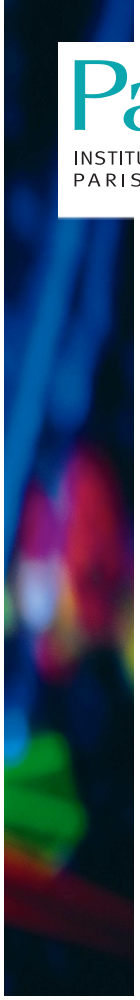
Apprentissage de représentations pour l'analyse de scènes sonores

Victor Bisot



ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUT OF TECHNOLOGY

TELECOM ParisTech

2018-ENST-0016

EDITE - ED 130

Doctorat ParisTech

T H È S E

pour obtenir le grade de docteur délivré par

Télécom ParisTech

Spécialité « Signal et Images »

présentée et soutenue publiquement par

Victor BISOT

le 16 Mars 2018

Apprentissage de représentations pour l'analyse de scènes sonores

Directeurs de thèse :

Slim ESSID et **Gaël RICHARD**

Jury :

M. Alain RAKOTOMAMONJY, Professeur, Université de Rouen, France	Rapporteur
M. Emmanuel VINCENT, Directeur de recherche, Inria Nancy, France	Rapporteur
M. Laurent DAUDET, Professeur, Université Paris-Diderot, France	Examinateur
Mme Annamaria MESAROS, Chercheure, Tampere University of Technology, Finlande	Examinatrice
Mme Jimena ROYO-LETELIER, Chercheure, Deezer, France	Examinatrice
M. Slim ESSID, Professeur, Télécom ParisTech, France	Directeur de thèse
M. Gaël RICHARD, Professeur, Télécom ParisTech, France	Directeur de thèse

THÈSE

Télécom ParisTech
École de l'Institut Mines-Télécom - Membre de ParisTech
46 rue Barrault 75013 Paris - (+33) 1 45 81 77 77 - www.telecom-paristech.fr

Résumé

Ce travail de thèse s'intéresse au problème de l'analyse des sons environnementaux avec pour objectif d'extraire automatiquement de l'information sur le contexte dans lequel un son a été enregistré. Ce domaine de recherche a connu un succès grandissant ces dernières années entraînant une rapide évolution du nombre de travaux et des méthodes employées. Nos travaux explorent et contribuent à plusieurs grandes familles d'approches pour l'analyse de scènes et événements sonores allant de l'ingénierie de descripteurs jusqu'aux réseaux de neurones profonds. Notre travail se focalise sur les techniques d'apprentissage de représentations par factorisation en matrices positives (NMF), qui sont particulièrement adaptées à l'analyse d'environnements multi-sources tels que les scènes sonores. Nous commençons par montrer que les spectrogrammes contiennent suffisamment d'information pour discriminer les scènes sonores en proposant une combinaison de descripteurs d'images extraits à partir des images temps-fréquence. Nous quittons ensuite le monde de l'ingénierie de descripteurs pour aller vers un apprentissage automatique des représentations. Nous entamons cette partie du travail en nous intéressant aux approches non-supervisées, en particulier à l'apprentissage de descripteurs par différentes variantes de la NMF. Plusieurs des approches proposées confirment l'intérêt de l'apprentissage de caractéristiques par NMF en obtenant des performances supérieures aux meilleures approches par extraction de descripteurs. Nous proposons ensuite d'améliorer les représentations apprises en introduisant le modèle TNMF, une variante supervisée de la NMF. Les modèles et algorithmes TNMF proposés se basent sur un apprentissage conjoint du classifieur et du dictionnaire de sorte à minimiser un coût de classification. Dans une dernière partie, nous discutons des liens de compatibilité entre la NMF et certaines approches par réseaux de neurones profonds. Nous proposons et adaptons des architectures de réseaux de neurones à l'utilisation de la NMF. Les modèles introduits nous permettent d'atteindre des performances état de l'art sur des tâches de classification de scènes et de détection d'événements sonores. Enfin nous explorons la possibilité d'entraîner conjointement la NMF et les paramètres du réseau, regroupant ainsi les différentes étapes de nos systèmes en un seul problème d'optimisation.

Abstract

This thesis work focuses on the computational analysis of environmental sound scenes and events. The objective of such tasks is to automatically extract information about the context in which a sound has been recorded. The interest for this area of research has been rapidly increasing in the last few years leading to a constant growth in the number of works and proposed approaches. We explore and contribute to the main families of approaches to sound scene and event analysis, going from feature engineering to deep learning. Our work is centered at representation learning techniques based on nonnegative matrix factorization, which are particularly suited to analyse multi-source environments such as acoustic scenes. As a first approach, we propose a combination of image processing features with the goal of confirming that spectrograms contain enough information to discriminate sound scenes and events. From there, we leave the world of feature engineering to go towards automatically learning the features. The first step we take in that direction is to study the usefulness of matrix factorization for unsupervised feature learning techniques, especially by relying on variants of NMF. Several of the compared approaches allow us indeed to outperform feature engineering approaches to such tasks. Next, we propose to improve the learned representations by introducing the TNMF model, a supervised variant of NMF. The proposed TNMF models and algorithms are based on jointly learning nonnegative dictionaries and classifiers by minimising a target classification cost. The last part of our work highlights the links and the compatibility between NMF and certain deep neural network systems by proposing and adapting neural network architectures to the use of NMF as an input representation. The proposed models allow us to get state of the art performance on scene classification and overlapping event detection tasks. Finally we explore the possibility of jointly learning NMF and neural networks parameters, grouping the different stages of our systems in one optimisation problem.

Table des matières

Liste des figures

Liste des tableaux

Abréviations

TFCT Transformée de Fourrier à court terme

CQT Transformée à Q constant (de l'anglais *Constant Q Transform*)

HOG Histogramme de gradient orienté (de l'anglais *Histogram of Oriented Gradients*)

SPD *Subband Power Distribution* en anglais

NMF Factorisation en matrices positives (de l'anglais *Nonnegative Matrix Factorization*)

SNMF Factorisation en matrices positives parcimonieuses (de l'anglais *Sparse Nonnegative Matrix Factorization*)

TDL *Task-driven Dictionary Learning* en anglais

TNMF de l'anglais *Task-driven Nonnegative Matrix Factorization*)

PCA Analyse en composantes principales (de l'anglais *Principal Component Analysis*)

KPCA Analyse en composantes principales à noyaux (de l'anglais *Kernel Principal Component Analysis*)

GMM Modèles de mélange gaussiennes (de l'anglais *Gaussian mixture model*)

HMM Modèles de Markov cachés (de l'anglais *Hidden markov models*)

SVM Machine à vecteurs supports (de l'anglais *Support vector machine*)

LR Régression logistique (de l'anglais *Logistic regression*)

DNN Réseau de neurones profond (de l'anglais *Deep neural network*)

MLP Perceptron multi-couches (de l'anglais *Multi-layer perceptron*)

CNN Réseau de neurones convolutifs (de l'anglais *Convolutional neural network*)

RNN Réseau de neurones récurrents (de l'anglais *Recurent neural network*)

LSTM *Long short-term memory* en anglais

GRU *Gated recurrent unit* en anglais

SGD Descente de gradient stochastique (de l'anglais *Stochastic gradient descent*)

ACSES Analyse computationnelle des scènes et événements sonores

KL Kullback-Leibler

IS Itakura-Saito

ER Taux d'erreur (de l'anglais *Error rate*)

Notations

a	Scalaire
a	Vecteur
A	Matrice
A[f,t]	Coefficient d'indice $[f, t]$ de la matrice **A**
$\mathbf{A}^T$	Matrice transposée de **A**
AB	Produit matriciel conventionnel : $(\mathbf{AB})[i, j] = \sum_r \mathbf{A}[i, r]\mathbf{B}[r, j]$
$\mathbf{a}_k$	Colonne d'indice k de la matrice **A**
$\mathbf{a}_{k:}$	Ligne d'indice k de la matrice **A**
$A \otimes B$, $\frac{A}{B}$, $A^{\otimes B}$	Produit, division et puissance matricielle terme-à-terme
$\|\|A\|\|_p$	Norme p : $(\sum_{f,t} \|A(f, t)\|^p)^{\frac{1}{p}}$
*	Produit de convolution
$\Pi_{\mathcal{A}}(x)$	Projection de x sur l'ensemble $\mathcal{A}$
$\nabla_{\mathbf{A}} f$	Gradient de la fonction f par rapport à **A**
$(f \circ g)(x)$	Fonction composée $f(g(x))$
$D_\beta(\mathbf{A}\|\mathbf{B})$	β-divergence entre **A** et **B**.
$\mathbf{a}_\Lambda$	Le vecteur dans $\mathbb{R}^{\|\Lambda\|}$ qui contient les entrées de $\mathbf{a} \in \mathbb{R}^F$ indexées par $\Lambda \subseteq \{1, ..., F\}$
Λ^c	Le complémentaire de l'ensemble Λ
$diag(\mathbf{a})$	Matrice diagonale formée par les éléments du vecteur **a**

Chapitre 1

Introduction

Sommaire

Imaginons-nous au restaurant avec des amis. La salle est remplie, le bruit des couverts et la superposition des conversations rendent déjà l'écoute de notre voisin difficile. Soudain, parmi les nombreuses sources sonores déjà présentes, nous distinguons un cri d'enfant à l'extérieur. Nous reconnaissons rapidement que ce cri est associé à un groupe d'enfants qui jouent et nous choisissons donc de l'ignorer sans pour autant arrêter notre conversation durant le processus. Ce genre de situation fréquente montre que nous pouvons, en un temps très court, identifier la nature de cette source sonore inattendue et analyser si elle constitue un danger. Malgré un environnement sonore très riche, le cerveau humain arrive à capter, identifier et analyser les différentes sources sonores. Nous sommes capables de détecter dans le bruit de fond, des événements d'intérêt qui vont conditionner notre prise de décision. Supposons maintenant que nous sommes au téléphone sans savoir où se trouve notre interlocuteur. Nous distinguons des bruits de pas, des conversations, une moto qui passe puis un klaxon. Nous supposons alors naturellement qu'il doit se trouver en centre ville. Par l'identification d'indices sonores, nous sommes capables en utilisant uniquement notre écoute, de caractériser la nature de l'environnement dans lequel nous nous trouvons.

On peut alors se demander s'il est possible de donner aux machines des capacités d'analyse de l'environnement sonore similaires, voire supérieures, à ce dont l'humain est capable. C'est une des grandes questions auxquelles tente de répondre le domaine de l'analyse computationnelle de scènes et événements sonores (ACSES), et auxquelles nous contribuons dans ce travail de thèse. Le principal objectif de l'analyse de sons environnementaux est d'extraire automatiquement de l'information sur le contexte et les objets sonores qui nous entourent à partir des signaux audio [Virtanen et al., 2017]. Avec des enregistrements comme seules données, nous tentons de répondre à deux grandes questions nous renseignant sur la nature et le contenu des environnements sonores. La première est : Dans quel contexte ce son a-t-il été enregistré ? En analysant le contenu audio, la machine va essayer d'identifier la nature de l'environnement dans lequel elle se trouve (dans une gare ou à la plage). La deuxième question est : Que se passe-t-il ? L'objectif est alors d'identifier les différents événements sonores présents dans un enregistrement ainsi que leur position temporelle. Pour résumer, on souhaite donner à la machine la capacité d'écouter et d'interpréter ce qu'elle écoute.

1.1 L'approche computationnelle de l'analyse de sons environnementaux

1.1.1 Idée générale et définition des problèmes

L'analyse computationelle des sons environnementaux est à l'intersection entre l'acoustique, le traitement du signal et l'apprentissage automatique. [1] L'interface entre la machine et le monde physique se fait par l'intermédiaire du microphone, chargé de capter et convertir l'ambiance sonore environnante en une forme d'onde. C'est à partir de cette forme d'onde, ou signal audio, que les méthodes d'ACSES extraient de l'information sur le contexte dans lequel ce son a été enregistré. Pour l'essentiel, cette information se caractérise par l'attribution de différentes étiquettes à un signal ou à plusieurs instants du signal. L'information recherchée se représente souvent sous la forme d'étiquettes sémantiques textuels décrivant différents concepts ou objets sonores caractérisant l'enregistrement. C'est dans la définition des catégories sonores constituant les étiquettes qu'apparaît la distinction importante entre *scènes* et *événements* sonores.

Un événement sonore fait référence à un son en particulier, produit par une source distincte. Un événement sonore possède généralement une durée limitée, souvent courte, définie par ses instants de début et de fin. Les étiquettes décrivant la nature d'un événement sonore sont souvent

1. *Machine Learning*

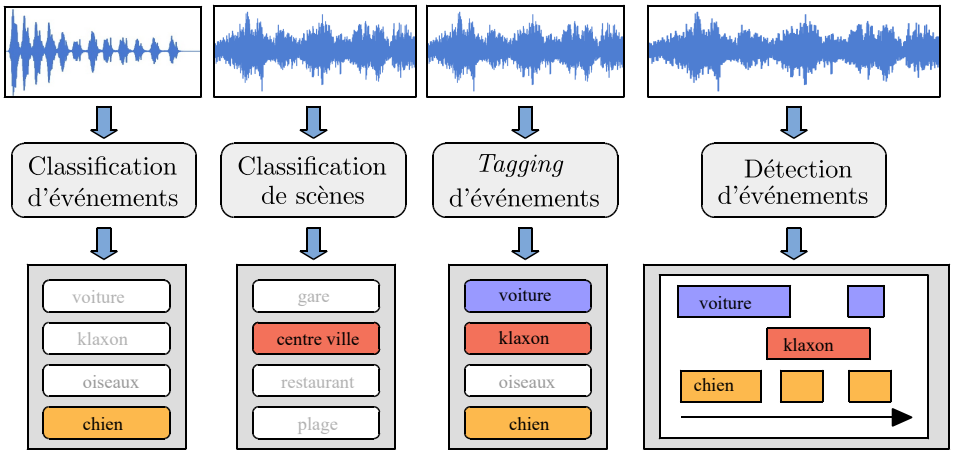


FIGURE 1.1 – Nature des prédictions recherchées pour les différentes tâches de l'analyse des sons environnementaux.

décomposés en deux parties, une première correspondant au type d'objet et une deuxième représentant l'action effectuée par ou sur cet objet. Cela donne des étiquettes d'événements telles que *voiture qui passe*, *personne qui marche* ou *eau qui coule*. Il peut arriver qu'une étiquette représentant un événement en regroupe en réalité plusieurs telle que *personne faisant la vaisselle* rendant la définition d'un événement utilisée en ACSES relativement large.

Une scène sonore est un mélange de plusieurs événements sonores provenant de différentes sources. On peut ainsi considérer un signal capté par un microphone dans un environnement multi-sources comme une scène sonore. En pratique les scènes sonores sont souvent associées à un type de lieu précis tel que *dans la rue* ou *au restaurant*. Par conséquent, les étiquettes décrivant les scènes sonores représentent des concepts qui englobent ou contiennent une variété d'événements de différentes natures.

Le choix et la définition de ces étiquettes est un des enjeux importants dans la définition des problèmes d'ACSES. Plusieurs travaux se sont attaqués à l'étude et à la proposition d'une taxonomie des événements et scènes sonores [Salamon et al., 2014; Gemmeke et al., 2017; Guastavino, 2018]. Si la variété des types d'événements dans la nature est immense, les méthodes d'ACSES se limitent souvent à la recherche d'un faible nombre de catégories, limité par le coût de construction de bases de données.

Par définition, l'analyse de sons environnementaux est une des nombreuses applications possibles de l'apprentissage automatique, et en particulier de la classification audio. L'ACSES se décompose en différentes sous-tâches de classification souvent séparées en tâches de *classification*, *tagging* et *détection*. Le point commun entre ces tâches étant l'attribution d'étiquettes à un enregistrement ou à différents instants d'un enregistrement. Le principe de ces trois tâches est illustré dans la figure 1.1.

Pour la classification, l'objectif est d'attribuer une seule étiquette à un segment audio. On a alors un problème de classification multi-classes standard où le signal est l'observation et les étiquettes sont définies parmi un ensemble d'étiquettes possibles. On fait souvent, à juste titre, la distinction entre classification de scènes et classification d'événements, en fonction des données et des étiquettes que l'on cherche à attribuer. Ensuite, les tâches de *tagging* peuvent être vues comme de la classification multi-labels, c'est-à-dire que l'on peut attribuer plusieurs étiquettes à une même observation. En pratique, pour l'ACSES, ces étiquettes correspondent à des événements présents dans le segment audio. Le *tagging* cherche uniquement à renseigner la présence ou non de certains événements dans le signal sans donner leur position précise dans le segment.

Enfin, les tâches de détection cherchent à fournir une séquence d'étiquettes donnant la catégorie ainsi que les instants de début et de fin de chaque événement d'intérêt dans l'enregistrement. On parle de détection d'événements monophonique lorsqu'un seul événement peut être détecté à chaque instant, et de détection polyphonique (ou avec recouvrement) lorsque plusieurs événements peuvent être présents à un même instant.

Nous illustrons sur la figure 1.2 les principales étapes qui constituent un système simplifié d'ACSES. Le premier grand bloc des systèmes d'ACSES est l'extraction de descripteurs dont le rôle est de représenter le signal brut par un ensemble de ses caractéristiques, dans l'objectif de faciliter son interprétation. Cette représentation peut être obtenue par des méthodes d'ingénierie ou d'apprentissage de descripteurs. Celle-ci est ensuite fournie à un classifieur qui a pour rôle d'apprendre une fonction associant la représentation des exemples sonores à leurs étiquettes respectives. Nous reviendrons plus en détails sur les différentes méthodes employées pour réaliser ces étapes dans le chapitre suivant.

1.1.2 Les grands défis de l'ACSES

Nous présentons ici, les deux difficultés de la recherche en ACSES qui ont particulièrement conditionné la direction de la recherche de la communauté et de ce travail de thèse.

Se distinguer du traitement de la musique et de la parole Les problèmes traités en analyse de sons environnementaux possèdent de nombreuses similitudes avec d'autres tâches établies du traitement automatique de l'audio. En effet, des domaines tels que le traitement de la parole et l'extraction d'information musicale (MIR), de l'anglais *music information retrieval*, traitaient des problèmes de classification et détection audio antérieurement à l'explosion de l'intérêt pour l'ACSES. Toutes ces approches possèdent alors de nombreux points communs, principalement dans le sens où l'on cherche à attribuer des étiquettes à des segments audio afin d'en décrire le contenu. Pour la parole, on trouve des tâches de classification similaires avec l'identification de locuteur, dont l'objectif est d'identifier la personne s'exprimant à partir du seul enregistrement. La reconnaissance de la parole est une forme de détection d'événements monophoniques, où l'on cherche à transcrire automatiquement la séquence de mots prononcés par le locuteur. De plus, on retrouve une importante variété de problèmes similaires en MIR, avec la classification d'instruments ou du genre musical, l'estimation de mélodie ou encore la détection du tempo. Motivées par ces ressemblances, les premières méthodes d'analyse de sons environnementaux se développent en s'inspirant des systèmes de l'état de l'art pour le traitement de la parole et de la musique. Cependant, il existe des différences clés entre ces différents domaines, qu'il convient de prendre en compte afin de proposer des systèmes d'ACSES efficaces.

