Reconnaissance des instruments dans la musique polyphonique par décomposition NMF et classification SVM

Instrument recognition in polyphonic music based on NMF decomposition and SVM classification

---

Alexey Ozerov
Slim Essid
Maurice Charbit

2009D014

Juillet 2009

Département Traitement du Signal et des Images
Groupe AAO : Audio, Acoustique et Ondes

Reconnaissance des instruments dans la musique polyphonique par décomposition NMF et classification SVM

Instrument recognition in polyphonic music based on NMF decomposition and SVM classification

Alexey Ozerov, Slim Essid et Maurice Charbit *

Institut Télécom; Télécom ParisTech; CNRS LTCI
37-39, rue Dareau, 75014 Paris, France.
{alexey.ozerov,slim.essid,maurice.charbit}@telecom-paristech.fr

*Ce travail a été financé par le projet ANR SARAH (StAndardisation du Remastering Audio Haute définition).

Résumé

Dans ce rapport nous présentons une nouvelle approche pour la reconnaissance des instruments dans la musique polyphonique multi-instrumentale. Ce travail est effectué dans le cadre du projet ANR SARAH ("StAndardisation du Remastering Audio Haute définition"), dont le but principal est de développer des méthodes de séparation de sources, qui soient performantes et applicables à une vaste classe d'enregistrements musicaux. Ainsi, dans le contexte de ce projet, le rôle du système de reconnaissance des instruments est d'identifier localement l'instrumentation d'une œuvre musical, afin de faciliter le choix de connaissances *a priori* (exprimées par exemple par des modèles probabilistes de sources) utilisées pour la séparation. L'approche, adoptée pour effectuer la reconnaissance des instruments dans la musique multi-instrumentale, est basée sur la suite d'opérations suivantes :

1. décomposition du signal de musique en composantes spectrales élémentaires à l'aide de la NMF (*Non-negative Matrix Factorization*),
2. estimation MAP (maximum *a posteriori*) des composantes "instrumentales" du mélange (filtrage de Wiener adaptatif),
3. extraction, à partir de ces composantes, d'attributs caractéristiques. Toutefois afin de rendre la caractérisation plus robuste, nous introduisons un mécanisme original de pondération des attributs qui vise à écarter les attributs peu vraisemblables, c'est-à-dire extraits de composantes élémentaires "mal séparées".
4. enfin classification à l'aide de SVM bi-classes (Support Vector Machine).

Plusieurs configurations du système de reconnaissance sont évaluées sur une base d'enregistrements de musique de jazz.

Mots-clés: Reconnaissance des instruments, musique polyphonique, factorisation en matrices à coefficients positifs, séparateurs à vaste marge.

Abstract

In this report we present a new approach for instrument recognition in polyphonic multi-instrumental music. This work was done in the context of French ANR project SARAH ("StAndardisation du Remastering Audio Haute définition"). The main goal of this project is to develop audio source separation methods that would be efficient and applicable to a wide variety of music recordings. Thus, in the context of the project, the role of instrument recognition system is to locally identify the instruments of a music piece, in order to simplify the choice of *a priori* knowledge (e.g., expressed by probabilistic models of sources) used for separation. Our original approach for instrument recognition in polyphonic multi-instrumental music is based on the following steps:

1. signal decomposition into spectral components using Non-negative Matrix Factorization (NMF),
2. maximum *a posteriori* (MAP) estimation of "instrumental" components of the mix (adaptive Wiener filtering),
3. extraction of features from estimated (separated) components. However, in order to increase robustness of this feature extraction step, we introduce an original feature weighing mechanism trying to exclude unlikely features, i.e., those extracted from "badly separated" elementary components.
4. Support Vector Machine (SVM) classification.

Several configurations of the proposed instrument recognition system are evaluated on a jazz music database.

Keywords: Instrument recognition, polyphonic music, non-negative matrix factorization, support vector machine.

Table des matières

1 Introduction **4**

2 Système de reconnaissance des instruments **5**
2.1 Présentation du système 5
2.2 Non-negative Matrix Factorization 6
2.3 Filtrage de Wiener 9
2.4 Extraction des attributs 9
2.5 Poids de la fiabilité et intégration précoce 10
2.6 Normalisation globale des attributs 11
2.7 Sélection des attributs 12
2.8 Apprentissage des SVMs bi-classes 12
2.9 Calcul des probabilités bi-classes 12
2.10 Calcul des probabilités multi-classes 12
2.11 Fusion des probabilités des composantes 12
2.12 Lissage temporelle des probabilités 12

3 Expérimentations **12**
3.1 Données expérimentales 12
3.1.1 Base des solos (apprentissage) 13
3.1.2 Base des mélanges (test et développement) 13
3.2 Tâches de la reconnaissance des instruments 13
3.3 Mesures de la performance 13
3.4 Résultats 14

4 Discussion **16**

5 Conclusion et pistes d'amélioration **16**

A Liste des attributs **17**

Liste des figures

1 Module de test du système de reconnaissance des instruments. 6
2 Module d'apprentissage du système de reconnaissance des instruments. 7
3 Matrice des confusions du meilleur système obtenu. 15

Liste des tables

1 Liste des instruments et des codes associés. 13
2 Performances moyennes des différents systèmes. 15
3 Performances détaillés du meilleur système obtenu. 15
4 Liste des attributs utilisés. Au total nous obtenons 284 attributs. 17

1 Introduction

Etant donné un enregistrement de musique polyphonique, le problème de reconnaissance des instruments consiste à identifier à chaque instant temporel l'ensemble des instruments présents dans l'enregistrement.

De nombreux travaux récents abordent le problème dans le cas de musique mono-instrumentale, c'est-à-dire quand il y a au plus un seul instrument présent dans l'enregistrement à un instant donné [1, 2, 3, 4, 5, 6, 7].

Cependant, à notre connaissance, il y a peu de travaux abordant le problème de reconnaissance des instruments dans la musique multi-instrumentale. La principale difficulté du cas multi-instrumental est précisément liée au fait que les instruments se recouvrent entre eux. Pour résoudre ce problème, Eggink et Brown [8] utilisent un modèle de mélanges de Gaussiennes (GMM) dans le contexte de la théorie des données manquantes [8]. Ils appliquent cette méthode pour la reconnaissance des duos, en supposant que les fréquences fondamentales sont connues. Cont *et al.* [9] présentent une approche originale pour la reconnaissance des instruments et l'identification du pitch conjointes. Cette approche, basée sur une NMF parcimonieuse, est appliquée pour la reconnaissance des duos du piano et du violon. Kitahara *et al.* [10] introduisent un système de reconnaissance des instruments basé sur une étape d'extraction des attributs à partir des notes identifiées auparavant et une étape de classification. De plus, les auteurs proposent d'estimer des poids (facteurs de pondération) qui représentent le degré de recouvrement des notes, ainsi ces poids représentent également la fiabilité des attributs extraites. Prendre en compte ces poids pendant la phase de classification mène à une augmentation significative des performances. L'approche proposée par Leveau *et al.* [11] consiste en une décomposition du signal par un algorithme de type *matching pursuit* basé sur un dictionnaire constitué d'atomes spécifiques aux instruments à reconnaître. Cette décomposition est ensuite utilisée pour calculer des saillances (similaire à un niveau de contraste) pour des sous-ensembles donnés d'instruments. Martins *et al.* [12] proposent une approche basée sur la séparation de sources utilisée comme un pré-traitement. Les sons provenant des différents instruments sont d'abord séparés, ensuite la classification est effectuée sur des signaux séparés en utilisant des modèles de timbres des instruments. Cette méthode est évaluée sur des mélanges des notes isolées. Rafii *et al.* [13] proposent de décomposer d'abord les spectrogrammes des instruments musicaux en utilisant la décomposition NMF, d'extraire ensuite des attributs à partir des coefficients de la NMF, et d'effectuer enfin la classification à l'aide des SVMs portant sur ces attributs. Essid [14] introduit une approche conceptuellement différente. Il propose de considérer un ensemble possible d'instruments comme un classe à reconnaître, sans essayer de pré-séparer (ou pré-décomposer) l'enregistrement traité. Son approche est basée sur l'extraction de nombreux attributs, suivies par une phase de sélection des attributs et une phase de classification par SVM.

