# Sélection de descripteurs audio pour la classification des sons environnementaux avec des SVMs mono-classe

Asma Rabaoui[1], Manuel Davy[2], Stéphane Rossignol[2], Zied Lachiri[1], Noureddine Ellouze[1]

[1]Unité de recherche Signal, Image et Reconnaissance des formes,
ENIT, BP 37, Campus Universitaire, 1002 le Belvédère, Tunis Tunisie
Tél : + (216) 71 87 68 00 – Fax : + (216) 71 87 68 00

[2]CNRS/Laboratoire d'Automatique, de Génie Informatique et Signal
INRIA Futur équipe SequeL,
BP 48 , 59651 Villeneuve d'Ascq cedex, France
Tél : 03 20 67 60 13 – Fax : 03 20 33 54 18
Asma.Rabaoui@enit.rnu.tn, Manuel.Davy@inria.fr

**Résumé** – Cet article traite de la classification supervisée de signaux sonores à des fins de télésurveillance. L'approche développée utilise des descripteurs audio originaux, basés en partie sur une analyse en ondelettes, et une procédure de classification en classes multiples en utilisant plusieurs SVM à une classe. L'originalité du travail réside essentiellement dans la proposition d'une procédure de sélection des meilleures combinaisons de descripteurs audio, ainsi qu'une nouvelle procédure de classification multiclasse. Ces nouvelles approches sont évaluées sur un corpus original et suffisamment important.

**Abstract** – This paper addresses a specific audio classification problem for surveillance and security applications. The developed approach uses original audio features including essentially wavelets-based features, and a procedure of multi-class classification using several One-class SVMs. The originality of our work lies primarily in the proposal of a selection procedure of the optimal audio features combinations, as well as a new method of multi-class classification. These new approaches are evaluated on an original and sufficiently large database.

## 1 Introduction

Cet article traite de la classification de sons de l'environnement. Il s'agit d'une tâche élémentaire intervenant dans la conception de systèmes de télésurveillance pour la sécurisation des transports urbains, l'assistance aux personnes âgées, etc.

L'architecture du système proposé est la suivante. Partant d'un enregistrement sonore, des descripteurs audio sont extraits, et concaténés sous la forme de vecteurs de longueur fixe, un par signal. Puis, dans une phase d'apprentissage, un classifieur multi-classe original est conçu, basé sur plusieurs algorithmes à vecteur support pour l'estimation de support de densité ( one-class SVM (1-SVM)). Enfin, ce classifieur est utilisé pour reconnaître les signaux sonores des classes ayant servi à l'apprentissage.

La section 2 décrit les descripteurs acoustiques utilisés. La section 3 décrit l'approche monoclasse suivie, tandis que la section 4 décrit brièvement les résultats expérimentaux obtenus.

## 2 Extraction des descripteurs audio

Le choix des descripteurs audio, et de leur combinaison, a une importance particulière lors de la conception d'un système de reconnaissance des sons. En effet, les descripteurs optimaux doivent contenir toute l'information nécessaire permettant à un classificateur la bonne discrimination des sons de classes différentes, sur la base de leurs différentes caractéristiques acoustiques.

Dans cet article, nous utilisons la famille standard des descripteurs audio, auxquels nous en ajoutons de nouveaux descripteurs issus d'une transformée en ondelettes dyadique. Puis, nous cherchons la façon optimale de les combiner, par optimisation sur les sons d'un ensemble d'apprentissage. Les descripteurs utilisés sont calculés pour chaque trame du signal (une trame est une portion de signal modulée par une fenêtre de type "Hamming" par exemple).

**Descripteurs temporels.** Nous considérons comme descripteurs temporels le nombre de passages par zéro (ZCR) noté $ZCR(t)$, et l'énergie totale $Energy(t)$ de la trame $t$.

ZCR est un paramètre déduit du nombre de fois où le signal change de signe dans la fenêtre d'analyse indiquant ainsi la fréquence dominante du signal. Il est calculé pour une trame $t$ de longueur $L$ comme suit :

$$ZCR(t) = \frac{1}{L}\sum_{\tau=1}^{L} |\text{sign}(\mathbf{s}_t(\tau+1)) - \text{sign}(\mathbf{s}_t(\tau))| \qquad (1)$$

Ce paramètre est étroitement lié au centroïde spectrale puisque il donne une mesure de la forme spectrale du signal dans une fenêtre.

