ADJUSTING THE SPECTRAL ENVELOPE EVOLUTION OF TRANSPOSED SOUNDS WITH GABOR MASK PROTOTYPES

Adrien Sirdey,
Laboratoire de Mécanique et d'Acoustique,
CNRS
Marseille, France
sirdey@lma.cnrs-mrs.fr

Olivier Derrien,
Laboratoire de Mécanique et d'Acoustique,
CNRS
Marseille, France
derrien@lma.cnrs-mrs.fr

Richard Kronland-Martinet,
Laboratoire de Mécanique et d'Acoustique,
CNRS
Marseille, France
kronland@lma.cnrs-mrs.fr

ABSTRACT

Audio-samplers often require to modify the pitch of recorded sounds in order to generate scales or chords. This article tackles the use of Gabor masks and their capacity to improve the perceptual realism of transposed notes obtained through the classical phase-vocoder algorithm. Gabor masks can be seen as operators that allows the modification of time-dependent spectral content of sounds by modifying their time-frequency representation. The goal here is to restore a distribution of energy that is more in line with the physics of the structure that generated the original sound. The Gabor mask is elaborated using an estimation of the spectral envelope evolution in the time-frequency plane, and then applied to the modified Gabor transform. This operation turns the modified Gabor transform into another one which respects the estimated spectral envelope evolution, and therefore leads to a note that is more perceptually convincing.

1. INTRODUCTION

The aim of this study is to improve the making of digital samplers, by extending the scale range wherein notes can be extrapolated from a single recording. With a classical phase-vocoder transposition, the color of the sound gets perceptually deteriorated as soon as the transposition range exceeds between 1.5 to 2 tones. Among the different causes of this phenomena lies the unmodified relationship between the amplitude and damping of the partials implied by the phase-vocoder algorithm, which is contradictory with the physics of vibrating structures. Basically, the relative amplitudes of partials is a characteristic of the structure that generated the sound and the damping coefficient usually increases with frequency, whereas applying the phase-vocoder transposition algorithm results in a simple displacement of the partials along the frequency axis, without any modification of amplitude or damping. It is shown here that suitable Gabor masks prototypes can be used in order to restore a distribution of energy that is more in line with the physics of the sound.

After a review of the theoretical aspects regarding the phase-vocoder algorithm and the Gabor transform, the result of a traditional phase-vocoder is compared to a real harmonic sound at the same pitch. This comparison is made by use of Gabor transforms and Gabor masks, through which some of the phase-vocoder limitations are highlighted in a next section. These observations justify the use of 'mask prototypes' which applications and results are presented in the last section.

2. THEORETICAL BACKGROUND

2.1. Gabor transforms and Gabor masks

Most of the mathematical results concerning the Gabor transform presented here are deeply reviewed in [1]. The theoretical fundaments of Gabor masks have been described in [2], and their application to audio sounds has already been investigated in [3].

The Gabor transform is a discrete version of the short time Fourier transform. Considering a signal $x(t)$ and a Gabor frame $f = \{g, \tau, \nu\}$ where $g(t)$ is the time-window and (τ, ν) are the time-frequency lattice parameters, the analysis operator is defined as:

$$\begin{aligned} \mathscr{C}_f x(m,n) &= c(m,n) \\ &= \int_{-\infty}^{+\infty} x(t) e^{-2i\pi m\nu t} \overline{g}(t-n\tau) \mathrm{d}t \end{aligned} \tag{1}$$

where $\overline{g}$ is the complex conjugate of g, m is the the discrete frequency index and n the discrete time index. (1) can also be written in terms of a scalar product of $x(t)$ with the so-called Gabor atoms $\mathscr{M}_{m\nu}\mathscr{T}_{n\tau}g = e^{2i\pi m\nu t}g(t-n\tau)$:

