IDENTIFICATION OF TIME-FREQUENCY MAPS FOR SOUNDS TIMBRE DISCRIMINATION

Anaïk Olivero

Laboratoire de Mécanique et d'Acoustique,
UPR 7051, CNRS
Marseille, France
olivero@lma.cnrs-mrs.fr

ABSTRACT

Gabor Multipliers are signals operator which are diagonal in a time-frequency representation of signals and can be viewed as time-frequency transfer function. If we estimate a Gabor mask between a note played by two instruments, then we have a time-frequency representation of the difference of timbre between these two notes. By averaging the energy contained in the Gabor mask, we obtain a measure of this difference. In this context, our goal is to automatically localize the time-frequency regions responsible for such a timbre dissimilarity. This problem is addressed as a feature selection problem over the time-frequency coefficients of a labelled data set of sounds.

1. INTRODUCTION

Given a pair of sound signals, our approach yields an estimate of a time-frequency transfer function to go from one sound to another. Our goal, in the present paper, is to further analyze this time-frequency transfer function. In particular we want to identify the regions in the time-frequency domain which carry most discriminant information, in the context of sounds categorization.

The approach proposed in [1] for the analysis and categorization of families of sound signals, exploits the transformation between signals in the family. In this method, the signals are supposed to be similar enough in the time-frequency domain so that these transformations can be modeled as Gabor multipliers, i.e. linear diagonal operator in a Gabor representation (subsampled version of Short Time Fourier Transform). Gabor multipliers are characterized by a *time-frequency transfer function*, hereafter called *Gabor mask*. Gabor Multipliers estimation have been studied in [2], [3], in the context of sounds transformation. In [1], Gabor masks were used to categorize sounds, by means of a corresponding complexity measure, on the basis of pairwise comparisons. Such estimated transfer functions can be viewed as a vector of features characterizing the differences between two signals. We have shown in [1] that a well chosen average the values of these features could yield sensible classifications within controlled musical signal families.

The timbre [4] is a relative notion defined as the difference between two sounds with same pitch, duration and loudness. We aim to automatically identify the time-frequency regions which have been responsible for a given timbre difference and propose a method for this task. In the context of harmonic sounds of musical instruments, it is well known [5] that the timbre can be characterized by time and spectral descriptors (such as attack time, spectral centroid, spectral flow,...). These sounds descriptors are implicitly captured in the time-frequency representation of a signal and so their differences are carried by the Gabor masks.

In the context of sounds synthesis, such a time-frequency significance map can be useful, as it gives up the time-frequency representation regions of interest for synthesizing a sound into another, as the authors in [6] who explain the importance of the control of the signals descriptors in the context of sounds morphing.

In Section 2, we present the general setting and describe the basic concepts of signal representation and time-frequency analysis we shall be working with. The feature selection problem is investigated in Section 3. Some examples of transfer functions between notes and the estimated time-frequency regions are given in Section 4 from three different instrument families.

2. GABOR FRAMES AND GABOR MULTIPLIERS

In the finite-dimensional situation $\mathbb{C}^L$, the Short Time Fourier Transform of the signal can be seen as the analysis map of a Gabor frame representation of the signal, as explained in [7]. A Gabor frame is an overcomplete family of time-frequency atoms generated by translation and modulation on a discrete lattice of a mother window, denoted by $g \in \mathbb{C}^L$. These atoms can be written as

$$\pi_{mn} g[l] = g_{mn}[l] = e^{2i\pi mb(l-na)} g[l-na] \,, \qquad (1)$$

where a and b are two positive integers, such that L is multiple of both a and b and (a, b) generates a time-frequency lattice. π_{mn} is a time-frequency shift operator. Here, all operations have to be understood modulo L. We set $M = L/b$ and $N = L/a$.

