Analyse et Catégorisation de sons par multiplicateurs temps-fréquence

Anaïk OLIVERO[1,2], Laurent DAUDET[3], Richard KRONLAND-MARTINET[1], Bruno TORRÉSANI[2]

[1]Laboratoire de Mécanique et d'Acoustique
31 Chemin Joseph Aiguier, 13331 Marseille, France

[2]Laboratoire d'Analyse, Topologie et Probabilités
CMI, 39 rue Joliot-Curie, 13453 Marseille Cedex 13, France

[3]UPMC Université Paris 06,
IJRLA/LAM, 11 rue de Lourmel, 75015 Paris - France

olivero@cmi.univ-mrs.fr, daudet@lam.jussieu.fr
kronland@lma.cnrs-mrs.fr, Bruno.Torresani@cmi.univ-mrs.fr

Résumé – On s'intéresse ici au problème de l'analyse et la catégorisation des sons par l'intermédiaire de la complexité des transformations permettant de les relier. Ces dernières sont modélisées au moyen de multiplicateurs de Gabor, opérateurs linéaires diagonaux dans une représentation de Gabor (complexe) et caractérisés par une *fonction de transfert temps-fréquence* (à valeurs complexes également). Une mesure de dissimilarité entre signaux est obtenue par mesure de la complexité de cette fonction de transfert. Cette dissimilarité est à son tour exploitée par une méthode de classification hiérarchique.

Abstract – We study here the problem of sound analysis and categorization through the complexity of transforms mapping sounds to each other. These transforms are modeled with Gabor multipliers, which are diagonal linear operator in the (complex) Gabor domain and caracterized by a time-frequency transfer function (with complex values too). A dissimilarity measure beetween signals is generated by evaluating the complexity of the so-obtained time-frequency transfer function. This dissimilarity is then used to obtain a hierarchical classification method.

1 Introduction

Les transformations temps-fréquence sont fréquemment utilisées pour fournir des représentations efficaces des signaux. L'efficacité est jugée en termes de parcimonie, c'est à dire de capacité à concentrer l'information sur un nombre contrôlé de coefficients significatifs, et d'interprétabilité. Cette seconde propriété signifie que les positions des coefficients significatifs dans le plan temps-fréquence, et leurs valeurs, contiennent des informations pertinentes sur la nature du signal analysé. Il est par conséquent pertinent de s'attendre à ce que des signaux proches se caractérisent par des représentations temps-fréquence proches, au sens où leur « morphologie » est similaire.

Parmi les approches usuelles pour la classification de signaux, on peut notamment mentionner les approches basées sur les comparaisons de spectres, par exemple des mesures de « distance » entre spectres calculées à partir d'une certaine « norme » du rapport des spectres. Le recours aux spectres suppose une hypothèse implicite de stationnarité, qui n'est généralement pas satisfaite. Nous montrons ici que des mesures similaires basées sur des « masques » estimés de multiplicateurs temps-fréquence sont plus pertinentes, y compris dans des situations où l'hypothèse de stationnarité n'est pas loin d'être satisfaite.

Nous décrivons dans cette contribution l'utilisation de cette approche pour la comparaison et la classification de signaux sonores constitués d'une note isolée, pour différents instruments. Nous comparons les résultats obtenus à partir de comparaison de spectres, et de comparaison de masques d'un multiplicateur de Gabor. Nous montrons en particulier que bien que ces sons soient relativement proches d'être stationnaires (au sens où leur attaque ne représente qu'une petite partie de leur énergie), les classifications obtenues sont bien plus pertinentes lorsque des outils temps-fréquence sont utilisés.

2 Multiplicateurs de Gabor

2.1 Repères de Gabor et multiplicateurs

Le cadre mathématique de notre approche est celui des transformations temps-fréquence linéaires, inversibles. Nous nous focaliserons sur la représentation de Gabor, et pour simplifier les notations nous nous placerons dans le cadre $L^2(\mathbb{R})$ des signaux d'énergie finie, en rappelant qu'une théorie identique existe pour des signaux de longueur finie (voir [7] par exemple).