La musique et la parole sont issues de phénomènes physiques bien connus que l'on est capable de modéliser afin d'en tirer parti pour guider les méthodes computationnelles. La structure harmonique des instruments de musique permet par exemple d'incorporer de la connaissance dans les méthodes d'apprentissage automatique classiques afin de les rendre plus performantes pour les tâches traitées. Or, le concept d'événements sonores englobe les signaux de musique et de parole en plus d'une immense variété d'autres catégories de sons de natures très différentes. S'il existe des sons aux procédés de production bien connus, la plupart d'entre eux sont soit la combinaison de plusieurs procédés, ou sont par nature plus similaires à du bruit ou à des impulsions. Il apparaît alors bien plus complexe de proposer des modèles caractérisant l'ensemble des événements sonores tout en capturant leur très grande variabilité. Il s'agit d'un des principaux enjeux et questions de recherche que pose l'ACSES. C'est-à-dire comment construire des systèmes capables de caractériser et discriminer des sons de natures et de procédés de production aussi variés ? Un des efforts de la communauté est de repenser l'extraction et l'apprentissage de représentations, inspi-

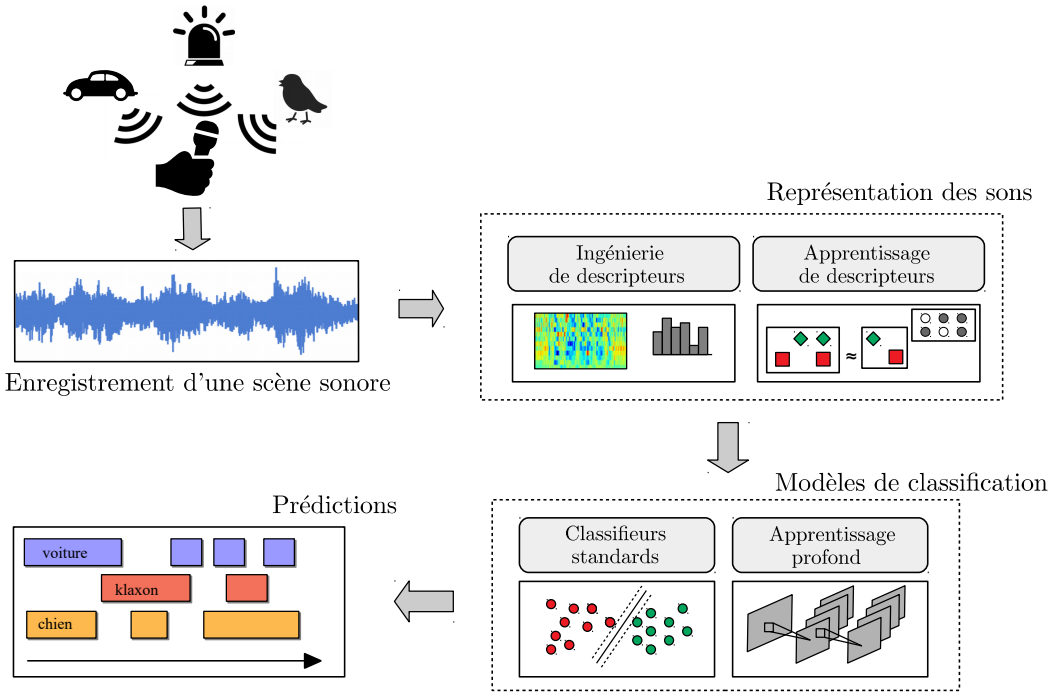


FIGURE 1.2 – Illustration des principales étapes d'un système d'ACSES.

rées du traitement de la parole et de la musique, afin de les adapter à l'analyse plus générale des scènes et d'événements sonores.

Concilier techniques avancées en apprentissage automatique et données disponibles Un des grands défis conditionnant les avancées en analyse de sons environnementaux est la difficulté que représente la construction de bases de données. La collection et l'annotation de données de scènes et événements sonores représentent un travail considérable limitant souvent la qualité et la taille des bases disponibles. En parallèle, les avancées récentes en apprentissage automatique, et en particulier sur l'apprentissage de modèles profonds,[2] sont souvent caractérisées par une efficacité conditionnée par le nombre de données d'apprentissage disponibles. Il devient alors crucial d'adapter les avancées en apprentissage automatique à la fois au problème mais aussi au type et à la taille des données traitées.

Un premier cas de figure est la collection et l'annotation manuelle de données de scènes et événements sonores. Il s'agit de la situation à laquelle la grande majorité des travaux d'ACSES se confrontent. Prenons l'exemple de la détection d'événements sonores. Si l'on veut avoir une base annotée en conditions réelles, une fois la scène enregistrée, un annotateur a la lourde tâche de placer manuellement les étiquettes ainsi que les instants de début et de fin de chaque événement remarquable présent dans la scène. Dans ce cas, la quantité d'audio annotée est linéaire dans le temps fourni par les annotateurs, ce qui a pour conséquences de fournir des bases de tailles relativement petites [Mesaros et al., 2016b, 2017a]. De telles bases limitent naturellement la taille et la complexité des modèles d'apprentissage pouvant être entrainés efficacement. Ce problème pose le défi de trouver des méthodes de classification et de détection performantes malgré la quantité

2. *Deep Learning*

limitée de données d'apprentissage.

L'alternative à la production manuelle de données audio annotées est la collection de données en masse sur des banques de sons en ligne. L'exemple à la fois le plus récent et le plus impressionnant est celui de la base Audioset [Gemmeke et al., 2017]. Il s'agit de plusieurs millions de segments audio extraits du site de vidéos en ligne *youtube* auxquels sont associés un ensemble d'étiquettes laissées par les utilisateurs. L'inconvénient d'une telle approche est que rien ne garantit que les étiquettes choisies par les utilisateurs du site soient la meilleure manière de décrire, ou représentent réellement le contenu audio de la vidéo. Contrairement au cas précédent, on sacrifie alors la qualité de l'annotation pour fortement augmenter la quantité de données. L'avantage de telles bases est qu'elles offrent des cas de figures nous rapprochant des ordres de grandeur des bases utilisées en traitement de l'image (avec Imagenet [3] par exemple). Cette récente évolution de la taille des bases rend alors plus réaliste l'application de modèles d'apprentissage profond particulièrement performants pour des tâches de vision par ordinateur. Ces avancées vont de pair avec de nouveaux défis. Comment adapter les modèles d'ACSES pour la prise en compte de données faiblement annotées ? Comment adapter les techniques modernes d'apprentissage automatique et en apprentissage profond aux particularités des problèmes de classification et détection de sons environnementaux ?

1.2 Applications

1.2.1 Applications

Donner aux machines la capacité d'écouter et d'analyser leur environnement sonore offre la possibilité de nombreuses applications industrielles et d'intérêt général. Un des principaux avantages de conditionner la prise de décision d'un système sur l'audio environnant est économique. Le coût d'un microphone ainsi que le coût de stockage de signaux audio est en majorité bien inférieur aux caméras de vidéo-surveillance. De plus, les propriétés de propagation des ondes sonores font qu'elles ne sont pas arrêtées par les surfaces. Si un mur se trouve entre le dispositif et la source sonore, le microphone peut tout de même capter le son, là où une caméra ne pourrait pas voir la source. Nous donnons quelques exemples d'applications concrètes des méthodes d'ACSES.

Tout d'abord, les avancées en ACSES s'avèrent prometteuses pour répondre aux défis des applications de fouille et d'indexation de données multimédias [Bugalho et al., 2009; Font et al., 2018]. La masse de contenu multimédia ajoutée sur le web chaque jour est immense et l'audio joue un rôle important souvent sous-estimé dans la caractérisation et l'indexation de ces contenus. Il existe de nombreuses banques de sons en ligne comptant sur l'annotation des utilisateurs afin de produire des méta-données aidant le parcours du contenu de la base. Les méthodes de classification de scènes et d'événements peuvent permettre dans ce cadre, d'ajouter des données non-annotées en plus grand nombre afin de créer des méta-données automatiquement. Ce genre d'outil peut également s'avérer très utile pour des applications policières [Serizel et al., 2016a; Geiger et al., 2013]. Avoir la possibilité de chercher un événement d'intérêt automatiquement (coups de feu, cri, explosion) dans une large collection de données permet de faciliter le travail des enquêteurs. De plus, on trouve des applications concrètes en bio-acoustique, où l'analyse des sons permet de faciliter l'archivage et la compréhension du comportement de certaines espèces [Stowell, 2018]. En particulier, des travaux ont montré la généralisation de certains systèmes d'ACSES à la classification d'espèces, en particulier pour les oiseaux [Salamon et al., 2017].

Une deuxième catégorie d'applications s'intéresse aux capteurs acoustiques dans l'objectif de donner la possibilité aux systèmes embarqués d'écouter leur environnement. Cela permet aux

3. www.image-net.org/

systèmes de réagir en fonction des sons qui les entourent en améliorant, par exemple, la qualité de la navigation et de l'interaction robotique [Chu et al., 2006]. Une autre part importante des applications concerne la surveillance acoustique pour des maisons et voitures connectées, ou tout simplement pour la surveillance dans les lieux publics. En complément des dispositifs d'alarme classiques pour le domicile, l'analyse de l'environnement sonore peut permettre de facilement détecter une vitre qui se brise ou une personne qui crie alertant ainsi le propriétaire d'un danger potentiel [Sigtia et al., 2016; Krstulović, 2018]. On peut imaginer le même genre de scénario pour une voiture intelligente, où entendre un klaxon ou un enfant au loin sans le voir doit avoir un impact sur la conduite d'une voiture autonome. D'importants dispositifs sont également mis en place dans la ville de New York afin de surveiller en temps réel les nuisances sonores et l'activité acoustique de la ville [Bello et al., 2018; Mydlarz et al., 2017; Salamon et al., 2014].

1.2.2 Un domaine en activité croissante

L'étendue des applications possibles ainsi que l'arrivée de bases de données de plus en plus larges attirent un nombre toujours plus important de chercheurs vers l'ACSES. Le nombre d'articles et de contributions en tout genre (bases de données, logiciel, livres) a très fortement augmenté ces trois dernières années, dans la même période où nous avons entamé ce travail de thèse. Cette attractivité provient en partie de l'approche par campagne d'évaluation suivant le succès de ces compétitions pour le traitement de la parole et des images, avec l'édition annuelle de multiples compétitions et du MIR avec MIREX. [4] Le domaine a connu une exposition importante dans la communauté de l'apprentissage automatique et du traitement du signal avec l'organisation des trois éditions des campagnes d'évaluations DCASE [Giannoulis et al., 2013; Mesaros et al., 2017a, 2016b]. [5] Le nombre d'équipes participant aux campagnes DCASE continue d'augmenter chaque année jusqu'à atteindre la participation de 75 équipes de recherche différentes pour l'édition 2017. Le DCASE permet de rassembler la communauté autour de nouvelles bases de données et d'un cadre d'application contrôlé facilitant une comparaison plus juste des différentes approches. Ce phénomène s'accompagne de l'organisation de sessions spéciales dans de multiples conférences du domaine (EUSIPCO, ICASSP, WASPAA), d'un workshop dédié au domaine (DCASE workshop) et de la sortie d'un livre [Virtanen et al., 2017] pour lequel nous avons contribué à l'écriture d'un chapitre [Serizel et al., 2018].

1.3 Structure du document et contributions

Notre travail de thèse s'articule principalement autour de l'extraction et l'apprentissage de représentations pour les scènes et événements sonores. Un des aspects clés pour traiter ce genre de tâche est de fournir une représentation adéquate à l'étape de classification des systèmes d'ACSES, de façon à faciliter la discrimination des différentes catégories de scènes ou d'événements. Nos travaux explorent plusieurs familles d'approches pour extraire des représentations adaptées et interprétables, allant de l'ingénierie de descripteurs jusqu'aux réseaux de neurones profonds, en s'arrêtant tout particulièrement sur les factorisations en matrices positives (NMF) [Lee et Seung, 1999], de l'anglais *nonnegative matrix factorization*. Les différentes approches proposées seront successivement évaluées sur plusieurs bases de données standards de classification de scènes et de détection d'événements. En particulier, la structure du document suit l'ordre chronologique de nos travaux de thèse mais aussi des familles d'approches mises en avant par la communauté au cours de ces dernières années. Ce manuscrit s'organise comme suit, où nous détaillons les contributions apportées dans chaque chapitre.

4. www.music-ir.org/mirex/wiki
5. Detection and Classification of Acoustic Scenes and Events

- Nous commençons, au chapitre 2, par une présentation des différentes approches en matière de détection et de classification des scènes et événements sonores. En partant de la compréhension humaine des sons environnementaux, nous détaillons les différentes familles d'approches pour l'extraction et l'apprentissage de représentations, avant d'introduire les stratégies de classification et de détection utilisées en ACSES.
- Le chapitre 3 constitue notre seul passage vers les techniques d'ingénierie de descripteurs dans ce manuscrit. Nous y discutons du choix important des représentations temps-fréquence de bas niveau utilisées comme matériel de base dans la majorité de nos approches. Ensuite, nous proposons comme première contribution, une combinaison de deux descripteurs d'images pour extraire les caractéristiques représentatives des scènes sonores. Nous combinons deux représentations complémentaires en regroupant les histogrammes de gradient orienté (HOG) [Rakotomamonjy et Gasso, 2015] et les histogrammes par bande de fréquences (SPD) [Dennis et al., 2013]. Nous justifions cette approche en clarifiant son intérêt pour décrire les aspects caractéristiques de certains environnements sonores ainsi qu'en la validant expérimentalement.
- Le chapitre 4 présente l'étude d'approches d'apprentissage de descripteurs par factorisation de matrices. Au lieu de compter sur des méthodes d'extraction de caractéristiques précises, nous allons directement apprendre les représentations en décomposant les représentations temps-fréquence de bas niveau. Nos contributions dans ce chapitre sont les suivantes : nous proposons un cadre d'application des techniques par factorisation de matrices simple et performant, en proposant des stratégies de mise en forme des spectrogrammes et de classification. Ensuite nous comparons l'intérêt de plusieurs variantes de la NMF et de l'analyse en composantes principales (PCA), de l'anglais *principal component analysis*, en discutant de leur capacité à traiter les problèmes que posent les tâches traitées et en les évaluant sur plusieurs bases de données.
- La chapitre 5 introduit une approche supervisée de la NMF pour la classification de scènes et la détection d'événements. L'objectif est d'apprendre automatiquement des dictionnaires à coefficients positifs qui vont minimiser un coût de classification associé au problème traité. Nous partons du modèle *Task-driven Dictonary Learning* (TDL)[Mairal et al., 2012] introduit pour des tâches de classification d'images, afin de l'adapter aux approches d'ACSES. Nous proposons le modèle TNMF comme variante positive de TDL, ainsi qu'un nouvel algorithme améliorant sa capacité de généralisation pour la classification de scènes. Le modèle TNMF garde les points forts des approches d'apprentissage de descripteurs par factorisation de matrices, tout en améliorant les performances de nos systèmes en adaptant les représentations au critère optimisé par le classifieur.
- Le chapitre 6 nous amène encore plus loin dans l'apprentissage supervisé en s'intéressant aux approches par réseaux de neurones profonds (DNN), de l'anglais *deep neural networks*. Nos contributions dans ce chapitre s'articulent autour de l'idée de se servir de représentations NMF comme entrées des DNN. Nous commençons par une discussion sur les points communs entre la NMF et les couches de DNN standards. Nous profitons des notions introduites pour montrer que le modèle TNMF est équivalent à un DNN à une couche cachée. Puis, nous proposons d'adapter des systèmes DNN état de l'art à la prise en compte de représentations NMF comme observations d'entrée des réseaux. Nous montrons que les systèmes proposés peuvent atteindre des performances état de l'art sur les bases de classification de scènes et de détection d'événements proposés. Enfin, nous montrons qu'il est possible d'apprendre conjointement un réseau de neurones et une NMF en repartant du modèle TNMF de façon à regrouper les différentes étapes de nos systèmes en un seul problème.
- En conclusion, nous revenons sur nos contributions afin de les mettre en perspective avec

l'état actuel du domaine. Nous suggérons ensuite quelques directions intéressantes que nous pourrions prendre pour continuer nos travaux.

1.3.1 Publications

Nous avons eu l'occasion durant cette thèse d'échanger avec la communauté en publiant et en présentant nos travaux dans plusieurs conférences internationales. Nous listons les différentes publications dans des journaux et conférences associées à ce travail de thèse.

Articles de revues et chapitres de livres

— V. Bisot, R. Serizel, S. Essid et G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," dans *IEEE Transactions on Audio, Speech, and Language Processing*, 2017.

— R. Serizel, V. Bisot, S. Essid et G. Richard, "Acoustic features for environmental sound analysis," dans *Computational analysis of sound scenes and events*, Springer, 2018.

Articles de conférences

— V. Bisot, R. Serizel, S. Essid et G. Richard, "Nonegative feature learning methods for acoustic scene classification," dans *Proc. IEEE International Workshop on Detection and classification of acoustic scenes and events (DCASE)*, 2017.

— V. Bisot, R. Serizel, S. Essid et G. Richard, "Leveraging deep neural networks with nonnegative representations for improved environmental sound classification," dans *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2017.

— V. Bisot, R. Serizel, S. Essid et G. Richard, "Overlapping sound event detection with supervised nonnegative matrix factorization," in *Proc. of International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

— R. Serizel, V. Bisot, S. Essid et G. Richard, "Supervised group nonnegative matrix factorization with similarity constraints and applications to speaker identification," dans *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

— R. Serizel, V. Bisot, S. Essid et G. Richard, "Machine listening techniques as a complement to video image analysis in forensics," dans *Proc. of International Conference on Image processing (ICIP)*, 2016.

— V. Bisot, R. Serizel, S. Essid et G. Richard, "Acoustic scene classification with matrix factorization for unsupervised feature learning," dans *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016.

— V. Bisot, S. Essid et G. Richard, "HOG and Subband Power Distribution Image Features for Acoustic Scene Classification," dans *Proc. European Signal Processing Conference (EUSIPCO)*, 2015.

Rapport technique

— V. Bisot, R. Serizel, S. Essid et G. Richard, "Supervised nonnegative matrix factorization for acoustic scene classification," dans *IEEE International Evaluation Campaign on Detection and Classification of Acousitc Scenes and Events (DCASE)*, 2016.

Chapitre 2

État de l'art

Sommaire

La quantité et la diversité des travaux d'ACSES ont fortement augmenté ces dernières années. Depuis le début de ce travail de thèse, le domaine a connu des transformations importantes, à la fois par le nombre de travaux, la diffusion de nouvelles bases de données, et dans la diversité des approches employées. La plus marquante de ces transformations est arrivée avec la généralisation des approches par apprentissage profond en traitement de l'audio, qui ont fortement modifié la manière de traiter les différentes tâches d'ACSES. Nous nous efforcerons dans ce chapitre de rendre compte des premières approches, qui sont nécessaires à la compréhension des difficultés associées à la tâche ainsi que pour comprendre l'état actuel du domaine. De plus, nous dégagerons certaines tendances majeures retrouvées dans les travaux de ces deux dernières années dans l'objectif de saisir les principales directions de recherche établies par la communauté.

2.1 Les étapes dans la construction d'un système d'ACSES

Les approches par apprentissage automatique constituent le cœur de l'activité en ACSES. Nous nous focalisons dans ce chapitre sur les tâches majeures du domaine : la classification et la détection de scènes et d'événements. Ces deux problèmes sont des cas particuliers parmi les nombreuses applications possibles de l'apprentissage automatique pour la classification. Les approches d'ACSES possèdent de nombreux points communs sur les architectures des systèmes et sur les techniques utilisées avec les autres applications de classification de données temporelles, en particulier pour le traitement de la parole et de la musique. A quelques exceptions près, les travaux et systèmes d'ACSES se décomposent en une série de grandes étapes computationnelles, allant du signal brut jusqu'à la prédiction des catégories. Nous proposons à la figure 2.1 un schéma décrivant l'enchaînement des grands blocs que l'on retrouve dans la majorité des systèmes d'ACSES. Dans cette section, nous présentons quelques généralités sur l'ordre et le contenu des systèmes de classification et de détection de sons environnementaux tout en indiquant l'organisation du reste du chapitre.

Tout d'abord, les différentes approches présentées incluent toutes une forme d'apprentissage supervisé et supposent donc la présence d'une base de données annotées. On se donne une base de données de N signaux audio $\{\mathbf{x}_1, .., \mathbf{x}_N\}$. Ces signaux peuvent être de longueur variable selon la tâche traitée. Ils peuvent être longs de plusieurs secondes pour la classification de scènes, ou de quelques dizaines de millisecondes pour la détection d'événements lorsque la décision se fait au niveau des trames temporelles. On se donne également un ensemble de C étiquettes dans $\{1, .., C\}$, où chaque entier est associé à unz étiquette textuelle représentant un type de scène ou d'événement. Alors, un vecteur d'étiquettes $y \in \{0, 1\}^C$ est associé à chaque signal dans la base, où $y_c = 1$ indique si l'étiquette c est présente dans le signal .

