Classification de signaux audio en temps-réel par un modèle de mélanges d'histogrammes

Maxime Baelde [1] & Christophe Biernacki [2] & Raphäel Greff [3]

[1,3] *A-Volute,* [1,2] *Université de Lille, CNRS, Inria*
[1] *maxime.baelde@a-volute.com*
[2] *christophe.biernacki@math.univ-lille1.fr*
[3] *raphael.greff@a-volute.com*

Résumé. La reconnaissance sonore consiste à attribuer un label à un signal audio inconnu. Celle-ci repose généralement sur des descripteurs audio ainsi que des modèles d'apprentissage statistique. Néanmoins les modèles actuels peinent à bien classer les sons dans un contexte temps-réel où ces derniers sont hétérogènes. Ce papier propose une nouvelle méthode basée sur un modèle de mélanges d'histogrammes représentant les spectres audio. La reconnaissance consiste à calculer la probabilité de chaque groupe puis à les agréger temporellement. Une étape de réduction du précédent modèle permet par ailleurs de passer au temps-réel. Cette méthode surpasse les algorithmes actuels, et peut atteindre 96,7% de bonne classification sur une base de 50 classes de sons en utilisant 0,5s de données audio.

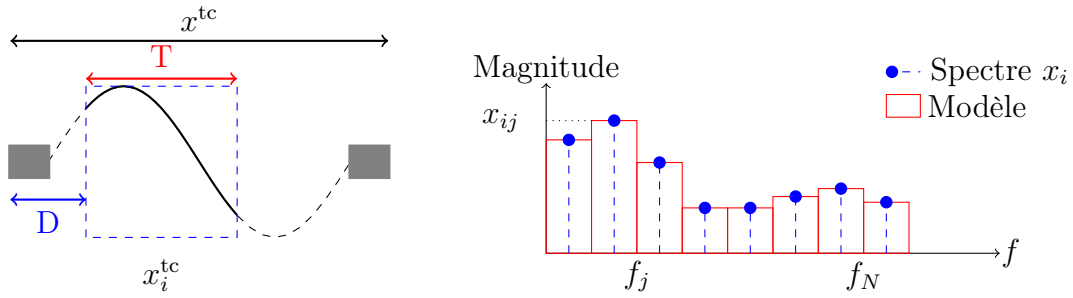
Mots-clés. temps-réel, classification, audio, modèle de mélanges, machine learning.

Abstract. Audio recognition consists in giving a label to an unknown audio signal. It relies on audio descriptors and machine learning algorithms. However, in a real-time context with heterogeneous sounds, the current models lack of performance to classify sounds. This article presents a novel method based on a model of histogram mixture representing audio spectra. The recognition consists in computing the probability of each group and aggregate them temporally. A reduction step of the models allows also to perform this algorithm in real-time. This method outperforms current state-of-the-art algorithms, and achieves an accuracy of 96,7% on a database of 50 classes, using only 0.5s of audio data.

Keywords. real-time, classification, audio, mixture models, machine learning.

1 Introduction

La classification supervisée consiste à attribuer un label à une observation à partir d'un modèle statistique. Appliquée aux technologies de l'audio, elle consiste notamment en la classification de genres musicaux, la reconnaissance vocale ou encore la classification de sons environnementaux



(a) Découpage de x^{tc} en sons partiels x_i^{tc} de taille T avec un décalage D.

(b) Représentation d'un spectre d'énergie x_i par un histogramme.

FIGURE 1 – Modélisation des spectres audio.

Les modèles statistiques de signaux audio reposent en général sur des descripteurs audio (Peeters (2011)) comme l'énergie du signal ou les MFCC (Mel-Frequency Cepstral Coefficients). Les modèles couramment utilisés dans la classification audio sont les mélanges de gaussiennes, les machines à vecteur de support, les modèles de Markov cachés ou les réseaux de neurones convolutifs profonds.