The time-frequency representations of signal x is given by

$$X[m, n] = \langle x, g_{mn} \rangle$$

In particular there are situations (called tight) where the inversion takes a particularly simple form, the analysis and synthesis windows are the same and the reconstruction is given by $x = \sum_{m,n} X[m, n] g_{mn}$. Gabor transforms give a frame framework to the time-frequency representations. In this context, a signal transformation can be constructed by pointwise multiplication between the analysis coefficients and a *transfer function*, followed by the reconstruction with the synthesis window. Such transformations are generically called *multipliers*. Denoting by $\mathbf{m}$ the transfer function, we shall denote by $\mathbb{M}_{\mathbf{m}}$ the corresponding multiplier such that

$$\mathbb{M}_{\mathbf{m}} x = \sum_{m,n} \mathbf{m}[m, n] X[m, n] g_{mn} \,. \qquad (2)$$

Let x_i and x_j denote the input and output signals, respectively. We assume the following model

$$x_j = \mathbb{M}_{\mathbf{m}} x_i + \epsilon \,,$$

where ϵ represent perturbations, modeled as additive gaussian noises, and $\mathbf{m}$ is an unknown Gabor mask, which we want to estimate. A possible solution is obviously $\mathbf{m} = X_j / X_i$, where X denote the Gabor transform of x, but such a solution is not bounded in general. We prefer to turn to a regularized least squares solution. More precisely, we seek $\mathbf{m} \in \mathbb{C}^{M \times N}$ which minimizes the expression

$$\Phi[\mathbf{m}] = \|x_j - \mathbb{M}_{\mathbf{m}} x_i\|^2 + \mu \, r[\mathbf{m}], \tag{3}$$

where $r[\mathbf{m}]$ is a regularization term, whose influence on the solution is controlled by the parameter μ.

The formulation (3) involves a non diagonal matrix, where the non diagonal terms arise from the correlations between the atoms of the representation. A first approach is to formulate the problem directly in the transform domain, or equivalently to a reduction of the problem (3) to its diagonal.

$$\tilde{\Phi}[\mathbf{m}] = \|X_j - X_i \mathbf{m}\|^2 + \mu \, r(\mathbf{m}), \tag{4}$$

Such an approximation has the advantage to admit a closed form expression for its unique minimizer. For example, we can choose $r(\mathbf{m}) = \|\mathbf{m} - \mathbf{m}^t\|_F^2$, where $\mathbf{m}^t$ is a given target time-frequency function that will help to design the estimated Gabor mask $\mathbf{m}$. The time-frequency function $\mathbf{m}^t$ can be useful in the context of sound morphing, where we aim to "interpolate" between two sound signals. Given $\mathbf{m}^t$, we obviously obtain a regularized solution for $\mathbf{m}$ which reads

$$\tilde{\mathbf{m}} = \frac{\overline{X_i} X_j + \mu \mathbf{m}^t}{|X_i|^2 + \mu} \,,$$

3. TIME-FREQUENCY CHARACTERIZATION OF A SOUNDS CATEGORIZATION

3.1. A divergence between two spectra

The Itakura Saito divergence is often used to compare two audio spectra in the context of speech processing [8]. This measure is expressed as

$$d_{IS}(|X_j|, |X_i|) = \sum_l \frac{|X_j|[l]}{|X_i|[l]} - \log \frac{|X_j|[l]}{|X_i|[l]} - 1$$

where $|X_i|$ and $|X_j|$ are the magnitude of signals spectrum or signals time-frequency spectrum and l denotes a frequency or a time-frequency bin. The Itakura Saito divergence is not symmetric and a symmetrized version [9] can be derived as

$$d_{SIS}(|X_j|, |X_i|) = \frac{d_{IS}(|X_j|, |X_i|) + d_{IS}(|X_i|, |X_j|)}{2} \tag{5}$$

We first denote that Equation (5) is not bounded in general and a way to avoid such a problem is to regularize it. If we denoted by $\mathbf{m}_{ij}$ the Gabor mask obtained by a diagonal approximation regularized with $r(\mathbf{m}) = \||\mathbf{m}| - 1\|_2^2$ between signals x_i and x_j, then $d_{SIS}(|\mathbf{m}|, 1)$ is a natural choice two compare two spectra as the masks are more stable than the quotient of two spectra.