Le premier bloc d'un système d'ACSES contient les pré-traitements appliqués directement à la forme d'onde. Ceux-ci peuvent être simplement des étapes de filtrage et de normalisation pour homogénéiser la base de données, ou encore des augmentations afin d'accroitre artificiellement la taille de la base (voir section 2.7). La deuxième étape, une des plus importantes, est l'extraction de descripteurs associés à chaque signal. L'objectif est de transformer l'ensemble des signaux en un ensemble de suites de vecteurs de descripteurs $\{\mathbf{V}_1, .., \mathbf{V}_N\}$ avec $\mathbf{V} \in \mathbb{R}^{P \times T}$ où P est la dimension de cette nouvelle représentation et T le nombre de trames temporelles. Cette étape a le rôle central de fournir une représentation interprétable par l'étape de classification. Les stratégies d'extraction de descripteurs se décomposent en plusieurs familles. Elles peuvent s'arrêter à l'extraction d'une représentation temps-fréquence du signal qui est passée directement à l'étape de classification (voir la section 2.3.1 et le chapitre 3). D'autres approches, souvent appelées ingénierie de descripteurs, cherchent à extraire des caractéristiques permettant de représenter certains aspects particuliers des signaux (voire section 2.3.2). Enfin, la dernière est l'apprentissage de représentations dont

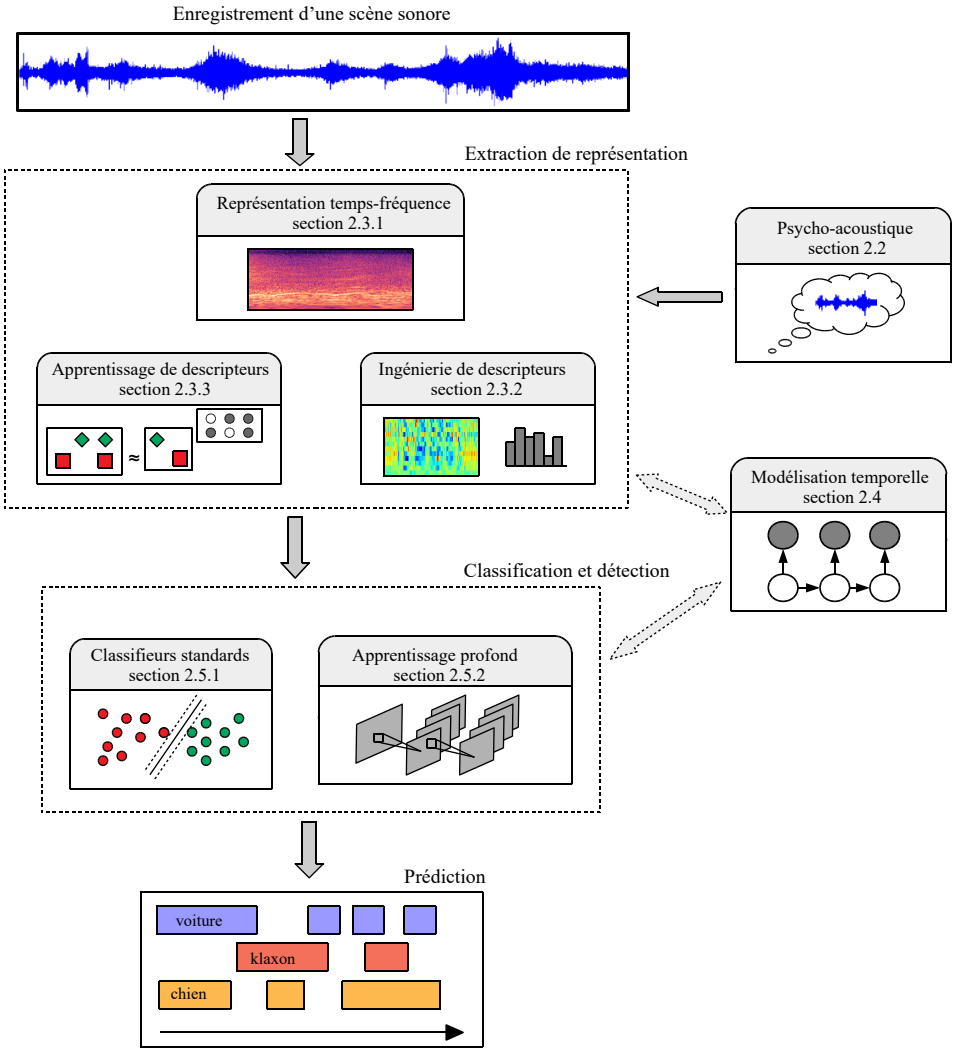


FIGURE 2.1 – Organisation du chapitre par principaux blocs des systèmes d'ACSES.

l'objectif est d'automatiquement apprendre une représentation latente facilitant l'interprétabilité des données pour les classifieurs (voir la section 2.4). Ensuite, nous discuterons dans la section 2.5 des différentes stratégies pour extraire ou modéliser l'information sur l'évolution et le contexte temporel.

Enfin, le dernier grand bloc computationnel que nous présentons dans la section 2.6 est l'étape de classification. Le rôle des classifieurs est, suivant une phase d'apprentissage, d'associer une ou plusieurs étiquettes à un segment du signal donné. Une première approche est de s'arrêter à l'utilisation de classifieurs standards dont le rôle est, pour les modèles discriminatifs, de séparer l'espace de la représentation d'entrée en régions associées aux différents étiquettes. La deuxième, plus récente, est l'apprentissage profond par l'utilisation de réseaux de neurones. Dans ce cas, par l'introduction de profondeur dans les modèles, l'étape de classification apprend par elle-même une succession de représentations intermédiaires amenant à une meilleure discrimination de l'ensemble des catégories. Pour finir, différentes stratégies agissant sur les données ou la nature des modèles sont souvent employées pour rendre les systèmes plus robustes. Les systèmes d'ACSES intègrent de plus en plus souvent des étapes d'augmentation de données, de fusion de classifieurs ou de prise en compte de l'imprécision des étiquettes (voir section 2.7).

2.2 S'inspirer de la compréhension humaine des sons environnementaux

2.2.1 Distinguer différents événements

On retrouve souvent des notions importantes de psycho-acoustique à la base de nombreux descripteurs audio fréquemment utilisés en analyse de sons environnementaux. L'idée étant d'essayer d'établir des liens entre notre perception des sons et des propriétés acoustiques du signal. Par exemple, les deux concepts parmi les plus utilisés à la base de la construction de systèmes et de descripteurs de traitement de l'audio sont la perception de la hauteur et de l'intensité des sons. Notre perception de la hauteur, ou le *pitch* en anglais, nous permet de placer les sons sur une échelle allant du grave à l'aigu. Pour des cas simples, ou monophoniques, le pitch est simplement lié à l'excitation de certaines zones de la membrane basilaire ainsi qu'à la périodicité de l'excitation du nerf auditif [Meddis et O'Mard, 1997; Oxenham et al., 2004]. Par exemple, les modèles expliquant notre compréhension de la hauteur sont à la base de nombreux algorithmes de détection de fréquence fondamentale [Boersma, 2006; De Cheveigné et Kawahara, 2002]. Ensuite, la sonie est ce qui nous permet de nous représenter la force des sons, de silencieux à fort. La sonie dépend à la fois de l'intensité du son et de son spectre, où le contenu spectral d'un son complexe influe sur notre perception de son intensité [Zwicker et al., 1991; Moore et al., 1997].

Toutefois, nous sommes tout de même capables de différencier deux événements sonores de même hauteur et de sonie similaire. On utilise alors la notion de timbre pour définir ce qui constitue la différence entre ces deux sons [Risset et Wessel, 1982]. Le timbre est une notion très riche, qui se définit par de multiples aspects tels que les modulations d'amplitudes, les répartitions de l'énergie en fréquence ou encore la durée d'attaque d'un son. Le timbre est depuis longtemps étudié pour comprendre comment nous différencions les instruments de musique à l'écoute [McAdams et al., 1995; Grey et Moorer, 1977]. Certains travaux similaires plus récents s'intéressent également au timbre des événements sonores urbains, en particulier pour étudier ce qui nous permet de qualifier la qualité de sons, tels que ceux émis par des portières de voiture ou des climatiseurs [Parizet et al., 2008; Susini et al., 2004; Lewicki, 2002]. En plus de la hauteur et de la sonie, ces notions ont inspiré de nombreux descripteurs de bas niveau utilisés en MIR mais aussi en analyse de sons environnementaux [Peeters, 2004] que nous mentionnons section 2.3.2.

En lien avec le timbre, des études se sont demandées quelles propriétés du signal nous permettent de percevoir certaines propriétés physiques de l'objet produisant le son. Par exemple, des travaux ont cherché des descripteurs acoustiques représentant le matériau de l'objet indépendamment d'autres propriétés de l'objet et de la nature de l'action produisant le son [Kunkler-Peck et Turvey, 2000]. Dans la même idée, d'autres se sont demandés si on peut prédire la taille et la forme des objets en interactions par la seule écoute du son qu'ils produisent. La majorité de ces travaux semble suggérer que les humains ont des capacités limitées pour percevoir de telles propriétés de l'objet rendant l'adaptation de ces études aux approches computationelles assez rares [Giordano et McAdams, 2006; Tucker et Brown, 2003]. En revanche, nous sommes beaucoup plus sensibles au type d'action produisant le son [Lemaitre et Heller, 2012]. A la différence des exemples précédents, caractériser l'action accorde plus d'importance à l'analyse de l'évolution temporelle du son, souvent informative quant à la nature de cette action. C'est entre autres pour cette raison qu'en détection d'événements sonores, une grande partie des bases de données se composent d'étiquettes représentant la combinaison de la nature de l'objet et de l'action responsable de la production du son.

Un autre aspect important de notre compréhension des événements sonores est notre capacité à reconnaître certains sons malgré des conditions très difficiles (filtrage, bruit, distance...). Ce phénomène attire la curiosité des chercheurs dans le but de trouver les caractéristiques acoustiques

dont nous nous servons pour catégoriser un son malgré de fortes dégradations. Une fois de plus, de telles notions ont surtout été étudiées pour la parole et la musique [Shannon et al., 1995] mais Gygi et al. [2004] proposent une première application aux sons environnementaux. Ces derniers montrent que les résultats sont hautement variables en fonction du type de son considéré, ne permettant pas de généralisation immédiate. Les sons se rapprochant plus du bruit (vent, pluie) ont eux aussi été étudiés et rassemblés sous la notion de "textures acoustiques" [Overath et al., 2010; McDermott et Simoncelli, 2011]. Il a été montré que notre perception de ces textures peut se modéliser facilement avec l'usage de seulement quelques statistiques [McDermott et Simoncelli, 2011], ce qui donne des premières pistes pour l'analyse de certaines scènes sonores.

2.2.2 Comprendre les scènes sonores

Nous sommes capables d'analyser et catégoriser la plupart des sons en fonction de plusieurs critères liés à notre système auditif. Mais que se passe-t-il quand une multitude de ces différents sons se superposent dans le temps et en fréquence pour former une scène sonore ? Même dans un environnement sonore très riche, nous sommes capables d'identifier certains événements d'intérêt nous permettant de caractériser la nature de la scène sonore (type de lieu, dangerosité, ambiance). L'analyse computationnelle de scènes sonores (CASA), de l'anglais *computational auditory scene analysis* [Wang et Brown, 2006], s'intéresse à la modélisation de notre compréhension des scènes sonores pour en isoler les différents éléments clés la constituant. Si idéalement nous serions capables de séparer et identifier chaque source sonore séparément, nous formons des groupes basés sur certains critères tels que le timbre, la position spatiale ou la nouveauté temporelle. Pour des applications d'ACSES comme la classification de scènes sonores, l'objectif est d'identifier la nature du lieu auquel correspond la scène sonore [Barchiesi et al., 2015]. Pour effectuer ce genre de tâches, il a été montré que l'Homme se base sur l'identification d'événements caractéristiques de la nature de la scène sonore (une moto pour une rue ou des vagues pour la plage) [Peltonen et al., 2001]. Durant notre écoute d'une scène sonore, nous collectons un certain nombre d'indices correspondant à l'occurrence de différents événements qui vont nous guider progressivement vers notre compréhension du contexte dans lequel nous nous trouvons.

2.2.3 Performances humaines comparées à celles de la machine

La plupart des résultats et hypothèses de psycho-acoustique que nous avons mentionnés sont appuyés expérimentalement par des tests d'écoute. Il est en revanche rare que l'Homme soit mis en compétition avec la machine sur le même jeu de données. Pour la classification de scènes, les premières études comparent les performances humaines sur des bases de données relativement petites [Peltonen et al., 2001; Eronen et al., 2006]. De plus, les systèmes auxquels sont confrontés les humains sont relativement simples comparés aux systèmes état de l'art actuels. Par exemple, Eronen et al. [2006] proposent une expérience où les humains atteignent un taux de reconnaissance d'environ 69% pour 64 catégories alors que la machine atteint seulement 58%.

Outre les comparaisons, ces études révèlent d'autres aspects intéressants et informatifs pour analyser notre compréhension des scènes sonores. En effet, la plupart des participants ont dit compter sur l'identification d'événements saillants pour qualifier le contexte audio. De plus, les participants ont eu besoin en moyenne de 13 secondes avant de prendre une décision, montrant que le temps nécessaire pour avoir suffisamment d'information pour caractériser la scène est relativement long comparé à la durée de la majorité des événements la constituant. Plus récemment une autre étude similaire a été conduite sur la base DCASE 2016 pour la classification de scènes [Mesaros et al., 2017b]. Cette base de données étant issue d'une campagne d'évaluation, les systèmes computationnels pris en compte sont parmi les approches récentes les plus efficaces. De plus, elle

avantage légèrement la machine par rapport aux études précédentes, car elle contient un nombre d'exemples significativement plus important. Les performances humaines atteignent seulement 54% en moyenne pour 15 catégories différentes, alors que les meilleurs systèmes obtiennent des taux de reconnaissance proche de 90%. Une des raisons expliquant cette différence est le faible nombre d'exemples d'apprentissage écoutés en moyenne par les sujets, 3 exemples par catégorie contre les 45 pour la machine. De même pour la classification d'événements isolés, Piczak [2015b] a montré que les performances humaines dépassaient la machine pour des systèmes de référence simples, pour discriminer 50 catégories d'événements.

2.3 Extraction de descripteurs

2.3.1 Représentations temps-fréquence

Les représentations temps-fréquence sont souvent utilisées en ACSES, tout comme pour la plupart des tâches de classification audio. Elles sont à la base d'une grande majorité de techniques d'extraction de descripteurs, de factorisation de matrices et d'apprentissage profond. Leur principal avantage par rapport au signal brut est de fournir une représentation plus parcimonieuse et plus aisément interprétable à la fois par l'humain et par la machine. L'intérêt et le choix de certaines de ces représentations temps-fréquence sont discutés plus en détails dans le chapitre 3, dans le cadre de l'extraction de descripteurs d'images.

Une majorité des représentations temps-fréquence s'obtient à partir de la transformée de Fourier à court terme (TFCT), représentant l'amplitude dans chaque bande de fréquence au court du temps en projetant le signal sur des ondes de Fourier. La TFCT possède par construction un axe fréquentiel espacé linéairement. Distribuer les fréquences ainsi rend l'interprétation de la TFCT parfois difficile, à la fois pour les Hommes, car elle retranscrit mal notre perception de la hauteur, et pour les machines, car elle nécessite un grand nombre de bandes de fréquence pour représenter de manière adéquate les sons en basses fréquences. Pour ces raisons, la majorité des travaux en analyse de sons environnementaux se base sur des représentations temps-fréquence motivées par notre perception auditive. Le premier exemple étant celui des bandes critiques [Fletcher, 1940] basées sur la largeur de bande de nos filtres auditifs dans la cochlée. L'échelle en bande rectangulaire équivalente (ERB), de l'anglais *equivalent rectangular bandwidth* [Glasberg et Moore, 1990], ou l'échelle de Bark [Zwicker et Terhardt, 1980] sont deux exemples d'échelles fréquentielles s'inspirant de la notion de bandes critiques.

Si ces échelles ont déjà été utilisées en ACSES, les représentations les plus largement plébiscitées sont les spectres Mel et la transformée à Q constant (CQT), de l'anglais *Constant-Q transform*. Les spectres Mel se basent sur notre sensation de la hauteur pour construire une échelle adaptée à notre perception des sons [Stevens et al., 1937]. Les CQT séparent l'échelle des fréquences géométriquement de sorte que le ratio Q entre la fréquence centrale d'une bande et la résolution fréquentielle soit constant [Brown, 1991]. Les spectres Mel et les CQT sont exploités pour plusieurs techniques d'extraction de descripteurs cepstraux ou de descripteurs d'images utilisés en ACSES [Dennis et al., 2013; Rakotomamonjy et Gasso, 2015; Battaglino et al., 2015]. De plus, ces deux représentations constituent le choix de représentation d'entrée pour une très grande majorité de systèmes par apprentissage de descripteurs ou réseaux de neurones profonds soumis aux éditions 2016 et 2017 du challenge DCASE. Il est par exemple rare de trouver un système de détection d'événements performant ne se servant pas du spectre Mel en entrée du système de classification.

2.3.2 Ingénierie de descripteurs

Lors des premières études sur la classification de scènes, l'objectif était souvent de trouver des descripteurs capables de caractériser les spécificités d'une scène sonore. Les premiers travaux se sont naturellement tournés vers les descripteurs audio performants pour d'autres applications (parole, musique...) dans le but de les combiner et de comparer leurs performances sur ce nouveau problème. Ces descripteurs sont conçus pour décrire des aspects précis du contenu temporel ou fréquentiel du signal, en supposant qu'ils permettent de différencier certaines catégories de sons présents dans les bases de données de sons environnementaux. Nous mentionnons ici quelques-unes des catégories les plus représentées.

Descripteurs temporels et fréquentiels Une large collection de descripteurs temporels et fréquentiels a été proposée par le passé. Il est usuel de faire référence à ces descripteurs comme descripteurs de bas-niveau. Parmi les descripteurs temporels les plus représentés on trouve l'enveloppe temporelle, le taux de passage par zéros, des coefficients d'auto-corrélation ou encore différents moments de la forme d'onde. Dans la même lignée, plusieurs descripteurs ont été proposés pour décrire des propriétés précises du contenu spectral du son, parfois reliés à des grandeurs perceptives. On trouve notamment l'enveloppe spectrale, les moments spectraux, la pente spectrale, le flux spectral et bien d'autres [Peeters, 2004]. Ces descripteurs de bas-niveau ont souvent été utilisés en ASCES, l'approche la plus répandue étant d'en combiner une large collection en complément d'autres types de descripteurs cepstraux [Chu et al., 2009; Geiger et al., 2013; Petetin et al., 2015].

Descripteurs cepstraux Les descripteurs cepstraux sont de loin les plus représentés en ACSES parmi les autres représentations issues de l'ingénierie de descripteurs. Ils permettent la décomposition du signal selon le modèle source-filtre pour modéliser le processus de production de la parole. Parmi les variantes de coefficients cepstraux disponibles, les plus populaires sont les coefficients cepstraux en bandes Mel (MFCC) [Davis et Mermelstein, 1980]. Ils sont calculés par la transformée en cosinus discrète inverse du logarithme de l'énergie des bandes Mel. La plupart des premières approches d'ACSES se basent sur l'extraction de MFCC pour caractériser les sons environnementaux [Peltonen et al., 2002; Aucouturier et al., 2007; Clavel et al., 2005]. Ce choix est principalement dû à la popularité des MFCC pour de multiples problèmes bien établis du traitement de la parole. Encore aujourd'hui, des travaux continuent de présenter des méthodes d'apprentissage automatique plus complexes s'apprenant à partir de coefficients cepstraux. Notamment, 7 parmi les 10 premiers systèmes à l'édition 2016 du challenge DCASE pour la classification de scènes incluent les MFCC dans leur représentation d'entrée [Eghbal-Zadeh et al., 2016; Marchi et al., 2016; Li et al., 2017]. La popularité des MFCC pour caractériser les sons environnementaux peut paraître relativement surprenante compte tenu du fait que certaines catégories d'événements sont très éloignées des propriétés de la parole. Les descripteurs MFCC, par le faible nombre de coefficients usuellement utilisés, ont des difficultés à représenter correctement les sons en plus haute fréquence en plus d'être une représentation relativement indépendante de la hauteur des sons. Avec la récente explosion des modèles d'apprentissage profond, les MFCC sont relégués au second plan. En effet, de moins en moins de travaux continuent de se servir des MFCC pour aller vers des systèmes apprenant les caractéristiques adéquates directement à partir des spectres Mel.