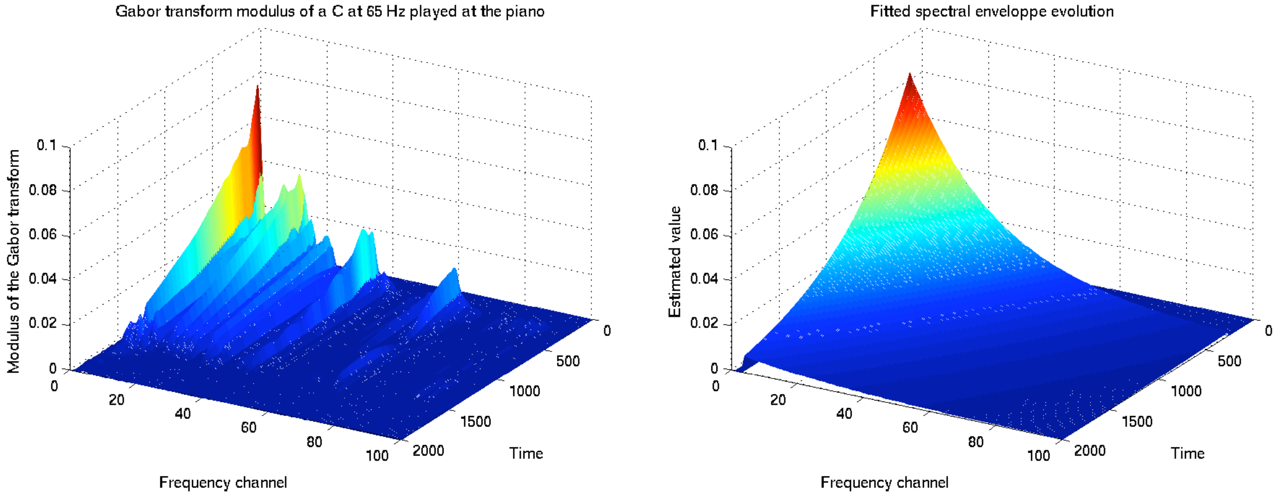
$$\mathscr{C}_f x(m,n) = \langle x, \mathscr{M}_{m\nu}\mathscr{T}_{n\tau}g\rangle_{L^2}$$

where $\mathscr{M}$ is the frequency-modulation operator and $\mathscr{T}$ is the time-translation operator.

The synthesis operator is given by:

$$\mathscr{D}_f c(t) = \sum_{m\in\mathbb{Z}}\sum_{n\in\mathbb{Z}} c(m,n)\mathscr{M}_{m\nu}\mathscr{T}_{n\tau}g \tag{2}$$

It is shown in [1] that suitable choices on g, τ and ν imply the relation $x(t) = \mathscr{D}_f\mathscr{C}_f x(t)$, which means that the Gabor transform can



(a) *Gabor transform modulus*

(b) *Fitted spectral envelope evolution*

Figure 1: *Gabor transform modulus of a real C (65 Hz) ([4]) played at the piano and the spectral envelope evolution fitted with its amplitude peaks.*

be inverted. The modulus of two Gabor transforms are presented on Figure (2a) and (2d).

Gabor multipliers are operators which modify signals through a point-wise multiplication on their Gabor transform. Given any $\mathbf{m} \in \ell^2(\mathbb{Z})$ named Gabor mask, the associated Gabor multiplier $\mathbb{M}_{\mathbf{m}}$ is defined as:

$$\mathbb{M}_{\mathbf{m}} x(t) = \mathscr{D}_f \mathbf{m} \mathscr{C}_f x(t)$$

Given two signals x_1 and x_2 called respectively the *source signal* and the *target signal*, if those two signals share close supports in the time-frequency plane (see [3] or [5] for more details), one can search for a Gabor multiplier which allows to pass from x_1 to x_2 by minimizing the cost function:

$$\Phi(\mathbf{m}) = \|x_2(t) - \mathscr{D}_f \mathbf{m} \mathscr{C}_f x_1(t)\|_2^2 + \lambda \|\mathbf{m} - 1\|_2^2$$

where the Lagrange parameter $\lambda \in \mathbb{R}^+$ is introduced in order to control the norm of $\mathbf{m}$. It is shown in [3] that under the approximation that $\mathscr{D}_f$ is an isometry, such a Gabor mask is given by:

$$\mathbf{m} = \frac{\overline{\mathscr{C}_f x_1} \mathscr{C}_f x_2 + \lambda}{|\mathscr{C}_f x_1|^2 + \lambda} \tag{3}$$

As λ increases, the mask values get closer to one; as λ decreases, (3) becomes more and more equivalent to a point-wise division in the time-frequency plane. Figure (2c) shows the modulus of the mask calculated between the two Gabor transforms displayed on Figure (2d) and (2b).

Thus, knowing two signals, Gabor masks provide an analysis tool that highlights the differences between their respective Gabor transforms. But they can also be seen as operators that allow the transformation of a Gabor transform into another one. However, their estimation requires the knowledge of both the source and target signals. In section 4 an *a priori* Gabor mask amplitude profile is elaborated (the so-called *mask prototype*) by only considering the expected behavior of the target sound.

2.2. The Phase-Vocoder Transposition Algorithm

The phase-vocoder theory is usually presented using the short-time discrete Fourier transform, but can as well be described using the Gabor transform formalism such as follows.

