MULTI-SCALE FRAME-BASED ANALYSIS OF AUDIO SIGNALS FOR MUSICAL TRANSCRIPTION USING A DICTIONARY OF CHROMATIC WAVEFORMS

Olivier DERRIEN

STD-ISITV, Université de Toulon
Av. G. Pompidou, BP 56 F-83162, La Valette du Var Cedex, France
E-mail: derrien@univ-tln.fr

ABSTRACT

A new method for analyzing audio signals in the context of musical transcription is described. It consists of a frame-based expansion of the signal over a multi-scale time-frequency dictionary with a set of logarithmic discrete frequencies. This method, based on the Matching Pursuit algorithm, provides the same frequency resolution as a constant-Q filter-bank, but with a better time resolution, especially in low frequencies, and an efficient noise rejection.

1. INTRODUCTION

Musical transcription can be defined as extracting musical information from an audio file and representing it by means of a musical notation, in terms of note starting time, duration, pitch, loudness and so on. It still remains an unresolved problem for current computer systems. However, some efficient solutions have been proposed for single instruments with a prior training period, for example by Marolt [1].

Concerning pitch estimation, some experimental results have shown that the human brain decomposes the signal in spectral components and tries to find a common reference to all the partials. Many transcription algorithms are designed according to this scheme: first, a time-frequency analysis is performed. Then, a partial-tracking and note-extraction stage tries to recognize elementary harmonic structures. In state-of-the-art transcription systems, time-frequency analysis often relies on a specific filter-bank with a logarithmic band-width progression: constant-Q [2] or gammatone [1]. Filter-banks are easy to implement as they require few computation time, but several drawbacks can be pointed out: no distinction is made between periodic (tonal) and a-periodic components. Thus, a-periodic components, considered as noise in a transcription point of view, can lead to false note-detection. Furthermore, constant-Q filter-banks are known to achieve a poor time-resolution in low-frequencies.

According to these considerations, a sinusoidal modeling of the audio signal seems more accurate. The Matching Pursuit algorithm (MP) proposed by Mallat and Zhang [3] decomposes a signal into a linear expansion of multi-scale time-frequency waveforms selected from a redundant dictionary. But in the standard implementation of the MP, the frequency scale is linear, while the transcription problem requires a logarithmic scale. One can note that MP decomposition differs form a wavelet-transform, where the frequency of each waveform is related to the scale-factor (for an application of a wavelet-transform to musical analysis, see [4]). The resulting time-frequency resolution of a wavelet-transform is hardly sufficient for transcription.

Besides the problem of time-frequency resolution, the practical implementation of a transcription system requires a frame-based signal analysis. A frame-based implementation of MP as been proposed, but with a single time-scale [5], which seriously reduces the interest of the algorithm. Gribonval and Bacry have proposed a specialized version of the MP algorithm for musical transcription [6], but their approach is focused on the use of harmonic waveforms, which allows to perform the time-frequency analysis and the partial-tracking in a single stage. However, this algorithm has a high computational complexity.

In this paper, we present a new method for time-frequency analysis of musical signals, based on the standard MP algorithm, but with an original dictionary of time-frequency waveforms featuring a chromatic frequency scale. The dictionary and the decomposition algorithm are described and our system is compared to a constant-Q filter-bank.

2. DICTIONARY OF CHROMATIC WAVEFORMS

To carry out a short-time decomposition of the audio signal, the time-discrete input signal is segmented in frames of N samples with an overlap factor M, and weighted with a Hanning window:

$$\omega[n] = \frac{1}{M}\left[1 - \cos\left(2\pi\frac{n}{N}\right)\right] \qquad (1)$$

This window was chosen because it achieves a local energy conservation after add-overlap, and because its simple

analytic expression allows further calculation, concerning inner-products and Wigner-Ville transform.