Un repère de Gabor est une famille surcomplète d'atomes temps-fréquence générés par translations et modulations sur un réseau discret d'une fenêtre de référence, notée g. Les atomes

sont de la forme

$$g_{mn}(t) = e^{2i\pi n\nu_0(t-mb_0)} g(t - mb_0),$$

où b_0 et ν_0 sont deux nombres fixant le réseau temps-fréquence utilisé. La transformation de Gabor associe à tout signal $x \in L^2(\mathbb{R})$ sa transformée $\mathcal{V}_g x$, définie par

$$\mathcal{V}_g x[m,n] = \langle x, g_{mn} \rangle = \int_{-\infty}^{\infty} x(t) e^{-2i\pi n\nu_0(t-mb_0)} \overline{g}(t-mb_0)\, dt.$$

Sous des hypothèses assez peu contraignantes sur la fenêtre g et si le produit $b_0\nu_0$ est assez petit, la transformation est inversible. Il est même possible de trouver des fenêtres g telles que l'on ait, $\forall x \in L^2(\mathbb{R})$,

$$x = \sum_{m,n} \mathcal{V}_g x[m,n] g_{mn}$$

(le repère de Gabor est alors dit ajusté). C'est dans ce cadre que nous nous placerons ici. Etant donnée une suite bornée $\mathbf{m} = \{\mathbf{m}[m,n], m,n \in \mathbb{Z}\}$ le multiplicateur de Gabor associé est alors défini comme l'opérateur linéaire $\mathbb{M}_{\mathbf{m}}$ suivant :

$$\mathbb{M}_{\mathbf{m}} x = \sum_{m,n} \mathbf{m}[m,n] \mathcal{V}_g x[m,n] g_{mn} \tag{1}$$

$\mathbf{m}$ est appelé *fonction de transfert temps-fréquence*, ou *masque* du multiplicateur. Les propriétés d'approximation d'opérateurs linéaires par masques de Gabor ont été étudiées dans [2].

2.2 Estimation de masques de Gabor et mesures de dissimilarité

Nous nous intéressons ici au problème d'approximation d'un système par multiplicateur de Gabor, et de l'estimation de sa fonction de transfert $\mathbf{m}$ à partir d'un jeu de signaux d'entrée et de sortie. Pour simplifier, plaçons-nous dans la situation d'un signal d'entrée x_0 et d'un signal de sortie x_1, supposés être liés par une relation de la forme $x_1 = \mathbb{M}_{\mathbf{m}} x_0 + \epsilon_1$, où ϵ_1 représente un bruit additif, supposé blanc Gaussien. On cherche à estimer le masque $\mathbf{m}$ qui minimise

$$\Phi[\mathbf{m}] = \|x_1 - \mathbb{M}_{\mathbf{m}} x_0\|^2 + \lambda \|\mathbf{m} - 1\|^2, \tag{2}$$

où $\lambda \in \mathbb{R}^+$ est un paramètre de régularisation [1]. L'optimisation de Φ conduit au problème matriciel

$$G\mathbf{m} = U,$$

où on a posé

$$U(m,n) = \mathcal{V}_g x_1(m,n) \overline{\mathcal{V}_g x_0}(m,n) + \lambda$$

et où G est la matrice Hermitienne

$$G = \mathcal{D}_g x_0 \mathcal{K}_g \overline{\mathcal{D}_g x_0} + \lambda \mathcal{I},$$

$\mathcal{D}_g x_0$ étant la matrice diagonale $\mathcal{D}_g x_0 = \mathrm{diag}(\mathcal{V}_g x_0)$, et $\mathcal{K}_g$ le noyau reproduisant $(\mathcal{K}_g(m,n,m_0,n_0) = \langle g_{m,n}, g_{m_0,n_0} \rangle)$.