Descripteurs d'images Des approches plus récentes ont introduit l'utilisation de descripteurs issus du traitement d'images afin d'extraire de l'information à partir des représentations temps-fréquence d'une scène. L'idée derrière les descripteurs d'images est de représenter une scène ou un événement par une image correspondant à son spectrogramme. Pour la classification de scènes,

les premiers descripteurs d'images introduits sont les histogrammes de gradient orienté (HOG) [Rakotomamonjy et Gasso, 2015], de l'anglais *Histogram of Oriented Gradients*. Les HOG se construisent en calculant un histogramme des directions du gradient pour chaque pixel dans différents blocs de l'image. L'objectif est donc de modéliser l'évolution de l'information temps-fréquence dans le spectrogramme de la scène, ce qui permet par exemple de directement caractériser les sons d'accélération souvent présents dans les environnements urbains (dans la rue ou dans le bus). Dans une approche similaire, d'autres travaux reprennent les motifs binaires locaux (LBP), de l'anglais *Local Binary Pattern* pour la classification de scènes [Battaglino et al., 2015; Yang et Krishnan, 2017]. Les LBP calculent un code binaire de différents patchs de l'image représentant l'activation des pixels de valeurs supérieures à un seuil fixé par le pixel central du patch. Il s'agit d'un autre moyen de modéliser la distribution temps-fréquence contenue dans les spectrogrammes des scènes, en s'inspirant de l'analyse de texture en traitement d'image. On trouve également l'utilisation d'autres descripteurs très répandus en traitement de l'image tel que les SIFT, de l'anglais *scale-invaraint feature transform*, ou d'autres descripteurs créés pour la classification d'événements tels que les histogrammes d'énergie par bandes de fréquence (SPD), de l'anglais *Subband power distribution* [Dennis et al., 2014, 2013]. Nous reviendrons plus en détail sur le fonctionnement de ces différents descripteurs dans le chapitre 3, tout en proposant une combinaison des HOG et des SPD pour la classification de scènes sonores.

Représentations multi-échelles et dictionnaires d'ondelettes Comme alternative aux représentations temps-fréquence et aux MFCC, des dictionnaires d'ondelettes ont été proposés afin de représenter les signaux environnementaux. Les ondelettes permettent une représentation temps-échelle des signaux par la dilatation et la translation de brèves oscillations [Mallat, 1989]. Des premiers travaux ont démontré le potentiel d'approches similaires en projetant des signaux de sons environnementaux sur des dictionnaires d'ondelettes de Gabor par *matching poursuit* [Chu et al., 2009]. Pour la classification de scènes, Ren et al. [2017] se sont intéressés à la construction d'alternatives aux TFCT pour apprendre des réseaux de neurones, en utilisant différentes transformées en ondelettes. Parallèlement, d'autres auteurs ont exploité la transformée en *scattering* en obtenant des performances très prometteuses pour la classification d'événements [Salamon et Bello, 2015a; Lostanlen et Andén, 2016]. La transformée en *scattering* construit des représentations invariantes en enchaînant des opérations de transformée en ondelettes, de modules et de filtrages passe-bas [Mallat, 2012]. Ces opérations s'avèrent efficaces pour caractériser les non-stationnarités du signal ainsi que les textures sonores, deux aspects importants de l'analyse des sons environnementaux.

2.4 Apprentissage de descripteurs

L'ingénierie de descripteurs s'appuie sur des connaissances expertes afin de construire des procédés d'extraction de certaines propriétés importantes des données. Cette approche est particulièrement adaptée pour des applications où l'on connaît plus ou moins les propriétés du signal à modéliser afin de traiter la tâche cible. Comme nous l'avons vu, l'analyse des sons environnementaux demande l'étude d'une grande variété de sons aux moyens de production et aux propriétés potentiellement très différents. Dans ce contexte, de nombreux travaux se sont tournés vers des approches d'apprentissage de descripteurs. L'apprentissage de descripteurs, en particulier l'apprentissage non-supervisé, est l'un des grands axes de recherche en apprentissage automatique. L'objectif est de construire des modèles capables d'apprendre par eux-mêmes des caractéristiques ou des concepts représentatifs des données, en faisant souvent abstraction des catégories ou des étiquettes. Pour des tâches de classification et de détection telles que celles que nous traitons, l'intérêt principal de l'apprentissage de descripteurs est de représenter les données dans un nouvel

espace facilitant la discrimination des catégories par les classifieurs.

La majorité des approches trouvées en analyse de sons environnementaux s'appuie sur la construction ou l'apprentissage d'un dictionnaire. Ce dictionnaire contient une collection de vecteurs de base, aussi appelés composantes de base. Ces vecteurs de base sont définis dans le même espace que les observations et représentent des directions caractéristiques dans l'espace des données, de manière similaire aux composantes principales issues d'une analyse en composantes principales. L'objectif est ensuite d'apprendre une représentation latente des données en les projetant dans un espace transformé défini par les vecteurs de base. Cette représentation, souvent appelée matrice d'activations, projection ou code, constitue alors les descripteurs appris et utilisés dans la suite de la chaîne de classification ou de détection. Il existe une grande variété d'approches à la fois pour construire et apprendre le dictionnaire et pour obtenir les activations. Nous discuterons dans ce chapitre de quelques-unes des approches les plus utilisées en ACSES.

Les techniques de décomposition en ondelettes présentées section 2.3.2 se trouvent à la frontière entre les techniques d'extraction et d'apprentissage de descripteurs. En effet, ces approches passent par la construction à la main d'un dictionnaire d'ondelettes, lesquelles sont utilisées par la suite pour représenter le signal. Dans ce cas, le dictionnaire n'est pas appris automatiquement à partir des données. En revanche l'objectif est le même : représenter les signaux traités dans un espace transformé facilitant leur interprétation. Les techniques que nous présentons dans cette section apprennent les vecteurs de base constituant le dictionnaire automatiquement à partir des données, soit avec des modèles probabilistes, soit par factorisation de matrices ou par réseaux de neurones profonds.

2.4.1 Approches par sacs de descripteurs

Les approches par sacs de descripteurs, de l'anglais *Bag of features*, ont été introduites et popularisées pour l'analyse de documents et d'images. Une variante avancée de ces descripteurs se base sur la modélisation des données par un modèle de mélange de gaussiennes (GMM), de l'anglais *Gaussian Mixtures Models*. Pour l'analyse de sons environnementaux, les GMM ont principalement été utilisés comme classifieur ou comme modèle pour les probabilités d'émission des modèles de Markov cachés (HMM). Plus récemment des travaux s'en servent également pour des techniques d'apprentissage de descripteurs telles que les sacs de descripteurs. Après avoir appris un GMM sur les observations, Ye et al. [2015] proposent d'extraire une nouvelle représentation par une étape d'encodage utilisant des vecteurs de Fischer. Plinge et al. [2014] apprenent un GMM par classe et construisent un *super-vecteur* représentant la probabilité d'une observation pour chaque gaussienne des mélanges appris. Dans la même idée, [Eghbal-Zadeh et al., 2016] introduisent l'utilisation des *I-vecteurs* pour la classification de scènes. Popularisés pour la reconnaissance de locuteurs, les *I-vecteurs* adaptent les données à un modèle du monde par GMM pour construire un super-vecteur avant de le réduire par analyse factorielle. Les *I-vecteurs* ont notamment obtenu des performances remarquables à l'édition 2016 du challenge DCASE [Mesaros et al., 2016b].

2.4.2 Factorisation en matrices positives

Les méthodes de factorisation de matrices sont parmi les approches les plus représentées pour l'apprentissage de descripteurs en analyse de sons environnementaux. On se donne une matrice de données $\mathbf{V} \in \mathbb{R}^{F \times N}$ de N exemples $[\mathbf{v}_1, ..., \mathbf{v}_N]$ de dimension F. De manière générale, les différentes techniques de factorisation de matrices peuvent souvent se résumer à l'approximation de $\mathbf{V}$ telle que :

$$\mathbf{V} \approx \mathbf{WH}, \tag{2.1}$$

où $\mathbf{W} \in \mathbb{R}^{F \times K}$ est le dictionnaire appris contenant K vecteurs de base et $\mathbf{H} \in \mathbb{R}^{K \times N}$ la matrice d'activations correspondant aux projections des données sur $\mathbf{W}$. Cette décomposition se fait la plupart du temps en cherchant la solution d'un problème d'optimisation ayant comme critère un coût d'attache aux données, dit aussi coût de reconstruction. C'est-à-dire que l'on cherche a minimiser une distance ou une divergence $D(\mathbf{V}\|\mathbf{WH})$ entre les données $\mathbf{V}$ et l'approximation $\mathbf{WH}$.

Pour le traitement de l'audio en général et particulièrement pour l'analyse de sons environnementaux, les factorisations en matrices positives (NMF) [Lee et Seung, 1999] constituent de loin l'approche la plus plébiscitée. En supposant que nous avons des données $\mathbf{V}$ dans $\mathbb{R}_{+}^{F \times N}$, le problème est alors le même que donné dans l'équation (4.1) dans laquelle on contraint $\mathbf{W}$ et $\mathbf{H}$ à être à coefficients positifs. La NMF permet de modéliser les données comme une superposition de vecteurs de base et fournit ainsi une décomposition additive des données. Cette propriété s'est montrée particulièrement intéressante pour traiter des données audio constituées de la superposition de plusieurs sources. En effet, la NMF est un des outils les plus efficaces pour la séparation de sources audio. Ainsi, sa capacité à fournir des représentations interprétables pour des données multi-sources en fait également un modèle attrayant pour l'analyse des sons environnementaux. La NMF comme technique d'apprentissage de représentations est à la base d'une grande partie de nos contributions présentées dans cette thèse. Nous présenterons le modèle et son optimisation plus en détail dans le chapitre 4, dans le cadre de notre étude de l'apprentissage non-supervisé de descripteurs pour la classification et détection d'événements sonores.

En analyse de sons environnementaux, la NMF joue la plupart du temps un rôle d'apprentissage de descripteurs. La décomposition NMF se fait majoritairement à partir de représentations temps-fréquence des scènes ou des événements sonores, le dictionnaire est alors constitué de représentations fréquentielles de base. Les premières applications de la NMF pour la classification de sons environnementaux ont la particularité de se servir des dictionnaires directement comme descripteurs. Dans ce cas, une NMF est appliquée au spectrogramme de chaque exemple individuellement. Chaque exemple est alors représenté par le dictionnaire obtenu lors de la décomposition de son spectrogramme [Cauchi, 2011; Benetos et al., 2012]. Depuis, il est plus fréquent de décomposer l'ensemble des données pour apprendre un dictionnaire commun à l'ensemble de la base ou à chaque catégorie. Les descripteurs NMF sont alors obtenus en projetant les données sur ce dictionnaire [Ghoraani et Krishnan, 2011; Komatsu et al., 2016b; Rakotomamonjy, 2017].

Cependant, la NMF est rarement directement utilisée dans sa formulation initiale. Il est commun, tel que nous l'avons fait dans nos travaux [Bisot et al., 2017b], d'ajouter des contraintes ou des régularisations sur la parcimonie des activations afin d'améliorer la qualité et l'interprétabilité des décompositions. On trouve également l'utilisation de la NMF convolutive [Smaragdis, 2004] permettant d'inclure du contexte temporel dans la décomposition en apprenant un dictionnaire de tranches temps-fréquence [Benetos et al., 2012; Komatsu et al., 2016a].

Certains travaux de détection d'événements sonores introduisent une utilisation particulière de la NMF, où la décomposition est utilisée à la fois pour l'apprentissage de représentation et comme outil de détection. Le dictionnaire est alors constitué de vecteurs de base, chacun associé à une catégorie d'événements. Puis, en projetant la séquence d'observation sur le dictionnaire, on obtient la séquence d'activations de ses différentes composantes au cours du temps. Après quelques étapes de post-traitement sur la matrice d'activations, on peut directement lire l'activation de chaque catégorie d'événements en seuillant les coefficients de la matrice [Dikmen et Mesaros, 2013; Mesaros et al., 2015; Benetos et al., 2016; Bui et al., 2016; Zhou et Feng, 2017; Benetos et al., 2017]. Une telle utilisation de la NMF se rapproche plus des applications de séparation de sources, où chaque catégorie d'événements à détecter est considérée comme une source différente dont la présence est obtenue par la matrice d'activations. Afin d'améliorer l'efficacité de la détection, l'évolution temporelle des activations peut aussi être modélisée par des HMM appris conjointement avec la décomposition résultant en une amélioration de la qualité de la détection [Benetos et al., 2016].

Une partie des travaux sur la NMF que nous venons de mentionner ne se limite pas au cadre non-supervisé. Il existe une variété d'approches pour intégrer de la connaissance sur les étiquettes afin d'améliorer les decompositions NMF [Rakotomamonjy, 2017; Komatsu et al., 2016a; Mesaros et al., 2015] ainsi que dans nos travaux [Bisot et al., 2017a]. Nous reviendrons sur les différentes approches et notamment sur celles exploitant des factorisations supervisées de matrices dans le chapitre 5.

2.4.3 Apprentissage de descripteurs par réseaux de neurones profonds

Les récentes avancées en termes d'apprentissage et de puissance de calcul ont donné lieu à la généralisation de l'utilisation des réseaux de neurones profonds. Ils constituent maintenant l'état de l'art dans de nombreuses applications du traitement du signal audio. Ces modèles profonds ont principalement le rôle de classifieur supervisé, comme nous le verrons section 2.6.3. Cependant ils peuvent aussi être détournés en outils d'apprentissage de descripteurs performants. L'approche la plus courante est d'apprendre un réseau de neurones profond intégrant une couche dite de *bottleneck*. Cette couche s'insère avant la couche de classification, et elle est souvent de taille inférieure au restant des couches du modèle. Une fois le réseau appris de manière supervisée, on retire la couche de classification et on se sert du réseau jusqu'à la couche *bottleneck* comme extracteur de descripteurs. Cette catégorie d'approche est à la base de quelques systèmes de classification de scènes sonores, où des réseaux de neurones convolutifs (CNN), de l'anglais *convolutional neural networks*, sont appris de manière supervisée pour l'extraction de descripteurs. Ces descripteurs sont par la suite combinés à d'autres représentations pour être traités par d'autres blocs de la chaîne de classification [Mun et al., 2017; Y.Hang et Park, 2017].

Il est également possible d'apprendre des réseaux de neurones profonds de manière non-supervisée grâce aux auto-encodeurs. Ces réseaux sont construits de façon à fournir en sortie une estimation des données d'entrée et sont appris en minimisant un coût de reconstruction. Les applications des auto-encodeurs sont moins fréquentes en analyse de sons environnementaux [Xu et al., 2017a; Amiriparian et al., 2017]. Par exemple, Amiriparian et al. [2017] ont construit un auto-encodeur à partir de réseaux récurrents de façon à apprendre une représentation vectorielle à partir de séquences d'observations.

2.5 Modélisation et intégration temporelle

La modélisation de l'évolution temporelle de l'information joue souvent un rôle central en analyse de sons environnementaux. En effet, la classification de scènes et d'événements sonores demande de classifier des séquences d'observations. Dans le même temps, la détection d'événements nécessite de trouver les instants de début et de fin des événements présents dans la scène. Dans l'organisation de ce chapitre, nous avons choisi de placer cette étape entre l'extraction de descripteurs et la classification, bien que la modélisation temporelle ne constitue pas forcement une étape à part entière. Elle est souvent intégrée et prise en compte directement dans les modèles de classification ou d'extraction de représentations.

La manière la plus simple de traiter l'aspect temporel des données est d'avoir recours à diverses statistiques représentant la distribution temporelle des descripteurs associés aux signaux à classifier [Joder et al., 2009]. Si on se contente souvent de la moyenne à travers tout l'exemple, certains travaux ont étudié l'ajout de moments d'ordres supérieurs [Geiger et al., 2013; Salamon et Bello, 2015a; Krijnders et Holt, 2013]. Pour aller plus loin, Roma et al. [2013] ont proposé avec succès l'application de la *Recurrence Quantification Analysis* (RQA) en anglais, une méthode inspirée de la théorie du chaos permettant d'analyser les motifs récurrents dans la séquence d'observations. Une autre alternative très courante est simplement de classifier les descripteurs

par trame et d'effectuer une intégration tardive. Par un système de vote ou d'opérations sur la séquence de probabilités en sortie de classifieur, l'intégration tardive permet d'obtenir une décision pour l'ensemble de la séquence d'observations.

L'évolution de l'information temporelle peut également se retrouver directement modélisée par certains descripteurs. En particulier, la majorité des descripteurs d'images se compose d'histogrammes de caractéristiques locales [Rakotomamonjy et Gasso, 2015; Dennis et al., 2013; Battaglino et al., 2015]. Ces approches se sont montrées surtout efficaces pour la classification de scènes où la distribution de l'information dans le temps est plus importante que l'ordre dans lequel les événements se produisent. De même, certaines techniques de factorisation de matrices prennent en compte une modélisation temporelle directement dans la décomposition. C'est le cas par exemple de la NMF convolutive et de ses équivalents probabilistes [Benetos et al., 2012; Komatsu et al., 2016a]. En revanche, les variations temporelles apprises par ces modèles restent relativement locales et demandent souvent une étape d'intégration temporelle supplémentaire. Ce n'est pas le cas des auto-encodeurs récurrents qui, par le choix d'unités appropriées dans les couches du réseau, permettent d'apprendre des vecteurs descripteurs représentant des séquences d'observations de longueur variables [Amiriparian et al., 2017].

Enfin, dans la majorité des cas, la modélisation de l'évolution temporelle de l'information dans les scènes et événements sonores se fait durant l'étape de classification. De nombreux modèles de détection ou de classification sont capables de tirer parti du contexte temporel pour prendre des décisions locales ou globales. Ces modèles incluent les HMM, les réseaux convolutifs ou encore les réseaux récurrents dont nous discuterons plus en détail dans la section suivante.

2.6 Classification et détection

Pour les systèmes récents d'analyse de sons environnementaux, l'étape de classification constitue de moins en moins souvent une étape à part entière. L'arrivée de réseaux de neurones profonds dans le domaine se traduit par la domination de modèles capables de conjointement apprendre des représentations, la modélisation temporelle et de classifier les observations. Cependant, avant l'arrivée de telles techniques, les premiers travaux se focalisaient davantage sur l'étude de représentations appropriées pour l'analyse de scènes et d'événements sonores. Dans ce cadre, le choix du classifieur constitue rarement une contribution mais il est vu comme un outil permettant d'évaluer la qualité des représentations proposées. Nous commençons par brièvement mentionner les stratégies de classification standards employées dans le domaine avant de discuter de la récente domination des réseaux de neurones profonds pour l'ensemble des tâches traitées.

2.6.1 Classification standard

De nombreux travaux s'arrêtent à l'utilisation de stratégies de classification relativement simples qui se focalisent principalement sur la représentation de sons environnementaux. Bien souvent il s'agit de variantes de classifieurs linéaires, dont l'objectif est simplement de trouver des frontières linéaires entre les différentes catégories dans l'espace des descripteurs. Lors de l'étude comparative de descripteurs, il est commun de retrouver l'utilisation de techniques telles que les K plus proches voisins (KNN), de l'anglais *K-nearest neighbors*, de régressions logistiques ou de machines à vecteurs supports (SVM), de l'anglais *support vector machine*, dans leur version à noyaux [Rakotomamonjy et Gasso, 2015; Geiger et al., 2013; Chu et al., 2009]. Plus rarement, certains travaux ont privilégié les arbres de décision [Li et al., 2013] ou les forêts aléatoires (*random forests*) [Olivetti, 2013; Salamon et Bello, 2015a]. Même à l'ère de l'apprentissage profond, il reste relativement fréquent de trouver des systèmes performants préférant des SVM [Mun et al.,

2017; Y.Hang et Park, 2017] pour classifier des représentations préalablement obtenues par des réseaux de neurones profonds. L'objectif est alors d'appuyer la qualité de la représentation proposée en montrant qu'elle amène à de bonnes performances même avec des classifieurs linéaires. C'est dans cette même logique que, pour une grande partie des travaux présentés, nous nous baserons sur une régression logistique linéaire pour classifier les descripteurs appris par les techniques de factorisation de matrices proposées.