The weighted signal in each frame is expanded over a redundant dictionary of complex waveforms localized in the time-frequency space:

$$g_\lambda[n] = \alpha_s \, h\left(\frac{n-p}{s}\right) \exp(j2\pi\nu n), \quad \lambda = \{s, p, \nu\} \quad (2)$$

s is the scale-factor, p the translation parameter and ν the modulation frequency. $h(t)$ is a real window function and α_s is a normalization factor. In standard MP, $g_\lambda[n]$ are time-discrete Gabor waveforms, which means $h(t)$ is a Gaussian function. In our case, we use compactly supported waveforms: $h(t)$ is adapted to the window function $\omega[n]$ in order to capture efficiently stationary tonal signals with a small number of components per frame.

$$h(t) = [1 + \cos(2\pi t)] \; . \; \mathbf{1}_{[0,1[}(t) \quad (3)$$

Parameters s and p are discretized in the following way:

$$s = 2^i, \quad i \in \{i_0 \cdots \log_2(N)\} \quad (4)$$

$$p = u\frac{s}{2}, \quad u \in \left\{1 \cdots \frac{2N}{s} - 1\right\} \quad (5)$$

Compared to the standard MP, we allow full-length atoms, i.e. $s = N$, and remove side-atoms, i.e. $p = 0$. This second condition can be justified by the fact that, after windowing, there is no energy left on both sides of the frame.

For frequency discretization, we propose:

$$\nu = \nu_0 \, 2^{\frac{k-k_0}{12}} \quad (6)$$

k is a frequency index corresponding to a half-tone in the musical chromatic scale, or equivalently to a MIDI note number. ν_0, at index k_0, is the calibration frequency. Usually, ν_0 is the frequency of A4, set to 440 Hz. The amplitude of k depends on the bounds of frequency analysis. Alternatively, a finer scale is obtained by adding side-frequencies around each half-tone:

$$\nu \in \bigcup \left(\nu_0 \, 2^{\frac{k-k_0}{12}}, \; \nu_0 \, 2^{\frac{k-k_0+\delta_k}{12}}, \; \nu_0 \, 2^{\frac{k-k_0-\delta_k}{12}}\right) \quad (7)$$

Setting the parameter $\delta_k \in \,]0, 1/3]$ is a trade-off between sensitivity and selectivity. In our experimentations, we chose $\delta_k = 0.2$. This finer frequency scale significantly improves partials detection when the calibration frequency is not optimal, or when a musical instrument is not perfectly in tune.

3. DECOMPOSITION ALGORITHM

The adaptive frame-based decomposition algorithm expands the audio signal $x[n]$ as:

$$x[n] = \sum_f \sum_w a_{f,w} \, g_{\lambda(f,w)}[n] \quad (8)$$

where f is the frame index, w the waveform index inside each frame and $a_{f,w}$ the decomposition coefficients. The intra-frame decomposition is performed independently with an algorithm similar to the MP. In fact, as the signal $x[n]$ is real, the intra-frame decomposition is only made of real atoms:

$$a \, g_{\{s,p,\nu\}}[n] + a^* \, g_{\{s,p,-\nu\}}[n] = 2 \, \mathrm{Re}\left(a \, g_{\{s,p,\nu\}}[n]\right) \quad (9)$$

Thus, the dictionary can be restricted to positive frequencies.

The intra-frame decomposition algorithm can be summarized as follows: at the beginning, the residual signal is equal to the signal itself. At each step, an atom is subtracted from the residual signal. This atom is co-linear to the waveform that maximizes the modulus of the inner-product with the residual signal. The decomposition is stopped when the energy of the residual signal becomes smaller that a predefined threshold. The exact description of this algorithm is:

