

Time-Frequency multipliers for sound synthesis

Philippe Depalle, Richard Kronland-Martinet, Bruno Torrésani

▶ **To cite this version:**

Philippe Depalle, Richard Kronland-Martinet, Bruno Torrésani. Time-Frequency multipliers for sound synthesis. SPIE annual Symposium Wavelet XII, Aug 2007, San Diego, United States. pp.670118-1 – 670118-15, 10.1117/12.732447 . hal-00287219

HAL Id: hal-00287219
https://hal.science/hal-00287219

Submitted on 11 Jun 2008

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Time-frequency multipliers for sound synthesis

Philipe Depalle[a,b], Richard Kronland-Martinet[b] and Bruno Torrésani[c]

[a]SPCL, McGill University, Montreal, Canada.
[b]LMA, Centre National de la Recherche Scientifique, Marseille, France.
[c]LATP, Université de Provence, Marseille, France.

ABSTRACT

Time-frequency analysis and wavelet analysis are generally used for providing signal expansions that are suitable for various further tasks such as signal analysis, de-noising, compression, source separation, ... However, time-frequency analysis and wavelet analysis also provide efficient ways for constructing signals' transformations. They are modelled as linear operators that can be designed directly in the transformed domain, i.e. the time-frequency plane, or the time-scale half plane. Among these linear operators, transformations that are diagonal in the time-frequency or time scale spaces, i.e. that may be expressed by multiplications in these domains, deserve particular attention, as they are extremely simple to implement, even though their properties are not necessarily easy to control.

This work is a first attempt for exploring such approaches in the context of the analysis and the design of sound signals. We study more specifically the transformations that may be interpreted as linear time-varying (LTV) systems (often called *time-varying filters*). It is known that under certain assumptions, the latter may be conveniently represented by pointwise multiplication with a certain time frequency transfer function in the time-frequency domain. The purpose of this work is to examine such representations in practical situations, and investigate generalizations. The originality of this approach for sound synthesis lies in the design of practical operators that can be optimized to morph a given sound into another one, at a very high sound quality.

Keywords: Time-frequency analysis, time-frequency multipliers, LTV systems, sound synthesis

1. INTRODUCTION

Time-frequency analysis is often used for studying signals, for example for analysis, denoising purpose. So far, little has been done at the level of time-frequency representation of systems, except the work done by the Vienna groups (see for example [8] for a mathematical approach, and [10] for signal processing developments).

A main point in time frequency analysis of systems is the fact that some of them may be well approximated by Gabor multipliers, which are defined by pontwise multiplication by some fixed **mask**, or **time-frequency transfer function** in the Gabor domain. To understand such approximations, one has to remember that an extremely wide class of systems can be expanded as (continuous) weighted sums of Doppler shifts (i.e. time-frequency shifts). When the latter are of small magnitude, Gabor multiplier approximations may be shown to provide very good approximations. However, when time-frequency shifts of larger magnitude are involved, one has to turn to more complex approaches. Among these, using combinations of Gabor multipliers with time-frequency shifts represent a relatively simple choice. Approximating linear systemls as sums of such objects leads to the so-called Multiple Gabor Multipliers, discussed in [3] and [4].

The purpose of this work is to investigate the practical potential of these approaches in the context of audio signal processing. More specifically, we consider the problem of estimating Gabor multipliers and multiple Gabor

Further author information:
Ph.D.: E-mail: depalle@music.mcgill.ca, Telephone: +1 514 398-4535, Ext.: 00317, Address: McGill University, Schulich School of Music, 555 Sherbrooke Ouest, Montreal (PQ), H3A 1E3, Canada
R. K-M: E-mail: kronland@lma.cnrs-mrs.fr, Telephone: +33 4 91 16 42 50, Address: LMA, CNRS, 31 ch. J. Aiguier, 13402 Marseille Cedex 20, France
B.T.: E-mail: Bruno.Torresani@cmi.univ-mrs.fr, Telephone: +33 4 91 05 46 78, Address: LATP, Université de Provence, 39 rue Joliot-Curie, 13453 Marseille Cedex 13, France

multipliers that transform a given sound signal into another one. After reviewing the main results in the Gabor multiplier theory that are of interest for the present work, we discuss the estimation problem, and propose simple estimation procedures for Gabor multipliers. Corresponding procedures in the multiple gabor multiplier case turn out to be extremely computationally intensive, and we propose a simple alternative, called the Masking pursuit algorithm. Numerical simulations on simple synthetic signals are described and commented, which illustrate the effectiveness of our approach. We conclude with an application to real sound transformation (saxophone to clarinet).

2. TIME-FREQUENCY OPERATOR REPRESENTATION AND APPROXIMATION

2.1. The time-frequency plane

Classical time-frequency signal analysis starts by mapping signals say, functions of one real variable, to a suitably chosen time-frequency representation (for example a short time Fourier transform), which is a function defined on the phase space, or time-frequency plane, denoted by $\mathbb{P}$.

The short time Fourier transform (STFT) of a (continuous time) signal $x \in L^2(\mathbb{R})$ is the function $\mathscr{V}_g x \in L^2(\mathbb{P})$, defined by

$$\mathscr{V}_g x(b,\nu) = \int_{-\infty}^{\infty} x(t) e^{-2i\pi\nu t} \overline{g}(t-b)\, dt \ = \langle x, g_{(b,\nu)} \rangle,$$

where g is a fixed **analysis window**, and $g_{(b,\nu)}(t) = \exp\{2i\pi\nu t\} g(t-b)$. If $g \neq 0$, the STFT is multiple of an isometry (i.e. $\|\mathscr{V}_g x\|^2 = C\|x\|^2$ for some constant C), and can be inverted in many different ways: for any **synthesis window** $h \in L^2(\mathbb{R})$ such that $\langle g, h\rangle \neq 0$, one has

$$x(t) = \frac{1}{\langle h, g\rangle} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathscr{V}_g x(b,\nu) e^{2i\pi\nu t} h(t-b)\, dbd\nu = \frac{1}{\langle h, g\rangle} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathscr{V}_g x(b,\nu) h_{(b,\nu)}(t)\, dbd\nu \ ,$$

the equality holding in the strong $L^2(\mathbb{R})$ sense.

It is important to realize that although the time-frequency plane $\mathbb{P}$ is topologically isomorphic to the usual Euclidean plane $\mathbb{R}^2$, it differs from it in several respects, in particular in terms of translation structure. In a few words, translating a function defined on $\mathbb{P}$ amounts to translating it in the usual way, and then "correcting" its argument, or phase. More precisely, the canonical time-frequency translation by (b,ν) on $\mathbb{P}$, denoted by $L(b,\nu)$, reads

$$[L(b,\nu)F](b_0,\nu_0) = e^{-2i\pi(\nu_0-\nu)b} F(b_0-b, \nu_0-\nu) \ . \tag{1}$$

As we shall see below, this remark has a strong importance when it comes to modelling finely transformations directly in the time-frequency plane.