The classical phase-vocoder transposition algorithm combines a re-sampling that brings the original sound to the desired frequency, with the so-called 'time-scaling' procedure that brings the sound back to its original duration. Considering a harmonic signal x_1, ω_1 its fundamental frequency, and ω_2 the frequency at which x_1 is to be transposed, and defining the ratio $r = \omega_2/\omega_1$, the phase-vocoder transposition algorithm runs as follows:

1. Re-sampling $\tilde{x}_2(t) = x_1(rt)$
2. Analysis with the Gabor frame $f_a = \{g_a, \tau_a, \nu\}$

   $$\tilde{c}_2(m,n) = \mathscr{C}_{f_a} \tilde{x}_2(t)$$

3. Time-scaling by a factor r

   This procedure corresponds to a change of Gabor frame. In the synthesis Gabor frame $f_s = \{g_s, \tau_s, \nu\}$, the time-scale is dilated/compressed, and the frequency scale is unchanged. So, the new window and time-step are respectively $g_s = g_a(t/r)$ and $\tau_s = r\tau_a$. The aim of the time-scaling procedure is to compute an estimation of the Gabor coefficients $c_2(m,n)$ corresponding to the transposed sound $x_2(t)$ at the original tempo. One assumes that the amplitudes are unchanged, while the phases are modified according to a so-called horizontal phase-coherence relationship:

   $$c_2(m,n) = |\tilde{c}_2(m,n)| e^{i\varphi_{c_2}(m,n)}$$

   For each frequency bin m, un-wrapping the phase of $\tilde{c}_2$ allows to estimate an instantaneous frequency ω_i which is used to compute the corrected phase $\varphi_{c_2}(m,n)$ at time n as a function of $\varphi_{c_2}(m,n-1)$:

   $$\varphi_{c_2}(m,n) = \varphi_{c_2}(m,n-1) + \omega_i(m,n) r \tau_a$$

Since the determination of the phase $\varphi_{c_2}(m,n)$ is a recursive process, it is necessary to set its value at a certain instant. Here, the original phase has been kept unchanged at the instant of attack, according to a 'phase-locking' procedure (see [6]). Since the considered sounds are recordings of single notes, they contain only one intensity peak, and thus no complex peak detection algorithm needs to be used.

4. Inverse Gabor transform with the frame f_s:

$$x_2(t) = \mathscr{D}_{f_s} c_2(m,n)$$

3. PHASE-VOCODER DRAWBACKS AND LIMITATIONS

3.1. Artefacts

First of all, it is important to note that in general, a modified Gabor transform is not the Gabor transform of any signal. Here, the estimated Gabor transform $c_2(m,n)$ computed with the phase-vocoder algorithm is generally not the Gabor transform of the synthesized signal $\mathscr{C}_{f_s} x_2(t)$. This is due to the strict interdependence conditions between the Gabor transform coefficients of a signal that belongs to $L^2(\mathbb{R})$ (see [7] for details). As a consequence, the use of $c_2(m,n)$ in the reverse Gabor transform process generates reconstruction artefacts. Other phase-vocoder drawbacks, such as aliasing, attack smoothing due to loss of vertical phase coherence have already been deeply reviewed (see for instance [6]).

3.2. Limitations of an exclusive signal processing approach

Although the artefacts named in 3.1 are a complex problem in the implementation of phase-vocoder algorithms, several improvements have been proposed to overcome them. In the present study are only taken into account the physical aspects that the pitch-shifting algorithm doesn't cover, for it is only based on signal processing considerations. In fact, an underlying hypothesis behind the phase-vocoder transposition is that the relationship between the partials (especially in terms of amplitude and damping) is the same for each pitch. This is obviously not the case for real-life harmonic sounds produced by physical structures (e.g. musical instruments). Figures (2d), (2b) provide an illustration of a typical issue when transposing a note to a higher pitch by use of the phase-vocoder: the high frequency energy content of the transposed note is much richer than the real one. This phenomena is one of the causes of the sound-color modification trough phase-vocoder transposition.

4. GABOR MASK PROTOTYPES

4.1. Elaboration of a Gabor mask prototype using the spectral envelope evolution

It is proposed to modify the Gabor transform of the transposed note so that the damping law of the partials remains unchanged. To do so, a mask prototype is elaborated by computing the ratio between an estimation of the spectral envelope that the ideal transposed sound should have, and the spectral envelope of the actual transposed sound. The mask prototype is then applied to the time-stretched Gabor transform. The protocol is summarized on Figure (3).

