Adaptive time-frequency representation for sound analysis and processing *

Florent Jaillet[1,2], Bruno Torrésani[2]

[1] GENESIS, Bât. Beltram, Dom. du Petit Arbois, BP 69, 13545 Aix-en-Provence cedex 4, France,

[2] LATP, CMI, Univ. de Provence, 39 rue Joliot-Curie, 13453 Marseille cedex 13, France.

Email: florent.jaillet@genesis.fr, Bruno.Torresani@cmi.univ-mrs.fr

Introduction

The readability of a time-frequency representation generally depend crucially on a priori choices of some analysis parameters, which are often chosen heuristically. We describe here a set of methods for automatically selecting optimal parameters, exploiting sparsity requirements. Using Shannon or Renyi entropies for defining sparsity, we show that the "optimal" representation unfortunately often depends on the criterion. Therefore, we present several ways out for correcting such a shortcoming. In particular, we exploit the idea of "local optimization" in the time-frequency plane. Two different approaches are presented. In a supervised approach, the user can manually select the time-frequency domain in which window optimization is to be performed. In the unsupervised case, an iterative algorithm yields an automatic time-frequency segmentation, together with a corresponding inversion algorithm. As a by-product, it also provides multilayered representations for signals as in [2] [4], using reconstructions from a given window type only.

Sparsity measure

We introduce in this section the main tools we shall be using thereafter. For the sake of simplicity, we shall avoid here discretization issues, and describe our approach in terms of the short time Fourier transform (STFT), defined as follows [1]. Given a finite energy window g (assuming that $\|g\| = 1$), associate with the finite energy signal x the function G_x^g on the time-frequency plane

$$G_x^g(\tau,\nu) = \int x(t)\overline{g}(t-\tau)\, e^{-2i\pi\nu(t-\tau)}\, dt \ . \qquad (1)$$

As is well known, the STFT is invertible:

$$x(t) = \int\int G_x^g(\tau,\nu) g(t-\tau) e^{2i\pi\nu(t-\tau)}\, d\tau\, d\nu \ , \qquad (2)$$

and preserves energy

$$\int\int |G_x^g(\tau,\nu)|^2 \ d\tau\, d\nu = \int |x(t)|^2\, dt \ . \qquad (3)$$

Clearly, different windows g yield different time-frequency representations of the same signal. We are concerned here with the problem of selecting adaptively the "optimal" time-frequency representation for a given signal. To do so, we have to introduce a criterion for optimality, and sparsity is a fairly natural choice: the idea is to find the representation which optimally concentrates the energy of the signal on a small number of time-frequency atoms. Unfortunately, there does not exist a unique way of measuring sparsity, and we shall see below that different sparsity measures yield different "optimal" representations. According to classical choices [6], we shall limit ourselves to the family of Renyi entropies for defining sparsity. Consider a finite energy signal x, its short time Fourier transform G_x^γ with window γ, and introduce its normalized spectrogram $\rho_\gamma(\tau,\nu) = |G_x^\gamma(\tau,\nu)|^2/\|x\|^2$. For $\alpha \in (0,1)$, the corresponding α-Renyi entropy reads

$$R_\alpha(\gamma) = \frac{1}{1-\alpha}\, \log\left(\int\int \rho_\gamma(\tau,\nu)^\alpha d\tau d\nu\right) \ , \qquad (4)$$

and the Shannon entropy is obtained as the limiting case $S(\gamma) = \lim_{\alpha\to 1} R_\alpha(\gamma)$.

Given a family of window functions $\gamma \in \Gamma$, the "optimal" one will be defined as the one which minimizes the chosen entropy. We shall more specially focus on window families Γ consisting of dilates of a single window: $\gamma_s(t) = s^{-1/2} g(t/s)$, therefore looking for an optimal time-frequency representation from libraries of time-frequency atoms that have been considered by several authors in the literature [3]. Although we have only presented here the discussion in the continuous case, the same argument goes through directly in the more practical situation where discrete Gabor expansions are considered instead of the STFT.

Supervised adaptation

Except for some specific cases, the size of the optimal window will depend highly on the entropy chosen to define the optimality. For example, using R_α with a signal containing time-frequency atoms of different characteristics, smaller value of α turn out to yield smaller optimal windows. This shows that the notion of optimal STFT representation is not well defined for signals containing components with various time-frequency characteristics. But for a signal containing a well localized time-frequency component, explicit calculation on simple cases and numerical experiments show that the influence of the criterion on the optimal choice is higly reduced or even negligible. We thus introduce a first step in the adaptation to insure the localization of the components of the signal, by limiting the optimization to a user-defined region of the time-frequency plane. More precisely, for a given region Ω, using a given analysis window g, the corresponding signal $x_{g,\Omega}$ is reconstructed with

$$x_{g,\Omega}(t) = \int_\Omega G_x^g(\tau,\nu)\, g(t-\tau)\, e^{i2\pi\nu(t-\tau)}\, d\tau d\nu \ . \qquad (5)$$

*This work was supported in part by the European Union's Human Potential Programme, under contratct HPRN-CT-2002-00285 (HASSIP)

The selection of the optimal window is then performed by minimizing the chosen entropy for this new signal $x_{g,\Omega}$.

Unsupervised adaptation

A natural extension of the latter approach consists in introducing a prior partitioning of the whole time-frequency plane and doing the adaptation in each subdomain. A simple example of such an approach can be built using only two windows, a "narrow" one g and a "wide" one h. This case is of special interest for the analysis and processing of audio signals. Indeed, these signals often contain two main class of components with different time-frequency characteristics. On one side, transient components are very localized in time and spread in frequency, and so best represented using a narrow window. On the other side, tonal components are slowly varying in time and well localized in frequency and so represented using a wide window. Therefore, as the automatic adaptation determines which components of the signal are best represented using a narrow or a wide window, it gives the opportunity to separate transient and tonal components and to represent each with suitable parameters. This algorithm could then for example be used as a precomputation to improve the quality of decompositions of the type described in [5].