L'énergie du signal, ou son volume, calculée sur une trame, est un indicateur pour détecter le silence et ensuite préciser la frontière d'un segment. C'est aussi un critère pour différencier par exemple un signal composé de plusieurs pics d'énergie d'un signal plus stable. Le volume dépend du gain du dispositif d'enregistrement et numérisation. Pour éliminer cette dépendance il est courant de normaliser sa valeur en fonction du volume maximum des trames adjacentes. Elle calculée pour une trame $t$ comme la puissance quadratique moyenne de l'amplitude du signal.

$$e(t) = \sqrt{\frac{1}{N_t} \sum_{i=0}^{N_t-1} s_t^2(i)} \quad (2)$$

avec $s_t(i)$ est l'amplitude du $i$ème échantillon de la trame $t$. $N_t$ est le nombre d'échantillons dans la trame $t$. Dans une échelle plus proche de la perception de l'oreille humaine, on obtient l'expression suivante de l'énergie :

$$e_{dB}(t) = 10 \cdot \log_{10} \sum_{i=0}^{N_t-1} s_t^2(i) \quad (3)$$

**Descripteurs fréquentiels.** Nous avons sélectionné le *Spectral Roll-off Point* noté $SRF(t)$ et Le centroïde spectral ($SC(t)$).

Le SRF est la fréquence de coupure au-dessous de laquelle se situe un certain pourcentage $TH$ de l'énergie du signal. Ce paramètre est considéré comme étant un indice de répartition du spectre de puissance du signal. Le Roll-off-point est plus grand pour les signaux ayant un spectre important au niveau des hautes fréquences. Il est calculé suivant la formule suivante :

$$SRF(t) = \max\{K \setminus \sum_{i=0}^{K} |S_t(f_i)|^2 < TH \sum_{i=0}^{N_t} |S_t(f_i)|^2\} \quad (4)$$

où $S_t(f_i)$ correspond à la composante spectrale de la trame $t$ à la fréquence $f_i$ et $N_t$ est le nombre d'échantillons dans la trame $t$. La valeur de $TH$ la plus utilisée est égale à 93%.

Le centroïde spectral (Spectral Centroïde (SC)) est le centre de gravité fréquentiel d'une densité spectrale de puissance. C'est la valeur de la fréquence qui partage le spectre de puissance du signal en deux parties d'égale énergie[1]. Il est formulé pour chaque trame $t$ comme suit :

$$SC(t) = \frac{\sum_{i=1}^{N_t} f_i S_t(f_i)}{\sum_{i=1}^{N} S_t(f_i)} \quad (5)$$

**Descripteurs cepstraux.** Nous avons considéré essentiellement les MFCCs (*Mel-Frequency Cepstral Coefficients*) qui sont extraits en appliquant la transformée en cosinus discrète sur le logarithme de l'énergie de banc de filtres Mel ; les LPCCs (*Linear Prediction Cepstral Coefficients*) calculés en utilisant une méthode d'autocorrélation ; et les PLPCCs (*Perceptual Linear Prediction Cepstral Coefficients*) qui prennent les caractéristiques de l'audition humaine en considération. Cette méthode projette le spectre de prédiction linéaire sur l'échelle nonlinéaire des fréquences de l'oreille.

[1]Son calcul se fait avec la même formule que le Roll-off point où $TH$=0.5.

**Descripteurs temps-échelle.** La transformée en ondelettes peut être adaptée à l'analyse de certains types de signaux parce que sa résolution fréquentielle diffère d'une bande de fréquences à l'autre. Cette transformation nous renseigne sur le contenu fréquentiel tout en préservant la localisation temporelle afin d'obtenir une présentation temps-échelle du signal. Les coefficients issus d'une décomposition dyadique en ondelettes discrètes (DWCs) seront utilisés directement comme descripteurs.