2.2. The spreading function representation

We shall work in this section in the setting of $L^2(\mathbb{R})$ (continuous-time signals with infinite support). We denote by $\mathscr{H}$ the class of Hilbert-Schmidt operator on $L^2(\mathbb{R})$. One of our starting points is the so-called *spreading function representation* for linear operators. The reference functional setting is of course the $L^2(\mathbb{R})$ setting. However, since even trivial operators can have extremely wild spreading functions, it is often necessary to turn to distributional settings. This may be done in the framework of the Feichtinger algebra. We shall denote by $\mathscr{S}_0(\mathbb{R})$ the Feichtinger algebra of functions whose short time Fourier transform (with a Gaussian window) is absolutely integrable, and by $\mathscr{S}_0'(\mathbb{R})$ the dual space (we refer to [9] and references therein for more details, and extensions to wider functional settings). In addition, we denote by $\mathscr{B}$ (resp. $\mathscr{B}'$) the family of continuous operators $\mathscr{S}_0'(\mathbb{R}) \to \mathscr{S}_0(\mathbb{R})$ (resp. $\mathscr{S}_0(\mathbb{R}) \to \mathscr{S}_0'(\mathbb{R})$).

Theorem 1.

1. Let $H \in \mathscr{H}$, the class of Hilbert-Schmidt operator on $L^2(\mathbb{R})$. Then there exists a function $\eta = \eta_H \in L^2(\mathbb{R}^2)$, called the *spreading function*, such that
$$H = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \eta(b,\nu)\pi(b,\nu)\, dbd\nu \ . \tag{2}$$
the integral being interpreted in the weak operator sense.
2. The relation $\eta \in L^2(\mathbb{R}) \leftrightarrow H \in \mathscr{H}$ extends to a Gelfand triple isomorphism $(\mathscr{S}_0(\mathbb{R}), L^2(\mathbb{R}), \mathscr{S}_0'(\mathbb{R})) \leftrightarrow (\mathscr{B}, \mathscr{H}, \mathscr{B}')$.

The second item of this result is important, as it allows one to cover the case of some very simple operators (such as the identity) whose spreading function is actually a distribution.

The spreading function representation turns out to be closely related to the *twisted convolution*, which is a convolution product suitably modified to account for the particular structure of the time-frequency plane, defined by
$$(F \natural G)(b,\nu) = \int_{\mathbb{R}^2} F(b',\nu')G(b-b',\nu-\nu')e^{-2i\pi b'(\nu-\nu')}\, db'd\nu' \ . \tag{3}$$
It was shown in [3, 4] that the spreading function representation of the operator $H \in \mathscr{H}$ actually takes the form of a twisted convolution in the (continuous) time-frequency domain. η be its spreading function, as follows.

Assume for the sake of simplicity that $g, h \in L^2(\mathbb{R})$ are such that $\langle g, h\rangle = 1$. Then H may be realized as a left twisted convolution in the time-frequency domain: for all $f \in L^2(\mathbb{R})$,
$$Hx = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (\eta_H \natural \mathscr{V}_g x)\,(b,\nu) M_\nu T_b h\, dbd\nu \ . \tag{4}$$
An immediate consequence is the fact, that the range of $\mathscr{V}_g$ is invariant under left twisted convolution.

Unfortunately, such an expression is of poor practical interest. Digital signals are in practice finite dimensional, and a corresponding finite-dimensional version may be derived. Unfortunately, the numerical evaluation of such discrete twisted convolutions turns out to be extremely time consuming ($O(N^4)$ complexity). In addition, subsampling the discrete twisted convolution results in very poor approximations. A better setting for discretizing such expressions is provided by (discrete) Gabor transforms.

2.3. Gabor frames and Gabor multipliers

The Gabor transform is a sampled version of the STFT. Given lattice constants b_0, ν_0 in the time-frequency domain, denote by $g_{mn} = g_{(mb_0,n\nu_0)}$ and $h_{mn} = h_{(mb_0,n\nu_0)}$ the corresponding time-frequency translates of the analysis and synthesis windows, hereafter termed **Gabor atoms**. It may be shown[2, 9] that for suitable choices of g and the sampling constants (b_0, ν_0), the corresponding Gabor atoms g_{mn} form a frame, which implies that there exists synthesis windows h such that for all $x \in L^2(\mathbb{R})$,
$$x = \sum_{m,n} \langle x, g_{mn}\rangle h_{mn} \ .$$
We denote by
$$\mathscr{V}_g : x \in L^2(\mathbb{R}) \longmapsto \mathscr{V}_g x \ , \quad \mathscr{V}_g x(m,n) = \langle x, g_{mn}\rangle$$
the analysis operator (we use the same notation for the STFT and the Gabor transform), and by
$$S : \alpha \in \ell^2(\mathbb{Z}^2) \longmapsto S\alpha = \sum_{m,n} \alpha_{mn} h_{mn}$$

the synthesis operator. When the frame is tight,[2] h may be chosen so that $h = Kg$ for some constant K, and $S = K^{-1}\mathscr{V}_g^*$, where $\mathscr{V}_g^*$ is the adjoint of $\mathscr{V}_g$.

Given a Gabor frame, and a linear operator H, the latter may of course be characterized by its matrix elements $\langle Hg_{mn}, g_{m'n'}\rangle$ in the frame. However, there are situations in which good quality approximations can nevertheless be obtained in a simple manner. These correspond to situations in which the spreading function η_H of H is sufficiently "concentrated" in the time-frequency space. The operator H can then be suitably approximated using the so-called *Gabor multipliers*, which we now describe.

Let consider the normalised analysis and synthesis windows $g, h \in L^2(\mathbb{R})$ ($\langle g, h\rangle = 1$). Given any $\mathbf{m} \in \ell^\infty(\mathbb{Z}^2)$, termed **time-frequency transfer function** (or **mask**) we define the associated **Gabor multiplier** $\mathbb{M}_\mathbf{m}$ by Gabor transform, followed by multiplication by the mask, and inverse Gabor transform:

$$\mathbb{M}_{\mathbf{m};g,h}x(t) = \sum_{m,n=-\infty}^{\infty} \mathbf{m}(m,n)\langle x, g_{mn}\rangle h_{mn}.$$

Gabor multipliers have been studied extensively (see [6], in particular the chapter by Feichtinger and Nowak). They can also be combined with other operators, for example time-frequency shifts to yield wider classes of operators.[3,4] For example, using a time-frequency shifted copy of g as a synthesis window h is a way of implementing time-frequency shifts in the multiplier.

2.4. Approximation by Gabor multiplier

Gabor multipliers provide a flexible class of operators that is fairly easy to use. Notice however that any linear operator cannot be expressed as a Gabor multiplier. Indeed, since a Gabor multiplier is a multiplication operator in the time-frequency domain, operators that involve time-frequency shifts of large magnitude can obviously not be written as Gabor multipliers, and can hardly be approximated as such. The following result gives more insights of what can be expected.

Proposition 1. The spreading function of the Gabor multiplier $\mathbb{M}_{\mathbf{m};g,h}$ is given by

$$\eta_{\mathbb{M}_\mathbf{m}}(b,\nu) = \mathscr{M}^{(d)}(b,\nu)\mathscr{V}_g h(b,\nu)\ , \tag{5}$$

where the $({\nu_0}^{-1}, {b_0}^{-1})$-periodic function $\mathscr{M}^{(d)}$ is the symplectic Fourier transform of the transfer function $\mathbf{m}$

$$\mathscr{M}^{(d)}(t,\xi) = \sum_{m=-\infty}^{\infty}\sum_{n=-\infty}^{\infty} \mathbf{m}(m,n)e^{2i\pi(n\nu_0 t - mb_0\xi)}\ .$$

It is useful to comment on this result. Let us assume that the analysis and synthesis windows have been fixed, as well as the lattice constants b_0 and ν_0. Then the following two remarks can be made.