Initialization : compute $\forall\lambda$ the inner-product $\langle x, g_\lambda \rangle$
set $w = 0$ and $R_0 = x$

while $E\{R_w\} > \varepsilon \, E\{x\}$

- Compute $\forall\lambda$ the inner-product:
  $\langle R_w, g_\lambda \rangle = \langle x, g_\lambda \rangle - \sum_{l=1}^{w} a_l \langle g_{\lambda(l)}, g_\lambda \rangle$
- Select the best waveform index:
  $\lambda(w) = \mathrm{Argmax} \, |\langle R_w, g_\lambda \rangle|$
- Subtract the corresponding real atom:
  $a_w = \langle R_w, g_{\lambda(w)} \rangle$
  $R_{w+1} = R_w - 2 \, \mathrm{Re}\left(a_w \, g_{\lambda(w)}\right)$
- Increment the waveform index: $w = w + 1$

end

$E\{x\}$ (resp. $E\{R_w\}$) is the energy of signal x (resp. R_w). The inner-products $\langle g_\mu, g_\lambda \rangle$ depend only on the dictionary and can be obtained from an analytic formula, and eventually pre-computed and stored if the dictionary is small enough. One can notice that most inner-products are equal to zero or very small, and thus, only the significant values need to be stored. The high-resolution version of the MP [7] has also been considered, but no significant improvement could be pointed out on real signals, while the computational complexity is higher.

If the frame-length is N samples, each step of the algorithm requires $O(N)$ operations. This is lower than a standard MP which requires $O(N \log_2 N)$ operations at each step, and lower that a FFT. The total amount of operations depends on the energy constant ε and on the input signal itself.

4. RESULTS

In our experimentations, we use monophonic signals downsampled from 44.1 to 11.02 kHz. The frequency scale defined by equation (7) starts at C2 (65.41 Hz) and stops at B6 (1,976 kHz). Each frame has $N = 2048$ samples, corresponding to approximately 0.4 s. With 5 scale-factors, i.e. $s \in \{256 \cdots 4096\}$, the dictionary is composed of 10260 waveforms (180224 for the standard MP). This relatively small value allows to pre-compute and store the inner-product values in order to save computation time. The overlap factor is set to $M = 32$. This value can be considered as a medium-high overlap factor. A high value gives a smooth time-frequency representation but requires more computation time. Computation time can se saved by choosing the minimum value, i.e. $M = 2$. The energy constant ε is set to a small value, 10^{-2}, in order to reject efficiently noise components.

We compare our method to a half-tone constant-Q filter-bank with 12 FIR filters per octave, derived from a quarter-tone bank [8], where only odd filters have been kept to improve the separation between sub-bands. A time-frequency image is obtained by smoothing the energy function at the output of each filter.

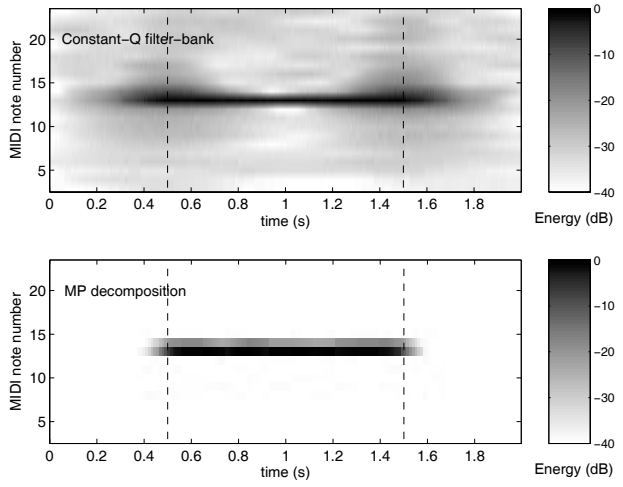

Fig. 1. Time-frequency image of a single sine (C3 + 1/16th of tone) in noise, for both methods. Dashed lines indicate the beginning and the end of the sine.

For our decomposition, we derive a time-frequency representation from the decomposition (8) with a time-discrete version of the Wigner-Ville transform [9]:

$$TF[n,\nu] = \sum_f \sum_w |a_{f,w}|^2 \; WV_{g_{\lambda(f,w)}}[n,\nu] \qquad (10)$$

The Wigner-Ville transform of each waveform depends on the transform of the window function $h(t)$:

$$WV_{g_{\{s,p,\mu\}}}[n,\nu] = \alpha_s^2 \, WV_h \left[\frac{n-p}{s}, s(\mu-\nu)\right] \qquad (11)$$

The analytic expression of WV_h was calculated from equation (3).

