TRANSIENT AND STEADY-STATE COMPONENT EXTRACTION USING NONLINEAR FILTERING

Ignacio Irigaray

Universidad de la República
IIE/FING, Montevideo, Uruguay
irigaray@fing.edu.uy

Luiz W. P. Biscainho

Universidade Federal do Rio de Janeiro
DEL/Poli & PEE/COPPE, Rio de Janeiro, Brazil
wagner@smt.ufrj.br

ABSTRACT

Recently, a fast and simple method for separation of the transient and steady-state components of a music signal was proposed in [4]. The technique involves the application of two median filters to the spectrogram: one along the frequency bins to eliminate steady-state components, and the other along the time frames to eliminate transient components. In the present paper, a modification of the filtering stage is proposed; the resulting algorithm is evaluated both regarding the perceptual quality of the attained separation and its use as a preprocessing stage for improving the performance of a beat-tracking technique. The results obtained for a reference data set of beat-annotated music excerpts are promising.

1. INTRODUCTION

Spectral modeling can be seen as the task of decomposing a signal in constituent components with some known behavior in time and frequency. In [7] a sinusoidal model for audio and speech waveforms was developed, and further improvements were presented in [10] and [12], extending the model to also include stochastic and transient components. This paper addresses the problem of separating transient and steady-state components, which finds application in remixing, adaptive audio effects, rhythm analysis, harmonic analysis, music information retrieval, audio coding, etc.

As an application example, we tackle the problem of beat tracking, where the previous elimination of steady-state components is expected to yield a more robust beat detection. The method was evaluated over a benchmark MIREX data set [1], showing promising results. Subjective listening tests were also conducted to evaluate the perceptual quality of the separation.

This paper is organized as follows. In Section 2 we present the definition of transient and steady-state components utilized in this work. Section 3 introduces the general procedure for their separation. Section 4 describes the Stochastic Spectrum Estimation filter. Section 5 describes the test methodology and the data sets utilized, and discusses the experimental results. Conclusions are drawn in Section 6.

2. TRANSIENT AND STEADY STATE MODELING

To precisely and meaningfully discriminate transient and steady-state components is not an easy task. Several works have addressed this problem. In [11] a feature-based classification of components extracted via Independent Component Analysis is presented. In [5], the authors propose a two-stage processing, involving a non-negative matrix factorization to decompose the spectrogram into components having fixed spectrum with time-varying gain, and a support vector machine to classify them as either pitched or drum components.

In this work, the model presented in [8] is adopted, where transients are considered broad-band components with highly concentrated energy in time, whereas steady-state sources are taken as discrete narrow-band components with smooth temporal behavior. These components can be seen in the spectrogram as vertical and horizontal ridges, respectively.

In Figure 1, a typical spectrogram, computed for an excerpt of a western popular song with piano, drums and vocal, is shown. It will be used to illustrate some paper results. In the first three seconds, when only piano and drums are present, one observes their respective steady-state and transient behaviors. Afterwards, the voice, presenting a deep vibrato, enters.

3. TRANSIENT AND STEADY-STATE SEPARATION

The departure point of the present work is the general procedure introduced in [4], which can be divided into the following four steps as shown in Figure 2 :

1. Obtain a time-frequency representation for the digital audio signal, typically a spectrogram computed via the Short-Term Fourier Transform (STFT):