1. The localization of the spreading function of a Gabor multiplier matches the localization of the cross ambiguity function $\mathscr{V}_g h$. For example, in the special case $h = g$, the spreading functionη_H is localized near the origin, and its decay matches that of $\mathscr{V}_g g$. If g is well localized in time and frequency, a corresponding Gabor multiplier can therefore not involve time-frequency shifts of large magnitude.

2. In order to involve time-frequency shifts of significant magnitude, one has to use a synthesis window which is sufficiently far away from the analysis window in the time-frequency domain.

3. The spreading function η_H of the Gabor multiplier has to be such that the function $\eta_H/\mathscr{V}_g h$ is $({\nu_0}^{-1}, {b_0}^{-1})$-periodic.

Let us mention that in addition, it is possible to derive necessary and sufficient conditions ensuring the existence of best Gabor multiplier approximations for all Hilbert-Schmidt operators.

2.5. Multiple Gabor multipliers

We just saw that Gabor multipliers are good at approximating linear operators whose spreading function is suitably well localized in the time-frequency domain. However, this is not the general situation. When the spreading function η_H of H is not well enough localized, it is still possible to seek approximations as sums of Gabor multipliers. The **Multiple Gabor Multipliers**, introduced in [3, 4] are linear combinations of Gabor multipliers, with fixed analysis window, and variable synthesis windows (with different time-frequency localizations) $h^{(j)}$ and masks $\mathbf{m}_j$, and are defined as follows:

$$\mathbb{M} = \sum \mathbb{M}_{\mathbf{m}_j;g,h^{(j)}} \tag{6}$$

In the particular case where the synthesis windows are defined by time-frequency shifts of a unique window, on a suitable time-frequency lattice, conditions were given to ensure the existence of the best MGM approximation,[3,4] together with an explicit expression for the latter, in terms of the spreading function of the operator under consideration. Unfortunately, the latter expression is quite complex, and it is not clear yet how to exploit it numerically in real world situations. We present below an alternative, based upon an iterative algorithm called *masking pursuit*.

3. ESTIMATING GABOR MULTIPLIERS

Let us now turn to the estimation problem. We consider the following problem: given an input signal $x_0 \in L^2(\mathbb{R})$, and a set of output signals $x_1, \dots x_k$, images of x_0 by some operator, find a good Gabor multiplier approximation of the operator. We describe below various situations, of increasing complexity. Throughtout this section, we assume for the sake of simplicity that the window $g \in L^2(\mathbb{R})$ generates a tight Gabor frame.

Let us first consider the simple case with one input and one output signals, linked by a linear operator H with spreading function η. The model is $x_1 = Hx_0 + \epsilon$, where ϵ is nome observation noise. One then has $\mathscr{V}_g x_1 = \eta \natural \mathscr{V}_g x_0 + \mathscr{V}_g \epsilon$. Obtaining an estimate of the spreading function η from x_0 and x_1 would require inversing this twisted convolution equation, which is not easy since

- the situation is not as simple as with usual convolutions, which are diagonal in the Fourier domain.
- The solution may be highly unstable.

However, the situation may change drastically if the spreading function of interest can be considered sufficiently "concentrated". Indeed, in such a case, the operator may be conveniently approximated by a Gabor multiplier, or a composition of the latter with a time-frequency shift.

3.1. The simple case

Suppose we have the input signal x_0, and an observation x_1 of the form

$$x_1(t) = \mathbb{M}_{g,g;\mathbf{m}} x_0(t) + \epsilon_1(t) \ ,$$

from which we try to estimate the mask $\mathbf{m}$. It is natural to seek a solution by minimizing

$$\Phi[\mathbf{m}] = \|x_1 - \mathbb{M}_{g,g;\mathbf{m}} x_0\|^2 + \lambda \|\mathbf{m}\|^2 \ , \tag{7}$$

where the Lagrange parameter $\lambda \in \mathbb{R}^+$ is introduced in order to control the norm of $\mathbf{m}$.

The latter quantity reads

$$\Phi[\mathbf{m}] = \|\mathscr{V}_g^* \left(\mathscr{V}_g x_1 - \mathbf{m} \mathscr{V}_g x_0\right)\|^2 + \lambda \|\mathbf{m}\|^2 \ ,$$

and is not obvious to minimize, because $\mathscr{V}_g^*$ is not one to one.

As an alternative, one may use the proxy defined by the minimizer of

$$\Psi[\mathbf{m}] = \|\mathscr{V}_g x_1 - \mathbf{m} \mathscr{V}_g x_0\|^2 + \lambda \|\mathbf{m}\|^2 \ , \tag{8}$$

which reads

$$\mathbf{m} = \frac{\overline{\mathscr{V}_g x_0}\, \mathscr{V}_g x_1}{|\mathscr{V}_g x_0|^2 + \lambda} \ . \tag{9}$$

The latter quantity is quite easy to evaluate numerically. Corresponding results are shown and commented in section 4 below.

3.2. Multiple realizations of the output signal

Let us now turn to the situation where several realizations of the output signal (with varying amplitude) are available, with a unique input. We assume that all output signals x_k to be of the form

$$x_k = a_k \mathbb{M}_{g,g;\mathbf{m}} x_0 + \epsilon_k \ ,$$

where ϵ_k is a perturbation, that may be conveniently modelled as random Gaussian white noise. To estimate the model parameters, we seek minimizers of the quantity

$$\Phi[\mathbf{m}, a_1, \dots a_K] = \sum_{k=1}^{K} \|\mathscr{V}_g x_k - a_k \mathbf{m} \mathscr{V}_g x_0\|^2 + K\lambda \|\mathbf{m}\|^2 \ ,$$

the second term of which is introduced to ensure boundedness of the mask.

The normal equations yield the system of two equations

$$\mathbf{m}(m,n) = \frac{\left(\sum_{k=1}^{K} a_k \mathscr{V}_g x_k(m,n)\right) \overline{\mathscr{V}_g x_0}(m,n)}{\left(\sum_{k=1}^{K} |a_k|^2 \left|\mathscr{V}_g x_0\right|^2\right) + K\lambda} \ , \tag{10}$$

and

$$a_k = \frac{\sum_{m,n} \overline{\mathbf{m}}(m,n) \overline{\mathscr{V}_g x_0}(m,n) \mathscr{V}_g x_k(m,n)}{\sum_{m,n} |\mathbf{m}(m,n)|^2 \left|\mathscr{V}_g x_0(m,n)\right|^2} \tag{11}$$

which suggests to use an iterative algorithm for the search of the mask $\mathbf{m}$ and the weights a_k: iterate evaluations of the mask and the weights.