L'approche, que nous avons adoptée pour effectuer la reconnaissance des instruments dans la musique multi-instrumentale, partage de nombreux points avec les travaux cités précédemment. Elle est basée sur la suite d'opérations suivantes :

- Décomposition du signal de musique en composantes spectrales élémentaires à l'aide de la NMF (*Non-negative Matrix Factorization*),
- Estimation MAP (maximum a posteriori) des composantes "instrumentales" du mélange (filtrage de Wiener adaptatif),
- Extraction, à partir de ces composantes, d'attributs caractéristiques. Toutefois afin de rendre la caractérisation plus robuste, nous introduisons un mécanisme original de pondération des attributs qui vise à écarter les attributs peu vraisemblables, c'est-à-dire extraits de composantes élémentaires "mal séparées".
- Enfin classification à l'aide de SVM (Support Vector Machine) bi-classes.

Le système est évalué sur un ensemble-test de signaux musicaux multi-instrumentaux. L'apprentissage est effectué sur un ensemble d'enregistrements solos des instruments qu'on souhaite reconnaître. Des décompositions NMF sont également effectuées sur les données d'apprentissage avant l'extraction des attributs, afin d'essayer de produire la même distorsion que sur les ensembles de test. D'autre part, nous avons également testé l'apprentissage à partir des solos non-décomposés par NMF.

Notons les points communs ainsi que les différences entre cette approche originale et celles envisagées dans les travaux existants. Comme dans [13] nous utilisons une décomposition NMF. Cependant contrairement à [13] nous extrayons des attributs, non pas à partir des paramètres de la NMF, mais à partir des composantes *ré-synthétisées* obtenues à partir des composantes NMF élémentaires. Toutefois nous avons aussi utilisé la première approche pour extraire des attributs complémentaires.

Martins *et al.* [12] utilisent la séparation de sources comme pré-traitement. A la différence, nous séparons des composantes NMF élémentaires qui ne sont pas censées représenter séparément les sources. Elles peuvent aussi représenter des "bouts de sources".

Comme dans [10] nous utilisons des poids de la fiabilité. Cependant, notre mécanisme d'estimation des poids basé sur la décomposition NMF est différent et ne se repose pas sur une étape de détection des notes.

2 Système de reconnaissance des instruments

2.1 Présentation du système

Le système de reconnaissance des instruments, présenté dans ce rapport, comporte deux modules principaux : un module de test représenté figure 1 et un module d'apprentissage représenté figure 2.

Le module de test se repose sur des étapes suivantes :

1. Décomposition NMF de l'enregistrement traité.

2. Ré-synthèse des composantes NMF élémentaires en calculant les gains à l'aide du filtrage de Wiener adaptatif. Dans un premier temps, on suppose qu'une composante NMF élémentaire fait partie d'une seule source musicale. En général, cette hypothèse n'est pas vérifiée strictement.

3. Estimation des poids de la fiabilité au cours du temps pour chaque composante NMF élémentaire. A chaque instant, le poids de la fiabilité représente le taux de confiance qu'on puisse accorder aux attributs extraits à partir de la composante NMF élémentaire. Ces poids de la fiabilité vont être pris en compte aux niveaux supérieurs du traitement (l'intégration temporelles précoce et la décision finale), enfin de diminuer l'influence négative de la séparation "non-parfaite" des composantes NMF élémentaires.

4. Extraction des attributs à partir de chaque composante NMF élémentaire.

5. Intégration précoce des attributs [7], qui consiste à calculer des moyennes des attributs sur plusieurs trames successives. En calculant ces moyennes nous utilisons les poids de la fiabilité, afin de privilégier des valeurs des attributs extraits à partir des composantes NMF élémentaires "bien séparées". Les valeurs moyennes des poids de la fiabilité sert comme des "nouveaux poids de la fiabilité" pour des attributs intégrés.

6. Sélection des trames pertinentes. Cette étape consiste à éliminer tous simplement les trames ayant un faible poids de la fiabilité, en comparant ce poids avec un seuil. Pour garder la structure temporelle ces trames ne sont pas éliminées complètement, mais marquées comme NFB (Non-FiaBle), par opposition des trames FB (FiaBle). Les valeurs des attributs des trames NFB ne sont pas gardées.

7. Normalisation globale des attributs par la soustraction de la moyenne globale et la division par l'écart type global (la moyenne et l'écart type sont calculés pendant la phase d'apprentissage (Fig. 2)).

8. Filtrage à partir des attributs sélectionnés pendant la phase d'apprentissage (Fig. 2).

9. Calcul des probabilités bi-classe, en utilisant les SVMs bi-classes appris pendant la phase d'apprentissage (Fig. 2), et la transformation des probabilités bi-class en probabilités multi-class.

10. Fusion des probabilités des composantes. Cette dernière étape consiste à estimer des probabilités de présence de chaque instrument dans le mélange, à partir des probabilités de présence de cet instrument dans des composantes NMF élémentaires. Pour évaluer les performances du système, nous effectuons la détection des instruments.

Le module d'apprentissage (Fig. 2) est constitué des étapes similaires.

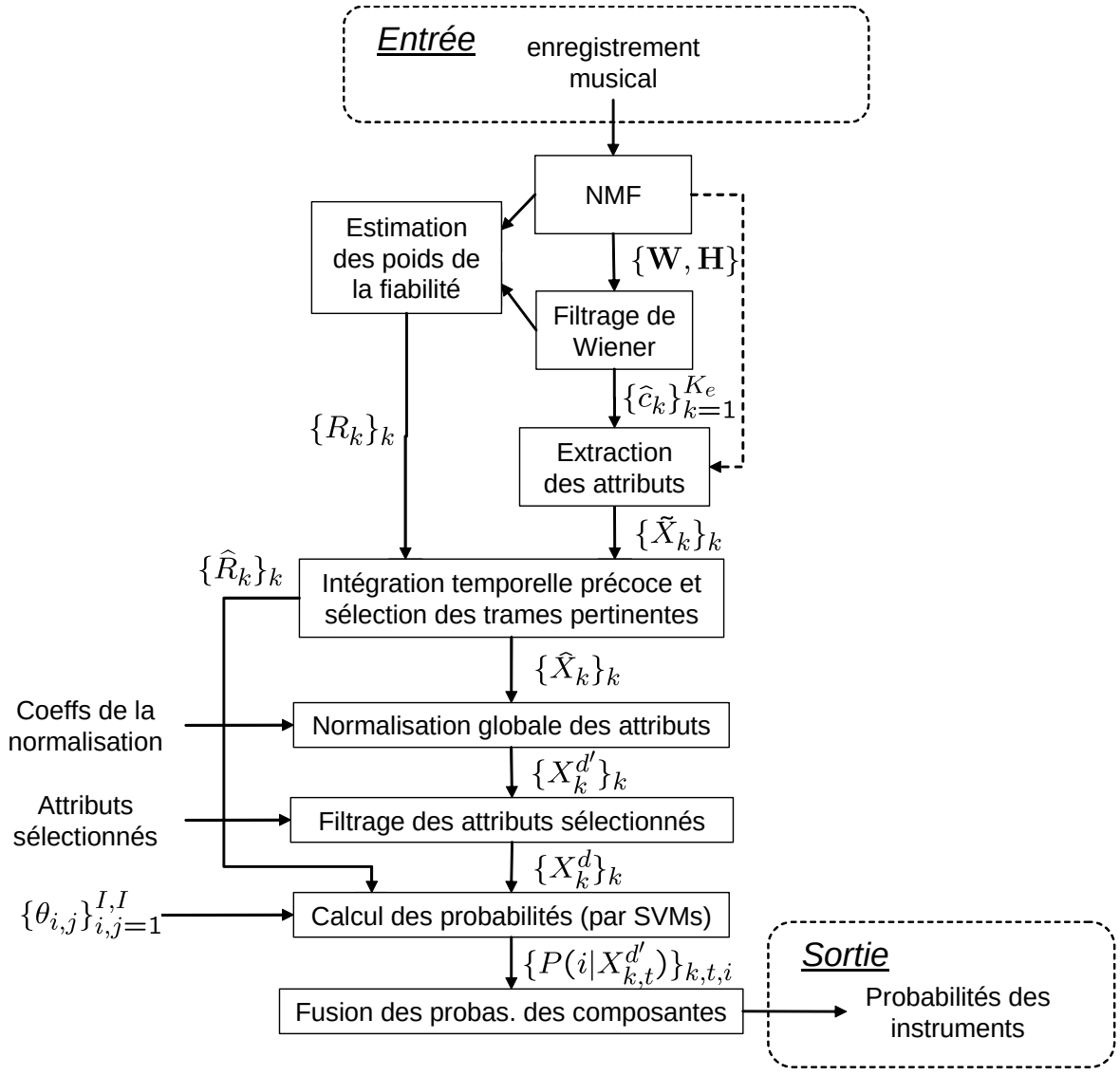

Figure 1: Module de test du système de reconnaissance des instruments.