## 2.1 Combinaison des descripteurs

Comme indiqué précédemment, le performance de classification des sons environnementaux dépend de l'ensemble des descripteurs sélectionnés. Nous proposons ici de construire un vecteur descripteurs global incluant les descripteurs définis sur plusieurs espaces de représentation précédemment décrits.

Notre approche diffère du processus de sélection des descripteurs fréquemment utilisé et qui consiste à trouver à partir d'un ensemble de descripteurs un sous-ensemble contenant le maximum d'informations sur le signal. Les méthodes statistiques, tel que l'analyse en composant principal (PCA) qui maximise la variance entre les descripteurs sont souvent appliquées dans ce contexte. Dans ce papier, notre approche consiste à appliquer non pas une sélection des descripteurs mais plutôt des vecteurs. Il s'agit d'évaluer plusieurs vecteurs descripteurs séparément et de choisir les meilleurs candidats comme descripteurs de base par validation croisée. Nous procédons de façon incrémentale. Un vecteur ayant été sélectionné, d'autres descripteurs sont ajoutés. En vue de faciliter la construction empirique de ce vecteur, les descripteurs "fondamentaux" sont les descripteurs cepstraux et/ou ceux issus de la décomposition en ondelettes. Ce choix est justifié par le fait que les autres descripteurs calculés dans le domaine temporels– $ZCR(t)$ et $Energy(t)$– et ceux issus du domaine fréquentiel ($RF(t)$ et $SC(t)$) n'apportent pas suffisamment d'information garantissant la discrimination entre des sons à caractère impulsif.

## 2.2 Normalisation des descripteurs

On suppose pour l'instant que la combinaison optimale de descripteurs est connue.

**Algorithme 1: Extraction du vecteur des descripteurs**

- Soit $s(t)$ ($t = 1, \ldots, T$) le signal sonore traité. Pour chaque instant $t$, extraire les descripteurs décrits dans la section 2, et les réunir dans une série de vecteurs $x(t)$, de dimension $n$. Pour $k = 1, \ldots, n$, normaliser la composante $k$ des vecteurs $x(t)$, de façon à ce qu'ils soient de moyenne nulle et d'écart-type 1.
- Découper les instants $t = 1, \ldots, T$ en trois tranches selon la proportion 3-4-3, et calculer la moyenne des $x(t)$ sur chaque portion. Regrouper ces trois vecteurs en un vecteur descripteur global de dimension $d = 3n$, noté $\mathbf{v}$.

On le voit, le problème de classification des signaux sonores se ramène à celui de vecteurs dans un espace de dimension fixe $d = 3n$.