3.3. Gabor multiplier composed with time-frequency shift

Let us now study the case where the output signal is obtained through a Gabor multiplier composed with some extra deformation (some linear operator), denoted by D. The model is then as follows

$$x_1(t) = D\mathbb{M}_{g,g;\mathbf{m}} x_0(t) + \epsilon_1(t) \ ,$$

which would naturally lead to the optimization of

$$\Phi[\mathbf{m}, D] = \|x_1 - D\mathbb{M}_{g,g;\mathbf{m}} x_0\|^2 + \lambda \|\mathbf{m}\|^2 \ ,$$

for which one faces the same difficulties as before.

In what follows, we shall limit ourselves to time-frequency translations $D = \pi_{k\ell} = M_{\ell\nu_0} T_{kb_0}$, which is a unitary transformation. Assume first that k and ℓ are known. Minimizing $\Phi[\mathbf{m}, \pi_{k\ell}]$ with respect to D and for fixed $\mathbf{m}$, is equivalent to maximizing $\Re(\langle x_1, \pi_{k\ell} \mathbb{M}_{g,g;\mathbf{m}} x_0(t)\rangle)$:

$$\min_{k,\ell} \Phi[\mathbf{m}, \pi_{k\ell}] \Leftrightarrow \max_{k,\ell} \Re(\langle x_1, \pi_{k\ell} \mathbb{M}_{g,g;\mathbf{m}} x_0\rangle) \ . \tag{12}$$

Then by noticing that

$$\langle x_1, \pi_{k\ell} y\rangle = \mathscr{V}_y x_1(k,\ell) \ ,$$

the search of minima with respect to D (here, k, ℓ) amounts to a search for the maxima of the cross-ambiguity function $\mathscr{V}_y x_1$.

Let us now assume that the deformation D is known, and turn to the estimation of the mask. We can proceed as before, and start by remarking that

$$\pi_{k\ell} \mathscr{V}_g^* = \mathscr{V}_g^* L_{k\ell} \ , \quad \text{with} \quad (L_{k\ell} G)(m,n) = e^{-2i\pi b_0 \nu_0 k(n-\ell)} G(m-k, n-\ell)$$

so that

$$\|x_1 - \pi_{k\ell}\mathbb{M}_{g,g;\mathbf{m}}x_0\|^2 = \|\mathscr{V}_g^*(\mathscr{V}_g x_1 - L_{k\ell}\mathbf{m}\mathscr{V}_g x_0)\|^2 \ .$$

Notice that $L_{k\ell}$ is unitary ($\|L_{k\ell}G\|^2 = \|G\|^2$ for all G), and the adjoint operator reads

$$(L_{k\ell}^* G)(m,n) = e^{2i\pi knb_0\nu_0} G(m+k, n+\ell) \ .$$

This suggests to replace the optimization of Φ with that of

$$\Psi[\mathbf{m}] = \|\mathscr{V}_g x_1 - L_{k\ell}\mathbf{m}\mathscr{V}_g x_0\|^2 + \lambda\|\mathbf{m}\|^2 = \|L_{k\ell}^*\mathscr{V}_g x_1 - \mathbf{m}\mathscr{V}_g x_0\|^2 + \lambda\|\mathbf{m}\|^2 \ ,$$

and leads to an expression very similar to (9)

$$\mathbf{m} = \frac{\overline{\mathscr{V}_g x_0}\ L_{k\ell}^*\mathscr{V}_g x_1}{|\mathscr{V}_g x_0|^2 + \lambda} \ . \tag{13}$$

In general, both D and $\mathbf{m}$ are unknown, and have to be estimated simultaneously. This may be done using an iterative strategy recursively optimizing with respect to the mask and the deformation

- The estimation of $\mathbf{m}$ is done via (9).
- The estimation of the deformation can be performed using either the above optimization of the ambiguity function (with fixed $\mathbf{m}$) $\mathscr{V}_{\mathbb{M}_{\mathbf{m}}x_0}x_1$, or restarting from the functional Ψ, and optimizing

  $$\langle \mathscr{V}_g x_1, L_{k\ell}\mathbf{m}\mathscr{V}_g x_1\rangle = \sum \mathscr{V}_g x_1(m,n)\overline{\mathscr{V}_g x_0}(m-k, n-\ell)e^{2i\pi kb_0\nu_0(n-\ell)} \ ,$$

  i.e. a kind of (squared modulus of) twisted correlation. Since the latter involves evaluation of oscillatory sums, a first guess may be obtained by optimizing a (possibly smoothed and subsampled) classical 2D correlation product of $|\mathscr{V}_g x_1|$ and $|\mathscr{V}_g x_0|$.

3.4. An algorithm for multiple Gabor multiplier estimation: masking pursuit

As stressed above, Gabor multipliers are often not sufficient for correctly approximating operators of interest in sound analysis. Multiple Gabor multipliers (see subsection 2.5) were proposed in [3, 4] as alternatives to Gabor multipliers. Numerical implementation of MGM is currently under study, but two difficulties can be mentioned:

- The best MGM approximation of a given operator with known spreading function requires the inversion of a system of twisted convolutions, which may turn out to be a difficult problem in real world situations.
- In addition, in the problem under consideration here, the MGM has to be estimated from data, which makes the task more complex.

We describe now a simple alternative, called **Masking Pursuit**, which estimates iteratively Gabor multipliers composed with time-frequency shifts, based upon the considerations of subsection 3.3

Let x_0 and x_1 denote respectively the input and output signals as before.

- `Initialization:` Set $r^{(0)} = x_1$.
- `Iteration:` for $n = 0, ..N_{max} - 1$,
  - Estimate a mask $\mathbf{m}_{n+1}$ and time-frequency shifts (k_{n+1}, ℓ_{n+1}) from x_0 and residual $r^{(n)}$, using the approach developed in subsection 2.5.
  - Update the residual $r^{(n+1)} = r^{(n)} - \pi_{k_{n+1}\ell_{n+1}}\mathbb{M}_{\mathbf{m}_{n+1};g,g}x_0$

This yields an estimate

$$H \approx \sum_{n=0}^{N_{max}-1} \pi_{k_n\ell_n}\mathbb{M}_{\mathbf{m}_n;g,g}$$

for a MGM approximation of the operator

3.5. Numerical experiments

In this section we present simple academic examples that demonstrate the most most salient behavior of the algorithm and the effect of the main parameters. It illustrates the practical potential of the Gabor masks. Considered signals are composed of a sum of L complex exponential sequences (CES) that are modulated in amplitude.

$$s(n) = \sum_{l=1}^{l=L} a_l(n) \exp(2i\pi f_l n) \tag{14}$$

where a_l and f_l are respectively the amplitudes and the normalized frequencies of the CESs. All the examples are generated as discrete-time signals at a sampling rate of $SR = 44100Hz$; this frequency remaining a standard in audio processing.

Let start with a simple case study where both reference signal x_0 and target signal x_1 are composed of a single CES at the same frequency ($f_0 = 2000Hz$), and which amplitude is modulated differently at the attack time. The attack shape is the shape of the beginning of the temporal envelope that represents the amplitude modulation of the CES. This is an important feature which has a strong influence on the perception of sound signals (e.g. changing only the attack portion of a sound signal might completely change the perceived nature of the sound). Here the aim is to study the design a Gabor multiplier which transforms a given temporal attack shape into a new one.