2.6.2 Particularités de la détection d'événements

Outre l'apprentissage profond, nous avons discuté une des familles d'approches très populaires pour la détection d'événements avec la NMF dans la section 2.4. L'autre tendance majeure parmi les premiers travaux de détection d'événements était de traiter le problème avec des HMM. Les HMM sont alors utilisés dans l'objectif de modéliser la dynamique des sons, par la construction d'un modèle probabiliste supposant des relations entre les trames successives de l'audio. Chaque observation dépend alors d'un état caché qui peut par exemple être la présence ou l'absence d'un événement pour la détection monophonique [Mesaros et al., 2010]. Ces états cachés peuvent aussi représenter la dynamique des événements sonores en séparant l'occurrence d'un son en trois phases : début, milieu et fin. Les HMM ont le premier défaut de ne pas être capables de correctement modéliser les dépendances à long terme. Certaines variantes des HMM existent pour pallier cette limitation mais n'ont pas été explorées pour la détection d'événements. Les HMM demandent également la création d'astuces pour traiter la détection d'événements polyphoniques, où plusieurs événements peuvent être présents simultanément à un instant donné. Cela peut se faire en effectuant plusieurs passages de modèles HMM sur les données [Heittola et al., 2010] ou en créant autant d'états que de combinaisons possibles d'étiquettess [Stowell et Clayton, 2015]. Les premiers modèles HMM, ayant des GMM appris sur des MFCC comme modèle pour les probabilités d'émissions, ont rapidement été surpassés par d'autres approches par NMF ou par apprentissage profond [Cakir et al., 2015b]. Cependant, les HMM restent utilisés dans certains systèmes, ils sont alors appris conjointement avec des modèles de réseaux récurrents [Hayashi et al., 2017] ou des factorisations de matrices [Benetos et al., 2012] afin de régulariser et modéliser la dynamique des sorties ou des activations de ces modèles.

Les inspirations du traitement de la parole amènent certains travaux à discuter de la possibilité d'apprendre des modèles de langage pour les sons environnementaux [Benetos et al., 2018]. Ces derniers permettraient de modéliser la périodicité de certains sons ou la co-occurrence de différents événements afin d'améliorer les systèmes de détection. Il existe certaines scènes sonores où la séquence d'événements peut être prévisible, comme la succession d'alertes sonores, de la fermeture des portes et de l'accélération du train dans une station de métro. En revanche pour des scènes plus riches comme en centre ville, il n'existe que très peu de liens entre les différents événements : par exemple il y a très peu de chances qu'un chant d'oiseau soit corrélé avec le passage d'une voiture. Les modèles acoustiques du langage ont surtout inspiré des approches modélisant la co-occurrence des différents événements pour la détection polyphonique par l'utilisation d'une analyse sémantique probabiliste latente [Mesaros et al., 2011] ou de systèmes linéaires dynamiques [Benetos et al., 2017].

2.6.3 Approches par réseaux de neurones profonds

L'arrivée en force des réseaux dans la majorité des applications de l'apprentissage automatique s'est également fait sentir en analyse de sons environnementaux. Les avancées algorithmiques, méthodologiques et pratiques des réseaux de neurones profonds ont fortement changé la nature des contributions du domaine durant ces deux dernières années. L'édition 2017 du challenge DCASE

illustre parfaitement cette récente transformation. En effet, le DCASE 2017 a rassemblé 75 équipes de recherche autour de différentes tâches pour lesquelles près de 90% des systèmes soumis se basent sur l'apprentissage de réseaux de neurones profonds. Cette famille d'approches facilite dans une certaine mesure l'entrée pour des chercheurs extérieurs au domaine, car les réseaux de neurones se sont montrés moins demandeurs en connaissances expertes pour commencer à obtenir des performances intéressantes. Les réseaux de neurones sont maintenant à la base de la plupart des systèmes état de l'art sur la majorité des bases de classification et détection de scènes et événements sonores. Nous présentons dans cette section certains des aspects et des avancées clés dans le succès de l'apprentissage profond pour le domaine. Nous reviendrons plus en détail sur le fonctionnement des différentes familles de couches de réseaux de neurones dans le chapitre 6.

Premiers réseaux pour l'analyse de sons environnementaux Les réseaux de neurones se définissent comme une succession d'opérations non-linéaires, organisées en différentes couches et appris selon un critère supervisé pour obtenir la sortie désirée. Ces différentes couches ont le rôle d'apprendre plusieurs niveaux de représentations intermédiaires à partir des données, permettant de modéliser et d'extraire une succession de concepts abstraits utiles à la caractérisation et à la discrimination des différentes catégories. Pour la classification de sons, les réseaux de neurones sont capables de regrouper les étapes d'apprentissage de représentation, de modélisation temporelle et de classification en un seul modèle supervisé. La forme la plus simple des réseaux de neurones est le perceptron multi-couches (MLP), de l'anglais *Multi-layer perceptron*. Ils sont constitués de plusieurs couches de neurones, chacune connectée aux neurones de la couche précédente, afin de fournir un signal de sortie pour la couche suivante.

Tout d'abord, les MLP peuvent être utilisés comme de simples classifieurs comme les SVM ou la régression logistique. Ils sont alors principalement utilisés, avec une approche similaire, pour classifier une combinaison de représentations issues de l'ingénierie de descripteurs. Dans ce sens, des premiers travaux ont montré le potentiel des MLP par rapport aux approches par SVM ou HMM pour la classification de scènes [Petetin et al., 2015] et pour la détection d'événements avec recouvrement [Cakir et al., 2015b]. Un des avantages des réseaux de neurones dans les cas polyphoniques est qu'ils permettent directement la prise en compte d'une classification multi-labels par l'apprentissage conjoint de la représentation et d'un modèle pour la détection de chaque étiquette [Cakir et al., 2015a].

Réseaux récurrents Ensuite, la prochaine avancée de performance majeure est arrivée avec les réseaux récurrents profonds (RNN), de l'anglais *reccurent neural networks*, pour la détection d'événements [Parascandolo et al., 2016]. Les couches récurrentes permettent la classification de séquences d'observations, où chaque neurone dépend de l'observation à l'instant courant et de sa sortie à l'instant précédent. Par l'utilisation d'unités LSTM [Hochreiter et Schmidhuber, 1997], de l'anglais *Long short-term memory*, ou GRU [Cho et al., 2014], de l'anglais *Gated recurrent unit*, les réseaux récurrents sont capables de modéliser des dépendances temporelles bien plus longues que les HMM, ce qui donne souvent lieu à de fortes améliorations des performances. Depuis, on retrouve des couches récurrentes dans la grande majorité des architectures de réseaux de neurones pour la détection d'événements, comme en témoigne les systèmes soumis au challenge DCASE 2017.

Réseaux convolutifs Le retour en force des réseaux de neurones profonds s'explique par les performances spectaculaires des réseaux convolutifs profonds (CNN) pour la classification d'images. Les couches convolutives effectuent des opérations de filtrage par la convolution de l'image avec des filtres locaux. L'enchaînement des opérations de filtrage et d'opérations dites de *pooling* per-

met d'apprendre des représentations globales de l'image à partir de l'extraction de caractéristiques locales. Ces opérations font des CNN des outils très performants pour apprendre des représentations de haut niveau des données, invariantes par rapport à leur position dans l'image. Les CNN sont également devenus l'état de l'art dans de nombreuses applications de traitement de l'audio en analysant directement les représentations temps-fréquence comme des images. Les premières applications à l'analyse des sons environnementaux sont apparues pour traiter des problèmes de classification d'événements [Piczak, 2015a]. Ensuite, l'édition 2016 du DCASE a été marquée par la proposition d'un nombre important de systèmes CNN pour la classification de scènes et d'événements [Valenti et al., 2016; Eghbal-Zadeh et al., 2016]. Cependant les performances obtenues en 2016 par les différents systèmes par CNN sont loin d'être aussi impressionnantes que pour d'autres tâches d'apprentissage automatique. D'autres approches d'apprentissage de descripteurs ou de simples MLP appris sur des descripteurs audio classiques restent compétitifs avec les meilleures approches CNN. On doit en grande partie ce phénomène à la limitation de la taille des bases de données ainsi qu'à la nouveauté des approches CNN dans la communauté. En revanche la domination des CNN est arrivée avec l'édition 2017 du DCASE. Les bases de données étant relativement similaires, ces améliorations s'expliquent surtout par un travail en amont et en aval des CNN, sur les pré-traitements, les représentations d'entrée et des stratégies de fusion de classifieurs. La sortie récente de la base de donnée *Audioset* [Gemmeke et al., 2017], plus de 200 fois plus grande que les bases actuelles, ouvre la voie vers l'exploration d'architectures plus complexes. Des premiers résultats ont été obtenus en comparant les performances des plus gros modèles CNN de classification d'image sur de la classification d'événements à grande échelle [Hershey et al., 2017].

Enfin, les modèles les plus performants à ce jour pour toutes les tâches de détection d'événements sont les réseaux de neurones convolutifs récurrents (CRNN), de l'anglais *convolutional recurrent neural network*. Les CRNN se définissent par la succession de couches convolutives et de couches récurrentes. Les couches convolutives ont le rôle d'apprendre une représentation modélisant les caractéristiques locales des représentations temps-fréquence, alors que les couches récurrentes sont responsables de la modélisation de l'évolution temporelle de l'information à plus long terme. Ces modèles sont à la base des systèmes les plus efficaces pour la détection d'événements [Cakir et al., 2017; Xu et al., 2017c] et offrent de fortes améliorations par rapport aux RNN ou CNN pris séparément.

2.7 Rendre les systèmes plus robustes

Nous présentons dans cette section d'autres axes de recherches et méthodologies introduits en plus des approches classiques pour construire les systèmes de classification et détection de sons environnementaux. L'objectif de ces approches est de rendre les systèmes plus robustes pour pallier certaines limitations des bases de données et des techniques employées. Les différentes stratégies présentées constituent des étapes essentielles pour construire des systèmes performants, moins sensibles à la taille et la qualité des bases de données ainsi qu'à la nature aléatoire de certains systèmes de classification.

2.7.1 Détection d'événements pour la classification de scènes et réciproquement

Les problèmes de détection d'événements et la classification de scènes sonores sont en majorité traités séparément. Pourtant, il existe des liens très forts entre ces deux tâches. En effet, avoir de l'information sur le contexte dans lequel nous nous situons peut nous informer sur la nature des événements pouvant être présents dans la scène sonore. Réciproquement, avoir un moyen d'identifier les occurrences de certains événements pourrait grandement faciliter la reconnaissance de la

nature de la scène sonore. Dans ce sens, quelques travaux ont proposé d'inclure un module de détection du contexte avant l'étape de détection d'événements. Ces approches utilisent par exemple un modèle HMM appris pour chaque contexte en supposant que l'occurrence temporelle des différents événements dépend du lieu dans lequel a été effectué l'enregistrement [Heittola et al., 2013a; Lu et al., 2015]. Dans l'autre sens, Heittola et al. [2010] ont proposé d'appliquer un système de détection d'événements dans l'objectif de construire un histogramme d'occurrence des événements dans la scène. Cet histogramme est ensuite utilisé comme représentation pour un système de classification de scènes. D'autres approches proposent de regrouper les événements sonores par thèmes puis de modéliser le contexte sonore comme une collection de thèmes, comme on peut le faire en analyse de documents [Kim et al., 2009; Imoto et al., 2013]. Ces deux types d'approches ont montré des améliorations de performances sur des bases de données de faible taille en se servant de modèles relativement simples. La taille des bases et les modèles de classification actuels ouvrent la possibilité d'une nouvelle exploration au goût du jour de ce type d'approches. Les modèles de classification de scènes et de détection d'événements étant de plus en plus performants séparément, il serait possible d'aller vers des modèles capables d'effectuer les deux tâches conjointement .

2.7.2 Pré-traitement et augmentations

Nous avons mentionné précédemment que la taille relativement faible de la plupart des bases de données du domaine rend difficile l'apprentissage de modèles de classification complexes. Le faible nombre de données disponibles à l'apprentissage n'est pas nécessairement représentatif de la variabilité de chaque catégorie. Une des réponses à ce problème est d'employer des méthodes d'augmentation ou de pré-traitement permettant d'augmenter artificiellement le nombre de représentations par exemple audio, se traduisant par une augmentation de la taille de la base d'apprentissage.

Une première approche, à la frontière entre l'extraction de descripteurs et l'augmentation, est de tirer parti des données multi-canal lorsqu'elles sont disponibles. De nombreuses bases de données sont enregistrées avec deux microphones fournissant des enregistrements binauraux. L'approche la plus répandue et la plus simple consiste à extraire de descripteurs ou spectrogramme séparément pour chaque canal. Cette stratégie a été retenue par un grand nombre de participants aux challenges DCASE. On peut également augmenter le nombre de canaux en ajoutant la somme ou la différence entre les canaux gauche et droite [Eghbal-Zadeh et al., 2016]. D'autres méthodes proposent l'extraction de descripteurs binauraux, modélisant les différences entre les canaux afin de rendre les systèmes de détection plus robustes [Adavanne et Virtanen, 2017; Adavanne et al., 2017]. Enfin, le nombre de canaux peut être augmenté artificiellement en utilisant des pré-traitements par séparation de sources. Par exemple, l'objectif peut être de séparer le bruit de fond des événements afin de faciliter leur détection [Heittola et al., 2011, 2013b; Y.Hang et Park, 2017].

D'autres approches utilisent des stratégies d'augmentations de données en effectuant de légères perturbations des signaux afin d'augmenter artificiellement le nombre et la variabilité des exemples. Pour l'analyse de sons environnementaux, l'objectif est de rendre les systèmes invariants à certaines modifications temporelles et fréquentielles. Les deux stratégies d'augmentations que l'on retrouve le plus fréquemment sont le décalage fréquentiel et l'étirement temporel, plus connus sous le nom de *pitch shifting* et *time-stretching* en anglais. L'idée est que deux événements d'une même catégorie peuvent avoir une structure fréquentielle similaire mais également avoir des spectrogrammes décalés de quelques fractions de tons ou étirés de quelques fractions de secondes l'un par rapport à l'autre. Les augmentations ont déjà fait leurs preuves dans certaines applications de MIR et de parole [Schlüter et Grill, 2015] et commencent à être de plus en plus appliquées

en analyse de sons environnementaux, en majorité pour améliorer la capacité de généralisation de modèles de type CNN [Lehner et al., 2017; Salamon et al., 2017; Salamon et Bello, 2017]. Dernièrement, Mun et al. [2017] ont exploré l'utilisation de modèle de réseaux génératifs adversaires (GAN), de l'anglais *generative adverserial networks*. Par leur capacité à générer de nouveaux exemples similaires aux données d'apprentissage, les modèles GAN ont été utilisés avec succès pour faire de l'augmentation de données pour la classification de scènes.

2.7.3 Fusion de classifieurs

De nombreux classifieurs, en particulier les réseaux de neurones profonds, sont sensibles à la fois à l'initialisation et aux choix des paramètres. De plus, même à initialisation et paramètres égaux, deux occurrences d'un même modèle peuvent donner des résultats très différents, s'expliquant par la difficulté et la non-convexité des problèmes d'optimisation. Les approches par fusion entraînent différents classifieurs séparément et définissent une stratégie prenant en compte la prédiction de chaque modèle afin de prendre la décision finale. Les approches les plus répandues sont de simplement prendre la somme, la moyenne ou un vote majoritaire à partir des sorties. D'autres méthodes vont avoir recours à l'apprentissage en entraînant une régression logistique pour obtenir une combinaison optimale des modèles. Ces approches permettent bien souvent d'augmenter la performance d'un système en combinant plusieurs variantes ou plusieurs occurrences d'un même classifieur. Ainsi, les stratégies de fusion sont très largement utilisées lors de participations à des campagnes d'évaluation, où l'objectif est simplement d'obtenir les meilleurs taux de reconnaissance possibles. Cette tendance s'applique aussi aux campagnes d'évaluation DCASE, où une grande majorité des meilleurs systèmes se base sur de telles stratégies de fusion [Eghbal-Zadeh et al., 2016; Bisot et al., 2016b; Mun et al., 2017; Lee et al., 2017; Xu et al., 2017a].

2.7.4 Apprentissage faiblement supervisé

L'apparition récente de bases de données de très grande taille s'accompagne de nouveaux défis. En particulier, des bases comme Audioset [Gemmeke et al., 2017] ou le *youtube 8M dataset* [1] se basent sur la collection d'un grand nombre de segments audio à partir de vidéos *youtube*. Le nombre très important d'exemples rend irréaliste leur annotation précise par des humains. Les annotations sont obtenues simplement par les étiquettes laissées par les utilisateurs sur la plate-forme ou sont grossièrement attribuées à des segments plus longs. On parle alors de données faiblement annotées. La taille de la base permet de s'affranchir dans une certaine mesure de l'imprécision sur les étiquettes par la possibilité d'apprendre des modèles de réseaux de neurones profonds complexes. Cependant, plusieurs travaux ont montré qu'adapter les systèmes au caractère faiblement annoté de la base permet de rendre les modèles plus robustes. En particulier, la plupart de ces travaux se sont concentrés sur des tâches de *tagging*, où un segment audio est associé à une ou plusieurs étiquettes sans connaissance de leur position temporelle exacte dans le segment.

L'approche la plus courante est de traiter cette difficulté comme un problème d'apprentissage à instances multiples (MIL), de l'anglais *multiple instance learning*. Si on a un ensemble ou sac de trames, les problèmes MIL supposent qu'une étiquette peut être associé à cet ensemble de trames sans savoir quel sous-ensemble d'entre elles contient effectivement l'étiquette en question. Des premières approches se sont intéressées au MIL pour des tâches de *tagging* d'événements sonores. Kumar et Raj [2016a,b] ont proposé des modèles MIL par SVM, GMM ou réseaux de neurones permettant de classifier et identifier la position des étiquettes dans chaque segment. D'autres approches construisent un mécanisme d'attention en apprenant simultanément deux réseaux en parallèle [Kong et al., 2017b; Xu et al., 2017b]. L'un des réseaux est responsable de la classification

1. research.google.com/youtube8m

de chaque trame, et l'autre pondère l'importance de la trame dans la décision, indiquant ainsi sur quelles trames le modèle doit porter l'attention. Plus récemment, des mécanismes similaires ont été appliqués avec succès à des bases plus volumineuses telles qu'Audioset, offrant une amélioration de performances par rapport aux modèles de CNN profonds pour la vision [Xu et al., 2017c; Kong et al., 2017a]. Ces premières contributions se situent probablement dans le début d'une nouvelle ère de travaux en analyse de sons environnementaux, s'attaquant pour la première fois à des bases de *tagging* de très grande taille.

Chapitre 3

Descripteurs d'images pour la représentation des sons environnementaux

Sommaire

Dans ce chapitre, nous présentons nos premiers systèmes de classification de sons environnementaux, basés sur l'extraction de descripteurs d'images à partir de représentations temps-fréquence. Nous commençons par introduire et discuter des représentations temps-fréquence que nous utiliserons tout au long du manuscrit afin de représenter les scènes et événements sonores. Ces représentations constituent notre matériel de base pour toutes les méthodes d'extraction et d'apprentissage de descripteurs proposées. Dans le même temps, nous présentons nos premiers systèmes de classification pour lesquels nous proposons une combinaison de deux descripteurs : les histogrammes de gradient orienté et les histogrammes par bandes de fréquences. Cette combinaison nous permet d'améliorer les performances par rapport aux autres approches par extraction de descripteurs sur plusieurs bases de données de classification de scènes sonores. Une partie des résultats présentés dans ce chapitre a fait l'objet d'une publication d'un article à la conférence EUSIPCO [Bisot et al., 2015].