Nous avons décidé de suivre une toute nouvelle approche basée sur des modèles de mélanges (McLachlan (2000)) d'histogramme représentant les spectres audio. C'est une extension du papier de Baelde (2017). L'intérêt de ces modèles est de pouvoir considérer des mélanges de sons qui sont alors modélisés comme des mélanges de mélanges. Une étape de réduction du dictionnaire permet de ne garder que les modèles essentiels et de pouvoir utiliser la technique en temps réel. La classification consiste alors à calculer les probabilités de chaque classe puis à les agréger temporellement. Le contexte dans lequel s'inscrit cet article est le suivant. On dispose de sons répartis en plusieurs classes. L'aspect temps-réel implique que le son n'est jamais connu en entier. C'est pourquoi les sons de chaque classe sont tronqués et décalés de manière à former des sons partiels. De plus, on ajoute du bruit blanc au début et à la fin de chaque son car celui-ci ne commence pas nécessaire au début d'un flux audio, ni ne se termine forcément à la fin.

Nous commençons par présenter dans la Partie 2 la création du modèle, le principe de la classification avec ce dernier ainsi qu'une méthode pour réduire le temps de calcul. La Partie 3 illustre les expériences menées et les résultats. Enfin une discussion ainsi qu'une conclusion sont proposées en Partie 4.

2 Modélisation de spectres sonores par histogrammes

2.1 Ensemble d'apprentissage

L'ensemble d'apprentissage est construit de la manière suivante. On dispose de plusieurs groupes de sons, numérotés $k = 1, ..., g$ (par exemple, le groupe 1 est celui des avions). Dans chaque groupe k, les sons sont numérotés $l = 1, ..., n_k$. Pour le son $x_l^{\text{tc}}(t) \in [-1, 1]$ $(t \in \mathbb{R})$ numéro l continu du domaine temporel, on fabrique les n_{kl} décalages de D unités temporelles et troncatures de taille T unités temporelles, noté $x_{li}^{\text{tc}}(t) = x_l^{\text{tc}}\,(t \in [iD, iD + T])$ (FIGURE 1(a)). Une unité temporelle est définie comme le pas d'échantillonnage temporel d'un signal sonore. Au total, on obtient $n = \sum_{k=1}^{g} \sum_{l=1}^{n_k} n_{kl}$ sons partiels qui constituent l'ensemble d'apprentissage $(x_i^{\text{tc}}, z_i)_{i=1}^{n}$. Les labels z_i associés à chaque x_i^{tc} indiquent l'appartenance du son partiel au groupe z_i (par exemple $z_i = 1$ indique que x_i^{tc} est un son partiel d'avion). Cependant, en pratique on n'a pas accès aux sons continus x_i^{tc}, qu'il faut discrétiser en $x_i^{\text{td}} \in [-1, 1]^T$ tel que $x_{ij}^{\text{td}} = x_i^{\text{tc}}((j-1)/f_e)$ avec f_e la fréquence d'échantillonnage. De plus, on se place dans le domaine fréquentiel en considérant la transformation en spectre d'énergie normalisée x_i tel que :

$$x_{ij} = \frac{\left| \left(\text{TFD}_T\left(x_i^{\text{td}}\right)\right)_l \right|^2}{\sum_{k=1}^{N} \left| \left(\text{TFD}_T\left(x_i^{\text{td}}\right)\right)_k \right|^2},\ j = 1, ..., N, \tag{2.1}$$

avec $\text{TFD}_T : \mathbb{R}^T \to \mathbb{C}^T$ l'opérateur de la Transformée de Fourier Discrète (TFD) d'un signal discret sur T points. Ce spectre d'énergie x_i est représenté par un histogramme normalisé (FIGURE 1(b)). On choisit de ne garder que $N < T$ composantes fréquentielles dans le spectre d'énergie car les hautes fréquences ne contiennent pas d'informations pertinentes dans notre cas. De plus, on considère les spectres d'énergies car ils conservent la propriété d'additivité lorsque deux sons décorrélés se mélangent. On se place dans le domaine spectral car le domaine temporel contient trop d'information spécifique à un son et ne permet pas d'extraire de caractéristiques générales communes à un groupe de sons. Les données considérées pour le modèle sont donc les couples $(x_i, z_i)_{i=1}^{n}$.