The choice of the regularization term r was motivated by the desire of maintain $\mathbf{m} = 1$ as reference, corresponding to "no transformation". However, given that Gabor transforms of real valued signals are complex valued, and that the phase of the Gabor transform is generally difficult to handle precisely, the reference choice may be $|\mathbf{m}| = 1$ rather than $\mathbf{m} = 1$. This suggests the use of a regularization term of the form $r(\mathbf{m}) = \||\mathbf{m}| - 1\|^2$. This leads to an explicit expression for the Gabor mask given by

$$|\mathbf{m}_{ij}| = \frac{|X_i X_j| + \mu}{|X_i|^2 + \mu} \,.$$

Then, the phase of the Gabor mask is given by the phase difference between X_j and X_i.

3.2. A time-frequency map of the information responsible for the categorization

First, the Itakura Saito divergence is separable and if we define

$$\begin{aligned} d_{ij}[m,n] &= \frac{1}{2}(|\mathbf{m}_{ij}[m,n]| - \log|\mathbf{m}_{ij}[m,n]| - 1 && (6) \\ &\quad + |\mathbf{m}_{ji}[m,n]| - \log|\mathbf{m}_{ji}[m,n]| - 1) && (7) \end{aligned}$$

then the symmetrized Itakura Saito divergence reads

$$d_{SIS}(|\mathbf{m}_{ij}|, 1) = \frac{1}{MN} \sum_{m,n} d_{ij}[m,n] \tag{8}$$

The dissimilarity matrix $d[m,n]$ represents the ability of a time-frequency bin to discriminate two given classes and gives us a dissimilarity measure between two sounds for each time-frequency bin. We see in Equation (8) that the information carried by $d_{SIS}(|\mathbf{m}|, 1)$ is drastically reduced, as we just consider the sum over all the time-frequency coefficients. We also propose to use a weighted Itakura Saito divergence as

$$d_{SIS}^{\alpha}(|\mathbf{m}|, 1) = \sum_{m,n} \alpha_{mn} d[m,n] \tag{9}$$

where the α are the weights, which indicate the relevance of each time-frequency bin and are are to be estimated from data. These weights are supposed to emphasize one subset of time-frequency bins over the others. Then, we impose the following properties : $\alpha_{mn} \geqslant 0$ and $\sum_{mn} \alpha_{mn} = 1$, so that the Equation (8) can be viewed as an uniform version of the Equation (9).

We propose to model our problem in the spirit of the Relief algorithm [10], a feature weighting algorithm that iteratively selects feature over a training data set. We suppose that we have a training data set of labelled signals $\{x_i, i = 1..N\}$ composed by two different classes of signals, where the first class contains N_1 signals, and the second contains N_2 signals. We denote by $\mathcal{C}_i$ the set of indices of the signals which are in the class of the signal x_i. We want to define a distance that discriminates the 2 classes as clearly as possible. We can formally model this problem as the maximization of a margin, where the margin is given by:

$$\rho(\alpha) = \sum_{mn} \alpha_{mn} \left(\sum_i \sum_{j \notin \mathcal{C}_i} d_{ij}[mn] - \sum_i \sum_{j \in \mathcal{C}_i} d_{ij}[mn] \right)$$

In other words, the margin is considered as measure of the ability of a set of weights to discriminate two classes of signals.

For the sake of clarity, let us define

$$\begin{aligned} z_{mn} &= \left(\sum_i \sum_{j \notin \mathcal{C}_i} d_{ij}[mn] - \sum_i \sum_{j \in \mathcal{C}_i} d_{ij}[mn] \right) \\ &= \left(\langle D^-, d[m,n] \rangle - \langle D^+, d[m,n] \rangle \right) \end{aligned}$$

The matrices D^+ and D^- are in $\{0,1\}^{N\times N}$ and are used to represent the repartition of the data in two different classes. $D^+_{ij} = 1$ if i and j are in the same class and 0 otherwise, whereas $D^-_{ij} = 0$ if i and j are in the same class and 1 otherwise. The maximization of the margin emphasize the data which are in the same class. Now, the problem takes the form of a minimization under constraints

$$\max_\alpha \ \alpha^T z \ \text{ s.t. } \ \|\alpha\|_2^2 = 1 \text{ and } \ \alpha \geqslant 0 \qquad (10)$$