Dans des sections qui suivent nous expliquons plus en détails certaines étapes de notre approche.

2.2 Non-negative Matrix Factorization

Les signaux considérés dans la suite sont réels. Ils sont découpés en blocs de longueur $2(F-1)$ avec un recouvrement de 50%. Typiquement F est égal à 129, 257 etc Du fait de la symétrie hermitienne, on peut limiter l'indice fréquentiel f entre 1 et F. Dans la suite n désigne l'indice temporel du bloc considéré.

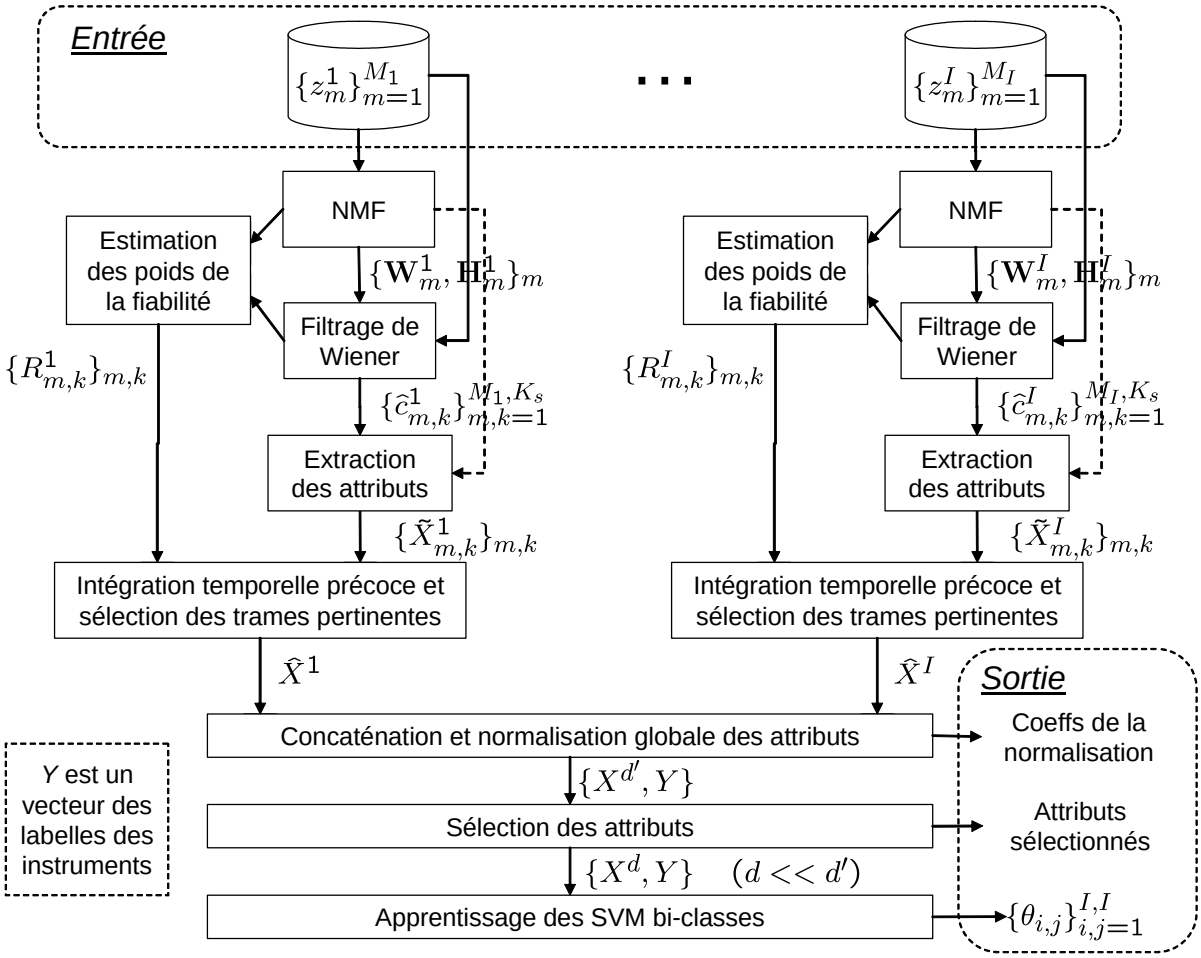

Figure 2: Module d'apprentissage du système de reconnaissance des instruments.

On note $\mathbf{Z} = \{z_{f,n}\}_{f,n=1}^{F,N}$ l'ensemble des Transformées de Fourier à Court Terme (TFCT) du signal considéré et $\mathbf{V} = \{v_{f,n}\}_{f,n=1}^{F,N}$ la matrice de dimensions $F \times N$ dont la colonne $\mathbf{v}_n = |\mathbf{z}_n|^2$ représente le spectre en puissance de la fenêtre temporelle d'indice n. Noter que les éléments de $\mathbf{V}$ sont non négatifs.

La décomposition NMF (Non-negative Matrix Factorization) consiste à approcher au mieux, en minimisant une fonction de "divergence" donnée, la matrice $\mathbf{V}$ par un produit de deux matrices $\mathbf{W}$ et $\mathbf{H}$ de dimensions respectives $F \times K$ et $K \times N$ où K est un entier donné. Tous les éléments de W et H sont supposés positifs, de là la désignation de la décomposition. Dans notre étude, l'entier K est typiquement compris entre 1 et 7 et son rôle est analysé de façon expérimentale. Il s'ensuit que le nombre d'observations est égal à FN alors que le nombre de degrés de liberté est égal à $(FK + KN)$. En pratique, si N est grand, $FN \gg (FK + KN)$. Le problème est alors sur-déterminé, dans le sens où on a plus d'équations que d'inconnues. Toutefois, même en l'absence d'écart i.e. si $V = WH$, rien ne garantit que la décomposition soit unique. En pratique, on a plutôt

$$\mathbf{V} \approx \mathbf{WH} \tag{1}$$

Dans ce cas pour déterminer l'approximation NMF (1) on minimise une certaine fonction de divergence $D(\mathbf{V}|\mathbf{WH})$, ce qui s'écrit

$$(\mathbf{W}, \mathbf{H}) = \arg \min_{(\mathbf{W}', \mathbf{H}')} D(\mathbf{V}|\mathbf{W}'\mathbf{H}'). \tag{2}$$