The basic hypothesis of the whole process presented in this paragraph is to consider that an approximation of a Gabor mask between two Gabor transforms can be given by the ratio of their respective spectral envelope evolutions. Keeping the same notations as in section 2.2, and calling $c_3(m,n)$ the ideal Gabor transform of the transposed sound, a mask prototype $\boldsymbol{m}_p$ is sought as:

$$\boldsymbol{m}_p(m,n) = \frac{\mathrm{Env}\{c_3\}(m,n)}{\mathrm{Env}\{c_2\}(m,n)} \tag{4}$$

where $\mathrm{Env}\{\}$ denotes the spectral envelope evolution expressed in the time-frequency plane. Given a convenient spectral envelope model, $\mathrm{Env}\{c_2\}(m,n)$ can be extracted by observing c_2, whereas $\mathrm{Env}\{c_3\}(m,n)$ is to be extrapolated from the spectral envelope evolution $\mathrm{Env}\{\mathscr{C}_{f_s} x_1\}$ of the source sound $x_1(t)$, for it contains information on the physical behavior that is wanted to be kept unchanged through transposition.

In order to compute $\mathrm{Env}\{\mathscr{C}_{f_s} x_1\}(m,n)$, it is possible to write the Gabor transform of $x_1(t)$ in f_s, and then apply a fitting algorithm over the Gabor transform modulus. But it is also possible to avoid the computational cost of another Gabor transform by using the specific relationship between the analysis and the synthesis frame, which allows to write:

$$\begin{aligned}\mathscr{C}_{f_s}[x_1(t)](m,n) &= r\mathscr{C}_{f_a}[x_1(rt)](rm,n)\\ &= r\mathscr{C}_{f_a}[\tilde{x_2}(t)](rm,n)\end{aligned} \tag{5}$$

Note that the notation rm can be misleading, for rm generally belongs to $\mathbb{R}$ whereas the time and frequency index (m,n) of a Gabor transform belong to $\mathbb{Z}^2$. But in the practical situations considered here, $\mathscr{C}_{f_a}[\tilde{x_2}(t)](rm,n)$ is never to be computed and this notation is only a calculus intermediary used to apply a similar equality to the spectral envelope evolutions. Indeed (5) implies that:

$$\begin{aligned}\mathrm{Env}\{\mathscr{C}_{f_s} x_1\}(m,n) &= r\mathrm{Env}\{\mathscr{C}_{f_a}\tilde{x_2}\}(rm,n)\\ &= r\mathrm{Env}\{\tilde{c_2}\}(rm,n)\end{aligned}$$

And since the time-stretching procedure doesn't modify the modulus of the Gabor transform coefficients, one finally has:

$$\mathrm{Env}\{\mathscr{C}_{f_s} x_1\}(m,n) = r\mathrm{Env}\{c_2\}(rm,n) \tag{6}$$