2.2 Règle de classement

Un nouveau son est traité suivant la même procédure que précédemment pour obtenir l'ensemble de test $(x_j, z_j)_{j=n+1}^{m}$. Dans le cadre de la classification supervisée, on cherche maintenant une règle de classement $R : \mathcal{X} \to \{1, ..., g\}$ qui à un son associe son label : $z = R\,(x_{n+1}, ..., x_m)$, avec $\mathcal{X} = \underbrace{[0, 1]^N \times ... \times [0, 1]^N}_{m-n}$. On va estimer une règle $\hat{R}$ à partir de l'ensemble d'apprentissage $(x_i, z_i)_{i=1}^{n}$. On considère le score $f\left(x_j \middle| z = k\right)$ suivant qui calcule la «distance» du spectre candidat x_j à la classe k :

$$f\left(x_j\middle|z=k\right)=\frac{1}{n_k}\sum_{\{i|z_i=k\}}e^{-H\left(x_j\middle||x_i\right)}, \tag{2.2}$$

avec $H\left(p||q\right)$ l'entropie croisée entre p et q : $H\left(p||q\right)=-\sum_{j=1}^{N}p_j\log\left(q_j\right)$.

La probabilité de la classe k est ensuite calculée en normalisant ces scores par la formule de Bayes :

$$p\left(z=k\middle|x_j\right)=\frac{f\left(x_j\middle|z=k\right)p\left(z=k\right)}{\sum_{h=1}^{g}f\left(x_j\middle|z=h\right)p\left(z=h\right)} \tag{2.3}$$

avec $p\left(z=k\right)=n_k/n$. Sous hypothèse d'indépendance conditionnelle des évènements $z=k|x_j$ $(m=1,...,M)$, les probabilités conditionnelles sont ensuite agrégées temporellement sur les $m-n$ sons partiels :

$$p\left(z=k\middle|x_{n+1},...,x_m\right)=\prod_{j=n+1}^{m}p\left(z=k\middle|x_j\right). \tag{2.4}$$

La règle de décision est donnée par le principe du maximum a posteriori (MAP) :

$$\hat{z}=\hat{R}(x_{n+1},...,x_m)=\underset{k}{\operatorname{argmax}}\,p\left(z=k\middle|x_{n+1},...,x_m\right). \tag{2.5}$$

2.3 Réduction du temps de traitement

La complexité de l'identification est $\mathcal{O}(\sum_k n_k)$, qui est potentiellement très grande (comme on le verra dans la Partie 3). Pour réduire cette complexité, une classification hiérarchique ascendante (CAH) sur les histogrammes est réalisée. La réduction consiste à mélanger les histogrammes présents dans les clusters induits par la CAH. Une CAH nécessite une distance entre les éléments à classer et un critère de regroupement. Notre donnée de base étant l'histogramme, on considère la distance de Hellinger. Le critère de regroupement choisit pour la CAH est le critère de Ward, aussi appelé critère de minimum de variance. Le résultat de la CAH dans chaque groupe k donne un arbre de classification des différents histogrammes. On choisit le nombre de clusters à extraire de cet arbre, noté n'_k. Pour chaque cluster $\mathcal{C}_{ki}\in\{\mathcal{C}_{k1},...,\mathcal{C}_{kn'_k}\}$, on note $b_{ki}=\{j\mid x_j\in\mathcal{C}_{ki}\}$ $(i=1,...,n'_k)$. L'histogramme réduit $\tilde{x}_i$ de label z_i associé au cluster $\mathcal{C}_{ki}$ est défini comme le mélange des différents histogrammes dans ce cluster :

$$\tilde{x}_i=\frac{1}{\operatorname{card}\left(b_{ki}\right)}\sum_{j\in b_{ki}}x_j. \tag{2.6}$$

Le nouvel ensemble d'apprentissage consiste donc en les couples $(\tilde{x}_i,z_i)_{i=1}^{n'}$ avec $n'=\sum_{k=1}^{g}n'_k$. Le score associé à un nouveau spectre x_j pour la classe k se calcule en remplaçant

TABLE 1 – Taux de bonne classification en % pour les différentes méthodes et bases de sons.

Base	A-Volute	ESC-10	ESC-50
Notre méthode	99,4	97,9	96,8
Méthode paramétrique	73,6	73,5	45,5
Méthode non-paramétrique	46,6	76,0	53,2
Humain	91,8	95,7	81,3

les x_i par $\tilde{x}_i$. On peut ainsi appliquer la procédure précédente pour obtenir la règle de décision.