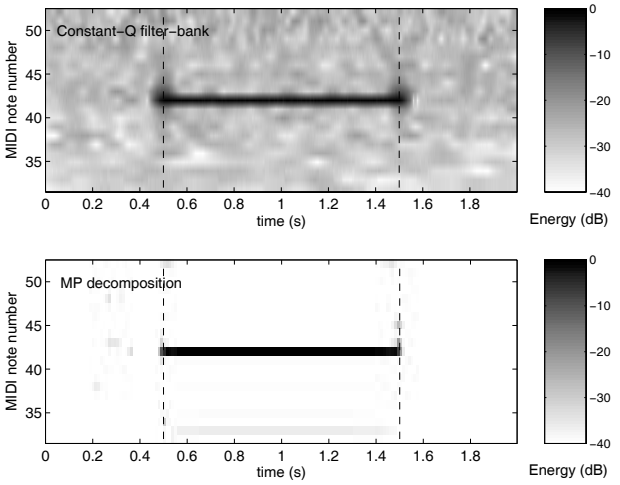

Fig. 2. Time-frequency image of a single sine (F5 + 1/16th of tone) in noise, for both methods. Dashed lines indicate the beginning and the end of the sine.

Figures 1 and 2 show the time-frequency images obtained from both methods when the signal is a single out-of-tune sine in white noise (RSB = 10 dB). The sine frequency corresponds neither to a half-tone, neither to side-frequencies defined in (7). Figure 1 represents a C3 + 1/16th of tone, and figure 2 a F5 + 1/16th of tone. We can see that our decomposition rejects efficiently the background noise. We can also notice that the time-resolution is significantly improved, especially on the C3. Both methods are characterized by a relative lack of frequency selectivity on the C3. Concerning our method, this is due to the fact that low-frequency waveforms are not orthogonal enough. One can notice a spurious sub-harmonic on the F5, but with a very low energy.

Figures 3 and 4 represent the analysis obtained from both methods on a real recording. We chose a piano piece with a fast tempo, W.A. Mozart's *Alla Turca* (see figure 5), which can be considered as a difficult material. We can see that our decomposition provides a cleaner time-frequency image than the constant-Q filter-bank, with a better contrast: with the filter-bank representation, some significant partials have the same energy-level as noise components, while with the MP decomposition, the most energetic partial of each note is clearly pointed up. Few spurious partials can be observed in the MP image, among which sub-harmonics, and their energy level is globally similar to the energy of noise components in the filter-bank image.

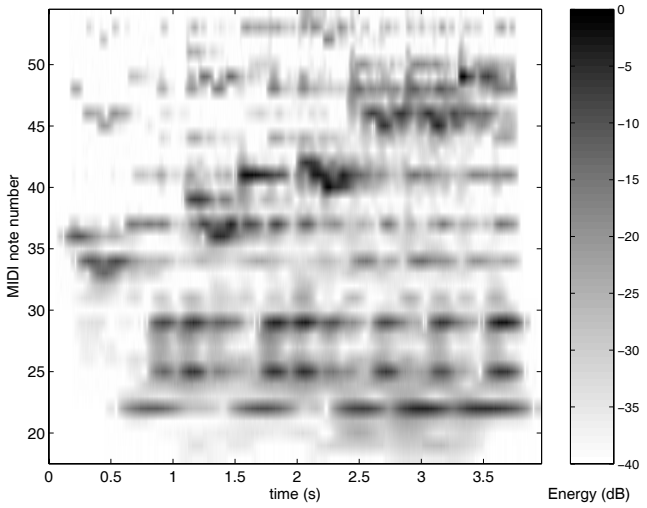

Fig. 3. Time-frequency image of a piano recording with the constant-Q filter-bank analysis.