The solution to this problem is given by : $\alpha = \frac{z^+}{\|z^+\|_2}$, where $z^+ = [\max(z_{mn}, 0)]_{mn}$

This problem implicitly contains sparsity constraints as it reduces the time-frequency information, by remonving the negative values of z. The removed time-frequency bins can also be viewed as irrelevant for the given classification task. The α values also give us the importance of a given time-frequency coefficient in our given task. Other algorithms performs such a feature selection and we refer to [11] for a review.

4. EXPERIMENTS

The time-frequency maps given by the coefficients $\{\alpha_{mn} : m, n\}$ allows to identify the time-frequency information responsible for a classification task. They provide an average map of the differences between each class, with less variability compared to the gabor masks obtained between individual pairs of sounds. This is also a generic way to automatically enlighten the time-frequency differences of timbre between two classes of harmonic sounds, as we suppose no signal model and no descriptors choices. Here, we argue that these time-frequency maps generalize the information contained in the Gabor Masks by pairwise comparison of two classes of sounds, which can be more useful in the context of sounds morphing, as they can be used to transform a sound from one class to another. As we will see below, the time-frequency differences will depend on the sounds classes we are comparing.

Some experiments are shown here. We used three classes of musical instrument sounds playing the same note, with fundamental frequency f_0 =196 Hz (G3) : 16 clarinets, 15 saxophones (8 alto and 10 tenors) and 13 trumpets. Prior to mask estimation, the signals are adjusted so that their onset coincide, as the onset time is not relevant in our task. Now, all the sounds are supposed to have a good time-frequency alignment, so that the Gabor mask capture a pertinent information. In each experiment, a data set contains the sounds of two different classes. We considered three different data sets : the clarinets and the saxophones, the clarinets and the trumpets, the trumpets and the saxophones. The spectrograms of one sound of each class are shown in Figure 1, obtained using a Hanning mother window and parameter values $M = 512$, $a = 64$ and displayed in a logarithmic amplitude scale.

The time-frequency maps for the three data sets are computed as explained in Section 3.2 and shown in Figure 2. As expected, we can see that the three instruments classes present some time-frequency differences at different locations and these differences can be interpreted physically. All time-frequency maps exhibit a harmonic structure supply by a formantic structure, which is coherent with our understanding of the acoustic of these musical instruments. Each map emphasize the differences between two classes of sounds. For example, the even harmonics (which are known to be a relevant clue for identifying the clarinets) appear strongly in the clarinets/saxophones and clarinets/trumpets maps. However, their importance in the classification process differs sightly when the clarinets are compared to trumpets or saxophones. The maps also reveal how the frequency content during the attack differs according to the two classes we are observing. This information can be particularly useful in practice to distinguish the trumpets from the clarinets and saxophones classes.

5. CONCLUSIONS AND PERSPECTIVES

We have described in this paper a method for better exploiting the information contained in time-frequency masks estimated from families of sound. Namely, the proposed approach is able to retrieve the sub-domains in the time-frequency plane that permit discrimination of two instrument sounds playing the same note, in other words the time-frequency information carrying the timbre differences. This goal is achieved by coupling mask estimation with using a feature selection method on a labelled class of sounds.

Further developments of this work will involve the construction of smoother versions of the time-frequency map, and applications in a context of sounds morphing.

6. ACKNOWLEDGMENTS

Many thanks to Richard Kronland-Martinet and Bruno Torrésani for their help in this work.

7. REFERENCES

[1] Anaïk Olivero, Laurent Daudet, Richard Kronland-Martinet, and Bruno Torrésani, "Analyse et catégorisation de sons par multiplicateurs temps-fréquence," in *XXIIe colloque GRETSI (Dijon)*, 8-11 septembre 2009.