$$S(n,k) = \sum_i x(i)w(i-nT)e^{-\frac{j2\pi ik}{N}}. \quad (1)$$

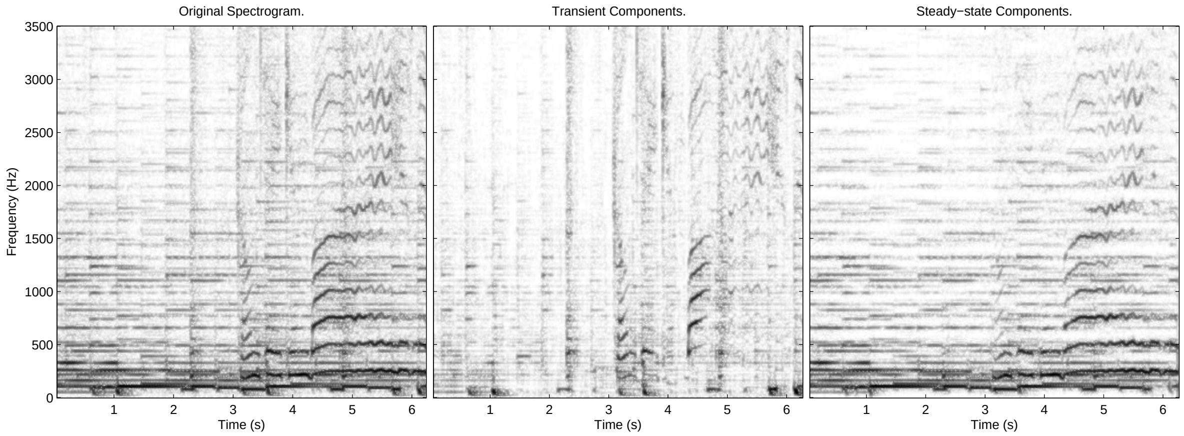

Figure 1. Left: Spectrogram of a excerpt from a popular music song. Middle: Spectrogram with transient components. Right: Spectrogram with steady-state components.

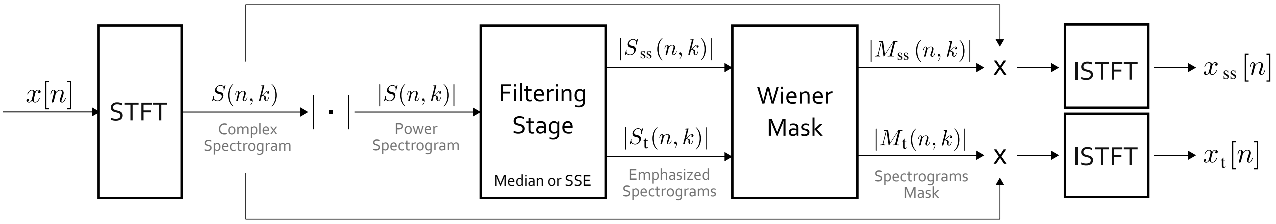

Figure 2. Diagram of the entire process.

2. Apply a median filter to the power spectrogram S along the frequency axis to eliminate steady-state components to obtain a "transient emphasized" spectrogram S_t, as well as along the time axis to eliminate transient peaks in order to obtain a "steady-state emphasized" spectrogram S_{ss}:

$$S_t(n.k) = \text{median}(|S(n-l:n+l,k)|), \qquad (2)$$

$$S_{ss}(n.k) = \text{median}(|S(n,k-l:k+l)|). \qquad (3)$$

3. From the emphasized spectrograms, calculate two soft masks based on the Wiener filter, given by:

$$M_t = \frac{S_t^2}{S_{ss}^2 + S_t^2}, \quad M_{ss} = \frac{S_{ss}^2}{S_{ss}^2 + S_t^2}. \qquad (4)$$

4. Multiply each mask with the original complex spectrogram, and compute the Inverse Short-Time Fourier Transform [1] (ISTFT) of the results to obtain, respectively, the transient signal x_t and the steady-state signal x_{ss}.

The method allows for perfect reconstruction ($x = x_{ss} + x_t$). Furthermore, the processes involved are very simple, thus allowing an efficient implementation.

4. STOCHASTIC SPECTRUM ESTIMATION

In this work we propose using an alternative non-linear filter in the second stage of the procedure described in Section 3. This filtering procedure was originally proposed in [6] for stochastic spectrum estimation in the modeling of musical sounds, leading to very good results.

Firstly, the reciprocal R of each element of the power spectrogram is calculated, turning the peaks of $S(n,k)$ into valleys of $R(n,k)$:

$$R(n,k) = S^{-1}(n,k). \qquad (5)$$

Then, a moving average (MA) filter is applied along the time axis to filter the transient components, and along frequency bins to eliminate the steady-state components. The MA applied to a valley in R (originally a peak in S) tends to make it disappear. The estimated reciprocals of the desired "transient emphasized" and "steady-state emphasized" spectra are given respectively by

$$\hat{R}_t(n,k) = \frac{1}{M+1} \sum_{i=-M/2}^{M/2} R(n,k+i), \qquad (6)$$

$$\hat{R}_{ss}(n,k) = \frac{1}{M+1} \sum_{i=-M/2}^{M/2} R(n+i,k). \qquad (7)$$

[1] The ISTFT is calculated via an Overlap-and-Add procedure.