The attack of the source sound starts at time $N_0 = 0.01 * SR$ and consists in a cosine-shape modulation for a duration $N = 0.05 * SR$, i.e.:

$$\begin{cases} a(n) = 0.0 & \text{for} \quad n < N_0 \\ a(n) = \frac{1}{2}(1 - cos(\frac{2\pi n}{N})) & \text{for} \quad N_0 < n \leq N + N_0 \\ a(n) = 1.0 & \text{for} \quad n > N \end{cases} \tag{15}$$

The attack of the target sound is a simple Heaviside function which discontinuity is located at time N. Figure1 represents the magnitude of the Gabor transform of the source sound, and of the target sound, while Figure 2 represents the magnitude of the mask between the source and the target. More specifically these representations have been calculated using 1000 modulations (also called channels in audio processing), and with a hop size of 8. A (very small) value of 10^{-12} was chosen for the regularisation parameter λ. The Gabor transforms of the source and the target exhibits clearly expected features. An horizontal stripe centered around $2000Hz$ that smoothly starts for the source, while it vertically spreads the power around N_0 for the target.

More interesting is the Gabor mask (cf. Figure 2), which highlights the transformation it represents. First, a vertical stripe with a high amplitude aiming at transforming the smooth start of the cosine-shape attack into a square-shape one. Second an horizontal patterns which starts at N_0, with an amplitude that monotonically decreases to stabilize at a value of 1.0 around time $N + N_0$. One may notice that the vertical stripe of the Gabor Mask remains of finite values, thanks to the benefit of the regularization provided by λ, and the presence of residual power due to the spreading of the spectral representation provided by the spectral shape of the window as well as some numerical noise.

The example number 2 uses the same source and target signals than in Example 1; the only difference is a $2000Hz$ shift in frequency and a $0.01s$ shift in time of the target signal. The pictures of Figure 3 and Figure 4 show the same main characteristics that were just described above. In addition the mask presents a few noticeable differences. First the horizontal stripe is now centred around $4000Hz$ in order to increase the weak power of the source available in these time-frequency region. Conversely, the mask comes near the zero value around the frequency of the source signal, i.e. $2000Hz$; this can be clearly seen when it clears the vertical stripe corresponding to the singularity.

The third example deals with the general case of the multiple Gabor multipliers and illustrates the *masking pursuit* algorithm described in section 3.4. While the source signal is still composed of a single CES, but at frequency $440Hz$, the target is a superposition of three equal-amplitude CES at the following frequencies

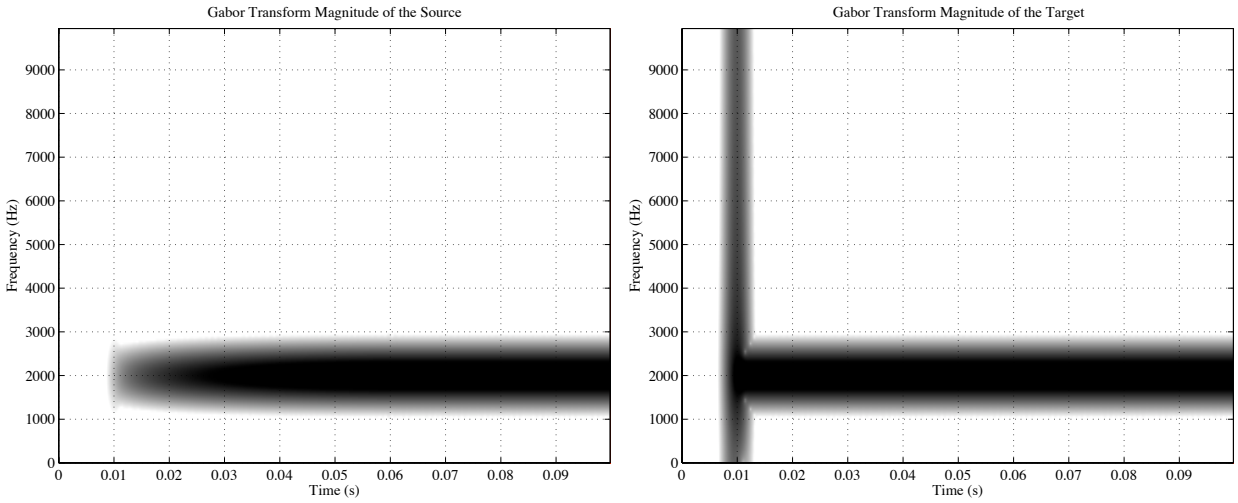


Figure 1. *Magnitude of the Gabor Transform of the source (left) and the target sounds (right).*

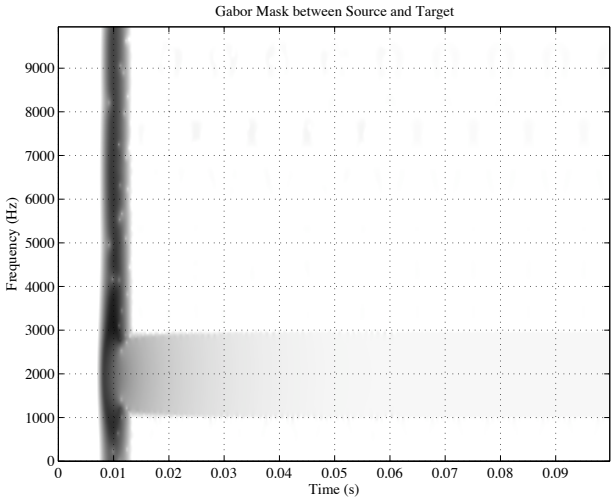


Figure 2. *Magnitude of the Gabor Mask.*

$1000, 2000$, and $3000Hz$. Both source and target components start at the same time, and with the same linear-shape attack. As discussed in section 3.4, the estimation of a time and frequency shift has to be performed at each iteration step. We use a classical 2D-correlation between the Gabor transforms of both signals. We then check for the maximum value of the correlation. At the first iteration, the estimated frequency shift is $1300Hz$, which is close to the expected value, i.e. $1560 = 2000 - 440Hz$. As a consequence the first iteration mainly focuses on the second component of the target ($2000Hz$), while it renders a fraction of the first one ($1000\ Hz$). This can clearly be seen on the left picture of Figure 5, which represents the magnitude of the Gabor Transform of the first reconstructed signal, noted u_1. Conversely the right pitcure of the same Figure shows the Magnitude of the Gabor Transform of the first residual signal r_1, as defined in 3.4.