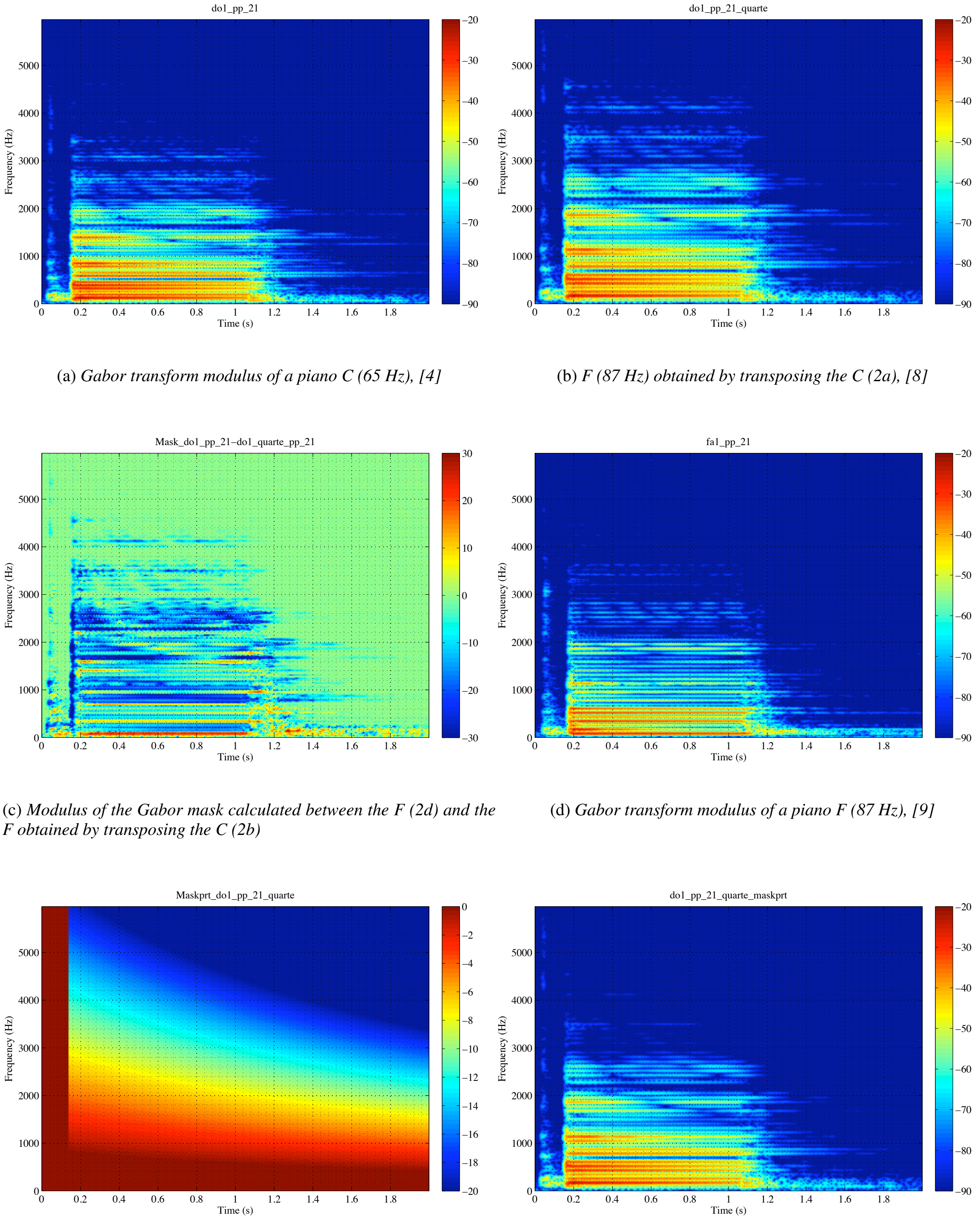
In the following, an explicit model of spectral envelope evolution is given to $\mathrm{Env}\{\}$. It is assumed that the energy of the manipulated sounds presents a log-linear decrease in frequency, an exponential decrease in time, and that the time-damping coefficient increases linearly with frequency. These assumptions are coherent with the free response of oscillating systems (such as the piano) in a linear approximation. Such hypothesis imply that the spectral envelope evolution of $c_2(m,n)$ can be written has:

$$\mathrm{Env}\{c_2\}(m,n) = ke^{\alpha n + (\beta n + \gamma)m} \tag{7}$$

The amplitude parameter k, and the damping parameters α, β, γ are estimated by use of a least-square method applied over the peaks of the time-stretched Gabor transform modulus $|c_2|$. α will be called the time-damping parameter, γ the frequency-damping parameter, and β the compound-damping parameter. Note that the estimation of these three damping parameters should logically lead to negative values. Equation (6) then leads to:

$$\mathrm{Env}\{\mathscr{C}_{f_s} x_1\}(m,n) = rke^{\alpha n + (r\beta n + r\gamma)m}$$

This shows that the transposition process presented on Figure (3) leads to a Gabor transform $c_2(m,n)$ that presents a frequency- and



(a) *Gabor transform modulus of a piano C (65 Hz), [4]*

(b) *F (87 Hz) obtained by transposing the C (2a), [8]*

(c) *Modulus of the Gabor mask calculated between the F (2d) and the F obtained by transposing the C (2b)*

(d) *Gabor transform modulus of a piano F (87 Hz), [9]*

(e) *Mask prototype*

(f) *Result of the point-wise multiplication between the transposed F (2b) and the mask prototype (2e), [10]*

Figure 2: *Mask prototype (2e) elaborated considering the transposition of a piano C a quart higher, and the result of its application (2f) to the modified Gabor transform (2b). Magnitude scales are in dB.*