1. Ce choix, différent de celui fait dans [1], est motivé par le fait de s'intéresser à des signaux suffisamment proches, et donc de rechercher un multiplicateur proche de l'identité

Il s'agit d'un système linéaire de grande dimension, et on montre facilement que G est inversible et diagonale dominante pour des choix raisonnables de la fenêtre g. Nous avons vérifié numériquement qu'approximer G par sa diagonale (approximation similaire à celle développée dans [1]) fournit une précision tout à fait satisfaisante. Ceci conduit à la solution simple

$$\widehat{\mathbf{m}}[m,n] = \frac{\mathcal{V}_g x_1[m,n]\, \overline{\mathcal{V}_g x_0}[m,n] + \lambda}{|\mathcal{V}_g x_0[m,n]|^2 + \lambda} \tag{3}$$

Pour comparaison, on définit des quantités similaires à partir des transformées de Fourier des signaux x_0 et x_1 (notées $\hat{x}_0$ et $\hat{x}_1$), ainsi que leurs spectres lissés obtenus en marginalisant en temps les transformées de Gabor (on note $\tilde{x}[n] = \langle \mathcal{V}_g x[\cdot,n] \rangle_m$), par

$$\widehat{\mathbf{m}}^{(S)} = \frac{\hat{x}_1 \overline{\hat{x}_0} + \lambda}{|\hat{x}_0|^2 + \lambda}, \qquad \widehat{\mathbf{m}}^{(M)} = \frac{\tilde{x}_1 \overline{\tilde{x}_0} + \lambda}{|\tilde{x}_0|^2 + \lambda},$$

Pour comparer deux signaux x_i et x_j au travers des trois « masques » définis ci-dessus, nous utilisons les trois mesures de dissimilarité (symétrisées) suivantes. Pour chacun des trois choix, en notant $\mathbf{m}$ le masque correspondant, on notera $\mathbf{m}_{ij}$ la fonction de transfert associée à la transformation $x_i \to x_j$.

$$\left\{ \begin{array}{lcl} d_1(x_i,x_j) & = & \frac{1}{2}\left(\|\mathbf{m}_{ij} - 1\|_1 + \|\mathbf{m}_{ji} - 1\|_1\right) \\ d_2(x_i,x_j) & = & \frac{1}{2}\left(\|\mathbf{m}_{ij} - 1\|_{\mathrm{fro}} + \|\mathbf{m}_{ji} - 1\|_{\mathrm{fro}}\right) \\ d_3(x_i,x_j) & = & \frac{1}{2}\big|\|\mathbf{m}_{ij}\|_1 + \|\mathbf{m}_{ji}\|_1 \\ & & -\|\log|\mathbf{m}_{ij}|\|_1 - \|\log|\mathbf{m}_{ji}|\|_1 - 2\big| \end{array} \right.$$

où on a noté $\|\cdot\|_{\mathrm{fro}}$ la norme de Frobenius, $\|\cdot\|_1$ la norme ℓ^1; d_3 est en fait la version symétrisée de la divergence d'Itakura-Saito. Ces trois mesures sont symétrisées, car $\mathbf{m}_{ij} \neq \mathbf{m}_{ji}$.
La fonctionnelle (2) est définie avec un terme de régularisation imposant au masque d'être le plus proche possible du masque unité, masque dont tous les coefficients sont égaux à 1. Signalons que la mesure d_2 donne un poids plus fort aux grandes valeurs prises par le masque ce que ne font pas les mesures d_1 et d_3. Dans la mesure où ces grandes valeurs apparaissent en particulier dans les zones du plan temps-fréquence où seul un des deux signaux présente une énergie significative. Le comportement de ces mesures ne va influencer que la structure fine de notre classification.

3 Pré-traitement de notre méthode

3.1 Recalage temporel des signaux

Le masque de Gabor capture l'information différenciant deux signaux en divisant point par point leurs images temps-fréquence. Il est donc nécessaire que leurs énergies déployées dans le plan temps-fréquence se superposent au mieux.

D'un point de vue « pratique », un décalage temporel (ou fréquentiel) entre deux signaux a d'importantes conséquences, car il génère un masque prenant de très grandes valeurs et de très petites valeurs, et les deux signaux sont alors décrétés très différents, même s'ils sont très similaires. Le masque peut également devenir très oscillant dans cette situation, ce qui rend

l'interprétation difficile. Pour s'abstenir de cette difficulté, il est important de procéder à une étape de « recalage » des signaux, pour s'assurer de la pertinence des masques éstimés.