In the same way, Figures 6 and 7 respectively show the same kind of data obtained after the second and third iteration of the adaptive algorithm. When adding the three shifted sources u_1, u_2, and u_3, we would get a generic source signal, which would provide us with a multiple Gabor multiplier simply composed by the sum of the three corresponding ones. As most audio signals are quasi-harmonic ones, or at least made of a sum of partials, this approach can become a general one to estimate masks enabling transformations of

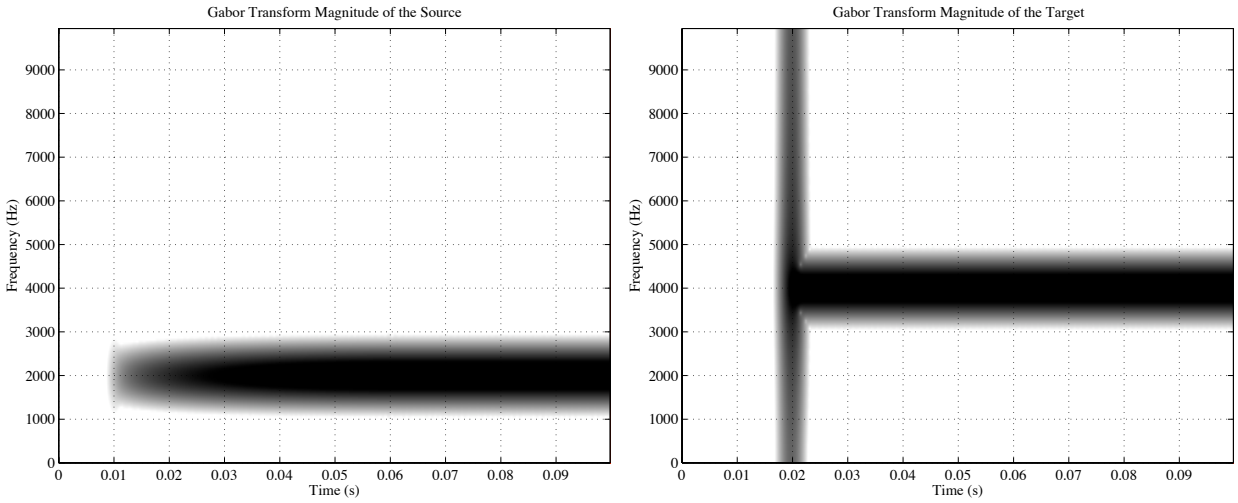


Figure 3. *Magnitude of the Gabor Transform of the source (left) and the target sounds (right).*

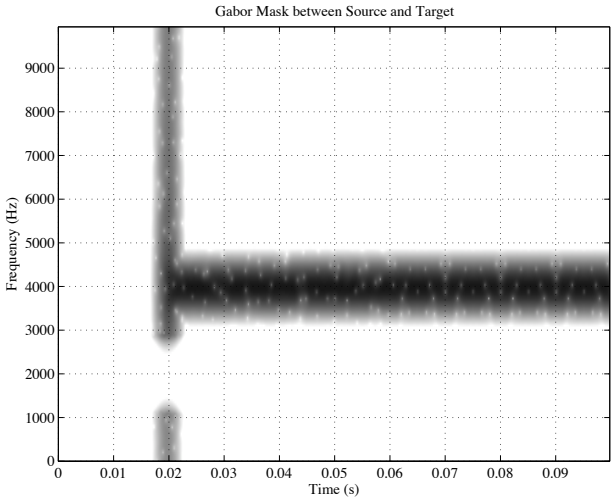


Figure 4. *Magnitude of the Gabor Mask.*

a sound of a given fundamental frequency to a sound with a different fundamental frequency. This iterative approach can be used to overcome the open problem of the Gabor multipliers and dilation.

Notice that we could have pushed the algorithm to focus on only one component for each iteration, by a proper adjustment of the λ parameter. λ might be interpretated in this context as a kind of *suppressor* that hides components, whose power are below the λ value. Thus a larger value of λ would have rejected the $1000Hz$ component at the first iteration. However the price for this simpler behavior would have been a slower convergence of the algorithm.

4. APPLICATION TO SOUND SIGNAL PROCESSING

In this section we describe what represents, to the best of our knowledge, the first application of Gabor multipliers to the processing of real audio signals. The aim of this audio application is two-fold. We first discuss the interest of using Gabor masks as a sound timbre differenciation tool, or let say it more precsiely as a time-frequency morphological differentiation of signals that can quantify the timbre perceptual differences. We then develop

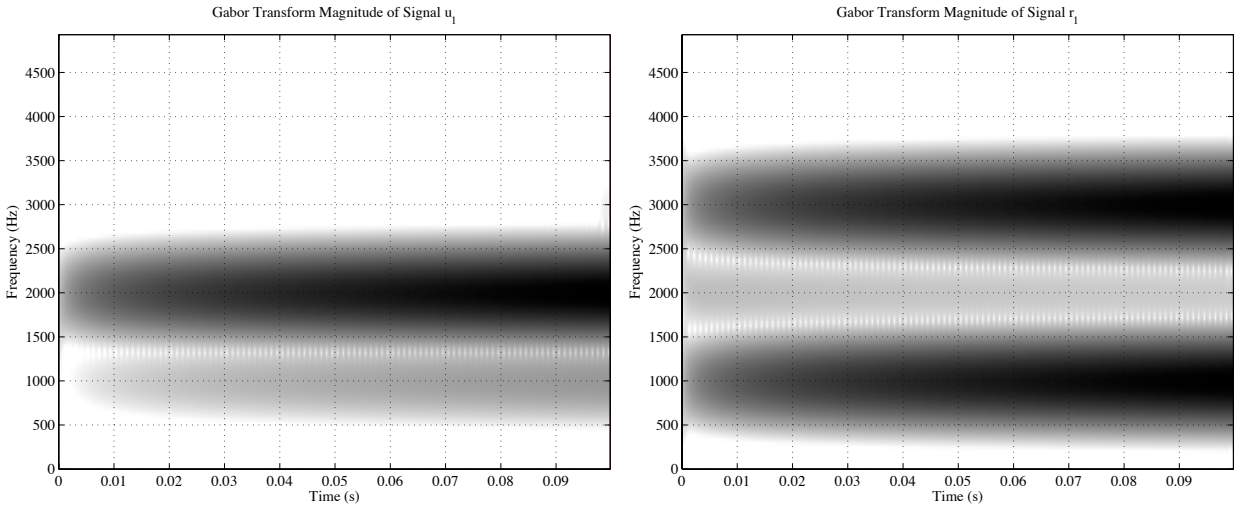


Figure 5. *First iteration of the Masking Pursuit algorithm. Magnitudes of the Gabor Transform of the first estimation (left) and the first residual (right).*

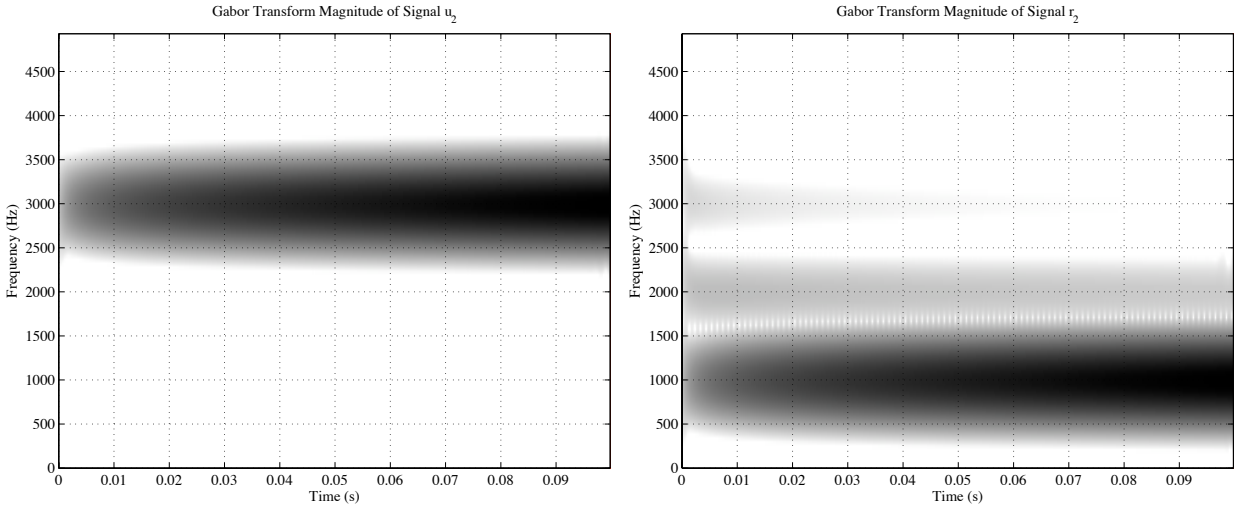


Figure 6. *Second iteration of the Masking Pursuit algorithm. Magnitudes of the Gabor Transform of the first estimation (left) and the first residual (right).*

the second idea, which is to validate the potential of Gabor multipliers for sound transformation as well as for synthesis by acting directly on the time-frequency domain.