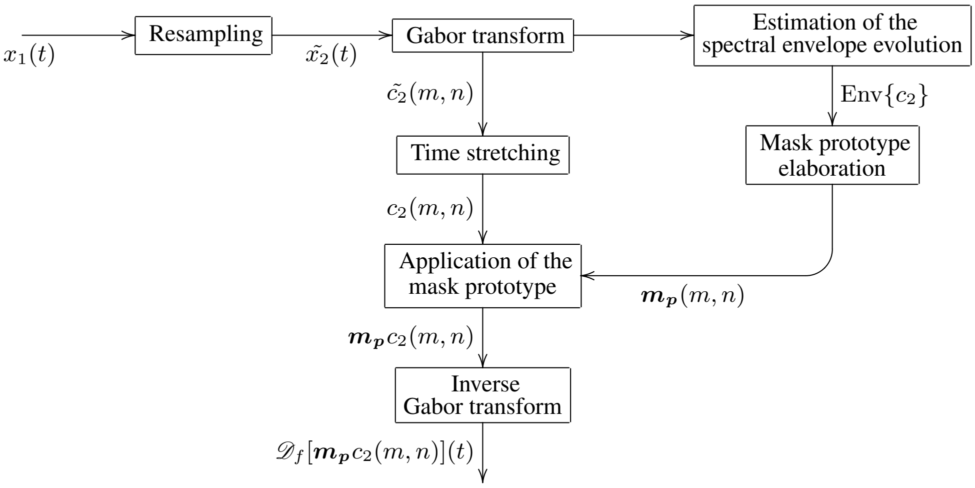
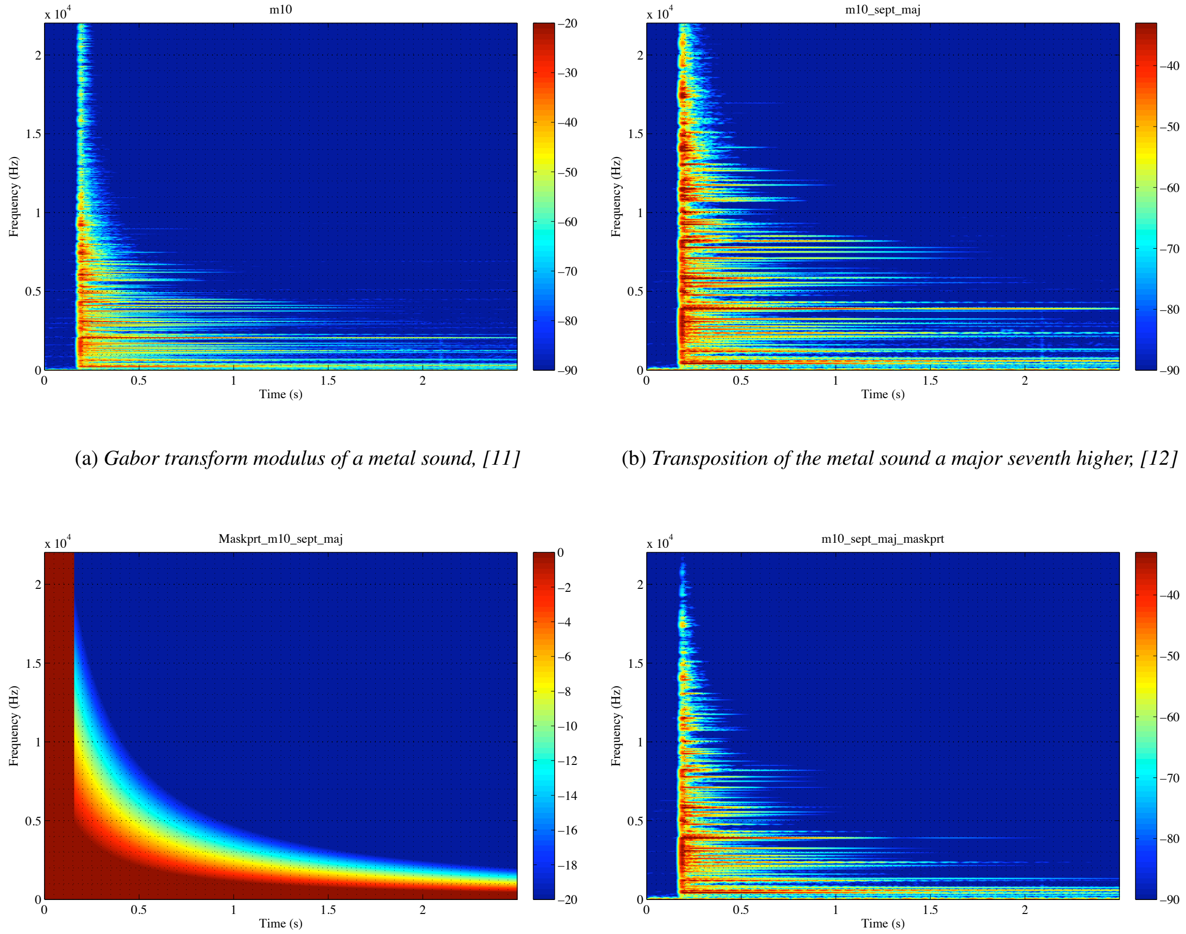

