COMBINING A HARMONIC-BASED NMF DECOMPOSITION WITH TRANSIENT ANALYSIS FOR INSTANTANEOUS PERCUSSION SEPARATION

Jordi Janer, Ricard Marxer

Music Technology Group
Universitat Pompeu Fabra, Barcelona

Keita Arimoto

Yamaha Corp.
Japan

ABSTRACT

Many recent approaches on musical source separation rely on model-based inference methods that take into account the signal's harmonic structure. To address the particular case of instantaneous percussion separation, we propose a method that combines a harmonic-based decomposition using a Non-negative Matrix Factorization (NMF) algorithm, with the transient analysis of spectral peaks from a single audio frame. The signal model allows the estimation of harmonic and non-harmonic sources. Later, as shown in the evaluation, adding transient peak information improves the Signal-to-Distortion Ratio (SDR). Compared to other existing methods, this approach achieves a comparable performance, being suitable at the same time for low-latency conditions.

Index Terms— source separation, harmonic analysis, transient analysis, NMF

1. INTRODUCTION

Recent techniques allow for separating instrumental sources from a musical mixture signal. This process may have various application areas including musical production (e.g remixes), entertainment (e.g. karaoke), music analysis (e.g. transcription) or cultural heritage (e.g. restoration). This paper addresses the particular case of separating non-harmonic percussion sources (e.g drums, cymbals) in musical mixtures.

In our scenario, we assume that the audio mixture contains one or more harmonic instrumental sources on top of the percussion to be extracted. The timbre structure of the percussion source is difficult to model, since it might comprehend a large variety of instruments. However, a more distinguishable trait is its time signature, consisting of a sharp attack followed by an exponential decay.

Algorithms such as Non-negative Matrix Factorization (NMF) decompose an input time-frequency representation into basis components without a prior knowledge, which allows a blind source separation. Most approaches impose additional constraints in the factorization process. For example, some authors [1] add temporal continuity constraints, while other approaches force a source/filter decomposition with a set of harmonic patterns and filter banks [2, 3]. More recently, Ozerov et al [4] have proposed a framework that combines spectral patterns (source/filter model) and temporal patterns (attack and decay envelopes of hundreds of milliseconds). In this case, the factorization step estimates the patterns activation gains.

We find methods that specifically address the problem of percussion separation. A two-step separation method, with NMF decomposition and SVM (Support Vector Machines) classification [5], classifies the separated components into drums or pitched. Another approach makes use of drum separation with NMF methods as a pre-process for its classification and transcription [6]. The Harmonic Percussion Sound Separation (HPSS) method [7] provides an efficient and effective two-dimensional filtering of the spectrogram, to separate harmonic component (temporal continuity but spectral discontinuity), and percussion (temporal discontinuity and spectral wide band energy). This method has proven effective as a pre-process for automatic music description tasks.

Our method combines transient estimation of spectral peaks with a model-based inference algorithm that decomposes the input signal into a harmonic and a non-harmonic magnitude spectra. The algorithm processes a single frame magnitude spectrum to estimate the two decomposed spectra. Compared to other approaches, our method is causal and therefore appropriate also for low-latency situations, which relates to the notion of instantaneous separation.

2. METHOD

The separation process involves various steps, as shown in figure 1. First, the input audio signal is windowed and represented as a sequence of complex spectra taking the Short-Time Fourier Transform (STFT). Next, we decompose the magnitude spectrum as a linear combination of basis spectral components with a NMF algorithm. Additionally, we extract temporal information by means of a transient analysis of spectral peaks in the current spectrum. Combining this information with the estimated non-harmonic spectrum, we can improve the separation of percussion sources from background noise and other harmonic components.

2.1. Signal model

The central part of the source estimation is the signal model, which is built from a set of spectral basis components. Our focus is on low-latency applications which require the decomposition of each spectrum frame to be done instantaneously.

For each frame, we assume that the spectrum magnitude Y can be decomposed in a linear combination of N_C elementary spectra, also named basis components. This can be expressed as $Y = BG$ where $Y \in \mathbb{R}^{N_S \times 1}$ is the spectrum at a given frame m, N_S being the size of the spectrum. $B \in \mathbb{R}^{N_S \times N_C}$ is the matrix whose columns are the basis components, it is also referred to as the basis matrix. $G \in \mathbb{R}^{N_C \times 1}$ is a vector of component gains for the current frame.