3.1 Représentations temps-fréquence

3.1.1 Représentations temps-fréquence et bancs de filtres

Interpréter un enregistrement sonore directement à partir du signal temporel, ou de la forme d'onde, est une tâche difficile tant pour l'Homme que pour les machines. La grande dimension de l'information contenue dans une forme d'onde rend l'identification des différents événements acoustiques composant un enregistrement sonore particulièrement compliquée. Les méthodes d'analyse de l'audio ont souvent recours à d'autres formes de représentations du signal audio afin de faciliter l'extraction de caractéristiques et la visualisation de sons. Les plus plébiscitées sont de loin les représentations fréquentielles du signal qui donnent la répartition de l'énergie du signal en différentes bandes de fréquences en projetant le signal sur une base d'ondes élémentaires (Fourrier, ondelettes..). Cette simple représentation fréquentielle n'est parfois pas suffisante pour représenter correctement certains signaux, en partie du à leur non-stationnarité. En effet, on souhaite être capable d'observer l'évolution de l'information dans le temps, en extrayant des caractéristiques localement par trames temporelles. C'est alors en projetant le signal fenêtré par trames sur un axe fréquentiel que l'on obtient ce que l'on appelle une représentation temps-fréquence. Elles offrent un moyen efficace de visualiser l'évolution du contenu fréquentiel du signal au cours du temps. La représentation temps-fréquence la plus répandue est la transformée de Fourrier à court terme (TFCT). Elle est la base de nombreuses variantes de représentations temps-fréquence et de méthodes d'extraction de descripteurs. La TFCT est obtenue par transformée de Fourier discrète de portions fenêtrées du signal. Ainsi nous appelons spectrogramme ou représentation temps-fréquence, une représentation en deux dimensions du signal où l'énergie dans chaque bande de fréquences est donnée en fonction du temps. Il existe de nombreuses autres formes de représentations temps-fréquence adaptées du signal notamment avec l'utilisation d'ondelettes [Mallat, 1999].

La TFCT, par construction, sépare l'axe des fréquences linéairement en plusieurs bandes de fréquences. Or, pour un grand nombre de tâches de traitement de l'audio il est souvent préférable de répartir les fréquences centrales autrement. Pour des signaux musicaux, on peut vouloir chercher à représenter l'énergie dans des bandes de fréquences spécifiques, comme par exemple des bandes représentant chaque note de musique grâce à la transformée à Q constant (CQT), de l'anglais *Constant-Q transform* [Brown, 1991]. De nombreux travaux ont également montré l'intérêt d'avoir des bandes de fréquences inspirées de la perception humaine de la hauteur en utilisant des représentations telles que les bandes Mel [Stevens et al., 1937]. La CQT et les spectres en bandes Mel sont les représentations temps-fréquence de bas niveau les plus répandues en analyse de sons environnementaux. Elles sont la base de la plupart des systèmes de détection et classification de

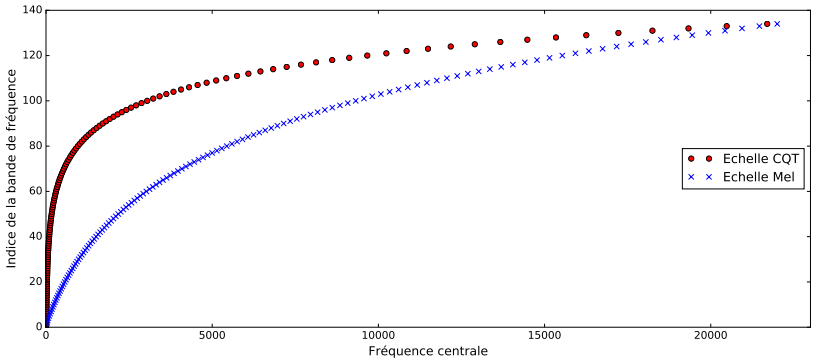


FIGURE 3.1 – Répartition des fréquences centrales pour la CQT et le spectre Mel avec 140 bandes dans les deux cas.

scènes et d'événements. On peut les retrouver comme étapes dans l'extraction de descripteurs comme les MFCC, les descripteurs d'images ou encore comme représentations d'entrée de bas niveau pour apprendre des réseaux de neurones profonds. De même, nos systèmes présentés dans ce travail de thèse ont tous la CQT ou le spectre Mel comme représentation de bas-niveau. Elles ont toutes les deux le point commun d'être définies sur une échelle de fréquence logarithmique, ayant l'avantage de s'approcher de la perception humaine de la hauteur mais aussi de compresser efficacement l'axe fréquentiel. Nous les décrivons brièvement avant de donner quelques visualisations de ces représentations pour des exemples de scènes et événements sonores.

Représentation temps-fréquence en bandes Mel L'échelle Mel correspond à une approximation de la sensation humaine de hauteur d'un son pur (d'une sinusoïde pure) [Stevens et al., 1937]. Il existe plusieurs expressions analytiques de l'échelle Mel, une des plus communes liant l'échelle Mel à l'échelle en Hertz donnée par [Fant, 1968] est la suivante :

$$\text{mel}(f) = \frac{1000}{\log 2} \log\left(1 + \frac{f}{1000}\right) \tag{3.1}$$

Transformée à Q constant La CQT est proche de la TFCT, toutefois l'échelle des fréquences est répartie géométriquement de sorte à ce que le ratio Q entre la fréquence centrale d'une bande et la résolution fréquentielle soit constante [Brown, 1991]. La fréquence centrale pour la bande d'indice k est donnée par :

$$f_k = f_0 \times 2^{\frac{k}{b}} \tag{3.2}$$

où f_0 est la fréquence centrale de la première bande et b le nombre de bandes de fréquences par octave. Cette échelle de fréquence permet notamment d'avoir chaque note de la gamme occidentale comme fréquence centrale, si l'on prend $b = 12$. Lors du calcul de la CQT, la taille des fenêtres dépend de la fréquence centrale, ce qui donne des fenêtres longues pour les basses fréquences et des fenêtres courtes pour les hautes fréquences.

Bien que les échelles Mel et CQT séparent toutes les deux l'échelle des fréquences logarithmiquement, elles ne sont pas réparties de la même manière. La répartition des fréquences centrales à nombre de bandes équivalent (140) pour les deux échelles est donnée dans la figure 3.1. Une des différences majeures entre ces deux représentations est que la CQT a une meilleure résolution en basse fréquence que l'échelle Mel mais une moins bonne en haute fréquence. Cela a son importance lors du choix d'une représentation adaptée à la tâche traitée. Enfin, nous donnons quelques exemples de représentations temps-fréquence calculées à partir de CQT ou de Mel pour différentes scènes et événements sur les figures 3.2 et 3.3. Ces exemples illustrent la moins bonne résolution

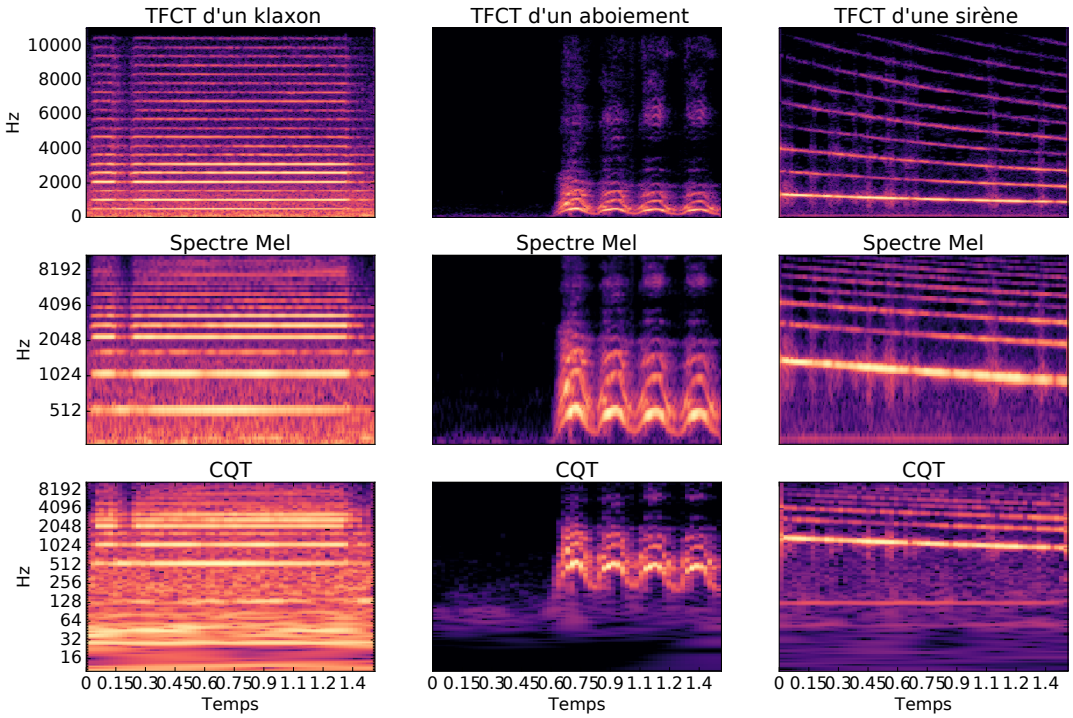


FIGURE 3.2 – Représentations temps-fréquence TFCT, Mel et CQT pour trois catégories d'événements de la base Urbansound.

fréquentielle en haute fréquence de la CQT, où seulement quelques bandes représentent les très hautes fréquences. De plus, en basse fréquence, l'information a tendance à être lissée dans le temps pour la CQT. Cela est dû à la taille variable des fenêtres. En effet, en basse fréquence les fenêtres peuvent être longues de plusieurs centaines de millisecondes.

3.1.2 Décrire un son par une image

La représentation temps-fréquence ou le spectrogramme d'un son le rend plus facilement visuellement interprétable par un humain. Le spectrogramme est souvent traité et interprété comme une image, où l'énergie est représentée par différents niveaux de gris des pixels. Cette image temps-fréquence contient suffisamment d'informations pour que les humains puissent reconnaître certains sons à l'œil nu comme des notes de musique ou des phonèmes en parole [Zue, 1985]. De manière analogue, il est également commun pour les méthodes computationnelles du traitement du signal audio de traiter la représentation temps-fréquence comme une image, en reprenant des méthodes inspirées de la vision par ordinateur. En effet, différentes catégories d'événements sonores sont souvent associées à des motifs temps-fréquence caractéristiques dans le spectrogramme. Ainsi, on espère détecter la présence de ces événements en traitant les motifs temps-fréquence qui leur sont associés de manière analogue à la détection d'objets dans une image. Le domaine du traitement des images a développé depuis de nombreuses années des méthodes très performantes pour identifier et caractériser différentes catégories d'objets. Cela a motivé l'application de techniques avancées issues de la vision par ordinateur sur des spectrogrammes pour de multiples applications de traitement de l'audio. Un exemple de cette tendance est l'utilisation de réseaux de neurones convolutifs (CNN). Les CNN ont été utilisés avec beaucoup de réussite dans nombreuses tâches de traitement de l'audio comme la reconnaissance de la parole [Abdel-Hamid et al., 2014], la détection de mélodie ou la classification d'événements sonores [Piczak, 2015a; Hershey et al.,

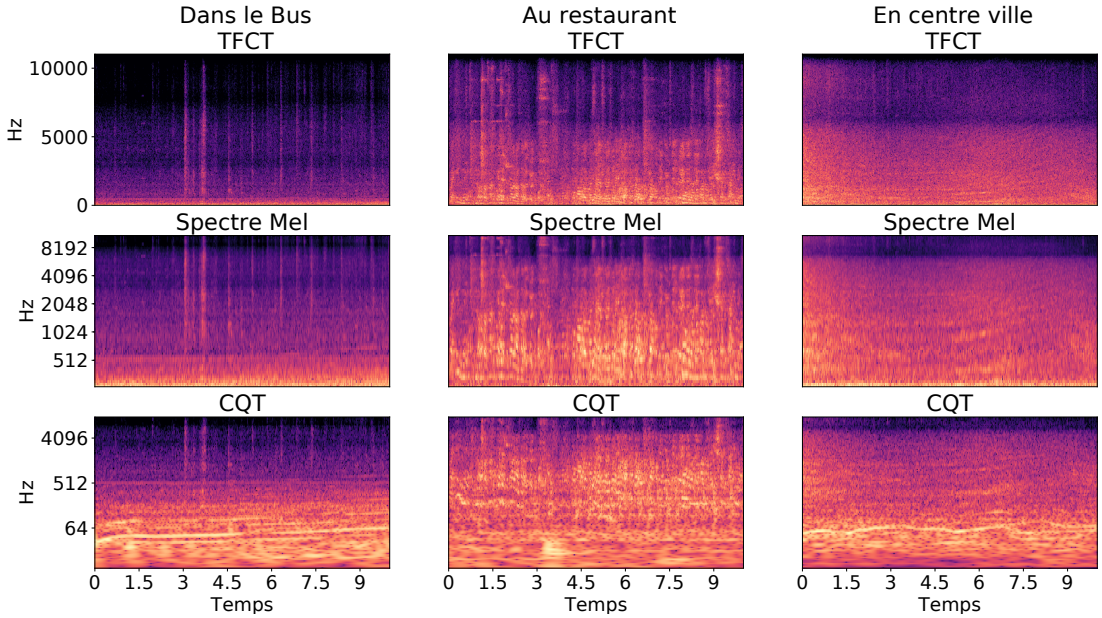


FIGURE 3.3 – Représentations temps-fréquence TFCT, Mel et CQT pour trois scènes sonores de la base DCASE 2017.

2017].

Il existe néanmoins des différences clés entre la vision et le traitement de l'audio. La différence principale est qu'on ne peut pas traiter les deux dimensions de l'image temps-fréquence de la même manière. En effet, en classification d'images, on cherche souvent des méthodes invariantes aux translations et à la rotation des images dans toutes les directions. En revanche, pour l'audio, on cherche la plupart du temps des invariances par translation sur l'axe temporel. Cela motive parfois l'adaptation des architectures CNN issues de la vision pour traiter la particularité des images temps-fréquence, en particulier pour des applications de détection d'événements [Cakir et al., 2017].

Les travaux présentés dans ce chapitre s'inscrivent également dans cette tendance, en proposant l'analyse et la combinaison de descripteurs extraits de l'image temps-fréquence pour la classification de scènes et événements sonores.

Pour la classification d'événements sonores, on cherche à attribuer une étiquette parmi une liste de catégories à un enregistrement contenant une ou plusieurs occurrences de cet événement. Une des approches possibles est de chercher un moyen d'identifier des motifs temps-fréquence spécifiques à chaque étiquette de sorte à pouvoir les classifier. Partant de cette hypothèse, des travaux ont proposé d'appliquer des descripteurs d'images permettant de caractériser les objets ainsi que leurs contours dans l'image. On trouve par exemple l'utilisation de descripteurs SIFT ou HOG pour la classification d'événements robustes au bruit [Dennis et al., 2014, 2013].

En revanche pour la classification de scènes sonores, on ne cherche pas forcement à identifier un seul motif temps-fréquence mais la combinaison de plusieurs motifs ou leur répétition, correspondant à des événements spécifiques à la scène. Certains travaux ont proposé l'utilisation de descripteurs de texture de l'image dans l'idée qu'un environnement sonore se caractérise par la densité de certains événements ou par la présence d'un bruit de fond constant. Par exemple, l'occurrence répétée de vagues à la plage ou le moteur d'un train peu se traduire par la présence de textures dans certaines bandes de fréquences du spectrogramme. Des descripteurs d'images comme les HOG [Rakotomamonjy et Gasso, 2015; Rakotomamonjy, 2017] ou les motifs binaires locaux (LBP) [Battaglino et al., 2015; Yang et Krishnan, 2017] ont exhibé de bonnes performances pour

la classification de scènes. Ils ont pour objectif de décrire les textures temps-fréquence de l'image en construisant des histogrammes de ses caractéristiques locales.

3.2 Combiner les HOG et les SPD

Nous proposons dans ce chapitre un moyen efficace et performant pour décrire le contenu des images temps-fréquence de sons environnementaux afin de mieux les discriminer. Pour cela, nous reprenons deux descripteurs ayant fait leurs preuves sur certaines tâches d'analyse de sons environnementaux. Les premiers sont les HOG, issus du traitement de l'image servant à décrire les directions des variations d'intensité des pixels dans une image. Les deuxièmes sont les SPD, introduits pour la classification d'événements, ils décrivent la distribution de l'intensité de l'information de chaque bande de fréquence. Nous estimons que ces deux représentations sont complémentaires car elles décrivent des aspects différents des images temps-fréquence nécessaires pour caractériser certaines catégories de sons environnementaux.

3.2.1 Utilisations précédentes

Les HOG ont été initialement utilisés pour la détection de personne en vision par ordinateur [Dalal et Triggs, 2005]. Ils ont été introduits avec l'idée qu'une forme peut être décrite localement par la direction du gradient de l'image. Les HOG ont plus récemment été proposés pour la classification audio. Ils ont notamment été utilisés pour la classification d'événements [Dennis et al., 2014], où certaines catégories d'événements présentent des formes caractéristiques dans le spectrogramme (un impact par un trait vertical ou une sirène par des diagonales dans le plan temps-fréquence) mais également pour de l'identification de locuteur [Muroi et al., 2009]. De plus, les HOG ont aussi démontré leur efficacité pour la classification de scènes sonores [Rakotomamonjy et Gasso, 2015; Rakotomamonjy, 2017]. En effet, certaines scènes peuvent contenir des variations temps-fréquence caractéristiques de l'environnement comme le bruit des vagues à la plage ou les accélérations fréquentes dans les moyen de transports.

Les SPD sont simplement des histogrammes par bandes de fréquences de l'image temps-fréquence d'un son. Certains descripteurs de bas niveau à l'utilisation très répandue sont basés sur une idée similaire en décrivant des caractéristiques du signal par bandes de fréquences [Peeters, 2004; Serizel et al., 2018]. Les SPD ont été introduits formellement pour la classification d'événements [Dennis et al., 2013] comme un moyen d'estimer la distribution de l'énergie par bandes de fréquences dans un spectrogramme. Ils ont également été utilisés comme première étape dans l'extraction de descripteurs d'images plus complexes pour la classification d'événements dans le bruit [Dennis et al., 2014].

3.2.2 Motivations

Les SPD ont initialement été proposés pour la classification d'événements dans le but d'avoir une représentation invariante par décalage temporel de l'objet sonore dans le segment à classifier [Dennis et al., 2013]. Nous faisons l'hypothèse que cette représentation est également particulièrement adaptée aux défis que posent la classification de scènes sonores. En effet, une scène sonore est souvent constituée d'une importante variété de sons de natures très différentes qui contribuent à former sa signature acoustique. La première supposition que l'on peut faire est que ces différents sons correspondent à des événements acoustiques (comme un klaxon) caractéristiques de certains environnements (comme une rue). On peut également supposer que chaque type d'événement possède une distribution temps-fréquence particulière. Ainsi avoir un moyen d'identifier l'occurrence

de ces distributions à travers les scènes aidera à mieux les décrire et les discriminer. C'est pour capturer ces occurrences que nous proposons l'utilisation des SPD. La représentation extraite par les SPD approxime la distribution de l'amplitude du spectrogramme dans chaque bande de fréquence en calculant leurs histogrammes. Les SPD vont donc nous permettre de modéliser la fréquence d'apparition et l'intensité de ces événements.

Bien que nous supposons que les SPD vont fournir de l'information importante pour décrire les scènes, ils ne sont peut-être pas suffisants. En effet, deux scènes sonores contenant des événements de natures différentes peuvent avoir des descripteurs SPD très similaires. Par exemple, la représentation temps-fréquence d'une scène contenant la répétition d'un événement périodique relativement court et un autre contenant un événement plus long suivi d'un silence ont possiblement des distributions d'amplitude par bandes de fréquences très similaires. Pour ce type d'exemple, avoir un moyen de caractériser l'évolution du contenu temps-fréquence aiderait à augmenter la capacité des systèmes à discriminer les différentes scènes. Nous proposons pour cela d'utiliser les SPD conjointement aux HOG. Les HOG ont déjà montré de très bons résultats sur certaines bases de données [Rakotomamonjy et Gasso, 2015; Barchiesi et al., 2015] mais caractérisent d'autres aspects du spectrogramme de la scène. En effet, ils modélisent l'évolution de l'information temps-fréquence au cours de la scène en s'appuyant sur le gradient de l'image. Cela permet par exemple de décrire des scènes contenant des sons variants dans le temps comme des accélérations. L'utilisation conjointe des HOG et des SPD nous permet alors à la fois de décrire les événements présents dans la scène par les SPD et leurs variations temporelles et fréquentielles avec les HOG.