$x_1(t)$
Resampling
$\tilde{x_2}(t)$
Gabor transform
Estimation of the spectral envelope evolution
$\tilde{c_2}(m,n)$
$\mathrm{Env}\{c_2\}$
Time stretching
Mask prototype elaboration
$c_2(m,n)$
Application of the mask prototype
$\boldsymbol{m_p}(m,n)$
$\boldsymbol{m_p}c_2(m,n)$
Inverse Gabor transform
$\mathscr{D}_f[\boldsymbol{m_p}c_2(m,n)](t)$

Figure 3: *Transposition process*



(a) *Gabor transform modulus of a metal sound, [11]*

(b) *Transposition of the metal sound a major seventh higher, [12]*

(c) *Mask prototype*

(d) *Result of the application of the mask prototype, [13]*

Figure 4: *Mask prototype (4c) elaborated considering the transposition of a metal sound a seventh higher and the result of its application (4d) to the modified Gabor transform (4b). Magnitude scales are in dB.*

compound-damping behavior modified by a factor r in comparison to the source sound, and that the temporal damping parameter remains unchanged. Thus the ideal envelope $\mathrm{Env}\{c_3\}(m,n)$ can be set to:

$$\mathrm{Env}\{c_3\}(m,n) = ke^{\alpha n + (r\beta n + r\gamma)m}$$

i.e. an envelope that has the amplitude parameter k of $\mathrm{Env}\{c_2\}$, but the damping parameters α, $r\beta$ and $r\gamma$ of $\mathrm{Env}\{\mathscr{C}_{f_s} x_1\}(m,n)$. The mask prototype (4) can therefore be written as:

$$\boldsymbol{m}_p = e^{(\beta n + \gamma)(r-1)m}$$

The values of the mask prototype before the attack and below the fundamental frequency after transposition are set to one.

It is important to note that:

- Such masks only concern amplitude modification, which means that they leave the phase of the modified Gabor transform unchanged. This is motivated by the fact that a simple model for phase information cannot be found.
- The model described above will only be used for transposition to higher tones, i.e with $r > 1$. Applying this model to a transposition to lower tones would lead to a diverging mask prototype as m and n grow. Furthermore, applying this mask would lead to an amplification of energy on non-harmonic portions of the time-frequency plane.

4.2. Applications

4.2.1. Piano notes

The method described in section (4.1) has first been applied to piano sounds. The result of the transposition has been compared to a real recorded note. Figure (2) provides support for the analysis of the method for the transposition of a C (65 Hz) a quart higher, at the frequency of an F (87 Hz). Comparing the Gabor transforms of the real C and the real F (resp. Figure (2a) and Figure (2d)), one can see that most of the energy is contained in the same frequency range, roughly below 3500 Hz. However, the Gabor transform of the F obtained by transposing the C (Figure (2b)) shows that some energy is present up to 4600 Hz. Indeed, by listening both F, one can note that the transposed F sounds more reedy that the natural one.

Important information can also be retrieved observing the Gabor mask modulus between the real F and the transposed one (Figure (2c)). First, the highest values of the mask are located along the frequency channels corresponding to the fundamental frequency of the F (87 Hz). This is due to the fact that, for the C, the fundamental is less energetic than the first harmonic (see Figure (1a)), whereas for the F, the highest energy is located on the fundamental. One can also note that the Gabor mask modulus contains values lower than one between 0 and 4500 Hz at approximately 0.15 s. This indicates that the real F and the transposed F were not perfectly aligned in time. But the main relevant observation for the present work is that most of the values of the Gabor mask modulus are lower than one, especially from 2000 Hz onwards. This is coherent with the observation of the two Gabor transforms made in the previous paragraph, and provides an empirical justification for the mask prototypes presented in this paper.

The mask prototype presented on Figure (2e) is elaborated by fitting the amplitude-peaks of the Gabor transform modulus of the transposed F presented on Figure (2c). As expected, the mask values get smaller as frequency and time grow, and its application to the Gabor transform of the transposed F (Figure (2f)) diminishes the energy at high frequencies. The resulting sound files are available at www.lma.cnrs-mrs.fr/~kronland/DAFx10/. One can note that although the 'masked' note is more perceptually convincing than the directly transposed one, it lacks some low frequency content that is present in the real F. This is probably due to the fact that the C has a low energy on the fundamental, which was already mentioned above. This characteristic is not compatible with the spectral envelope model used to elaborate the mask prototype. This example allows to point out the benefits of the method as well as its limitations: the mask prototypes can only be used to erase an excess of energy in the time-frequency plane, but cannot correct a lack of energy.