3 Expériences

Afin de quantifier les performances de la méthode, plusieurs expériences ont été réalisées. Trois bases de données sont considérées : la base A-Volute (constituée de 704 sons répartis en 9 classes), et les bases ESC-10 et ESC-50 (Piczak (2015b)) (constituées de 400 et 2000 sons environnementaux répartis en 10 et 50 classes respectivement). Toutes ces données audio sont rééchantillonnés à $f_e = 44,1$kHz et centrés. La taille de fenêtre est réglée à $T = 2048$ et le décalage temporel à $D = 512$ unités temporelles. La taille de la TFD est de T et on considère $N = T/5 = 410$ composantes fréquentielles dans le spectre d'énergie. Le nombre de sons partiels du son de test est fixé à $m - n = 10$; si en pratique le nombre possible de sons partiels dépasse $m - n$, on considère plusieurs blocs de $m - n$ sons partiels et on réitère la procédure sur ces blocs. Ce nombre de sons partiels correspond à 464ms de données audio. Les ensembles d'apprentissage associés aux bases de sons sont divisés suivant le schéma v-fold pour réaliser une procédure de validation croisée, avec $v = 5$. La métrique de performance est le taux de bonne classification en validation croisée. La méthode a également été comparée à d'autres techniques de classification audio : celle de Clavel (2005) (mélange de gaussiennes et descripteurs classiques), ainsi que celle de Piczak (2015a) (réseaux de neurones convolutifs profonds utilisant un spectrogramme). Enfin, des tests d'écoutes ont été réalisés pour avoir les performances d'humains sur les bases de sons considérés. Les résultats pour les bases ESC-10 et ESC-50 étant connus (Piczak (2015b)), seule la base A-Volute a été étudiée par nos soins. Au total, 21 participants ont classé 10 sons choisis aléatoirement dans les 9 classes de la base. Une estimation grossière du score de classification a été obtenu sur cette base.

Les résultats sur le dictionnaire complet sont disponibles dans la TABLE 1. Notre méthode surpasse largement les méthodes concurrentes considérées (toujours supérieur à 95% peu importe la base). Pour un fold de la base A-Volute ($n = 70000$), la décision prend 232ms pour être calculée sur un processeur Intel®Core™i7 @2.7 GHz et 13ms sur

carte graphique NVidia TitanX. Sachant que la durée d'un son partiel est de 46,4ms, le temps de calcul sur processeur est largement trop grand. Sur carte graphique ce temps est considérablement réduit mais ce matériel est coûteux et peu répandu. On considère les résultats du dictionnaire réduit sur la base A-Volute. Pour une réduction très faible ($n'_k = n_k/2$), le taux de bonne classification est de 99,3% et le temps de calcul est de 120ms. Pour une réduction très forte ($n'_k = n_k/400$), le taux de bonne classification est de 82,8% et le temps de calcul est de 0,4ms. On remarque que la construction de ce modèle réduit permet de contrôler le compromis précision - temps de calcul.

4 Discussion et conclusion

La méthode développée dans ce papier a pour but de réaliser de la classification supervisée de signaux audio en temps réel. Elle repose sur une nouvelle approche à base de modèles de mélanges d'histogramme. Dans le but de pouvoir utiliser l'algorithme en temps réel, une étape de réduction des modèles est nécessaire, basée sur une classification hiérarchique des modèles. Les performances de la méthode sont bien supérieures aux autres méthodes de l'état de l'art considérées dans cet article (mélange de gaussiennes et deep learning). Pour le moment, cette méthode permet de faire de l'identification mono-source (c'est-à-dire une source audio active à la fois). Néanmoins par construction, on peut étendre ce procédé pour identifier plusieurs sources présentes en même temps (contexte multi-sources) en considérant des mélanges de mélanges présents dans les modèles.

Bibliographie

[1] Baelde M., Biernacki C. et Greff R. (2017), A mixture model-based real-time audio sources classification method, *The 42nd IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP2017.*
[2] Clavel C., Ehrette T. et Richard G. (2005), Events Detection for an Audio-Based Surveillance System, *2005 IEEE International Conference on Multimedia and Expo*, 1306–1309.
[3] McLachlan G. et Peel D. (2000), Finite Mixture Models, *Wiley.*
[4] Peeters G., Giordano B. L., Susini P., Misdariis N. et McAdams S. (2011), The Timbre Toolbox : extracting audio descriptors from musical signals, *The Journal of the Acoustical Society of America*, 130, 5, 2902-2916.
[5] Piczak K. J. (2015a), Environmental sound classification with convolutional neural networks, *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, 1-6.
[6] Piczak K. J. (2015b), ESC : Dataset for Environmental Sound Classification, *ACM Press*, 1015-1018.