[2] P. Depalle, R. Kronland-Martinet, and B. Torrésani, "Time-frequency mutlipliers for sound synthesis," in *Proceedings of the Wavelet XII conference, SPIE annual Symposium*, San Diego, 4-8 September 2007, pp. 221–224.

[3] Anaïk Olivero, Bruno Torrésani, and Richard Kronland-Martinet, "A new method for gabor mulipliers estimation : Application to sound morphing," in *EUSIPCO 2010*, Alborg, Danemark, August 2010, pp. 507–511.

[4] Acoustical Terminology, *ASA, American Standards Association*, New York, 1960.

[5] G. Peeters, S. McAdams, and P. Herrera, "Instrument sound description in the context of mpeg 7," in *Proc. ICMC*, Berlin, Germany, August. 27- Sept. 1 2000, pp. 203–206.

[6] Marcelo Caetano and Xavier Rodet, "Automatic timbral morphing of musical instruments sounds by high-level descriptors," in *Proc. ICMC 2010*, pp. 11–21.

[7] Peter Søndergaard, *Finite Discrete Gabor Analysis*, Ph.D. thesis, Vienna University, 2007.

[8] R. Gray, A. Buzo, Jr. Gray, A., and Y. Matsuyama, "Distortion measures for speech processing," aug 1980, vol. 28, pp. 367 – 376.

[9] B. Wei and J. D. Gibson, "Comparison of distance measures in discrete spectral modeling," in *Proc. 9th DSP Workshop 1st Signal Processing Education Workshop*, Oct. 15-18, 2000.