Diverses distances ou divergences ont été étudiées dans la littérature (voir pour une étude détaillée [15]). La distance Euclidienne

$$D_{EUC}(\mathbf{V}|\mathbf{WH}) = \frac{1}{2} \sum_{n=1}^{N} \sum_{f=1}^{F} \left(v_{f,n} - \sum_{k=1}^{K} h_{n,k} w_{f,k} \right)^2 \tag{3}$$

et la divergence de Kullback-Leibler

$$D_{KL}(\mathbf{V}|\mathbf{WH}) = \sum_{n=1}^{N} \sum_{f=1}^{F} \left(v_{f,n} \log \frac{v_{f,n}}{\sum_{k=1}^{K} h_{n,k} w_{f,k}} - v_{f,n} + \sum_{k=1}^{K} h_{n,k} w_{f,k} \right) \tag{4}$$

sont deux fonctions de coût couramment utilisées. La divergence de Itakura-Saito

$$D_{IS}(\mathbf{V}|\mathbf{WH}) = \sum_{n=1}^{N} \sum_{f=1}^{F} \left(\frac{v_{f,n}}{\sum_{k=1}^{K} h_{n,k} w_{f,k}} - \log \frac{v_{f,n}}{\sum_{k=1}^{K} h_{n,k} w_{f,k}} - 1 \right) \tag{5}$$

a été introduite récemment comme une fonction de coût pour la NMF [15]. Cette fonction de coût a deux propriétés qui la rendent particulièrement attrayante. Premièrement, la divergence de IS est invariante par rapport à une mise à échelle (c'est-à-dire multiplication du signal par un gain) [15]. Deuxièmement, selon [15] le critère d'estimation de $\mathbf{W}$ et $\mathbf{H}$ avec $D_{IS}(\mathbf{V}|\mathbf{WH})$ est équivalent à l'estimateur du Maximum de Vraisemblance (MV) associée aux N observations issues du modèle probabiliste:

$$\mathbf{z}_n = \sum_{k=1}^{K} \sqrt{h_{n,k}} \mathbf{x}_{k,n} \tag{6}$$

où $\mathbf{z}_n = [z_{1,n}, z_{2,n}, \ldots, z_{F,n}]^t$ et où on suppose que les K composantes vectorielles $\mathbf{x}_{k,n}$ de la somme sont des vecteurs aléatoires de dimension F, indépendants et distribués suivant une loi Gaussienne complexe *circulaire*, centrée et blanche, ce qui s'écrit

$$\mathbf{x}_{k,n} \sim \mathcal{N}_c \left(\bar{0}, \mathrm{diag}(\mathbf{w}_k) \right). \tag{7}$$

Noter que $\mathbf{w}_k$ ne dépend pas de n. Il s'ensuit que la log-vraisemblance de $\mathbf{Z}$ s'écrit

$$\log p_{\mathbf{Z}}(\mathbf{z}; \mathbf{W}, \mathbf{H}) = -NF \log(\pi) - \sum_{n=1}^{N} \sum_{f=1}^{F} \left(\log \sum_{k=1}^{K} h_{k,n} w_{f,k} + \frac{|z_{f,n}|^2}{\sum_{k=1}^{K} h_{k,n} w_{f,k}} \right) \tag{8}$$

On vérifie aisément que la maximisation de (8) par rapport $\mathbf{V}$ et $\mathbf{H}$ est équivalente à la minimisation de (5) par rapport $\mathbf{V}$ et $\mathbf{H}$.

Dans notre étude, nous avons principalement considéré la divergence de IS et la divergence de KL. Pour ces deux divergences les formules itératives de mise à jour de $\mathbf{W}$ et $\mathbf{H}$, permettant d'optimiser (2) avec $D_{IS}(\mathbf{V}|\mathbf{WH})$ ou avec $D_{KL}(\mathbf{V}|\mathbf{WH})$, peuvent être trouvées dans [15].

Dans la suite on pose $\mathbf{c}_{kn} = \sqrt{h_{nk}}\mathbf{x}_{kn}$ et donc $\mathbf{z}_n = \sum_{k=1}^{K} \mathbf{c}_{kn}$.

2.3 Filtrage de Wiener

On montre aisément que la distribution *a posteriori* de $\mathbf{c}_{kn}$ sachant $\mathbf{z}_n$ a pour expression:

$$\mathbf{c}_{k,n}|\mathbf{z}_n \sim \mathcal{N}_c\left(\hat{\mathbf{c}}_{k,n}, \Sigma_{\mathbf{c}_{k,n}}^{post}\right) \tag{9}$$

où l'espérance conditionnelle s'écrit

$$\hat{\mathbf{c}}_{k,n} = \mathbb{E}\left[\mathbf{c}_{k,n}|\mathbf{z}_n; \mathbf{W}, \mathbf{H}\right] = \begin{bmatrix}\hat{c}_{k,n,1} & \cdots & \hat{c}_{k,n,f} & \cdots & \hat{c}_{k,n,F}\end{bmatrix}^t \tag{10}$$

avec

$$\hat{c}_{k,n,f} = \frac{h_{k,n}w_{f,k}}{\mathbf{w}_f\mathbf{h}_n} z_{f,n} \quad \text{et} \quad \mathbf{w}_f\mathbf{h}_n = \sum_{k=1}^{K} w_{f,k}h_{k,n}$$

et où la covariance *a posteriori* vérifie:

$$\begin{aligned}
\Sigma_{\mathbf{c}_{k,n}}^{post} &= \mathbb{E}\left[(\mathbf{c}_{k,n} - \hat{\mathbf{c}}_{k,n})(\mathbf{c}_{k,n} - \hat{\mathbf{c}}_{k,n})^H|\mathbf{z}_n; \mathbf{W}, \mathbf{H}\right] \\
&= \mathbb{E}\left[(\mathbf{c}_{k,n} - \hat{\mathbf{c}}_{k,n})(\mathbf{c}_{k,n} - \hat{\mathbf{c}}_{k,n})^H|\mathbf{W}, \mathbf{H}\right] \\
&= \text{diag}\left\{\left[\left(1 - \frac{h_{kn}w_{f,k}}{\sum_{k'=1}^{K} h_{k',n}w_{f,k'}}\right) h_{k,n}w_{f,k}\right]_{f=1:F}\right\} \qquad (11) \\
&= \text{diag}\left\{\left[\frac{\mathbf{w}_f^{-k}\mathbf{h}_n^{-k}}{\mathbf{w}_f\mathbf{h}_n} h_{k,n}w_{f,k}\right]_{f=1:F}\right\} \qquad (12)
\end{aligned}$$

où $\mathbf{w}_f^{-k}\mathbf{h}_n^{-k} = \sum_{l=1\neq k}^{K} w_{f,l}h_{l,n}$.

Rappelons que, par définition, l'espérance conditionnelle $\hat{\mathbf{c}}_{k,n}$ est, parmi toutes les fonctions de $\mathbf{z}_{k,n}$, celle qui minimise l'erreur quadratique avec $\mathbf{c}_{k,n}$. Du fait du caractère gaussien, cette espérance conditionnelle est une fonction linéaire de $\mathbf{z}_{k,n}$. Dans ce contexte, $\hat{\mathbf{c}}_{k,n}$ est désigné, dans la littérature, sous le terme de filtrage de Wiener.

2.4 Extraction des attributs

La caractérisation des instruments est faite à partir d'attributs présentés et étudiés de façon détaillée dans [14]. La liste complète des attributs retenus dans notre étude est présentée dans l'annexe A. Ces attributs sont calculés à partir du signal *obtenu dans le domaine temporel* à partir de chaque composante NMF de vecteurs spectraux $\{\mathbf{c}_{k,n}\}_{n=1:N}$.

Il faut noter que certains attributs sont calculés en utilisant une fenêtre d'analyse courte (32 ms avec 50 % de recouvrement) et d'autres en utilisant une fenêtre d'analyse longue (960 ms avec 50 % de recouvrement). Ainsi, à partir de chaque composante NMF élémentaire, nous obtenons un ensemble de suites de vecteurs d'attributs noté $\tilde{X}_k$ (voir Fig. 1).

2.5 Poids de la fiabilité et intégration précoce

Lors de l'estimation des attributs de la composante k, il est souhaitable de prendre en compte la distribution de $\{\mathbf{c}_{k,n}\}_{n=1:N}$ (et non pas seulement l'estimation $\{\hat{\mathbf{c}}_{k,n}\}_{n=1:N}$), en particulier la suite des covariances $\{\Sigma^{post}_{\mathbf{c}_{k,n}}\}_{n=1:N}$.