Spectral basis components B are constant and fixed a priori. It consists of a set of N_P single pitch multiple-harmonic spectra. In order to model different timbres we must allow different spectral envelopes. This is done by filtering the single pitch components by a bank of N_F filters. To cope with all possible observed spectra (e.g. in presence of percussive events or noise), we add a set of

This research is partially funded by Yamaha Corp. (Japan)

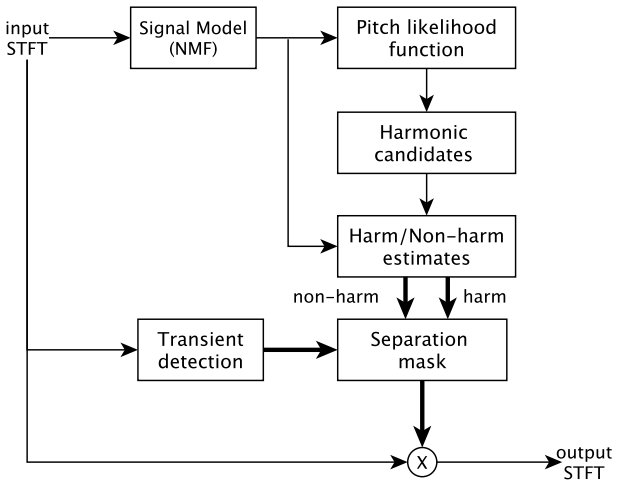

Fig. 1. The separation mask combines the non-harmonic estimation and the transient peak analysis. Thick arrows represent spectral masks.

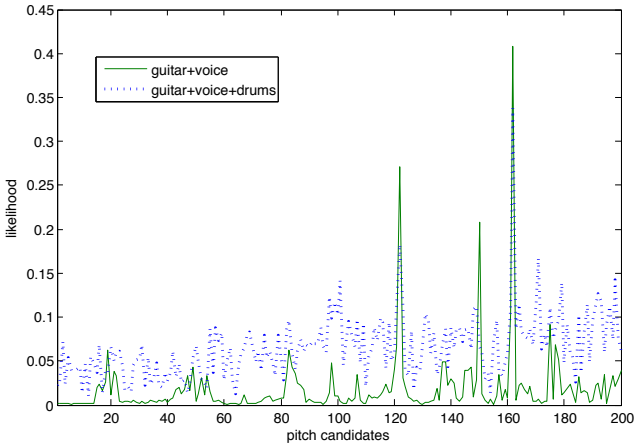

Fig. 2. Pitch likelihood curves in two different time instants of an audio excerpt containing: vocals and guitar (solid green); and vocals, guitar and drums (dashed blue).

filters as wideband components similar to [8]. This results in a total of $N_C = (N_P + 1) \cdot N_F$ basis components. Detailed information about the creation of these spectral basis components can be found in [9].

Solving our spectrum decomposition problem with NMF consists in finding the best non-negative component gains $\hat{G}$ that minimizes a given objective function. In this case we use the Itakura-Saito divergence function, well known in source separation applications [10]. The solution $\hat{G}$ can be computed iteratively by means of a multiplicative update rule. Apart from reconstructing the input spectrum, we can use $\hat{G}$ to compute a pitch likelihood function by summing the individual gains corresponding to a given pitch candidate in the basis components matrix B .

2.2. Harmonic and non-harmonic source estimation

In the decomposition solution, we expect an harmonic instrumental source to contribute principally to specific candidates in the pitch likelihood function. In contrast, we expect percussion source contributions to be distributed over several candidates, both pitched and wide-band filter candidates. The consequence is that non-harmonic sources will show energy spread over the pitch likelihood function and not exclusively localized in individual candidates.

Then to reconstruct the harmonic component, we select K pitch candidates from the pitch likelihood function by means of a peak picking algorithm. Candidates with a likelihood value below an empirically defined threshold τ_1 are discarded. Figure 2 shows two pitch likelihood curves, corresponding to different time instants of a polyphonic audio mixture, one with percussion and one without.