Dans ce qui suit, nous allons comparer des sons issus d'instruments de musique de même fréquence fondamentale, même intensité et même durée. L'objectif est de caractériser les aspects du timbre permettant de discriminer deux sons. Le recalage temporel est effectué suivant la méthode développée dans [4]. Les signaux ont été tronqués pour être de la même durée et les transformées de Gabor des signaux ont été normalisées.

3.2 Débruitage temps-fréquence

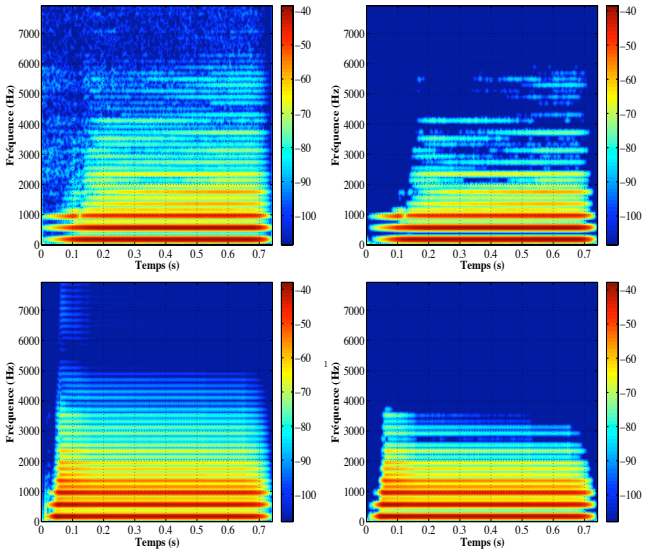

FIG. 1 – *En haut: à gauche une clarinette de synthèse, et à droite sa version débruitée. En bas: à gauche une clarinette réele, et à droite sa version débruitée.*

Nous avons vu que la transformation de Gabor fournit une représentation pertinente des classes de signaux sur lesquels nous nous focalisons ici. Ceci étant, il peut arriver que la totalité de l'information contenue dans ces représentations ne soit pas pertinente pour la classification, voire même qu'elle soit préjudiciable lorsqu'il s'agit d'information variable à l'intérieur d'une même classe de signaux. On pense notamment aux composantes « bruitées » des sons, que l'on qualifie aussi d'aléatoires. Il est alors utile de simplifier la représentation avant catégorisation. Une approche possible, que nous suivons ici, utilise l'approximation temps-fréquence en ne retenant que les coefficients les plus pertinents. Dans cette étude, nous utilisons la méthode du *basis pursuit denoising*, aussi appelée régression LASSO, mise en oeuvre dans le paquet LTFAT [6] via un algorithme de Landweber (seuillage doux itératif, voir par exemple [5]). Compte tenu de la décroissance fréquentielle classique des sons étudiés, nous utilisons ici un seuil variant avec la fréquence, afin que le seuillage ne se traduise pas par un simple filtrage passe-bas.

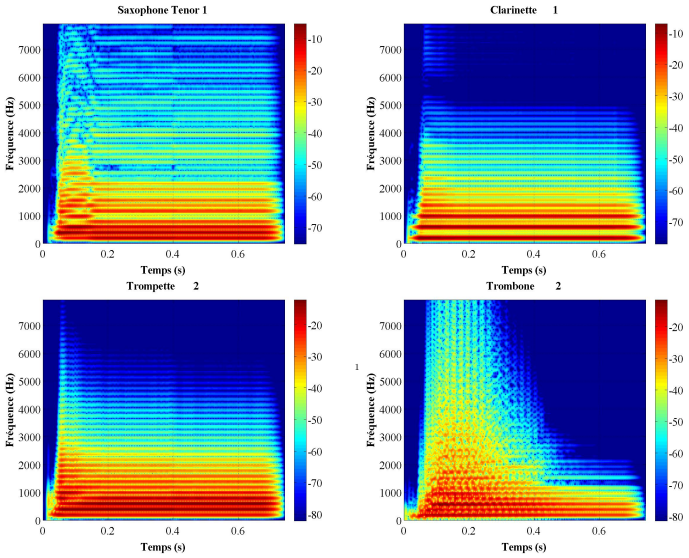

FIG. 2 – *En haut: à gauche un saxophone tenor, et à droite une clarinette. En bas: à gauche une trompette, et à droite un trombone.*

La figure 1 illustre le débruitage obtenu, en haut à partir d'une clarinette de synthèse issue du modéle de synthèse décrit dans [3] et en bas, d'une clarinette réelle. Les transformées de Gabor originales sont présentées sur la gauche de la figure, et les transformées de Gabor débruitées sont présentées sur la droite de la figure. On peut observer que les parties stochastiques des signaux sont largement atténuées et que l'on voit apparaître plus nettement les harmoniques impaires, ainsi que les formants et anti-formants.