Une façon d'aborder le problème est de considérer un modèle simplifié d'extraction des attributs. Celui-ci s'appuie sur le fait que beaucoup d'attributs sont calculés à partir des log-spectres du signal. Ce qui s'écrit:

$$\mathbf{a}_{k,n} = Q \cdot \log |\mathbf{c}_{k,n}|^2 \tag{13}$$

où la fonction $\log|\cdot|^2$ est appliquée au vecteur $\mathbf{c}_{k,n}$ *élément par élément*, où Q est une transformation orthogonale, et où $\mathbf{a}_{k,n}$ est le vecteur d'attributs correspondant. Noter que dans l'équation (13) nous ne revenons pas au signal temporel.

Déterminer la loi de $\mathbf{a}_{k,n}$ conditionnellement à $\mathbf{z_n}$ est un calcul complexe même si on sait que la loi de $\mathbf{c}_{k,n}$ conditionnellement à $\mathbf{z_n}$ est un vecteur gaussien. Pour simplifier le problème nous effectuons une approximation de la fonction log par un développement de Taylor au premier ordre au voisinage de $\hat{c}_{k,n,f}$, ce qui s'écrit:

$$\log |c_{k,n,f}|^2 \approx \log |\hat{c}_{k,n,f}|^2 + \mu(\hat{c}_{k,n,f})\,(c_{k,n,f} - \hat{c}_{k,n,f}), \quad \text{où} \quad \mu(c) = \partial_c \log cc^* = \frac{1}{c} \tag{14}$$

De cette approximation et du fait que la transformation Q est supposée orthogonale, on déduit aisément que *conditionnellement à* $\mathbf{z_n}$ on a:

$$\mathbf{a}_{k,n}|\mathbf{z}_n \sim \mathcal{N}_c\left(\hat{\mathbf{a}}_{k,n}, \Sigma^a_{k,n}\right) \tag{15}$$

où nous avons posé

$$\hat{\mathbf{a}}_{k,n} = Q \cdot \log |\hat{\mathbf{c}}_{k,n}|^2 \tag{16}$$

expression qui est à rapprocher de l'expression (13) et

$$\Sigma^a_{k,n} = Q\,\mathrm{diag}\left\{\left[\sigma^a_{k,f,n}\right]_{f=1:F}\right\} Q^H \quad \text{où} \quad \sigma^a_{k,f,n} = \frac{\mathbf{w}_f^{-k}\mathbf{h}_n^{-k}}{w_{f,k}h_{k,n}} \cdot \frac{w_{f,k}h_{k,n} + \mathbf{w}_f^{-k}\mathbf{h}_n^{-k}}{|z_{f,n}|^2} \tag{17}$$

Il est utile de rappeler que la moyenne de la loi conditionnelle est, par définition, l'espérance conditionnelle. Par conséquent l'approximation qui apparaît dans (15) signifie que l'on "confond" le log de l'espérance conditionnelle $\log|\mathbb{E}\{\mathbf{c}_{k,n}|\mathbf{z}_n\}|^2$ avec l'espérance conditionnelle du log $\mathbb{E}\{\log(|\mathbf{c}_{k,n}|^2)|\mathbf{z}_n\}$. Toutefois, d'après l'inégalité de Jensen et la convexité de la fonction log, on a:

$$\log|\mathbb{E}\{\mathbf{c}_{k,n}|\mathbf{z}_n\}|^2 \geq \mathbb{E}\{\log(|\mathbf{c}_{k,n}|^2)|\mathbf{z}_n\}$$

Le terme de variance d'erreur dans l'expression (17) admet une interprétation assez intuitive. En effet, il se compose du produit de deux facteurs : le premier $\mathbf{w}_f^{-k}\mathbf{h}_n^{-k}/w_{f,k}h_{k,n}$ peut être vu comme l'inverse du rapport de puissance entre la source utile et les sources perturbatrices, et le second comme le niveau de vraisemblance des paramètres du modèle, à savoir $\mathbf{W}$ et $\mathbf{H}$, par rapport aux observations, à savoir $\mathbf{Z}$. En effet, $\mathbf{w}_f\mathbf{h}_n/|z_{f,n}|^2$ est directement relié à la fonction vraisemblance (8).

Intégration précoce

Il a été proposé dans les approches de type SVM de faire appel à des techniques de concaténations des données qualifiées de précoce, intermédiaire et tardive. Dans notre étude, nous reprenons la technique dite précoce, déjà utilisée dans [7] dans le contexte de la classification d'instruments de musique, et qui consiste à former un seul ensemble de vecteurs d'entrée du système SVM en prenant la moyenne de plusieurs attributs obtenus sur des traces successives, ce qui s'écrit:

$$\hat{b}_{k,\ell,f} = \sum_{j=1}^{L} \omega_{k,f,j}\hat{a}_{k,\ell L+j,f} \tag{18}$$

où ℓ dénote le nouvel indice temporel. Tout se passe comme si on supposait que, *conditionnellement à* $\mathbf{z_n}$, on a pour tout j allant de 1 à L

$$\hat{a}_{k,\ell L+j,f} = b_{k,\ell,f} + \epsilon_{k,\ell L+j,f}$$

où $\epsilon_{k,\ell L+j,f}$ sont des variables aléatoires, centrées, indépendantes.

Dans [7], les auteurs utilisent $\omega_{k,f,j} = 1/L$. Cette pondération considère implicitement que les variances des $\epsilon_{k,\ell L+j,f}$ sont identiques, ce qui n'est pas le cas, pour nous, d'après l'expression (15) qui dit que les variances des $\epsilon_{k,\ell L+j,f}$ sont égales à $\sigma^a_{k,\ell L+j,f}$. En passant à une notation vectorielle en concaténant les L valeurs, on a

$$\mathbf{a}_{k,\ell,f} = \mathbf{u} b_{k,\ell,f} + \epsilon_{k,\ell,f}$$

où u est un vecteur comportant L composantes toutes égales à 1. Les résultats classiques sur l'estimation par la méthode des moindres carrés pondérés donne alors comme estimateur de $b_{k,\ell,f}$:

$$\begin{aligned} \hat{b}_{k,\ell,f} &= (\mathbf{u}^t R_\epsilon^{-1} \mathbf{u})^{-1} \mathbf{u}^t R_\epsilon^{-1} \mathbf{a}_{k,\ell,f} \\ &= \frac{1}{\sum_{j'=1}^{L} (\sigma^a_{k,\ell L+j',f})^{-1}} \sum_{j=1}^{L} (\sigma^a_{k,\ell L+j,f})^{-1} a_{k,\ell L+j,f} \end{aligned} \tag{19}$$

On montre aisément que cet estimateur est sans biais et a comme variance:

$$\begin{aligned} \mathrm{var}(b_{k,\ell,f}) &= (\mathbf{u}^t R_\epsilon^{-1} \mathbf{u})^{-1} \\ &= \frac{1}{\sum_{j'=1}^{L} (\sigma^a_{k,\ell L+j',f})^{-1}} \end{aligned} \tag{20}$$

L'expression (19) montre que les poids de pondération optimaux au sens des moindres carrés dans (18) s'écrivent

$$\omega^{\mathrm{opt}}_{k,f,j} = \frac{(\sigma^a_{k,\ell L+j,f})^{-1}}{\sum_{j'=1}^{L} (\sigma^a_{k,\ell L+j',f})^{-1}}$$

D'autre part, partant de l'expression (20), la quantité

$$\bar{R}_{k,\ell,f} = \frac{1}{\mathrm{var}(b_{k,\ell,f})} = \sum_{j'=1}^{L} (\sigma^a_{k,\ell L+j',f})^{-1} \tag{21}$$

peut être considérée comme un niveau de fiabilité pour la valeur intégrée d'attribut $\hat{b}_{k,\ell,f}$. Nous en avons déduit la technique de seuillage suivante: les trames ayant des bins en fréquence dont le niveau de fiabilité $\bar{R}_{k,\ell,f}$ est inférieur à un seuil (fixé de façon expérimentale) sont dites NFB comme "Non-FiaBles". Leurs valeurs d'attributs ne sont pas retenues.