# 3 Classification en classes multiples par 1-SVM

## 3.1 SVM monoclasse

L'approche basée sur les 1-SVM monoclasse (1-SVM) [4] a été appliquée avec succès dans plusieurs problèmes d'apprentissage. Cette approche consiste à trouver une région de l'espace de volume minimal qui englobe la plupart des éléments d'un ensemble d'apprentissage. Plus précisément, soit $\mathcal{X} = \{x_1, \ldots, x_m\}$ un ensemble d'observations dans $\mathbb{R}^d$ où chaque $x_i$ est le vecteur descripteur d'un signal $\mathbf{s_i}$. L'objectif des 1-SVMs est d'estimer à partir des données d'apprentissage une fonction $\mathtt{f}_{\mathcal{X}} : \mathbb{R}^d \mapsto \mathbb{R}$ vérifiant que la plupart des éléments de $\mathcal{X}$ appartiennent à un ensemble $\mathcal{R}_{\mathcal{X}} = \{x \in \mathbb{R}^d \setminus \mathtt{f}_{\mathcal{X}} \geq 0\}$ de volume minimal. Ce problème est appelé estimation du volume minimal d'un ensemble de données (minimum volume set estimation (MVS)) [2], et la relation d'appartenance de $x$ à $\mathcal{R}_{\mathcal{X}}$ indique si cet elément présente oui ou non une similarité avec les éléments de $\mathcal{X}$. Ainsi, apprendre les régions $\mathcal{R}_{\mathcal{X}_i}$ de chaque classe ($i = 1, \ldots, N$), revient à apprendre les $N$ fonctions d'appartenance $\mathtt{f}_{\mathcal{X}_i}$. Etant donnée les $\mathtt{f}_{\mathcal{X}_i}$, un élément $x$ sera assigné à une classe suivant une certaine mesure de similarité. Le problème de l'estimation MVS est résolu par les SVMs-1 de la façon suivante. Premièrement, une fonction définie-positive appelée noyau $k(\cdot,\cdot)$ ; $\mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}$ doit être sélectionnée [4]. On a considéré un noyau RBF gaussien tel que $k(x,x') = \exp[-\|x-x'\|/2\sigma^2]$, où $\|\cdot\|$ désigne la norme euclidienne dans $\mathbb{R}^d$. Ce noyau induit un espace de redescription $\mathcal{H}$ à travers la fonction de projection $\phi : \mathbb{R}^d \mapsto \mathcal{H}$ définie par $\phi(x) \triangleq k(x,\cdot)$, où $\mathcal{H}$ est un espace de fonctions de Hilbert à noyau reproduisant (RKHS) muni d'un produit scalaire $\langle\cdot,\cdot\rangle_{\mathcal{H}}$. La propriété du noyau reproduisant implique que $\langle\phi(x),\phi(x')\rangle_{\mathcal{H}} = \langle k(x,\cdot),k(x',\cdot)\rangle_{\mathcal{H}} = k(x,x')$ ce qui rend l'évaluation de $k(x,x')$ est une opération linéaire dans $\mathcal{H}$. Dans le cas du noyau RBF gaussien, on a $\phi(x) \triangleq \langle\phi(x),\phi(x)\rangle_{\mathcal{H}} = k(x,x) = 1$, donc les données projetées seront placées dans une hypersphère $S_{(o,R=1)}$ de rayon un et de centre l'origine de $\mathcal{H}$.

L'approche à base de SVM-1 consiste à déterminer dans l'espace de redescription l'hyperplan $\mathcal{W}$ qui sépare la majorité des points de l'origine de l'hypersphère en se situant aussi loin que possible de cette origine. Etant donné que, l'image par $\phi$ de $\mathcal{R}_{\mathcal{X}}$ dans $\mathcal{H}$ est incluse dans la portion de l'hypersphère délimitée par $\mathcal{W}$, l'estimation MVS sera alors appliquée [2]. En pratique, soit $\mathcal{W} = \{h(\cdot) \in \mathcal{H} \backslash \langle h(\cdot), w(\cdot)\rangle_{\mathcal{H}} = 0 - \rho\}$, les paramètres $w(\cdot)$ et $\rho$ seront déterminés à partir du problème d'optimisation suivant :

$$\min_{w,\xi,\rho} \frac{1}{2}\|w(\cdot)\|^2_{\mathcal{H}} + \frac{1}{\nu m}\sum_{j=1}^{m}\xi_j - \rho \qquad (6)$$

avec (pour $i = 1, \ldots, m$)

$$\langle w(\cdot), k(x_j,\cdot)\rangle_{\mathcal{H}} \geq \rho - \xi_j, \text{ et } \xi_j \geq 0 \qquad (7)$$

où $\nu$ est un paramètre qui détermine la fraction des données qui peuvent se placer du mauvais côté de $\mathcal{W}$ (se sont les points appelés *outliers* qui n'apprtiennent pas à $\mathcal{R}_{\mathcal{X}}$) ; et les $\xi_j$ sont les variables de relâchement.

La solution de (6)-(7) est $w(\cdot) = \sum_{j=1}^{m} \alpha_j k(x_j,\cdot)$, telle que les $\alpha_j$ sont la solution du problème d'optimisation (sous sa formule duale) suivant :

$$\min_{\alpha} \frac{1}{2}\sum_{j=1}^{m}\sum_{j'=1}^{m}\alpha_j\alpha_{j'}k(x_j,x_{j'}) \qquad (8)$$

$$\text{avec} \quad 0 \leq \alpha_j \leq \frac{1}{\nu m}, \quad \sum_j \alpha_j = 1 \qquad (9)$$