3.2.3 L'extraction des HOG et SPD

Extraction des HOG

Il existe plusieurs variantes pour l'extraction des HOG, nous reprendrons celle proposée par [Rakotomamonjy et Gasso, 2015] pour la classification de scènes. En général l'extraction des HOG se compose toujours des étapes suivantes :

1. Calculer le gradient de l'image pour chaque pixel
2. Calculer l'angle de chaque gradient
3. Séparer l'image en cellules sans recouvrement
4. Compter les différents valeurs des angles du gradient dans chaque cellule pour construire un histogramme

Dans notre cas, la première étape est de construire l'image temps-fréquence en utilisant une des deux représentations (Mel ou CQT). Comme proposé par [Rakotomamonjy et Gasso, 2015] l'image est échantillonnée par interpolation bi-cubique de sorte à être de taille 512×512. Cela permet, pour les bases où les segments sont de tailles variables, de ramener tous les exemples à une image carrée de même taille. On applique alors le procédé listé ci-dessus en utilisant des cellules 8×8. Afin d'obtenir le vecteur de descripteurs final pour chaque exemple, nous moyennons les histogrammes à travers l'axe temporel de sorte à avoir un histogramme d'orientation pour les différentes bandes de fréquences à travers le temps. Cette stratégie s'est avérée être la plus efficace parmi celles discutées et évaluées dans [Rakotomamonjy et Gasso, 2015] car elle suit la recherche d'invariance à la position temporelle des objets sonores dans les segments à classifier. Elle permet également d'avoir un seul vecteur de descripteurs pour tout le segment permettant l'utilisation de classifieurs classiques.

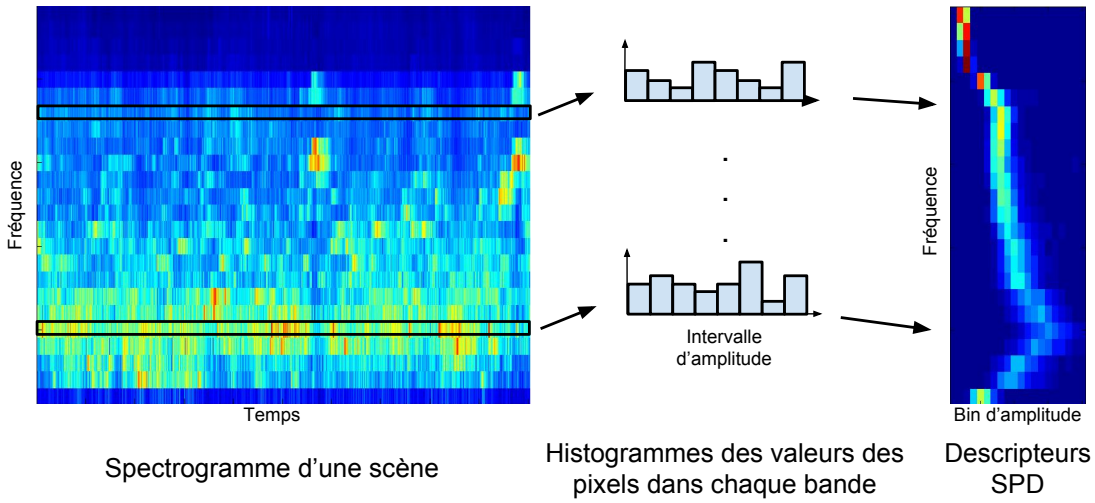


FIGURE 3.4 – Schéma d'extraction des SPD à partir d'un spectrogramme d'une scène.

Extraction des SPD

Les SPD sont calculés à partir de la représentation temps-fréquence sans ré-échantillonnage. Ils sont obtenus en calculant un histogramme de la valeur des pixels de l'image à travers le temps dans chaque bande de fréquences. Pour cela on sépare linéairement l'échelle d'amplitude des pixels dans l'image et on compte le nombre de pixels compris dans chaque intervalle d'amplitude. On obtient finalement autant d'histogrammes que de bandes de fréquences (voir la figure 3.4), que l'on concatène pour former les descripteurs utilisés pour décrire la scène. Le processus d'extraction de l'image SPD tel qu'introduit dans [Dennis et al., 2013] est légèrement différent. Il inclut le calcul des histogrammes à partir de MFCC en remplacement de la représentation temps-fréquence.

Nous proposons également une légère modification dans le schéma d'extraction des SPD afin d'améliorer les performances que nous désignons dans la suite par L-SPD pour Log-SPD. Afin d'extraire les L-SPD, nous commençons par compresser la dynamique de l'image temps-fréquence en lui appliquant un logarithme comme proposé par [Dennis et al., 2013]. Cette compression permet de se rapprocher de notre perception de l'intensité des sons et évite aux histogrammes de trop concentrer l'information dans les premiers bins de basse énergie. De plus nous proposons d'adapter les intervalles des bins d'histogramme à chaque bande de fréquences à travers toute la base. Jusqu'à maintenant, les bins d'histogrammes étaient répartis linéairement entre les valeurs extrêmes des pixels de l'image traitée pour chaque bande de fréquence. A la place, nous proposons d'adapter les bornes des intervalles en prenant en compte toute la base de données. Pour cela nous cherchons la valeur maximum dans chaque bande de fréquence à travers toute la base d'apprentissage et séparons linéairement l'intervalle gardant la même séparation pour chaque image. Dans ce cas, pour une bande de fréquence donnée, un même bin d'histogramme pour deux images différentes représente le même intervalle, rendant les SPD extraits de deux scènes différentes plus facilement comparables pour les classifieurs.

3.3 Classification

3.3.1 Classification

Le schéma d'extraction des descripteurs d'images présenté ci-dessus nous permet d'avoir un vecteur de descripteurs par segment audio à classifier. Lors de l'utilisation de méthodes basées

sur l'extraction de descripteurs, il est commun en classification audio d'utiliser des machines à vecteurs supports (SVM). Les SVM ont l'avantage de permettre d'obtenir des performances raisonnables même avec assez peu de données et ont donc souvent été employées dans les premiers travaux sur l'analyse de sons environnementaux. De plus, les SVM restent le classifieur majoritairement utilisé pour les systèmes basés sur des descripteurs d'images. Plus récemment les SVM ont été remplacées par des réseaux de neurones profonds de type perceptron multi-couche pour des méthodes de type extraction de descripteurs, nous discuterons cet aspect dans le chapitre 6.

3.3.2 Choix du noyau pour la classification

Il est commun lors de l'emploi de SVM d'utiliser un noyau non-linéaire afin de mieux représenter des données complexes, parfois non séparables linéairement. On retrouve le plus souvent l'usage du noyau gaussien, comme c'est également le cas pour certaines méthodes de classification de scènes et d'événements [Rakotomamonjy et Gasso, 2015; Battaglino et al., 2015]. Pour $\mathbf{x}$ et $\mathbf{y}$ deux vecteurs de descripteurs, le noyau Gaussien se définit par la fonction,

$$k(\mathbf{x}, \mathbf{y}) = e^{-\frac{\|\mathbf{x}-\mathbf{y}\|_2}{2\sigma^2}}, \tag{3.3}$$

où σ^2 est le paramètre d'échelle. Nous nous limitons dans ce chapitre au cas du noyau gaussien. Bien que nous reprenons dans l'Annexe C l'idée d'une construction du noyau par la distance de Sinkhorn [Cuturi, 2013], une approximation de l'*Earth Mover's Distance* (EMD). Cependant, à la fois la complexité et les performances obtenues par le noyau de Sinkhorn dans nos premiers travaux dans [Bisot et al., 2015] ne justifient pas son utilisation pour la suite de notre étude.

3.4 Validation expérimentale

3.4.1 Objectifs et bases de données

Nous choisissons d'évaluer les méthodes d'extraction de descripteurs d'images sur trois bases de données de classification de scènes dans l'objectif de confirmer l'intérêt de la combinaison pour ce type de tâches. De plus, en tant que méthode d'extraction de descripteurs performante, le système proposé nous servira de référence dans la suite du manuscrit, afin d'avoir un point de départ pertinent pour comparer les contributions suivantes. Nous proposons également d'évaluer les méthodes sur une base de données de classification d'événements afin de discuter des différences entre les deux tâches. L'objectif est de confirmer les particularités de la tâche de classification de scènes acoustiques qui ont motivé l'emploi de descripteurs issus du traitement des images. En plus des bases de données, nous étudions également l'impact du choix de la représentation temps-fréquence sur quelques cas particuliers.

Nous introduisons brièvement l'ensemble des bases utilisées dans ce chapitre (une présentation plus détaillée est donnée en Annexe A). Nous précisons pour chaque base de données, le système de référence correspondant au premier système publié par les auteurs de la base, afin de faciliter la comparaison des systèmes pour les premiers travaux.

LITIS Rouen Il s'agit d'une des premières grandes bases de données de classification de scènes publiques sortie en 2014 [Rakotomamonjy et Gasso, 2015]. Elle contient 25 heures d'enregistrement de scènes sonores segmentées en 3026 exemples de 30 secondes répartis en 19 catégories. Les systèmes de référence pour cette base utilisent des descripteurs MFCC modélisés par une analyse quantitative récurrente (RQA) Roma et al. [2013] et les HOG classifiés avec une SVM à noyau gaussien Rakotomamonjy et Gasso [2015]. Ces systèmes ont obtenu parmi les meilleures

performances à l'édition 2013 du challenge DCASE pour la classification de scène sonores [Giannoulis et al., 2013]. Les résultats présentés correspondent à la moyenne sur les 20 ensembles de validation croisée fournis par les auteurs.

DCASE 2016 La base DCASE 2016 correspond à l'ensemble de développement de la tâche 1 pour la classification de scènes sonores de l'édition 2016 du challenge DCASE [Mesaros et al., 2016b]. Elle contient 10 heures d'enregistrement de scènes sonores segmentées en 1170 exemples de 30 secondes répartis équitablement en 15 catégories. Le système de référence pour cette base utilise des Mel spectres à 40 bandes modélisés par une GMM par classe [Mesaros et al., 2016b]. Les résultats présentés correspondent à la moyenne sur les 4 ensembles de validation croisée fournis par les auteurs.

DCASE 2017 La base du DCASE 2017 est l'extension de la version 2016 pour l'édition 2017 du challenge [Mesaros et al., 2017a]. Elle est composée de la version de développement et de l'ensemble d'évaluation de la base 2016. Elle contient 13 heures d'enregistrement de scènes sonores segmentées en 4680 exemples de 10 secondes répartis équitablement en 15 catégories. Le système de référence pour cette base utilise des spectres Mel à 40 bandes classifiés avec un perceptron multi-couches Mesaros et al. [2017a]. Les résultats présentés correspondent à la moyenne sur les 4 ensembles de validation croisée fournis par les auteurs.

Urbansound Il s'agit d'une base de classification d'événements urbains contenant 8120 exemples de 1 à 4 secondes répartis en 10 catégories différentes [Salamon et Bello, 2015b]. Chaque segment contient de une à plusieurs occurrences du même événement avec la possibilité de présence de bruit de fond. Pour cette base, le système de référence correspond à des descripteurs MFCC ainsi que leurs premières et secondes dérivées classifiés avec des forêts aléatoires [Salamon et Bello, 2015b]. Les résultats présentés correspondent à la moyenne sur les 10 ensembles de validation croisée fournis par les auteurs.

3.4.2 Protocole expérimental

Métriques d'évaluation Les métriques principales pour la classification de scènes et d'événements durant tout le manuscrit sont le score F1 et le taux de reconnaissance (*accuracy* en anglais). Cependant, les travaux ayant introduit l'utilisation des HOG pour la classification de scènes sur la base du LITIS ont été présentés en termes de précision, afin de faciliter la comparaison nous inclurons lorsque nécessaire la précision et le rappel. La précision, le rappel et le score F1 sont calculés par classe avant d'être moyennés sur l'ensemble des classes puis sur les ensembles de validation croisée. Dans ce chapitre, nous privilégierons le score F1 pour comparer les performances des systèmes car il permet de mieux prendre en compte le déséquilibre entre les classes. En particulier pour les bases du LITIS et Urbansound, certaines classes sont bien moins représentées, leur contribution au taux de reconnaissance est donc plus faible.

Représentation temps-fréquence Les représentations temps-fréquences sont extraites en utilisant la toolbox YAAFE [Mathieu et al., 2010]. La première étape est toujours de normaliser le signal audio de sorte à ce que ses valeurs restent dans intervalle $[-1, 1]$. Les CQT sont extraites en utilisant 12 bandes par octave de 5 à 22kHz donnant 134 bandes de fréquences en utilisant des fenêtres de 60 ms sans recouvrement. Les Mel spectres sont extraits en l'intervalle de 5 à 22kHz en 80 bandes Mel avec des fenêtres de 60ms avec 50% de recouvrement.

Extraction des HOG Les HOG sont extraits de manière similaire aux travaux de Rakotomamonjy et Gasso [2015], le spectrogramme (CQT ou Mel) est ré-échantilloné pour former une image 512×512 pour les scènes et 256×128 pour les événements par interpolation bi-cubique. L'image est choisie volontairement plus petite pour les événements car leur durée moyenne est beaucoup plus courte que les scènes. Les histogrammes sont calculés dans des cellules de taille 8×8 sans recouvrement. Nous gardons les orientations signées du gradient et nous moyennons les histogrammes d'orientation obtenus à travers le temps, pour les $512/8 = 64$ bandes de fréquences.

Extraction des SPD Lors de l'extraction des SPD à partir de CQT, les bandes de fréquences sont regroupées par deux sans recouvrement en les moyennant par paire. Cela permet d'approcher le nombre de bandes de fréquences des CQT (134/2=73 bandes de fréquences) de celui des spectres Mel (80 bandes de fréquences). Les histogrammes sont toujours calculés en espaçant linéairement l'intervalle d'amplitude des pixels de l'image en 20 segments. Lors de l'extraction des L-SPD, le maximum de l'intervalle est choisi pour chaque bande de fréquence à travers toute la base de données.

Classifieur Nous rappelons que nous utilisons une SVM à noyau gaussien comme classifieur. Le paramètre d'échelle σ^2 est choisi dans $\{1, 5, 10, 20, 50, 100\}$ et le paramètre de régularisation du classifieur dans $\{0.1, 1, 10, 100\}$.

3.4.3 Résultats

Impact de la représentation temps-fréquence

Nous commençons par nous intéresser à l'impact de la représentation temps-fréquence sur les performances des descripteurs d'images pour les deux tâches. En plus de la base de classification d'événements Urbansound, nous sélectionnons la base DCASE 2017 parmi les trois bases de classification de scènes, étant celle qui contient le plus d'exemples. Nous rappelons que due au processus d'extraction des SPD présenté précédemment, la dimension des descripteurs est similaire lors de l'utilisation des spectres Mel et des CQT. Les score F1 sur les deux bases sont indiqués dans le tableau 3.1.

Pour la classification de scènes, à dimension équivalente, les HOG ainsi que les SPD et leur combinaison sont bien plus performants lorsqu'ils sont extraits à partir des CQT. Comme nous l'avons vu précédemment, la différence principale entre les spectres Mel et les CQT est la répartition des filtres sur l'axe des fréquences, les CQT étant plus précises en basse fréquence mais moins en haute fréquence que les spectres Mels. Cela laisserait supposer que l'information basse fréquence a son importance pour discriminer les environnements sonores. En effet, on peut reconnaitre les scènes en identifiant les événements qu'elles contiennent mais également en caractérisant la nature du bruit de fond. Par exemple, des environnements tels que *dans le train* ou *dans la voiture* peuvent être caractérisés par la présence d'un bruit de fond constant correspondant au moteur du véhicule. La différence de performance pour la classification de scènes souligne bien l'importance d'un choix pertinent de représentation temps-fréquence, base de la plupart des systèmes de classification audio. On ne peut cependant trop généraliser ces résultats car ils dépendent de la base et de la méthode d'extraction ou d'apprentissage de descripteurs. Par exemple, selon la liste des catégories d'événements présents dans la base de données, il peut être nécessaire d'adapter la représentation temps-fréquence pour l'adapter au cas particulier traité.

	Urbansound		DCASE 2017	
Système de référence	**67.5**		73.5	
	Mel	CQT	Mel	CQT
HOG	65.7	65.5	72.3	78.6
L-SPD	53	52.1	70.6	77.7
HOG + L-SPD	66.3	65.3	78.0	**80.1**

TABLEAU 3.1 – Scores F1 pour les descripteurs d'images sur deux représentations temps-fréquence différentes.

En revanche, sur la base de classification d'événements, le choix de la représentation temps-fréquence n'a que peu d'impact sur les performances. Il semblerait que l'importance d'une bonne résolution en basses fréquences pour la classification de scènes ne se retrouve pas dans la même mesure pour la classification d'événements. En effet, dans la plupart des bases de classification d'événements, les catégories décrivent des événements de natures très différentes. Pour certaines étiquettes telles que *Kids playing*, l'énergie aura tendance à être concentrée en haute fréquence, alors que pour d'autres tels que *air conditioning*, l'énergie est plutôt concentrée en basse fréquence. La base Urbansound contient également des événements se caractérisant par la production de sons harmoniques parcourant la majorité du spectre tel que *car horn* ou *sirene*.

Résultats sur l'ensemble des bases

Nous présentons maintenant les performances des descripteurs d'images sur les 4 bases de données. Les scores F1 sont tous exposés dans le tableau 3.2, ils sont obtenus en utilisant le noyau Gaussien et la représentation CQT. Nous incluons les scores pour les systèmes de référence mentionnés lors de la description des bases de données. Pour la base du LITIS, nous présentons également les résultats en utilisant notre premier procédé d'extraction des SPD [Bisot et al., 2015] afin de le comparer aux améliorations proposées section 3.2.3.

On peut commencer par remarquer que les L-SPD permettent de significativement améliorer les performances sur la base du LITIS. L'utilisation du schéma d'extraction initial des SPD offre des scores F1 faibles relativement aux autres approches sur les bases DCASE. Ensuite, nous pouvons remarquer que la combinaison HOG+L-SPD permet d'obtenir des performances supérieures aux méthodes de référence sur les bases de classification de scènes. Cependant, cette amélioration reste faible pour la base du DCASE 2016, ce qui peut s'expliquer par la mauvaise performance des L-SPD due à la présence d'un nombre d'exemples 3 à 4 fois inférieur aux autres bases. En effet, les SPD décrivent la distribution de l'amplitude des pixels par un histogramme séparé en 20 intervalles pour chaque bande de fréquences, ce qui en fait une représentation en relativement grande dimension. Les résultats semblent suggérer qu'un plus grand nombre de données d'apprentissage est nécessaire afin d'éviter le sur-apprentissage dû à la grande dimension des SPD.

Les résultats restent inférieurs au système de référence pour la base Urbansound de classification d'événements. En particulier, les résultats pour les L-SPD sont très inférieurs à ceux des HOG. Cela suggère que ce genre de descripteurs est bien moins adapté pour la classification d'événements. Cela semble être confirmé par le fait que les HOG et autres descripteurs capturant la texture de l'image tels que les LBP [Yang et Krishnan, 2017; Battaglino et al., 2015] ont surtout été utilisés pour la classification de scènes. En effet, ayant pour but de caractériser la distribution de l'information dans les images temps-fréquence, ces descripteurs permettent essentiellement de décrire la texture de l'image spectrogramme. Ils nécessitent donc d'avoir suffisamment d'information pour estimer la distribution de l'information ce que ne permettent pas des segments de 1 à 4 secondes

	Urbansound	**LITIS**	**DCASE 2016**	**DCASE 2017**
Référence	**67.0**	91.7*	70.9	73.5
SPD	49.2	89.7	63.0	71.0
HOG	65.7	90.5	**77.3**	78.6
L-SPD	53.0	91.8	69.4	77.7
HOG + L-SPD	66.3	**94.0**	**77.7**	**80.1**

TABLEAU 3.2 – Scores F1 pour les descripteurs d'images sur les représentations CQT pour les 4 bases de données de classification de scènes et événements. (*) Ce résultat correspond à la précision et est donné à titre indicatif.