To perform the unsupervised adaptation we use an iterative algorithm to control the reconstruction error. We first define a tiling of the time-frequency plane into rectangular "super-tiles" defined by

$$\square_{m,n} = [m\Delta_\tau, (m+1)\Delta_\tau) \times [n\Delta_\nu, (n+1)\Delta_\nu) \ , \quad (6)$$

with $\Delta_\tau > 0$ and $\Delta_\nu > 0$ the time and frequency widths of the tiles. We introduce localized versions of the previously defined entropies. For a given $\alpha \in (0,1)$, for a signal y and a window γ, we define

$$C_y^{m,n}(\gamma) = \frac{1}{1-\alpha} \log\left(\int_{\square_{m,n}} \left(\frac{|G_y^\gamma(\tau,\nu)|^2}{E_\gamma^{m,n}(y)} \right)^\alpha d\tau d\nu \right) \quad (7)$$

with $E_\gamma^{m,n}(y) = \int_{\square_{m,n}} |G_y^\gamma(\tau,\nu)|^2 d\tau d\nu$ the energy in $\square_{m,n}$.

The algorithm is initialized by defining $r^{(0)} = x$. At each iteration, starting from $k = 1$, the set of tiles for which the window g is better than the window h

$$S_k = \{(m,n) \in \mathbb{Z}^2 |\, C_{r^{(k-1)}}^{m,n}(g) < C_{r^{(k-1)}}^{m,n}(h)\} \quad (8)$$

is determined. From this, the corresponding partial reconstruction $x_g^{(k)}$ of the residual $r^{(k-1)}$ is obtained via

$$x_g^{(k)} = \sum_{(m,n)\in S_k} \int_{\square_{m,n}} G_{r^{(k-1)}}^g(\tau,\nu)\, g(t-\tau)\, e^{2i\pi\nu(t-\tau)}\, d\tau d\nu, \quad (9)$$

and from the complementary subset $\mathbb{Z}^2 \setminus S_k$ the other partial reconstruction $x_h^{(k)}$ of $r^{(k-1)}$ reads

$$x_h^{(k)} = \sum_{(m,n)\in \mathbb{Z}^2\setminus S_k} \int_{\square_{m,n}} G_{r^{(k-1)}}^h(\tau,\nu)\, h(t-\tau)\, e^{2i\pi\nu(t-\tau)}\, d\tau d\nu. \quad (10)$$

$x^{(k)} = x_g^{(k)} + x_h^{(k)}$ is then an approximation of $r^{(k-1)}$ which is used to construct the new residual $r^{(k)} = r^{(k-1)} - x^{(k)}$ for the next iteration. Finally, we introduce the two layers

$$x_{g,K} = \sum_{k=1}^{K} x_g^{(k)} \ , \quad x_{h,K} = \sum_{k=1}^{K} x_h^{(k)} \ . \quad (11)$$

$x_{g,K}$ contains components that are best represented with the narrow window, i.e. mostly the transient part of x, and $x_{h,K}$ contains components that are best represented with the wide window, i.e. mostly the tonal part of x. The approximation of x at iteration K is $x_K = x_{g,K} + x_{h,K}$. Numerical tests show that the approximation error decreases (and become very small) as K grows.

An example of separation obtained with this algorithm is shown on figure 1.

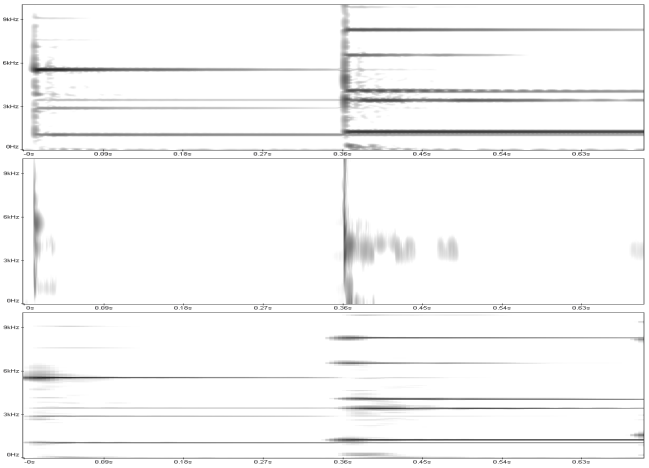

Figure 1: Example of separation on a glockenspiel sound : spectrogram of the original sound (top), spectrogram of the transient part computed with the narrow window (middle), spectrogram of the tonal part computed with the wide window (bottom).

References

[1] R. Carmona, W.L. Hwang, and B. Torrésani. *Practical Time-Frequency Analysis*, volume 9 of *Wavelet Analysis and its Applications*. Academic Press, San Diego, 1998.

[2] L. Daudet and B. Torrésani. Hybrid representations for audiophonic signal encoding. *Signal Processing*, 82(11):1595–1617, 2002. Special issue on Image and Video Coding Beyond Standards.

[3] S. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41:3397–3415, 1993.

[4] S. Molla and B. Torrésani. An Hybrid Audio Scheme using Hidden Markov Models of Waveforms, Preprint, Sept. 2003, submitted to *Appl. and Comp. Harm. Anal.*

[5] X. Serra. *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic plus Stochastic Decomposition*. PhD, Stanford University (1989).

[6] M. V. Wickerhauser. *Adapted Wavelet Analysis from Theory to Software*. AK Peters, Boston, USA, 1994.