The respective stochastic spectrum estimates (SSE) are then computed as

$$S_{\mathrm{t}}(n,k) = \hat{R}_{\mathrm{t}}^{-1}(n,k), \quad S_{\mathrm{ss}}(n,k) = \hat{R}_{\mathrm{ss}}^{-1}(n,k). \qquad (8)$$

Figure 3 illustrates the process for a fixed frequency bin.

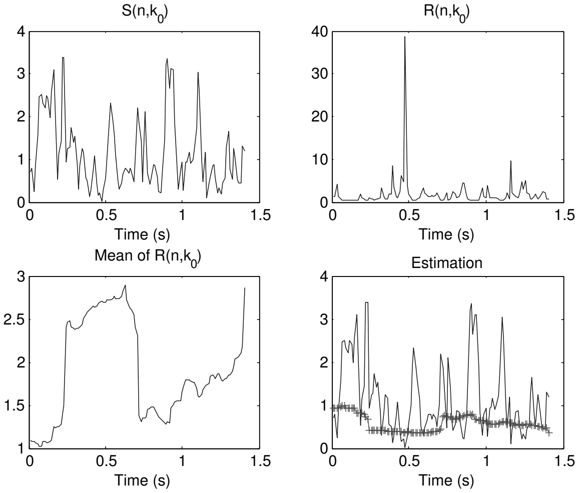

Figure 3. Steps of the Stochastic Spectral Estimation applied along the time axis.

5. TEST AND RESULTS

We evaluate the performance of the modified algorithm in two ways: 1. Systematic listening tests are conducted to compare the original and proposed methods as to their separation performances; 2. An application-based evaluation is carried out regarding the beat-tracking problem.

5.1. Data set description

Two data sets were utilized in this work, one for the evaluation of beat tracking and another one for the subjective tests. Both data sets have a sampling rate of 44.1 kHz with 16-bit resolution.

5.1.1. Subjective Test Data set

The data sets for the subjective tests consists of excerpts of thirteen pieces of North American popular music (rock, folk and blues) with a length of ten seconds. It exhibits multiple combinations of transient and steady-state components in the sense of perceptual presence in the mix.

5.1.2. Beat-Tracking Data set

To measure the performance of a beat-tracking algorithm the benchmark MIREX 2006 data set [1] was utilized. This data set is composed of twenty excerpts of western popular music with thirty-second duration. Each recording has been annotated by 40 different listeners.

5.2. Subjective test

In order to measure the perceptual difference in performance of the proposed and the original nonlinear filters, a set of formal subjective tests was designed and conducted following the recommendations suggested in [13]. Each participant should listen and compare the separated steady-state and transient components produced by the original and modified algorithms.

For this purpose, a graphical user interface specifically designed to comply with the requirements of this test was adapted from [9]. For each song in the data set, the interface presents the original signal as reference in conjunction with the processed signals to be compared. The order of the compared signals is randomized to assure the blindness of the test. The interface implements audio controls to play/stop any of the signals, and the listener can also define loop points to allow a detailed listening of the signals.

In this test, ten participants answered the next three questions for each output signal:[2]

* Q1: How much of the desired components has been properly separated?

* Q2: How much of the undesired (residual) components has been left?

* Q3: How would you rate the integrity (in the sense of naturalness) of the separated signal?

Table 1 summarizes the results of the transient separation subjective test, quantifying the preferred results for each signal and question: both methods were considered virtually equivalent regarding perceptual quality.

	$Q1_{\mathrm{t}}$	$Q2_{\mathrm{t}}$	$Q3_{\mathrm{t}}$	$Q1_{\mathrm{ss}}$	$Q2_{\mathrm{ss}}$	$Q3_{\mathrm{ss}}$
Median	23.1	30.8	27.9	24.0	21,2	25.0
SSE	26.0	25	23.1	30.8	25	20.2
Equals	50,9	44.2	49	45,2	54.8	54.8

Table 1. Result of subjective test. The subscripts (t) and (ss) indicates transient and steady-state respectively. “Equals” means that the signals were not distinguishable.

5.3. Beat-tracking problem

Most of the beat information present in music is contained in its transient components. Thus, beat-tracking algorithms could potentially be favored by a preprocessing stage after which only transient components are left. Such hypothesis is tested by evaluating the performance of a state-of-the-art beat-tracking algorithm (presented in [3]) over the 2006 MIREX audio beat-tracking practice database [1].