4 Application à la catégorisation d'instruments de musique

Les mesures de dissimilarité ci-dessus ont été calculées sur une banque de données de sons, puis exploitées dans un algorithme de classification hiérarchique (basé sur la méthode de Ward) développé par P. Kleiweg [8].

Les tests ont été effectués sur deux banques de sons. La première banque, assez homogène, est composée de différentes réalisations d'une note (LA2) jouée par 4 intruments (saxophone tenor, clarinette, trombone et trompette, 10 réalisations par instrument) issus du modèle physique de synthèse sonore de [3]. Ces réalisations ne sont pas identiques, et présentent un degré contrôlé de variabilité. Dans la deuxième banque, quatre sons réels ont été ajoutés à la première banque : deux clarinettes (notée `/cla9/` et `/cla10/`) et deux trompettes (`/trp8/` et `/trp10/`). On augmente ainsi progressivement la variabilité entre les sons de la banque, ce qui permet de tester la robustesse de notre méthode.

Plusieurs mesures de dissimilarité ont été testées, et nous donnons ici un exemple de résultat en figures 3, 4 et 5, où sont représentés les dendogrammes obtenus avec les sons de la

banque 2 en utilisant la distance d_3. La classification est obtenue à partir des masques $\hat{\mathbf{m}}^{(M)}$ sur la figure 3, $\hat{\mathbf{m}}$ sur la figure 4, puis $\hat{\mathbf{m}}$ obtenu à partir des transformées de Gabor débruitées sur la figure 5. Tout d'abord, nous avons observé que le choix de la mesure de dissimilarité n'influence que peu la forme globale de l'arbre obtenu, c'est-à-dire la position des classes les unes par rapport aux autres.

Pour la première banque (faible variabilité intra-classe), les deux approches sont équivalentes et sont capables de différencier clairement les classes d'instruments. L'approche stationnaire (spectres lissés) différencie clairement les clarinettes, ce qui est naturel compte tenu de la forme particulière de leur spectre (dominance des harmoniques impaires). L'approche non-stationnaire (transformées de Gabor) différencie quant à elle les trombones. Là encore, ceci s'explique simplement par inspection des images temps-fréquence de la figure 2, sur lesquelles l'on voit que les sons de trombone présentent un caractère instationnaire très marqué. Notons que seuls les masques « temps-fréquence » (figures 4 et 5) sont à même de différencier clairement les réalisations des sons réels des réalisations des sons de synthèse. C'est en cela que les méthode temps-fréquence se démarquent des méthodes fréquentielles. Nous avons observé que les trompettes réelles sont mal placées dans les figures 3 et 5, ce qui s'explique par le fait leur spectre présente des différences significatives avec les trompettes de synthèse, et que celles-ci ressortent lors de l'étape de débruitage.

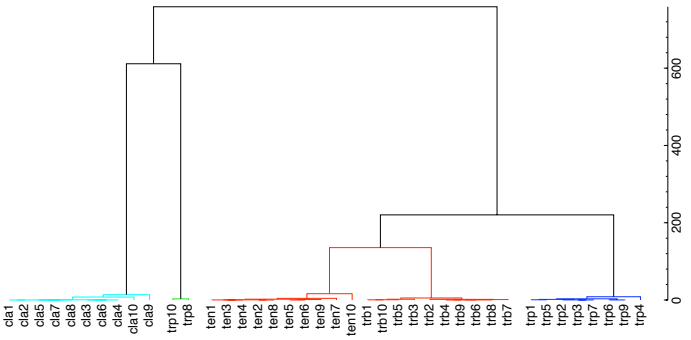

FIG. 3 – *Arbre obtenu en utilisant d_3 et $\hat{\mathbf{m}}^{(M)}$ pour la banque 2.*