Dans notre étude, nous avons considéré des attributs à fenêtre courte et des attributs à fenêtre longue. Les attributs à fenêtre courte sont associés à une intégration sur $L = 20$ trames successives, ce qui correspond à une durée de $20 \times 32/2 = 320$ ms qui est aussi égale à la durée moyenne d'une note de musique [7]. Les attributs à fenêtre longue sont associés à une intégration sur $L = 40$ trames. Afin de synchroniser en temps ces deux longueurs d'attributs, nous répétons simplement les valeurs obtenues pour les attributs à fenêtre longue.

2.6 Normalisation globale des attributs

Les différents attributs étant de natures disparates, leurs valeurs absolues ne sont pas toujours comparables entre elles. Aussi, pour faciliter la sélection des attributs et l'apprentissage des SVMs nous avons adopté une procédure de normalisation globale des attributs qui consiste (i) en une soustraction de la moyenne globale et (ii) en une division par l'écart-type global. La moyenne et l'écart type sont calculés pendant la phase d'apprentissage (voir Fig. 2) sur les données d'apprentissage de toutes les classes (par classe il faut entendre instrument).

2.7 Sélection des attributs

Une réduction de la dimension du vecteur d'attributs est effectuée par sélection de J' attributs (les plus pertinents dans leurs ensemble pour la tâche de classification visée) parmi les J attributs pré-calculés ($J' < J$). Dans notre étude nous avons fixé $J' = 40$ parmi $J = 284$ attributs. L'algorithme de sélection utilisé est l'algorithme IRMFSP (*Inertia Ratio Maximization using Feature Space Projection*) (voir [14] pour plus de détail).

2.8 Apprentissage des SVMs bi-classes

Pour l'apprentissage des SVMs bi-classes (un SVM par couple d'instruments), nous avons utilisé la librairie `LibSVM` [16]. Pour chaque SVM, les paramètres de régularisation et la largeur σ de la gaussienne des noyaux Radial Basis Function (RBF) ont été réglés à l'aide d'une procédure classique de validation croisée k-fold[1] avec $k = 10$. Noter que le fait que chaque attribut soit normalisé entre -1 et $+1$, justifie le choix d'une valeur unique pour les largeurs σ des gaussiennes.

2.9 Calcul des probabilités bi-classes

Pour pouvoir calculer, pendant la phase de test, des probabilités bi-classes, au lieu de décisions "en dur" données par SVMs, nous probabilisons chaque SVM en ajustant une fonction sigmoïdale sur sa frontière de décision (voir [14] pour les détails). Ainsi, pendant la phase de test, les valeurs de la fonction de décision sont d'abord calculées à l'aide de LibSVM, puis ces valeurs sont transformées en probabilités en utilisant la fonction sigmoïdale correspondante.

2.10 Calcul des probabilités multi-classes

Les probabilités bi-classes sont à leurs tours transformées en probabilités multi-classes à l'aide de la méthode proposée par Hastie et Tibshirani [17] (voir [14] pour les détails).

2.11 Fusion des probabilités des composantes

Cette dernière étape consiste à estimer des probabilités de présence de chaque instrument dans le mélange à partir des probabilités de présence de cet instrument dans des composantes NMF élémentaires. Pour l'instant nous utilisons une procédure très simple qui consiste à calculer cette probabilité comme le maximum des probabilités des composantes NMF. Les trames des composantes NMF marquées Non-FiaBles ne sont pas prises en compte dans le calcul de ce maximum.

2.12 Lissage temporelle des probabilités

Enfin, pour lisser un peu les probabilités en temps, nous calculons les moyennes des logarithmes des probabilités sur des fenêtres de 4 secondes avec un recouvrement de 2 secondes. Pour évaluer le système, une décision finale "en dure" peut être toujours prise en seuillant ces probabilités, ainsi qu'en intégrant dans cette décision d'éventuelles connaissances complémentaires.

3 Expérimentations

3.1 Données expérimentales

Le système que nous avons proposé est appliqué à la reconnaissance, dans des enregistrements de jazz, de combinaisons d'instruments pris parmi les 8 instruments présentés au tableau 1.

Tous les enregistrements utilisés sont convertis en mode monophonique (si nécessaire) et ré-échantillonnés en 32000 Hz (si nécessaire).

[1] En SVM, une procédure simple d'évaluation de l'erreur est la validation croisée dite k-fold. Elle consiste à diviser l'ensemble des données en k sous-ensembles de taille approximativement égale. L'apprentissage est effectué en utilisant $(k-1)$ sous-ensembles et le test est effectué sur le sous-ensemble restant. Cette procédure est répétée k fois et chaque sous-ensemble est utilisé une fois pour le test. La moyenne des k taux d'erreur obtenus estime l'"erreur globale.

Instrument	Code
piano	Pn
guitare acoustique	Gt
trompette	Tr
saxophone ténor	Ts
contrebasse-*pizzicato*	Bs
batterie	Dr
saxophone alto	As
voix (féminine ou masculine)	Vv (Vf ou Vm)

Table 1: Liste des instruments et des codes associés.

3.1.1 Base des solos (apprentissage)

La base des solos est constituée des enregistrements solos de tous les instruments du tableau 1. Pour chaque instrument la durée totale du signal est comprise entre 669 et 2280 secondes.

3.1.2 Base des mélanges (test et développement)

La base des mélanges est composée de 101 enregistrements de jazz qui contiennent des instruments du tableau 1. Chaque enregistrement de la base est fourni avec une annotation manuelle décrivant son contenu instrumental.

Cette base est utilisée pour le test. Elle peut être également utilisée en développement pour une validation croisée k-fold.

3.2 Tâches de la reconnaissance des instruments

Nous distinguons les deux tâches de la reconnaissance des instruments suivantes :

1. *Identification d'instruments.* Le but de cette tâche est d'identifier si (oui ou non) un instrument particulier est présent dans l'enregistrement à un instant temporel donné.

2. *Reconnaissance d'ensembles.* Cette tâche a pour but d'identifier à instant donné l'ensemble des instruments présents dans l'enregistrement parmi des ensembles possibles.

En principe, la première tâche semble être plus facile que la deuxième, dans le sens où elle devrait donner de meilleurs scores dans les mêmes conditions. Il y a, en effet, plus de chance de se tromper en identifiant un ensemble d'instrument plutôt qu'un seul instrument. Toutefois, dans la deuxième tâche, on utilise *a posteriori* des connaissances supplémentaires sur les ensembles possibles.

3.3 Mesures de la performance

Pour chacune des deux tâches de reconnaissance envisagées, nous adoptons des mesures de performance qui sont celles généralement utilisées dans des travaux de ce type. Pour l'identification d'instruments nous utilisons :

- le *Miss Error Rate* (MER)[2] et le *False Alarm Rate* (FAR)[3] [18].

- la *F-mesure* définie [19] par

$$F = \frac{2pr}{p+r}$$

[2]Le MER est le taux de trames où l'instrument n'a pas été identifié, alors qu'il est présent
[3]Le FAR est le taux de trames où l'instrument a été identifié, alors qu'il n'est pas présent.

où la *précision*

$$p = \frac{1}{n}\sum_{i=1}^{n} p_i, \quad \text{avec} \quad p_i = \frac{\text{nombre d'éléments correctement attribués à la classe } i}{\text{nombre d'éléments attribués à la classe } i}$$

et le *rappel*

$$r = \frac{1}{n}\sum_{i=1}^{n} r_i, \quad \text{avec} \quad r_i = \frac{\text{nombre d'éléments correctement attribués à la classe } i}{\text{nombre d'éléments appartenant à la classe } i}$$

Pour la reconnaissance d'ensembles nous utilisons :

- la *Matrice des confusions* qui est une matrice de taille $E \times E$ (E étant le nombre total des ensembles possibles) dont l'élément $c_{i,j}$ est le pourcentage des trames de l'ensemble i reconnues comme des trames de l'ensemble j [14].
- le *Taux de reconnaissance* (RA pour *recognition accuracy*) est le pourcentage de trames pour lesquelles l'ensemble a été reconnu correctement [14].