In order to perform an objective comparison of the beat-tracking algorithm the results obtained with and without the

[2] The subjective tests are available at `http://iie.fing.edu.uy/~irigaray/maestria/TestSubjetivo/main/main.html`

preprocessing stage were evaluated with the methods for performance measure utilized in the MIREX 2012 contest. For a detailed description of these methods see [2]. To avoid unfair comparisons, we searched for the optimum performance of each algorithm via a grid search over its respective parameters.

	Original	Median	SSE
F-Measure (%)	48	48,5	51.4
Cemgil (%)	35,6	35,5	37,5
Goto (%)	6,75	7,25	7,24
PScore (%)	50,8	50,9	53
cmlC (%)	10,2	10,3	11,2
cmlT (%)	18,3	18,4	20,7
InfGain (bits)	1,2	1,3	1,4

Table 2. Beat-tracking performance measure.

The results of the evaluation are presented in Table 2. If the median-based pre-processing did not improve the performance of the beat-tracking algorithm in general, the SSE-based procedure was always beneficial to the application, thus suggesting some additional investigation may produce interesting results.

6. CONCLUSION AND FUTURE WORK

In this work, a modification of an existing method for separating transient and steady-state components of a musical signal was proposed and evaluated. Subjective listening tests were conducted to compare the perceptual quality of the separation. Additionally, the impact of its application as a pre-processing stage on the performance of a beat-tracking algorithm was evaluated.

If the subjective tests indicate that original and modified method performances are perceptually equivalent, experiments conducted in the context of the beat tracking shows that removing the steady-state signal components using the proposed method improves the overall beat-detection score.

Of course, several improvements can be envisaged to ameliorate the attained perceptual quality of the separation. For instance, multi-resolution techniques can be utilized in the time-frequency representation to allow more flexibility.

The promising results obtained for the beat-tracking problem, which will be further addressed in future work, also suggest the idea of applying a similar procedure to remove the transient signal components as a pre-processing stage for a pitch tracking algorithm.

7. REFERENCES

[1] "Music information retrieval evaluation exchange (mirex) competition," 2006, http://www.music-ir.org/mirex/wiki/2006:Audio_Beat_Tracking.

[2] M. Davies, N. Degara, and M. Plumbley, "Evaluation methods for musical audio beat tracking algorithms," Queen Mary University, Centre for Digital Music, Technical Report C4DM-TR-09-06, Oct. 2009.

[3] D. P. Ellis, "Beat tracking by dynamic programming," *Journal of New Music Research*, vol. 36, no. 1, pp. 51–60, Jul. 2007.

[4] D. Fitzgerald, "Harmonic/percussive separation using median filtering," in *Proc. of the DAFx-10*, Graz, Austria, Sept. 2010.

[5] M. Helén and T. Virtanen, "Separation of drums from polyphonic music using non-negative matrix factorization and support vector machine," in *In: Proc. of the EUSIPCO'2005*, Antalya, Turkey, Sept. 2005, pp. 1091–1094.

[6] N. Laurenti, G. De Poli, and D. Montagner, "A nonlinear method for stochastic spectrum estimation in the modeling of musical sounds," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 15, no. 2, pp. 531–541, Feb. 2007.

[7] R. McAulay and T. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, Aug. 1986.

[8] N. Ono, K. Miyamoto, J. Le 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 EUSIPCO 2008*, Lausanne, Switzerland, Aug. 2008.

[9] K. Sebastian. (2013, Mar.) mushrajs @ONLINE. [Online]. Available: https://github.com/seebk/mushraJS

[10] X. Serra and J. Smith III, "Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition," *Computer Music Journal*, vol. 4, no. 14, pp. 12–24, Winter 1990.

[11] C. Uhle, C. Dittmar, and T. Sporer, "Extraction of drum tracks from polyphonic music using independent subspace analysis," in *In Proc. of the ICA2003*, Charleston, USA, Mar. 2003, pp. 843–848.

[12] T. Verma, S. Levine, and T. Meng, "Transient modeling synthesis: a flexible analysis/synthesis tool for transient signals," in *Proc. of the ICMC 1997*, Thessaloniki, Greece, Sept. 1997, pp. 164–167.

[13] S. Zielinski and F. Rumsey, "On some biases encountered in modern audio quality listening tests-a review," *Journal of the Audio Engineering Society*, vol. 56, pp. 427–451, Jun. 2008.