From the estimated vector $\hat{G}$, we create a new vector $\hat{G}_h$ containing non-zero values only at those selected candidates. Therefore, we can compute the harmonic signal estimation as $S_h = B\hat{G}_h$. In a complementary fashion, the reconstruction of the non-harmonic part takes a gains vector $\hat{G}_{nh}$ containing non-zero values for the unselected pitch candidates plus the wideband filter banks. The non-harmonic source estimation is computed as $S_{nh} = B\hat{G}_{nh}$

With the estimated magnitude spectra S_h and S_{nh}, we can recover a separated output complex spectrum by means of Wiener filtering, as used extensively in recent approaches [11]. Equation 1 contains the spectral mask M_{nh}, which is then multiplied element-wise by the input complex spectrum $Y(\omega, m)$ to reconstruct the non-harmonic signal.

$$M_{nh} = \frac{S_{nh}^2}{S_h^2 + S_{nh}^2} \tag{1}$$

By an informal listening to the separated non-harmonic signal, we realized that in the presence of a percussion event, the separated signal is weak and lacks of clarity. The rationale behind is that a percussion attack increases the spectrum's energy in form a wide-band noise. The parameter estimation, instead of representing it exclusively with the wideband (unpitched) filter candidates, it also assigns energy to the selected pitch candidates, to represent the percussion spectrum. To help the identification and separation of the percussion contribution in the spectrum, we propose to include transient analysis.

2.3. Transient analysis

Our aim is to detect transient events in the signal, which should reveal the presence of percussion sources. This analysis can be achieved in the spectral domain by means of the temporal center of gravity of spectral peaks. Given a magnitude spectrum, spectral peaks are detected by localizing local maxima, and neighboring local minima, which determine the spectral peak width.

Röbel [12] suggested to compute the temporal center of gravity (COG) to treat transient events in a phase vocoder algorithm. The COG of an isolated spectral peak can describe how the energy of a given frequency is localized inside the temporal window. It is based on the group delay and it can be computed directly from the bins of a spectral peak in the complex STFT, written as $A(\omega, t_m)$ in equation 2. If a spectral peak is part of a transient event, its energy will be concentrated at the rightmost part of the window, and will have a high COG value. A spectral peak that corresponds to a sustained sound will have its energy spread over the whole window, having a COG value near 0.

$$t_{COG} = \frac{\int -\frac{\partial\phi(\omega,t_m)}{\partial\omega} A(\omega,t_m)^2 d\omega}{\int A(\omega,t_m)^2 d\omega} \quad (2)$$

Similar to the transient detection in [13], which computes statistical measures of COG values of individual peaks t_{COG}, our approach defines 14 bands with a bandwidth of 1500Hz. For each band i, we compute the average of the COG value of individual spectral peaks, referred to as c_i. We create a transient spectral mask M_t, in which the bins corresponding to all spectral peaks in a frequency band i are set to one if $c_i > \tau_2$ and set to zero otherwise.

However, the decay of a percussion sound can typically extend over hundreds of milliseconds. To handle the decay, our method keeps a history of N frames (e.g. covering 250 ms) of each band's COG average c_i. First, for a given frame m we compute the time derivative as $\Delta c_i[m] = c_i[m] - c_i[m-1]$ of all band's COG average. Then, for a band i, a binary transient decay value $d_i[m]$ is set to one if two conditions are fulfilled:

if $\max_n c_i[n] > \tau_2$ **and** $\Delta c_i[m] < 0$ **then**
  $d_i[m] = 1$
else
  $d_i[m] = 0$
end if

The leftmost condition in the above pseudo-code requires the presence of a transient event in that past N history frames ($m - N \leq n \leq m$). At the same time, by forcing a negative derivative value $\Delta c_i[m]$, we assure that the transient "is shifting to the left of the window". In order to take the transient decay into account, we compute a decay mask M_d, in which the bins of all spectral peaks comprised in a frequency band i are set to the binary value d_i. It is worth remarking that the transient analysis does not distinguish between harmonic and non-harmonic transients. Therefore the masks M_t and M_d would let through transients corresponding to harmonic instrumental sources.

2.4. Separation mask

Finally, to build the final percussion separation mask M_p for the current frame, we combine the partial masks previously computed (M_t, M_d, M_h and M_{nh}). We have to take into account that, on the one hand, a percussion source will contribute largely to the estimated harmonic mask M_h. Therefore, we cannot achieve the separation only from the estimated non-harmonic mask M_{nh}. On the other hand, applying only the transient masks M_t and M_d based on spectral peaks analysis, we would effectively separate the percussion but leaks from other harmonic transients (e.g. bass guitar, piano) will be equally present.