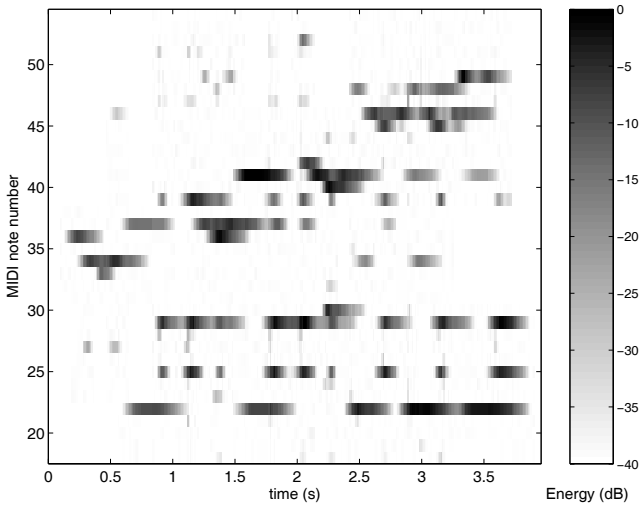

Fig. 4. Time-frequency image of a piano recording with the MP decomposition.

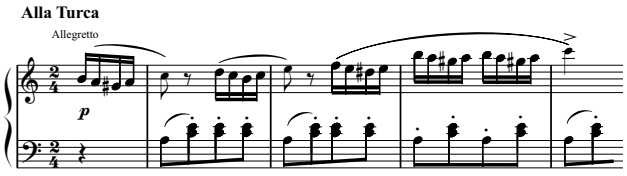

Fig. 5. Musical score of the piece of piano used for tests.

5. CONCLUSION

In this paper, we have presented a new method for frame-based time-frequency analysis of audio signal in the context of musical transcription. Our decomposition algorithm have been compared to a constant-Q filter-bank, which remains a classical solution for musical transcription. We have pointed out that frequency selectivity is similar in both techniques, but our algorithm provides a better time-resolution while the background noise is efficiently rejected. We have also shown that our method is characterized by a relatively low computational complexity, compared to other adaptive decompositions, in $O(N)$ where N is the frame-length. We believe that this analysis algorithm could be efficiently associated with a partial-tracking and note-extraction stage, using neural networks for example. This point will be investigated in further studies.

6. REFERENCES

[1] M. Marolt, "A connectionnist approach to automatic transcription of polyphonic piano music," *IEEE tr. on Multimedia*, vol. 6, no. 3, pp. 439–449, 2004.

[2] Y.R. Chien and S.K. Jeng, "An automatic transcription system with octave detection," *Proceedings of ICASSP'02*, vol. 2, pp. 1865–1868, 2003.

[3] S.G. Mallat and Z. Zhang, "Matching pursuits with time-frequency dictinaries," *IEEE tr. on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[4] B. Su and S.K. Jeng, "Multi-timbre chord classification using wavelet transform and self-organized map neural network," *Proceedings of ICASSP'01*, vol. 5, pp. 3377–3380, 2001.

[5] T.S. Verma and T.H.Y. Meng, "Sinusoidal modeling using frame-based perceptually weighted matching pursuits," *Proceedings of ICASSP'99*, vol. 2, pp. 981–984, 1999.

[6] R. Gribonval and E. Bacry, "Harmonic decomposition of audio signals with matching pursuit," *IEEE tr. on Signal Processing*, vol. 51, no. 1, pp. 101–111, 2003.

[7] R. Gribonval, E. Bacry, S. Mallat, Ph. Depalle, and X. Rodet, "Analysis of sound signals with high resolution matching pursuit," *IEEE Symp. Time-Freq. and Time-Scale Anal.*, pp. 125–128, 1996.

[8] J.C. Brown, "Calculation of a constant-Q spectral transform," *J. Acoust. Soc. Am.*, vol. 89, no. 1, pp. 425–434, 1991.

[9] J. Jeong and W.J. Williams, "Alias-free generalized discrete-time time-frequency distributions," *IEEE tr. on Signal Processing*, vol. 40, no. 11, pp. 2557–2765, 1992.