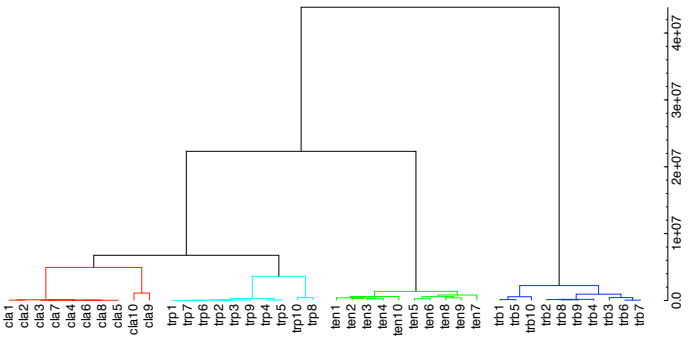

FIG. 4 – *Arbre obtenu en utilisant d_3 et $\hat{\mathbf{m}}$ pour la banque 2.*

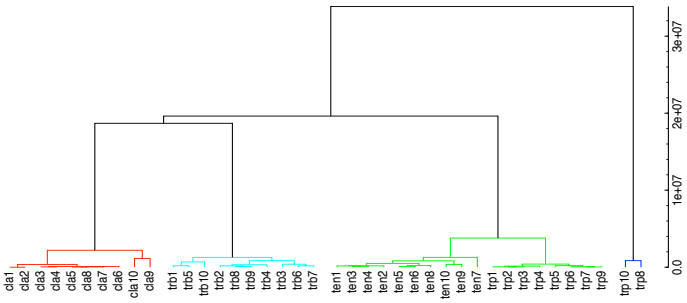

FIG. 5 – *Arbre obtenu en utilisant d_3 et $\hat{\mathbf{m}}$ calculé sur des représentations temps-fréquence débruitées pour la banque 2.*

5 Conclusions, perspectives

Les résultats obtenus montrent l'intérêt de l'utilisation de masques de Gabor pour la comparaison de sons non-stationnaires, car ils tiennent compte du déploiement energétique temps-fréquence des signaux. Nous avons également montré que l'introduction d'une procédure de simplification des signaux (appelée ici débruitage) permet également d'améliorer la catégorisation, même si elle soulève encore de nombreuses questions, parmi lesquelles la détermination des seuils.

Des prolongements de ce programme concernent l'estimation de multiplicateurs de Gabor « moyens » entre classes de sons, la synthèse de sons « prototypes » pour les catégories, l'approfondissement de l'étude de la variabilité à l'intérieur d'une classe, toujours au moyen de multiplicateurs, et l'extension à des multiplicateurs généralisés (voir [2]).

Ce travail est partiellement financé par le programme PEPS-ST2I du CNRS, projet MTF&Sons:
`http://www.latp.univ-mrs.fr/MTFetSons/`

Références

[1] P. Depalle, R. Kronland-Martinet et B. Torrésani. *Wavelet XII*, *SPIE annual Symposium*, San Diego (2007).

[2] M. Dörfler et B. Torrésani. *Sampling Theory and Applications* (SAMPTA'07), Thessaloniki, Juin 2007.

[3] P. Guillemain, J.Kergomard, T. Voinier. *J. Acoust. Soc. Am.* **118**:1 (2005), pp. 483-494.

[4] J. P. Bello, C. Duxbury, M. Davies et M. Sandler, *IEEE Signal Processing Letters*, vol. 11, no. 6 (June 2004), pp. 553-556.

[5] I. Daubechies, M. Defrise et C. De Mol. *Communications in Pure and Applied Mathematics*, 57:1413-1457, 2004.

[6] P. Soendergaard, Linear Time-Frequency Analysis Toolbox: `http://sourceforge.net/projects/ltfat/`

[7] H. G. Feichtinger. et T. Strohmer. Gabor Analysis and Algorithms: Theory and Applications. ISBN: 0817639594, *Birkhauser Boston*, 1997.

[8] P. Kleiweg, Data Clustering software: `http://www.let.rug.nl/~kleiweg/indexs.html`