To tackle this problem, for those spectral peaks classified as transients in M_t, we compare the values of the harmonic mask M_h to a given threshold τ_3 at the center frequency of each spectral peak. Typically, when a spectral peak in the input spectrum corresponds to a harmonic source frequency partial, the estimated value in the harmonic mask at this specific frequency will be high. Hence, we can identify those harmonic transient peaks and not separate them as percussion. During the percussion decay, we proceed in a similar manner, but adding at the same time the estimated non-harmonic mask M_h to the final percussion separation mask M_p. This process can be written as mask operations.

$$M_\tau = \begin{cases} 1, & \text{if } M_h < \tau_3 \\ 0, & \text{if } M_h \geq \tau_3 \end{cases} \quad (3)$$
$$M_p = \lfloor (M_\tau \otimes M_t) + (M_\tau \otimes M_d \otimes M_{nh}) \rfloor_{0,1}$$

In equation 3, a binary matrix M_τ is computed by thresholding the harmonic mask M_h. The operator $\otimes$ denotes Hadamard's (element-wise) product and $\lfloor\ \rfloor_{0,1}$ indicates a clipping of the mask values between 0 and 1. Finally, the separated percussion source $\hat{Y}_p(\omega, m)$ is computed from the input complex spectrum Y and the separation mask as $\hat{Y}_p(\omega, m) = M_p \otimes Y(\omega, m)$. The time-domain signal is recovered by means of the inverse Fourier transform and an overlap-add mechanism.

3. EVALUATION

Source separation algorithms can be objectively evaluated if the original multi-track sources are available. We use the same measurements employed in the community evaluation campaigns such as SiSEC [14]: SDR (Signal to Distortion Ratios), ISR (Image to Spatial distortion Ratios), SIR (Source to Interference Ratios) and SAR (Sources to Artifacts Ratios). Evaluation material consists of a dataset of 17 multi-track recordings with presence of drums, compiled from publicly available resources (MASS[1], SiSEC[2] and BSS Oracle[3]).

Table 1 shows the results of three variants of our method Transient Harmonic Percussion Separation (THPS). Different masks are used, M_{nh} for the non-harmonic separation (THPS-NH), $\lfloor M_t + M_d \rfloor_0^1$ for the transient separation (THPS-T) and M_p for the final percussion separation (THPS). We include also to two state-of-the-art methods: a custom implementation of the HPSS method [7] and the publicly available implementation of FASST[4] [4].

Error measures in table 1 are the difference between the measures obtained by the oracle estimator [15], used here as a baseline, and the measures obtained by each algorithm. Values are the average of all examples in the dataset.

	SDR	*ISR*	*SIR*	*SAR*
THPS-NH	15.17	15.31	25.49	10.67
THPS-T	15.42	7.90	23.87	10.82
THPS	14.29	15.23	23.16	13.93
HPSS	13.82	19.22	23.43	14.40
FASST	14.96	20.20	27.39	13.95

Table 1. Average error measures for various algorithms.

For the experiments with the THPS algorithm, we have performed an STFT analysis with a Blackman-Harris window 92ms long ($F = 4096$ for signals at sample rate $S_r = 44100$), a hop size of 11ms ($H = 512$) and a DFT size of 8192 which results in $N_S = 4097$. Regarding the parameters of the B matrix we have set the number of filters $N_F = 12$, the lowest pitch frequency $f_l = 35$ Hz, 40 pitches per octave covering a total of 5 octaves ($N_P = 40 \cdot 5 = 200$). This leads to a total number of components $N_C = 2412$. The number of NMF iterations is set to 15, and harmonic candidate threshold is $\tau_1 = 0.05$, the transient mask threshold is $\tau_2 = 0.3$, and the harmonic mask threshold is $\tau_3 = -30$dB.

The HPSS implementation separates the input signal into two sources: harmonic and percussion. The frame size in this process was set to 1024, which offered a good trade-off between audio quality and vocals/percussion separation. Regarding the FASST frame-

[1] http://www.mtg.upf.edu/static/mass
[2] http://sisec.wiki.irisa.fr/
[3] http://bass-db.gforge.inria.fr/bss_oracle/
[4] http://bass-db.gforge.inria.fr/fasst/

work, we used its default configuration that separates the input signal into four sources: lead melody, bass, drums and other. In our experiment, we consider only the separated *drums* as percussion source.

From the results, we show that the performance of our THPS algorithm is comparable to both state-of-the-art methods. It outperforms both partial configurations THPS-NH and THPS-T, demonstrating the hypothesis of the proposed combination. Additionally, figure 3 illustrates the SDR error for the individual audio examples in the dataset[5]. It shows how depending on the audio example, one approach may work better than the others, explaining also the similar average results. A perceptual-based evaluation, either by subjective listening tests or using perceptual software toolkits (e.g. PEASS), was not possible to carry out for the current experiment.