For this purpose we choose a simple example that involves two harmonic instrumental sounds played at the same pitch, a F#4 at 370 Hz. The target signal is a sustained Eb clarinet note, while the source is a soprano saxophone note with no vibrato. In practice the actual fundamental frequencies might be slightly different for the two signals. This sounds have been downloaded from the Musical Instrument Samples database from the University of Iowa.

Before describing the shape of the Gabor Mask, let us first describe in a somewhat detailed way the main features of these two tones. This will be helpful to demonstrate how well a Gabor Mask is able to extract the appropriate differences between these two sounds.

The left part of Figure 8 is the magnitude of the Gabor Transform of the Saxophone tone. It shows most of the typical time-frequency features of this family of instruments. Among them, there is a relatively sharp attack,

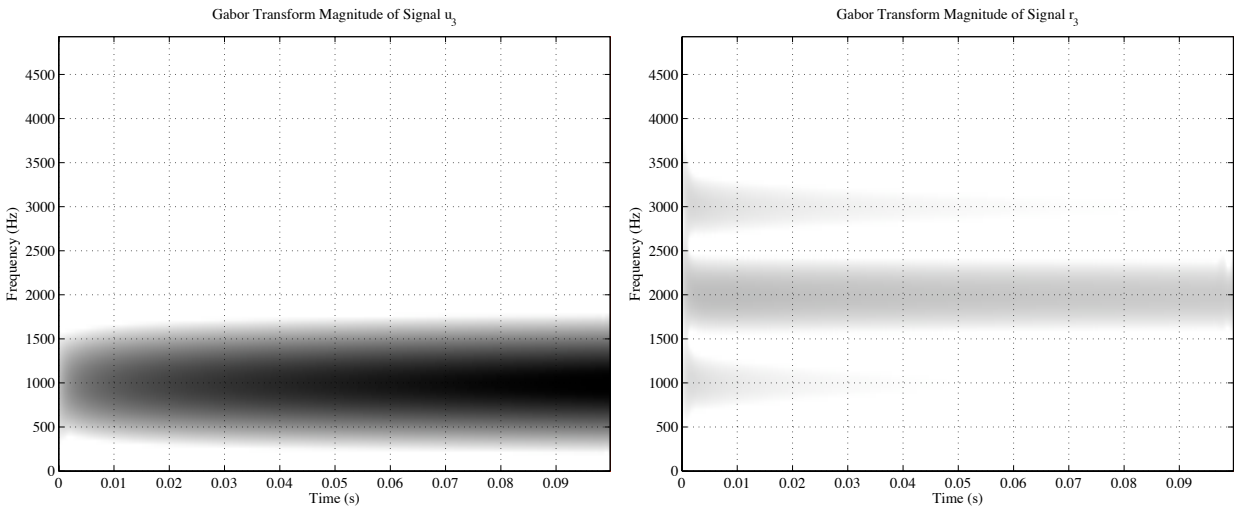


Figure 7. *Third iteration of the Masking Pursuit algorithm. Magnitudes of the Gabor Transform of the first estimation (left) and the first residual (right).*

with most of its harmonic components starting at the same time. A second feature is a relatively smooth spectral envelope, i.e. a smooth variation of the amplitude of the harmonic components with respect to the frequency. A so-called formantic structure is meanwhile responsible for an increase of power around $800Hz$ and $3500Hz$. A decrease of power can be observed around $3000Hz$; this is know as an anti-formant. As regards the release of the temporal envelope one can see a faster decrease of the high-frequency components when compared to the low-frequency ones. Finally the whole saxophone signal includes some noisy components as this is the case for most of wind instruments. However the noisy components appear in this case at a rather weak power level.

The right part of Figure 8 is the magnitude of the Gabor Transform of the clarinet tone. Contrarily to the saxophone note, the attack is smoother and the harmonic structure exhibits only a few noticeable components. Furthermore the noise level across the entire time-frequency spectrum is higher. A very specific feature of the clarinet sound is the strong difference between the power level of the odd and the even components. This is coherent with the physics (opposite boundary conditions at the limits of the clarinet tube of the instrument, see [13]), and is also the most significant perceptual feature for the recognition of clarinet tones. The release of the temporal envelope is, in the case of this clarinet sound, similar for each of spectral component. One may notice also a specific noise which appears after the release of the harmonic components (after $2s$). This might be due to the noise generated by the release of a keypad.

The Gabor mask remarkably make the differences between the two sounds explicit as shown on Figure 9. Given that the two sounds have very close fundamental frequencies, we chose to use the simple procedure described in 3.1. The mask was then computed with no frequency and no time shifting. One can notice that the slight differences in frequencies between the two sounds lie within the phase of the Gabor mask. As far as time-shift is concerned, the Gabor mask shows a black vertical stripe around the time origin. This compensate for the lack of power at the very beginning of the source sound du to the presence of a time offset, which does not appear in the clarinet sounds. An appropriate time shift estimation would have made this discrepancy vanish.

The Gabor mask highlights a strong difference between the attack-shape of the two instrumental sounds. This can be observed between time 0.01 and $0.04s$ and is somewhat similar to what appears in example 1 (see Figure 2). Nevertheless actual sounds are more complicated and the vertical stripe can be decomposed in two parts. The first one localized between $0.01s$ and $0.02s$ aims at increasing the power of the saxophone attack in order to match the one of the clarinet. The second part, between $0.02s$ and $0.04s$ shows a much stronger decrease of the power right after the attack time (usually called the decay) in the case of the clarinet.