Finalement, on va considéré la fonction de décision proposé dans [1] :

$$d(\mathcal{X},x) \;=\; -\log[\sum_{j=1}^{m}\alpha_j\, k(x,x_j)] + \log[\rho] \qquad (10)$$

et $\rho$ est calculé en utilisant le fait que $\mathtt{f}_{\mathcal{X}}(x_j) = 0$ pour les valeurs de $x_j$ qui se trouvent exactement sur la frontière, c-à-d, qui vérifient $\alpha_j \neq 0$ et $\alpha_j \neq 1/\nu m$. Ici, il est important de signaler que la majorité des $\alpha_i$ sont nuls (Ils correspondent aux $x_j$ qui sont à l'intérieur de la région $\mathcal{R}_{\mathcal{X}}$ et vérifiant $\mathtt{f}_{\mathcal{X}}(x_j) > 0$).

## 3.2 Classification multiclasse avec plusieurs SVM à une classe

Le principe de la classification multiclasse de signaux sonore peut être résumé comme suit.

**Algorithme 2: Classification de signaux sonores**

**Etape 1** Apprentissage du modèle
- Soit $N$ le nombre de classe considérées, et $m_i$, $i = 1, \ldots, N$ le nombre de signaux d'apprentissage de la classe $N$. Pour $i = 1, \ldots, N$,
  - Extraire les vecteurs de descripteur de chacun de ces signaux. Cela fournit un ensemble de données $\mathcal{V}_i = \{\mathbf{v}_{i,1}, \ldots, \mathbf{v}_{i,m_i}\}$.
  - Apprendre le 1-SVM à partir de $\mathcal{V}_i$.

**Etape 2** Classification d'un nouveau signal
- Extraire le vecteur de descripteurs $\mathbf{v}$ de dimension $d = 3n$.
- Calculer la distance $d(\mathbf{v},\mathcal{V}_i)$ pour $i = 1, \ldots, N$ en utilisant (10).
- Affecter le signal à la classe $\hat{i}$ vérifiant la condition $\hat{i} = \arg\max_{i=1,\ldots,N} d(\mathbf{v},\mathcal{V}_i)$

# 4 Validation expérimentale

La plupart des échantillons sonores utilisés dans nos expériences sont récupérés à partir de CD commerciaux [8]. Ils concernent la surveillance et sont regroupés en 9 classes (appels au secours, explosions, bris de verre, chutes, etc.) comme détaillé dans TAB. 2. La plupart sont impulsifs. Une caractéristique importante du corpus considéré est la grande variété inter-classe et intra-classe. Tous les signaux ont une résolution de 16 bits et une fréquence d'échantillonnage de 44100 Hz se caractérisant ainsi par une bonne résolution temporelle et une large bande fréquentielle.

En se basant sur des tests préliminaires, Nous avons remarqué que l'utilisation des dérivées première et seconde

TAB. 1 – Taux de reconnaissance pour plusieurs combinaisons des descripteurs.

| Descripteurs | Dimension $ns$ | Bonne classif. (%) |
|---|---|---|
| MFCC + Energy + Log_energy | 14 | 92.89 |
| MFCC + Energy + Log_energy + SRF + SC | 15 | 93.73 |
| PLPCC + Energy + Log_energy | 14 | 93.99 |
| LPCC + Energy + Log_energy + SRF + SC + ZCR | 17 | 93.33 |
| DWC + Energy + SRF + SC + ZCR | 115 | 86.80 |
| DWC + MFCC + Energy + Log_energy + SRF + SC + ZCR | 117 | 96.89 |

TAB. 2 – Classes de sons et nombres d'échantillons pour chaque classe dans la base utilisée pour l'évaluation du système

| Classes | Entraînement | Test | Totale |
|---|---|---|---|
| Cris de secours | 48 | 25 | 73 |
| Coups de fusils | 150 | 75 | 225 |
| Bris de verre | 58 | 30 | 88 |
| Explosions | 41 | 21 | 62 |
| Claquements de portes | 209 | 105 | 314 |
| Aboiement de chiens | 36 | 19 | 55 |
| sonneries de téléphones | 34 | 17 | 51 |
| Voix d'enfants | 58 | 29 | 87 |
| Machines | 40 | 20 | 60 |
| Totale | 674 | 341 | 1015 |

des vecteurs descripteurs a un effet négligeable sur les performances. De même, les descripteurs pris séparément sont peu discriminants.