4.2.2. Metal sound

The method described above is not only suitable for notes obtained with a musical instrument. It can also be applied to any sound for which the damping model is consistent. In the following, the transposition of a sound obtained by impacting a metal plate with a drumstick is considered. Although the resulting sound is not harmonic, a pseudo chromatic scale can be obtained by use of the transposition algorithm. The several Gabor transform modulus involved in the transposition of this metal sound a major seventh higher are presented on Figure (4). On Figure (4a), one can observe that the original sound contains a high concentration of partials up to 8500 Hz. After transposition, these partials are displaced up to 17000 Hz, which significantly changes the sound color. The Gabor transform on Figure (4d), obtained with the mask prototype displayed on Figure (4c), exhibits an energy distribution that is much more coherent with the source sound. The corresponding audio files are available at [11], [12] and [13].

5. CONCLUSION AND PERSPECTIVES

It has been shown here that a presumption on the physical model that describes the transposed sounds can be efficiently used to improve the realism of the phase-vocoder transposition, by use of Gabor mask prototypes.

The spectral envelope estimation being based on a peak detection algorithm and a fitting algorithm, it is necessary, in order to compute it, to set the value of several parameters such as the minimal amplitude of the peaks or their maximal thickness. These choices are of importance for they influence the shape of the mask prototype, and consequently the color of the resulting sound. A deep investigation on the influence of the underlying algorithms parameters over the resulting sound color would provide more stability to the transposition algorithm, as well as more flexibility.

For more complex sounds, which spectral envelope evolution can not be described by the model (7) used in this paper, a more sophisticated spectral-envelope model could be developed. This again, would allow to use the transposition algorithm in a wider range of situations.

6. ACKNOWLEDGMENT

The authors would like to thank Benoît Djerigian for providing the piano sounds database, and Mitsuko Aramaki for providing an environmental sound database. The Linear Time/Frequency Toolbox written by Peter Soendergaard has also been of great use for the numerical implementations of the protocol.

7. REFERENCES

[1] Karlheinz Gröchenig, *Foundations of Time-Frequency Analysis*, Birkhäuser, 2001.

[2] Monika Dörfler and Bruno Torrésani, "Representation of operators in the time-frequency domain and generalized Gabor multipliers," *Journal of Fourier Analysis and Applications*, vol. 16, n° 2, pp. 261–293, April 2010.

[3] Philippe Depalle, Richard Kronland Martinet, and Bruno Torrésani, "Time-frequency multipliers for sound synthesis," in *Wavelet XII SPIE annual Symposium*, SPIE, Ed., San Diego USA, 2007, vol. 6701 of *Proceedings of the SPIE*, pp. 670118–1 – 670118–15, OR 20.

[4] "piano_pp_21_65_do1.wav," http://www.lma.cnrs-mrs.fr/~kronland/Dafx10/.

[5] Monika Dörfler and Bruno Torrésani, "Spreading function representation of operators and Gabor multiplier approximation," http://hal.archives-ouvertes.fr/hal-00146274/en/, may 2007.

[6] Jean Laroche and Mark Dolson, "Improved phase vocoder time-scale modification of audio," *IEEE Transactions on Speech and Audio Processing*, vol. ASSP-32, n° 2, pp. 236–242, April 1984.

[7] Daniel W. Griffin and Jae S. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Transactions on Speech and Audio Processing*, vol. 7, n° 3, pp. 323–332, May 1999.

[8] "do1_pp_21_quarte.wav," http://www.lma.cnrs-mrs.fr/~kronland/Dafx10/.

[9] "piano_pp_21_87_fa1.wav," http://www.lma.cnrs-mrs.fr/~kronland/Dafx10/.

[10] "do1_pp_21_quarte_maskprt.wav," http://www.lma.cnrs-mrs.fr/~kronland/Dafx10/.

[11] "m10_expe.wav," http://www.lma.cnrs-mrs.fr/~kronland/Dafx10/.

[12] "m10_sept_maj.wav," http://www.lma.cnrs-mrs.fr/~kronland/Dafx10/.

[13] "m10_sept_maj_maskprt.wav," http://www.lma.cnrs-mrs.fr/~kronland/Dafx10/.