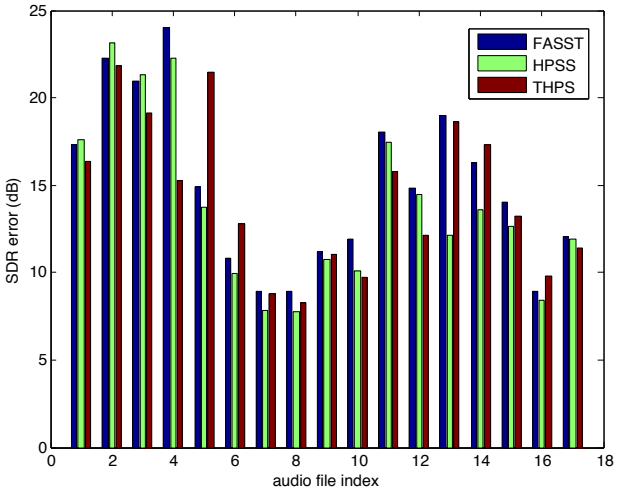

Fig. 3. SDR error measures of individual audio examples for three approaches FASST, HPSS and THPS.

4. CONCLUSIONS

This paper presents a musical source separation approach specifically adapted to isolate percussion sources. It combines transient analysis with a NMF spectrum decomposition based on a harmonic model. We show that the combination of these two strategies improve the separation quality.

In contrast to other state-of-the-art methods, our method features instantaneous separation from a single audio frame, which makes this approach suitable for low-latency situations. A quantitative evaluation show that it obtains very similar performance as other offline methods, which means that the quality is not sacrificed with the instantaneous processing constrains. Nevertheless, the method still presents some limitations. In presence of vocals in the mix, the separated percussion source contains residues of fricative phonemes. Also the separated percussion decay loses fidelity. Apart from a further exploration of algorithm's parameters (e.g. NMF iterations, threshold values), we think that in both cases the quality can be refined by including statistical modeling of these particular signals.

[5] Audio examples are available online: `http://www.mtg.upf.edu/~jjaner/presentations/icassp12`.

5. REFERENCES

[1] T. Virtanen, "Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066–1074, 2007.

[2] A. Klapuri, T. Virtanen, and T. Heittola, "Sound source separation in monaural music signals using excitation-filter model and em algorithm," in *IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP)*. IEEE, 2010, pp. 5510–5513.

[3] J.L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David, "Main instrument separation from stereophonic audio signals using a source/filter model," *Proc. European Signal Processing Conference*, 2009.

[4] A. Ozerov, E. Vincent, and F. Bimbot, "A general modular framework for audio source separation," in *9th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA'10)*, Saint-Malo, France, Sept. 2010.

[5] M. Helén and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," in *Proc. of the European Signal Processing Conference*, 2005, vol. 2005.

[6] O. Gillet and G. Richard, "Transcription and separation of drum signals from polyphonic music," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 3, pp. 529–540, 2008.

[7] N. Ono, K. Miyamoto, H. Kameoka, and S. Sagayama, "A real-time equalizer of harmonic and percussive components in music signals," in *Proc. of the 9th International Conference of Music Information Retrieval ISMIR*, 2008, p. 139.

[8] Jun Wu, E. Vincent, S.A. Raczynski, T. Nishimoto, N. Ono, and S. Sagayama, "Multipitch estimation by joint modeling of harmonic and transient sounds," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, may 2011, pp. 25 –28.

[9] R. Marxer, "Signal decomposition by a joint pitch, timbre and wideband model," Tech. Rep., Universitat Pompeu Fabra, 2011.

[10] Févotte C., Bertin N., and Durrieu J.-L., "Nonnegative Matrix Factorization with the Itakura-Saito Divergence: With Application to Music Analysis," *to appear in Neural Computation*, 2008.

[11] L. Benaroya, F. Bimbot, and R. Gribonval, "Audio source separation with a single sensor," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 191–199, 2006.

[12] A. Röbel, "Transient detection and preservation in the phase vocoder," in *Proc. Int. Computer Music Conference (ICMC)*, 2003, pp. 247–250.

[13] A. Röbel, "Onset detection in polyphonic signals by means of transient peak classification," *International Symposium for Music Information retrieval (ISMIR/MIREX'05)*, 2005.

[14] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech & Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[15] E. Vincent, R. Gribonval, and M.D. Plumbley, "Oracle estimators for the benchmarking of source separation algorithms," *Signal Processing*, vol. 87, no. 8, pp. 1933–1950, 2007.