3.4 Résultats

Avec le système de reconnaissance des instruments présenté section 2, nous avons testé les points suivants :

- la décomposition NMF avec la divergence de IS ou avec la divergence de KL.
- le nombre de composantes NMF utilisées pour la décomposition des solos lors de la phase d'apprentissage. Nous avons testé $K_s = 1$, 3 et 7 composantes NMF. Pour les ensembles, lors de la phase de test, nous avons uniquement utilisé $K_e = 12$ composantes NMF.
- deux modes de calcul des poids de la fiabilité et de la sélection des trames :
  - le mode *silence* où l'intégration temporelle précoce est effectuée comme dans [7], c'est-à-dire il n'y a pas de pondération, et la sélection des trames pertinentes est faite à l'aide d'un détecteur de silence simple.
  - le mode *avancé* où l'intégration temporelle précoce et la sélection des trames pertinentes sont effectuées comme cela a été expliqué section 2.5.

Les performances moyennes, en termes de MER, de FAR, de F-mesure et de RA, de tous ces systèmes sont regroupées Tableau 2.

Les meilleurs performances sont obtenues avec le système utilisant $K_s = 7$ composantes pour décomposer les solos lors de la phase d'apprentissage, la divergence de IS, et le calcul des poids de la fiabilité en mode *avancé*. Ces résultats correspondent à nos attentes. Cependant, les différences de performances entre IS et KL d'une part, et mode *silence* et mode *avancé* d'autre part, ne sont pas très significatives.

On observe, par contre, que l'augmentation du nombre de composantes NMF pour décomposer les solos mène à une amélioration très importante des performances. Les plus mauvaises performances Pour $K_s = 1$, cad l'absence de décomposition NMF lors de la phase d'apprentissage, les performances sont alors équivalentes au pur hasard : MER $\approx$ FAR $\approx$ 50 %. La conclusion qu'on tire de ces observations, c'est qu'il est très important de rapprocher au mieux le traitement fait lors de la phase d'apprentissage à celui fait lors de la phase du test.

Tableau 3 et Figure 3 donnent les performances détaillées du meilleur système obtenu. En regardant la matrice des confusions de la Figure 3 [4], on peut constater que les performances de la reconnaissance des ensembles sont satisfaisants pour des solos, sont parfois satisfaisantes pour des mélanges à deux instruments, et ne sont pas du tout satisfaisantes pour des mélanges de plus de deux instruments.

Enfin, ces performances sont significativement mois bonnes que celles obtenues avec d'autres systèmes de reconnaissance (voir par exemple [14]).

[4] "ex_num" de la Fig. 3 indique le nombre de trames (de durée 4 secondes, voir Sec. 2.12) de l'ensemble correspondant.

NM-NMF (appr.).	Poids de Fiab.	Divergence de IS				Divergence de KL			
		MER	FAR	F-mes	RA	MER	FAR	F-mes	RA
1	Silence	*49.6*	*50.0*	*36.0*	*9.7*	-	-	-	
3	Silence	34.9	34.3	51.3	22.2	-	-	-	
7	Silence	30.2	29.8	56.6	33.6	29.9	30.3	56.4	29.5
7	Avancé	**27.2**	**27.6**	**59.6**	**35.7**	30.6	30.0	56.2	31.0

Table 2: Performances moyennes des différents systèmes.

	MER	FAR	F-mesure
Total	27.2	27.6	59.6
Pn	10.7	39.0	86.8
Gt	3.9	74.9	9.4
Tr	19.7	9.9	58.4
Ts	39.4	21.9	30.7
Bs	47.1	10.1	66.5
Dr	15.6	20.6	80.2
As	41.3	25.0	7.2
Vv	47.8	8.7	56.8

Table 3: Performances détaillés du meilleur système obtenu.

ensemble	ex_num	N1	N2	N3	N4	N5	N6	N7	N8	N9	N10	N11	N12	N13	N14	N15	N16	N17	N18	N19
N1 : Pn	1291	87	2	0	3	0	7	0	0	0	0	0	0	0	0	0	0	0	0	0
N2 : Dr	195	2	78	0	0	2	0	0	0	0	0	0	11	7	0	0	0	0	0	0
N3 : Bs	792	2	5	61	6	20	0	0	1	1	3	0	0	0	1	0	0	0	1	0
N4 : BsPn	614	27	1	37	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N5 : BsDr	239	3	36	26	1	22	0	0	0	0	3	0	2	7	0	0	0	0	0	0
N6 : PnTr	328	26	0	2	1	0	72	0	0	0	0	0	0	0	0	0	0	0	0	0
N7 : PnVv	509	68	7	1	0	0	1	11	12	0	0	0	0	0	0	0	0	0	0	0
N8 : GtVv	201	26	12	4	0	7	0	17	33	0	0	0	0	0	0	0	0	0	0	0
N9 : BsVv	97	9	0	66	12	0	0	0	0	8	1	0	0	1	0	2	0	0	0	0
N10 : BsDrPn	1010	28	39	12	8	2	1	0	1	0	8	0	0	0	0	0	1	0	0	0
N11 : BsDrTr	153	1	76	3	1	0	16	0	0	0	1	2	0	0	0	0	0	0	0	0
N12 : AsBsDr	57	0	91	2	0	0	0	0	7	0	0	0	0	0	0	0	0	0	0	0
N13 : BsDrTs	219	6	92	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
N14 : BsDrVv	58	2	74	7	0	0	0	0	10	2	0	0	0	5	0	0	0	0	0	0
N15 : BsPnVv	92	10	7	59	5	11	0	0	4	2	0	0	0	0	1	1	0	0	0	0
N16 : BsDrPnTr	202	22	43	4	5	4	15	0	0	0	4	1	0	0	0	0	0	0	0	0
N17 : BsDrPnTs	397	39	44	4	0	0	0	1	1	0	4	0	0	4	0	0	0	4	1	0
N18 : BsDrPnVv	537	20	39	3	1	1	1	8	11	5	3	0	0	1	1	0	0	0	4	2
N19 : AsBsDrPn	64	5	55	11	0	9	0	0	0	0	0	0	20	0	0	0	0	0	0	0

Figure 3: Matrice des confusions du meilleur système obtenu.

4 Discussion

Nous pensons que des performances médiocres qu'on obtient sont dues aux problèmes suivants :

1. Notre mécanisme de lutte contre les erreurs de la pré-séparation n'est pas très performant, étant supposé que le modèle simplifié d'extraction des attributs est vérifié. En effet, ce mécanisme permet de diminuer l'influence d'erreurs en pondérant les observations lors de l'intégration temporelle, ainsi qu'en rejettent certaines observations avec trop d'erreurs. Cependant, la répartition des erreurs en fréquence n'est pas du tout prise en compte, ainsi que les erreurs ne sont pas prises en compte lors de la classification par les SVMs.

2. L'espérance conditionnelle, donnée par l'équation (10), est un bon estimateur du spectre d'amplitude des composantes NMF. Toutefois il est mal adapté au cas de l'estimation du log-spectre qui est, comme nous l'avons indiqué, une grandeur plus appropriée au calcul des attributs (13).

3. Il y a toujours des différences entre le traitement de la phase d'apprentissage et celui de la phase du test, puisque pendant la phase d'apprentissage des solos sont décomposés par la NMF, tandis que pendant la phase de test ce sont des mélanges.

4. Enfin, la pré-séparation des mélanges par la NMF ne donne pas toujours de résultats satisfaisants, puisque certaines composantes NMF représentent parfois plusieurs instruments à la fois.

5 Conclusion et pistes d'amélioration

Nous avons proposé une nouvelle approche pour la reconnaissance des instruments basée sur une pré-séparation du contenu sonore en composantes NMF élémentaires. Nous avons développé le système de reconnaissance correspondant et nous l'avons évalué (dans différentes configurations) sur une base d'enregistrements de jazz. Une conclusion majeure que nous avons tirée de ces expérimentations est qu'il est très important de rapprocher au mieux le traitement fait lors de la phase d'apprentissage à celui fait lors de la phase du test. Malheureusement, les meilleures performances qu'on obtient sont bien au-dessous des performances rapportées pour d'autres systèmes de reconnaissance d'instruments ou d'ensemble d'instruments.