A more obvious fact that is clearly represented in this Gabor mask concerns the difference of the power distribution along the frequency components. The Gabor Mask exhibits clear white horizontal stripes, located

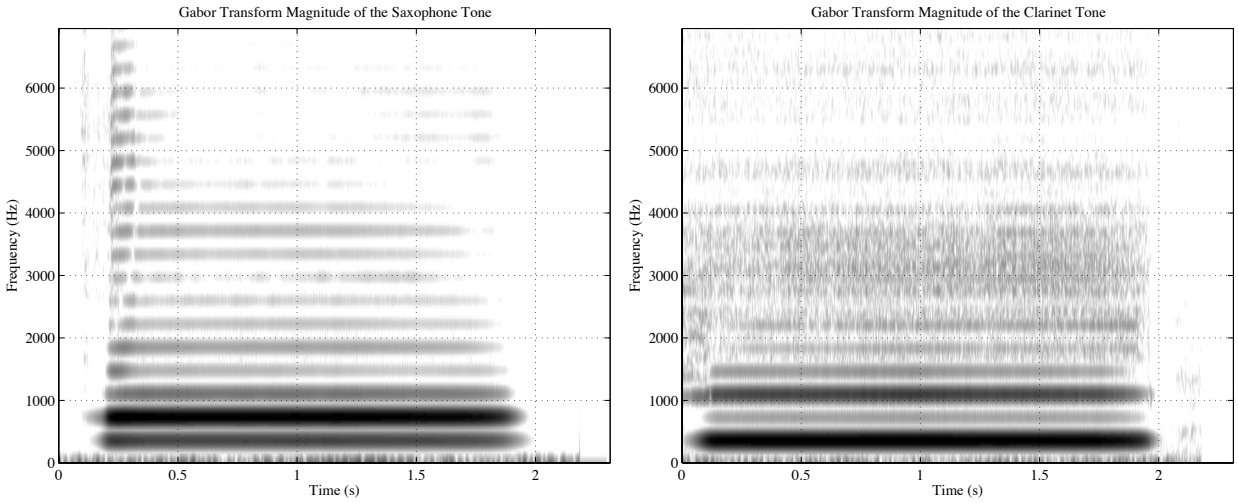


Figure 8. *Magnitude of the Gabor Transform of the saxophone tone (left) and of the clarinet tone (right).*

around even harmonics. Moreover the differences in the formantic and anti-formantic structure is well represented by respectively darker and lighter horizontal regions (around $3000Hz$ and $4000Hz$).

Finally the differences during the release part of the temporal envelope also clealry appears in the Gabor mask starting around $1.5s$ for high-frequency components. Notice that the presence of the keypad noise in the clarinet tone is taken into account and corresponds to an increase of the mask at low frequencies after $2s$.

In addition to the application we just described, the Gabor mask allows for a very high quality synthesis and transformation of sound. This is due to two facts. The first one is that the Gabor mask takes into account the phase relationship between source and target while keeping it coherent in terms of a time-frequency representation. The second one is that it explicitly design the transfer function of a system that maps a given sound into one another. In the case of the saxophone and the clarinet tone, very high quality sounds have been obtained providing an appropriate choice of the regularisation λ parameter.

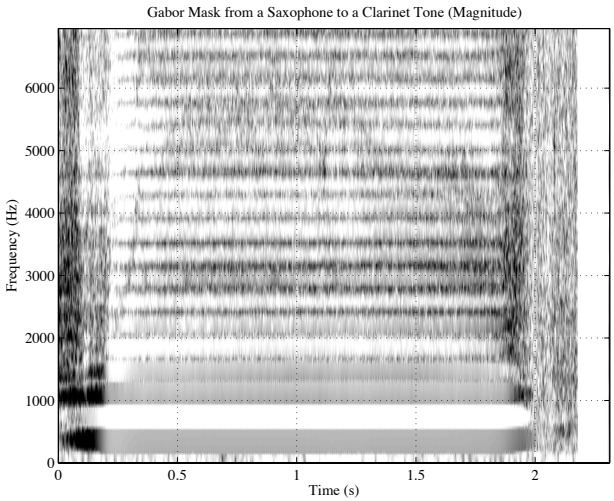


Figure 9. *Example 4.b*

CONCLUSION

We have described in this presentation some first steps towards application of Gabor multipliers to audio signal transformations. We have sticked to simple estimation procedures, which nevertheless allowed us to illustrate the interest of the proposed approach. However, lot still remains to be better understood, for example the tuning of the regularization parameter λ.

It is clear that for processing more complex signals, Gabor multipliers are generally not sufficient, and multiple Gabor multipliers will have to used, the estimation problem becoming extremely complicated. However, the masking pursuit approach we have proposed in this contribution represents a simple and efficient alternative. The previous examples have demonstrated the great potential of the Gabor mask for audio applications. We have focused on the properties of the magnitude of these masks. Nevertheless the experimental study of the phase of the Gabor masks, will likely help for a better understanding of the micro frequency variations between target and source sounds, such as a vibrato discrepancies, and subtle frequency modulations. A natural future work is to extend the use of these technique to generate a sequence of intermediate morphed sound signals. A future work has also to addresses specific strategies for the estimation of the time and frequency shift depending on the features of both source and target sounds.

ACKNOWLEDGMENTS

Numerical results were obtained using the LINEAR TIME-FREQUENCY ANALYSIS toolbox,[12] developed mainly by Peter Søndergaard, and using some specific pieces of software developed at LMA by Gregori Robinson. BT wishes to thank Monika Dörfler and Peter Balazs for stimulating discussions.

REFERENCES

1. R. Carmona, W. Hwang, and B. Torrésani. *Practical Time-Frequency Analysis: continuous wavelet and Gabor transforms, with an implementation in S*, volume 9 of *Wavelet Analysis and its Applications*. Academic Press, San Diego, 1998.
2. I. Daubechies. *Ten lectures on wavelets*. SIAM, Philadelphia, PA, 1992.
3. M. Dörfler and B. Torrésani. Spreading function representation of operators and Gabor multiplier approximation. In *Sampling Theory and Applications (SAMPTA'07), Thessaloniki, June 2007*.
4. M. Dörfler and B. Torrésani. On the time-frequency representation of operators and generalized Gabor multiplier approximations. preprint (2007), submitted
5. Y. C. Eldar, E. Matusiak, and T. Werther. A constructive inversion framework for twisted convolution. *Monatshefte fuer Mathematik* **150**:4, pp. 297-380 (2007).
6. H.G. Feichtinger and T. Strohmer Eds., *Gabor analysis and algorithms*, Series in *Applied and Numerical Harmonic Analysis*, Birkhäuser Boston Inc. (1998).
7. H. G. Feichtinger. Spline type spaces in Gabor analysis. In D. Zhou, editor, *Wavelet analysis: twenty years' developments*, Singapore, 2002. World Scientific.
8. H. G. Feichtinger and K. Nowak. A first survey of Gabor multipliers. In H. G. Feichtinger and T. Strohmer, editors, *Advances in Gabor Analysis*, Boston, 2002. Birkhauser.
9. K. Gröchenig. *Foundations of Time-Frequency Analysis*. Birkhaüser, Boston, 2001.
10. F. Hlawatsch and G. Matz. Linear time-frequency filters. In B. Boashash, editor, *Time-Frequency Signal Analysis and Processing: A Comprehensive Reference*, page 466:475, Oxford (UK), 2003. Elsevier.
11. M. Kahrs and K. Brandenbourg, *Applications of Digital Signal Processing to Audio and Acoustics*, Kluwer Academic Press, Dortrecht, The Netherland, (1998).
12. P. Søndergaard, LTFAT, the Linear Time-Frequency Analysis Toolbox, avaliable online at `http://sourceforge.net/projects/ltfat`
13. N. H. Fletcher and T.D. Rossing. *The physics of musical instruments*, Springer-Verlag, New-York, (1991).