider a default one. As proposed in the paper, V1 and V2 synthesize an estimate of HT using the NMF vectors not classified as drums. On contrary, V3 uses another approach, which consists in synthesizing an estimate of DT using the NMF basis vectors classified as drums and subtracting this estimate to MT in order to obtain an estimate of HT.