Dans la discussion, nous indiquons plusieurs problèmes qui, à notre avis, sont à l'origine des performances médiocres obtenues. Pour y remédier nous proposons les pistes suivantes [5] :

1. Utiliser la théorie des données manquantes [20], ou bien ses extensions [21, 22]. Ces approches permettent de classifier des observations partiellement manquantes et (ou) bruitées. Dans ce cas il faudrait probablement utiliser des GMMs au lieu des SVMs, puisque à l'heure actuelle la théorie des données manquantes n'est pas très au point pour des SVMs.

2. remplacer l'espérance conditionnelle du spectre par l'espérance conditionnelle du log-spectre (voir par exemple [23]). Le calcul des poids de la fiabilité devra être revu.

3. Rapprocher le traitement de la phase d'apprentissage à celui de la phase du test. Par exemple, pour effectuer l'apprentissage sur les mélanges (comme dans le test) les solos de la base d'apprentissage peuvent être artificiellement mélangés avec des accompagnements musicaux.

4. Améliorer la décomposition NMF pour qu'elle sépare mieux des instruments et/ou considérer certaines combinaisons d'instruments, difficilement séparables par la NMF, comme des classes à part entière.

[5] La numérotation des pistes d'amélioration correspond à celle des problèmes dans la discussion.

A Liste des attributs

Descripteur ou paquet d'attributs	Taille	Synopsis
$Cp = [Cp1, ..., Cp11]$, $(\delta,\delta 2)[Cp0, ..., Cp10]$	33	Coefficients cepstraux à partir de 30 sous-bandes MEL et dérivées temporelles.
$Cc = [Cc1, ..., Cc11]$, $(\delta,\delta 2)[Cc0, ..., Cc10]$	33	Coefficients cepstraux à partir de 11 sous-bandes MEL et dérivées temporelles.
uCq, $(\delta,\delta 2)uCq$	27	Coefficients cepstraux à partir d'une CQT avec résolution d'une octave.
dCq, $(\delta,\delta 2)dCq$	30	Coefficients cepstraux à partir d'une CQT avec résolution d'une demi-octave.
tCq, $(\delta,\delta 2)tCq$	30	Coefficients cepstraux à partir d'une CQT avec résolution d'un tiers d'octave.
qCq, $(\delta,\delta 2)qCq$	30	Coefficients cepstraux à partir d'une CQT avec résolution d'un quart d'octave.
Sx=[Sc,Sw,Sa,Sk]+δ+$\delta 2$	12	Moments spectraux et dérivées temporelles.
ASF= [$A1$,...,$A23$]	23	Platitude spectrale (MPEG-7).
SCF=[$SCF1$,...,$SCF23$]	23	Facteur de crête spectrale.
AR=[$AR1$,$AR2$]	2	Coefficients LPC.
[Ss,Sd,Sv,So,Fc]	5	Pente, décroissance, variation temporelle, platitude du spectre, fréquence de coupure.
Si=[$Si1$,...,$Si21$]	21	Irrégularité spectrale.
$OBSI$=[$O1$,...,$O8$]	8	Intensités en sous-bandes d'octaves.
$OBSIR$=[$OR1$,...,$OR7$]	7	Rapports d'intensité en sous-bandes d'octaves.

Table 4: Liste des attributs utilisés. Au total nous obtenons 284 attributs.

Bibliography

[1] K. Martin and Y. Kim, "Musical instrument identification: a pattern-recognition approach," in *Proc. 136th Meeting of the Acoustical Society of America*, 1998.

[2] J. C. Brown, "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," *The Journal of the Acoustical Society of America*, vol. 105, no. 3, p. 1933, 1999.

[3] A. Eronen and A. Klapuri, "Musical instrument recognition using cepstral coefficients and temporal features," in *Acoustics, Speech, and Signal Processing, 2000. ICASSP '00. Proceedings. 2000 IEEE International Conference on*, vol. 2, June 2000, pp. 753–756.

[4] H. Hacıhabiboğlu and N. Canagarajah, "Musical instrument recognition with wavelet envelopes," in *Proc. EAA Convention, Forum Acusticum Sevilla*, Sevilla, Spain, 16-20 September 2002.

[5] A. A. Livshin and X. Rodet, "Musical instrument identification in continuous recordings," in *Proceedings of the 7th International Conference on Digital Audio Effects*, 2004, pp. 222–226.

[6] E. Benetos, C. Kotropoulos, T. Lidy, and A. Rauber, "Testing supervised classifiers based on non-negative matrix factorization to musical instrument classification," in *Proc. 14th European Signal Processing Conf.*, September 2006.

[7] C. Joder, S. Essid, and G. Richard, "Temporal integration for audio classification with application to musical instrument classification," *Submitted to IEEE Transactions on Audio, Speech, and Language Processing*, 2008.

[8] J. Eggink and G. Brown, "A missing feature approach to instrument identification in polyphonic music," in *Applications of Signal Processing to Audio and Acoustics, 2003 IEEE Workshop on.*, 19-22 Oct. 2003, p. 49.

[9] A. Cont, S. Dubnov, and D. Wessel, "Realtime multiple-pitch and multiple-instrument recognition for music signals using sparse non-negative constraints," in *Digital Audio Effects (DAFx)*, Bordeaux, 2007.

[10] T. Kitahara, M. Goto, K. Komatani, T. Ogata, and H. G. Okuno, "Instrument identification in polyphonic music: Feature weighting to minimize influence of sound overlaps," *EURASIP Journal on Applied Signal Processing*, 2007.

[11] P. Leveau, D. Sodoyer, and L. Daudet, "Automatic instrument recognition in a polyphonic mixture using sparse representations," in *8th International Conference on Music Information Retrieval (ISMIR 2007)*, Vienna, Austria, September 23-27 2007.

[12] L. G. Martins, J. J. Burred, G. Tzanetakis, and M. Lagrange, "Polyphonic instrument recognition using spectral clustering," in *8th International Conference on Music Information Retrieval (ISMIR2007)*, Vienna, Austria, September 2007.

[13] Z. Rafii, R. Blouet, and A. Liutkus, "Discriminant approach within non-negative matrix factorization for musical components recognition," DMRN+2: Digital Music Research Network One-day Workshop, Dec. 2007, poster presentation.

[14] S. Essid, "Classification automatique des signaux audio-fréquences : reconnaissance des instruments de musique," Ph.D. dissertation, Université Pierre et Marie Curie, 2005.

[15] C. Févotte, N. Bertin, and J.-L. Durrieu, "Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis," *Neural Computation*, vol. 21, no. 3, pp. 793–830, Mar. 2009.

[16] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[17] T. Hastie and R. Tibshirani, "Classification by pairwise coupling," in *The Annals of Statistics*. MIT Press, 1998, pp. 507–513.

[18] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki, "The det curve in assessment of detection task performance," in *European Conf. on Speech Communication and Technology (EuroSpeech'97)*, 1997, pp. 1895–1898.

[19] C. J. van Rijsbergen, *Information Retrieval*. Butterworths, London, 1979.

[20] M. P. Cooke, P. D. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, pp. 267–285, 2001.

[21] J. P. Barker, M. P. Cooke, and D. P. W. Ellis, "Decoding speech in the presence of other sources," *Speech Communication*, vol. 45, no. 1, pp. 5–25, 2005.

[22] L. Deng, J. Droppo, and A. Acero, "Dynamic compensation of HMM variances using the feature enhancement uncertainty computed from a parametric model of speech distortion," *IEEE Trans. Speech and Audio Proc.*, vol. 13, pp. 412– 421, 2005.

[23] A. Ozerov, R. Gribonval, P. Philippe, and F. Bimbot, "Choix et adaptation des modèles pour la séparation de voix chantée à partir d'un seul microphone," *Traitement du signal*, vol. 24, no. 3, pp. 211–224, 2007.

Dépôt légal : 2009 – 3ème trimestre
Imprimé à l'Ecole Nationale Supérieure des Télécommunications – Paris
ISSN 0751-1345 ENST D (Paris) (France 1983-9999)

TELECOM ParisTech

Institut TELECOM - membre de ParisTech

46, rue Barrault - 75634 Paris Cedex 13 - Tél. + 33 (0)1 45 81 77 77 - www.telecom-paristech.frfr

Département TSI