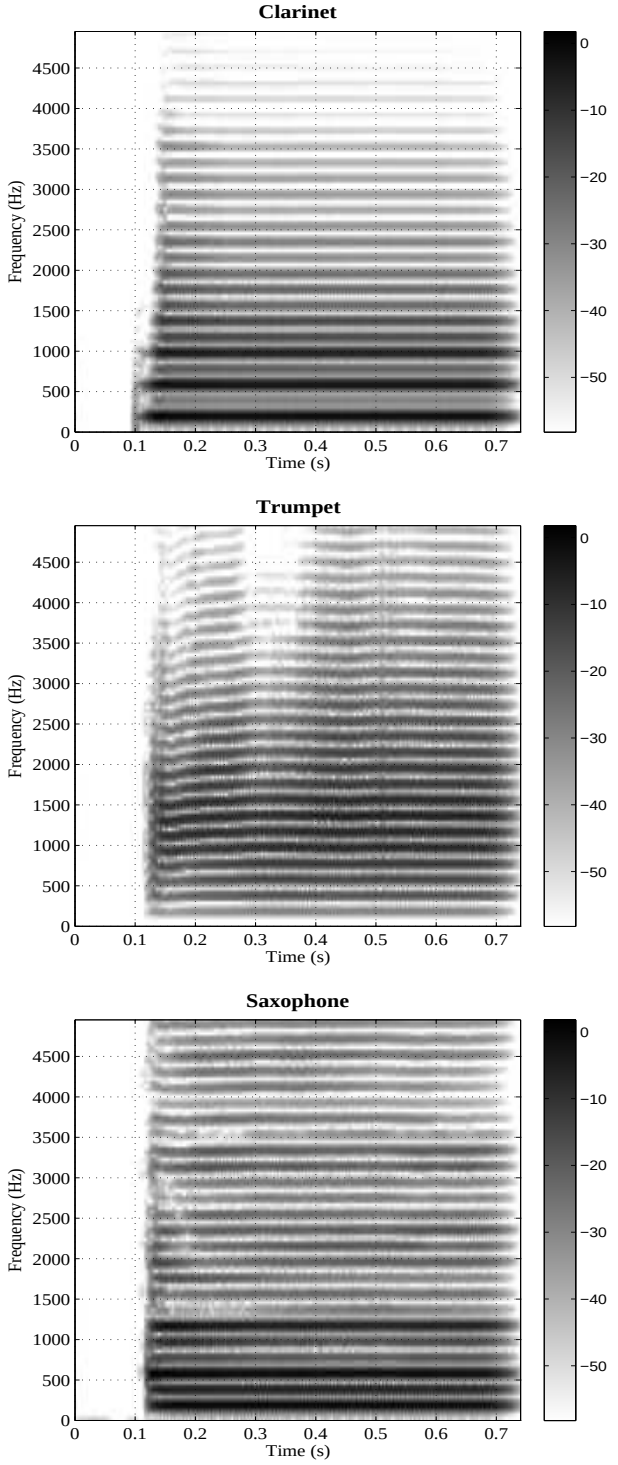

Figure 1: Three spectrograms of our sounds data set

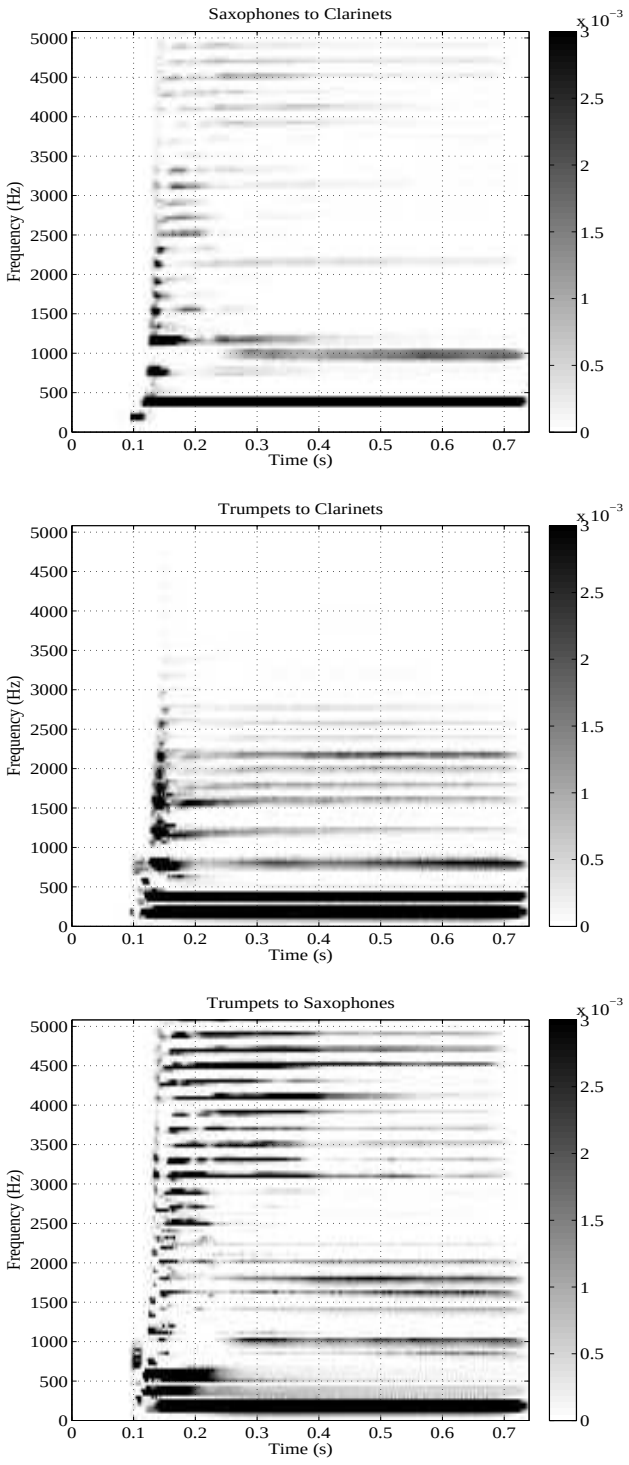

Figure 2: The three time-frequency maps α obtained from our data set by pairwise comparison of three classes

[10] Yijun Sun, "Iterative relief for feature weighting: Algorithms, theories, and applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, No 6, pp. 1035–1051, June 2007.

[11] A. Bagherjeiran and C.F. Eick, *Studies in Computational Intelligence (SCI) 73*, chapter Distance Function Learning for Supervised Similarity Assessment, Springer-Verlag, 2008.