Dans [3], il a été démontré que l'information représentée par des coefficients LPCCs est déjà saisie par les coefficients MFCCs qui sont plus expressifs. Nos expériences préliminaires ont confirmé cette conclusion. Ceci est également vrai pour des coefficients PLPCs. Par conséquent, nous n'incluons pas les coefficients LPCCs et PLPCs dans la combinaison quand les MFCCs sont déjà inclus. Pour les groupes de paramètres fortement redondants nous choisissons seulement ceux qui sont les plus représentatifs quand ils sont utilisés individuellement. [7] montre que l'ajout de paramètres temporels peut améliorer la classification. Ainsi, nous avons ajouté ZCR et l'énergie moyenne qui sont des représentations unidimensionnels. ZCR est étroitement lié à la fréquence fondamentale dans une trame. Dans le cas des sons environnementaux la fréquence fondamentale peut être semblable pour différentes classes ; c'est pourquoi le ZCR ne convient pas à la classification quand il est utilisé seul comme descripteur. A cause de leur faible dimension, les descripteurs temporels (ZCR et l'énergie moyenne) et les descripteurs fréquentiels (SRF et SC) ne permettent pas de représenter toute l'information utile dans les signaux à reconnaître. Néanmoins, ces descripteurs peuvent améliorer la qualité de la reconnaissance en les combinant avec les descripteurs de base présélectionnés. En général, les combinaisons impliquant les descripteurs contenant des informations sur les représentations spectrales et temporelles sont utiles, puisqu'elles combinent des informations provenant de deux domaines complémentaires.

En combinant plusieurs descripteurs, le vecteur composé tend à avoir une grande dimension surtout quand il s'agit d'introduire une centaine de coefficients issus de DWT. Ceci n'est plus un problème avec les SVMs qui sont des classifieurs peu sensibles à la dimension des vecteurs de descripteurs.

Il est important de signaler qu'un certain nombre de descripteurs ne permettent pas de distinguer entre certaines classes avec succès quand ils sont utilisés de façon individuelles. Les coefficients issus de la DWT n'assurent pas une bonne discrimination entre quelques classes. Ceci peut être en partie expliqué par le fait que les coefficients de DWT contiennent des informations sur les basses fréquences, et ils tendent à négliger les hautes fréquences. Ceci justifie notre utilisation de 12 coefficients MFCCs en plus des coefficients issus de DWT. Comme on peut le constater dans TAB. 1, ceci améliore de manière significative les résultats de classification.

# 5 Conclusion

Les résultats présentés montrent que le système proposé peut atteindre de très hautes performances de classification. En outre, contrairement à d'autres classificateurs statistiques classiques, il est capable de gérer des vecteurs descripteurs de grande dimension assurant ainsi une meilleure représentation d'une catégorie bien spécifique de sons.

# Références


[1] M. Davy, F. Desobry, A. Gretton, et C. Doncarli, "An online Support Vector Machine for Abnormal Events Detection," *Signal Processing*, vol. 86, no. 8, pp. 2009–2025, 2006.

[2] M. Davy, F. Desobry et S. Canu, Estimation of Minimum Measure Sets in Reproducing Kernel Hilbert Spaces and Applications. *ICASSP, Toulouse, France*, 2006.

[3] D. Mitrovic. *Discrimination and Retrieval of Environmental sounds.* Thèse, Vienna University of Technology, Décembre 2005.

[4] B. Scholkopf et A. J. Smola, *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press, Cambridge, MA, 2002.

[5] T. Hastie, R. Tibshirani et J. Friedman. *The Elements of Statistical Learning.* Springer, New-York, USA, 2001.

[6] D. M. J. Tax, *One-class classification.* Thèse, Delft University of Technology, Juin 2001.

[7] A. Dufaux, *Detection and recognition of Impulsive Sounds Signals.* Thèse, Faculté des sciences de l'Université de Neuchâtel, Switzerland, 2001.

[8] Leonardo Software, *http ://www.leonardosoft.com.* Santa Monica, CA 90401.