FIGURE 3.5 – Scores F1 par classe sur la base de classification de scènes DCASE 2017.

pour la classification d'événements.

Enfin, c'est la combinaison HOG + L-SPD qui permet d'obtenir les meilleures performances sur les 3 bases de classification de scènes. Cela confirme l'intuition que ces deux descripteurs peuvent être complémentaires dans la description des scènes sonores. De plus, lorsque la base de données est de taille suffisante, ils permettent d'obtenir des performances largement supérieures à celles de systèmes plus classiques de classification audio tels que ceux utilisés comme référence. Nous présentons les scores F1 par classe sur la base du DCASE 2017 dans la figure 3.5 afin de mieux comprendre l'intérêt des deux descripteurs et de leur combinaison. Les HOG affichent de meilleurs scores F1 pour des environnements contenant d'importantes variations du contenu temps-fréquence pour des catégories comme *train*, *tram* ou *station de métro*. Ce genre de scènes se compose souvent d'accélérations facilement identifiables dans l'image temps-fréquence et aisément capturées par les HOG. Les SPD exhibent de meilleures performances pour les catégories *residential area*, *Beach* ou *Grocery store* qui ont en commun d'être caractérisées par la densité de certains événements clés spécifiques à la scène. En effet, une scène à la plage peut être facilement identifiée par une forte densité de sons de vagues ou un magasin peut être identifié par les alertes sonores du passage des articles en caisse. Finalement, la combinaison des deux descripteurs permet d'améliorer les scores F1 pour la majorité des catégories. Cela confirme que chaque catégorie de scènes est caractérisée à la fois par la distribution du contenu fréquentiel et de ses variations.

3.5 Conclusion

L'étude des descripteurs d'images effectuée dans ce chapitre a permis d'exhiber la complémentarité de deux d'entre eux, les HOG et les SPD. Les HOG capturent la direction des variations dans le plan temps-fréquence tandis que les SPD donnent la répartition de l'énergie par bandes de fréquences. La combinaison proposée s'est montrée particulièrement efficace sur des problèmes de classification de scènes sonores en offrant de meilleures performances que les précédentes approches par ingénierie de descripteurs. En revanche, étant basées sur des histogrammes représentant la distribution temporelle de l'information, ces descripteurs se sont montrés moins performants pour classifier des segments d'événements plus court. Toutefois, ils confirment dans une certaine mesure, que représenter la distribution temporelle l'information à partir de spectrogrammes constitue un moyen efficace de décrire les scènes sonores.

Chapitre 4

Apprentissage non-supervisé de descripteurs par factorisation de matrices

Sommaire

Dans ce chapitre, nous entamons notre étude des méthodes d'apprentissage de représentations par factorisation de matrices en nous intéressant au cas non-supervisé. Nous nous servons des factorisations de matrices comme technique d'apprentissage de descripteurs en décomposant directement les représentations temps-fréquence de scènes sonores. Nous proposons l'étude de différentes variantes de l'analyse en composantes principales et des factorisations en matrices positives qui incluent des versions avec parcimonie, à noyaux ou intégrant du contexte temporel dans la décomposition. Après avoir discuté leur intérêt pour les tâches traitées, nous proposons une comparaison de ces approches sur plusieurs bases de données de classification de scènes et de détection d'événements. Nous effectuons cette étude en proposant un cadre simple et efficace facilitant la comparaison des différentes approches. Les systèmes introduits dans ce chapitre constituent la base sur laquelle sont construites les approches présentées dans le reste de cette thèse tout en proposant une meilleure alternative en termes de performance aux techniques par descripteurs d'images. Les résultats de ce chapitre ont été présentés dans deux publications à la conférence ICASSP [Bisot et al., 2016a, 2017a] et dans un article de journal [Bisot et al., 2017b].

4.1 Apprentissage non-supervisé de descripteurs

Dans le chapitre précédent, nous avons étudié et contribué aux approches par extraction de descripteurs pour la classification de scènes et d'événements acoustiques. Nous avons entre autres montré que caractériser la distribution et l'évolution du contenu des images temps-fréquence permet une bonne discrimination des différentes scènes sonores. Les approches que nous avons étudiées et proposées rentrent dans la catégorie de l'ingénierie de descripteurs, c'est-à-dire la construction de procédés d'extraction d'information à partir des données. Ces approches par extraction de descripteurs ont à la fois l'avantage et l'inconvénient de caractériser uniquement certaines propriétés précises des signaux. En effet, ce genre de procédé peut s'avérer très efficace pour des applications de classification où l'on connaît la nature de l'information permettant de discriminer les différentes catégories. En revanche, les problèmes d'analyse de sons environnementaux peuvent nécessiter l'analyse d'une grande variété d'événements sonores, de natures et de moyens production très différents. La grande variabilité inter et intra-classes présente naturellement dans l'analyse de scènes sonores rend plus difficile la définition de procédés universels d'extraction de caractéristiques discriminatives. De multiples travaux ont montré que l'obtention de bonnes performances avec des approches par extraction de descripteurs nécessitait souvent la combinaison d'une large variété d'entre eux.

Depuis quelques années, la tendance à travers de nombreuses tâches de classification de signaux est de se tourner vers des techniques d'apprentissage de descripteurs. L'apprentissage de descripteurs permet au système, souvent en résolvant un problème d'optimisation, d'apprendre automatiquement l'information caractéristique ou discriminative directement à partir des données. L'apprentissage de représentation est un domaine relativement vaste, regroupant plusieurs grandes familles, telles que les méthodes probabilistes, par factorisation de matrices ou réseaux de neurones profonds. Le principal avantage de ces approches est leur flexibilité et leur relative indépendance au jeu de données traités. En effet, elles permettent d'apprendre automatiquement un procédé d'extraction de représentations caractéristiques s'adaptant au type de données exploitées par les systèmes de classification. Ces méthodes sont également présentes sous différentes formes en analyse de sons environnementaux. Cependant elles ne sont devenues une des tendances dominantes que très récemment. Alors que les premières approches étaient basées sur des méthodes par modèles de gaussiennes ou par factorisation de matrices, on trouve actuellement des systèmes d'apprentissage de descripteurs supervisés de plus en plus avancés basés sur des réseaux de neurones profonds.

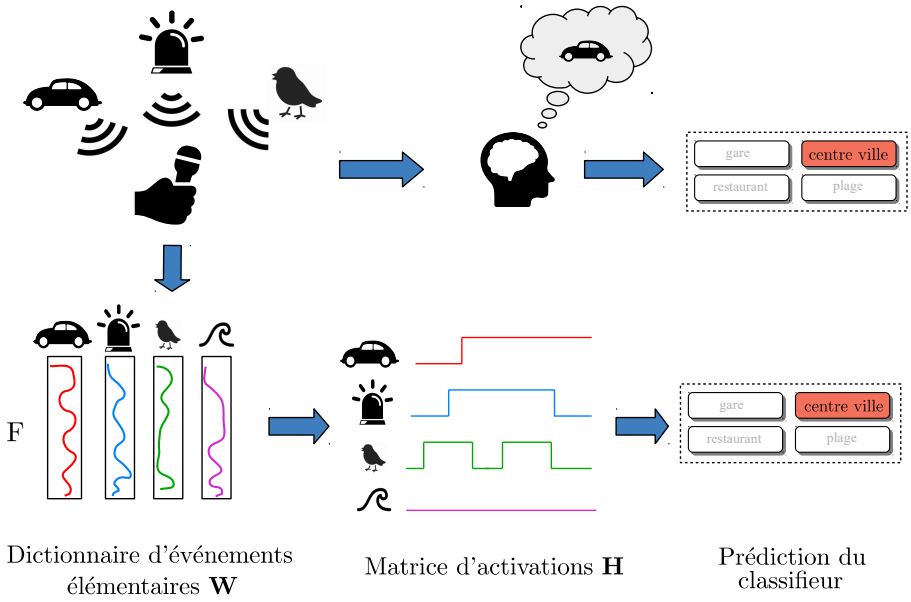


FIGURE 4.1 – Illustration des similarités entre la reconnaissance humaine des scènes sonores et les approches par apprentissage de dictionnaires des méthodes computationnelles.

Comme première entrée dans le vaste domaine de l'apprentissage de représentations nous nous intéressons aux techniques d'apprentissage non-supervisé. L'apprentissage non-supervisé connaît un regain d'intérêt ces dernières années avec pour objectif de répondre à certaines limitations des modèles supervisés. Malgré leur domination, les approches supervisées d'apprentissage profond gardent quelques défauts majeurs. Elles permettent uniquement la prise en compte de données annotées, potentiellement coûteuses à produire. De plus, elles sont plus facilement sujettes au sur-apprentissage, due à leurs difficultés à traiter et à s'adapter aux données non-observées par le système. En faisant abstraction des étiquettes, l'apprentissage non-supervisé offre à la machine la capacité d'apprendre sa propre représentation de l'ensemble des données à sa disposition. Cette nouvelle représentation possède à la fois l'avantage d'offrir de meilleurs descripteurs et de faciliter les capacités de généralisation du système. Ainsi, nous nous demandons dans ce chapitre si, au lieu de se baser sur des schémas d'extraction déterminés, des descripteurs peuvent être appris automatiquement de manière non-supervisée à partir des représentations temps-fréquence. En particulier, nous nous intéressons aux techniques de factorisation de matrices. Nous présentons le principe et l'intérêt de ces approches section 4.2 avant d'introduire la manière dont nous les utilisons dans nos systèmes de classification de scènes section 4.3 et de détection d'événements section 4.5.

4.2 Approches par factorisation de matrices

4.2.1 Factorisation de matrices pour l'apprentissage de descripteurs

Nous commençons par rappeler les principes de base des techniques de factorisation de matrices présentées au chapitre 2. On se donne une matrice de données $\mathbf{V} \in \mathbb{R}^{F \times N}$ de N observations de dimension F. L'objectif des techniques de factorisation de matrices est d'approximer la matrice de données $\mathbf{V}$ par le produit de deux matrices telle que :

$$\mathbf{V} \approx \mathbf{WH}, \tag{4.1}$$

où $\mathbf{W} \in \mathbb{R}^{F \times K}$ représente le dictionnaire et $\mathbf{H} \in \mathbb{R}^{K \times N}$ la matrice d'activation. Le dictionnaire est construit à partir d'une collection de K vecteurs de base définis dans l'espace des observations. Ils capturent les aspects les plus représentatifs de la variabilité des données décomposées. La matrice d'activation s'obtient par la projection des données sur ce dictionnaire. Elle nous donne la pondération à appliquer dans la combinaison linéaire de vecteurs de base pour reconstruire chaque observation. Cette décomposition s'obtient par la résolution d'un problème d'optimisation en minimisant un terme d'attache aux données, représentatif de la qualité de l'approximation. De manière générale nous noterons le terme d'attache aux données par $D(\mathbf{V}\|\mathbf{WH})$. Dans beaucoup de cas D est simplement une distance euclidienne entre $\mathbf{V}$ et son approximation par $\mathbf{WH}$ mais de nombreuses autres formes de distances, divergences ou l'opposé de la log-vraisemblance peuvent être utilisées comme critère cible.

Les factorisations de matrices sont utilisées dans de multiples domaines de l'apprentissage automatique et du traitement du signal comme le *clustering*, la réduction de dimension ou la séparation de sources. Dans nos travaux, nous nous focalisons sur leur utilisation pour l'apprentissage de descripteurs. L'objectif est alors d'apprendre un dictionnaire de vecteurs de base représentant certains aspects caractéristiques des données. Approximer les données par une combinaison de ces éléments de base fournit alors une représentation intermédiaire servant de descripteur pour la classification. Les représentations obtenues par factorisation de matrices ont souvent l'avantage d'être plus facilement interprétables, ce qui peut se traduire par une amélioration de la capacité de généralisation des classifieurs.

Les approches d'apprentissage de descripteurs par factorisation de matrices s'interprètent particulièrement bien pour l'analyse de sons environnementaux. En effet, la façon dont nous les utilisons pour traiter ce genre de tâches possède quelques points communs avec la manière dont les humains comprennent et classifient les scènes sonores. Une scène sonore se compose, par nature, de la superposition d'une multitude d'événements indépendants. Pour identifier ces scènes sonores, les humains se focalisent sur certains événements saillants, pouvant être informatifs quant au contexte dans lequel nous nous trouvons. Ainsi, dans une certaine mesure, on peut interpréter les systèmes par factorisation de matrices pour la classification de scènes comme se rapprochant de la compréhension humaine des scènes. L'objectif est alors de se servir de ces techniques pour apprendre un dictionnaire composé de vecteurs de base qui peuvent s'interpréter comme des "événements de base" ou des "événements élémentaires". Si l'on effectue la factorisation de matrices sur des représentations temps-fréquence, le dictionnaire contient donc une collection de représentations fréquentielles des événements les plus représentatifs des différentes scènes sonores. En projetant les données sur ce dictionnaire, on obtient une matrice d'activations, nous indiquant les occurrences de ces différents événements de base au cours de l'enregistrement. Le classifieur aura alors le rôle d'associer et de pondérer la contribution des différents événements à la reconnaissance du contexte dans lequel le son a été enregistré. Nous illustrons cette interprétation figure 4.1, en faisant le parallèle entre les modèles de factorisation de matrices et les procédés que nous employons pour reconnaître les scènes.

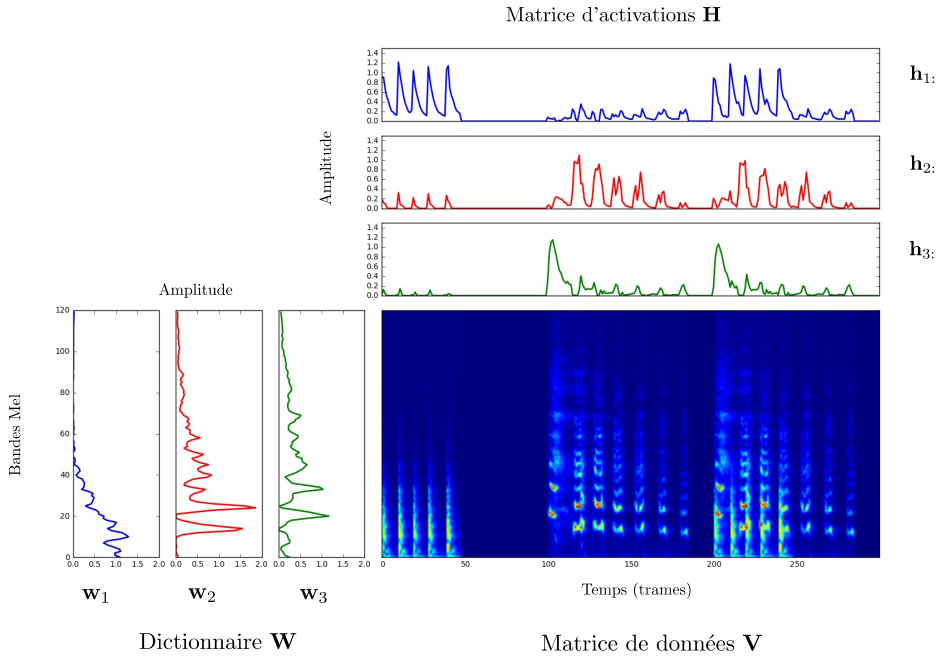


FIGURE 4.2 – Illustration des composantes du dictionnaire et activations obtenues par la décomposition d'un spectre Mel par NMF avec la distance euclidienne. L'exemple contient l'enchainement d'une personne frappant à la porte, d'un éclat de rire et des deux événements simultanés.

En pratique, l'apprentissage de descripteurs par factorisation de matrices se fait en deux étapes. On commence par apprendre le dictionnaire en décomposant uniquement les données de l'ensemble d'apprentissage. Une fois le dictionnaire appris, il est gardé fixe et les données de l'ensemble d'apprentissage et de test sont ensuite projetées sur le même dictionnaire. Cette projection est effectuée selon le même problème d'optimisation mais en fixant $\mathbf{W}$ afin d'obtenir la matrice d'activation $\mathbf{H}$ sur les deux ensembles.

4.2.2 Factorisation en matrices positives

Les factorisations en matrices positives (NMF), de l'anglais *nonnegative matrix factorisation*, sont une des familles de factorisation de matrices les plus utilisées en traitement de l'audio. Initialement, la NMF a été popularisée pour la décomposition de données positives telles que les images [Lee et Seung, 1999] ou les données textuelles [Pauca et al., 2004; Xu et al., 2003]. Elle est également à la base de nombreuses approches de traitement du signal audio en particulier pour des applications de séparations de sources et de débruitage [Virtanen, 2007; Wang et Plumbley, 2005; Wilson et al., 2008] mais aussi pour la transcription de partitions [Smaragdis et Brown, 2003].

La NMF reprend le principe de base des factorisations de matrices donné équation (4.1), en supposant la décomposition d'une matrice de données $\mathbf{V} \in \mathbb{R}_+^{F \times N}$ à coefficients positifs. La NMF revient alors à chercher une factorisation qui approxime $\mathbf{V}$ telle que $\mathbf{V} \approx \mathbf{WH}$ avec $\mathbf{W} \in \mathbb{R}_+^{F \times K}$ et $\mathbf{H} \in \mathbb{R}_+^{K \times N}$, le dictionnaire et la matrice d'activations tous deux à coefficients positifs. La décomposition NMF est obtenue en résolvant le problème d'optimisation suivant :

$$\min_{\mathbf{W},\mathbf{H}} D(\mathbf{V}\|\mathbf{WH}) \text{ t.q. } \mathbf{W}, \mathbf{H} \geq 0 \tag{4.2}$$

où D est une divergence ou une distance entre les données et leur approximation **WH**. L'attrait principal de la NMF est la contrainte de positivité imposée dans la décomposition qui rend la factorisation interprétable intuitivement. En effet, la NMF cherche à expliquer les données par addition de composantes de base à coefficients positifs, nous permettant de représenter les données positives comme une addition d'objets élémentaires. La NMF rentre alors dans la catégorie de ce que l'on appelle des modèles *compositionels* de l'audio [Virtanen et al., 2015]. Ce type de modèles permet de représenter les données comme une combinaison linéaire positive de différentes *parties* à coefficients positifs. Ainsi, cette combinaison n'implique aucune soustraction ou annulation des différentes parties facilitant grandement l'interprétation du modèle obtenu. Le son peut être vu comme de nature compositionelle. En effet, un signal sonore se compose de l'addition de plusieurs sons élémentaires s'annulant rarement entre eux. De plus, pour le traitement de l'audio, les modèles compositionels tels que certains modèles NMF permettent de modéliser l'additivité des spectrogrammes de puissance ou d'amplitude des différentes sources [Févotte et al., 2009; Virtanen et al., 2008].

Ces interprétations et modélisations des signaux se traduisent également par une amélioration de la performance, en particulier pour des applications de séparation de sources audio où la NMF est largement plébiscitée. De plus, ces propriétés de la NMF en font également un outil particulièrement adapté à l'analyse de sons environnementaux. En effet, les scènes sonores correspondent en majorité à des environnements multi-sources où les différentes sources sont associées aux événements constituant la scène. La NMF nous permet de reprendre l'interprétation entamée dans la section 4.2.1 sur l'intérêt des factorisations de matrices pour représenter les sons environnementaux. Avec la NMF, chaque scène sonore est représentée par une addition de représentations fréquentielles des événements de base à coefficients positifs, facilitant l'identification de la contribution de chacun de ces événements de base la caractérisant. Nous illustrons 4.2 le résultat d'une décomposition NMF sur la représentation temps-fréquence en spectre Mel de deux événements sonores joués séparément puis simultanément. Dans cet exemple simplifié, on peut noter que chaque composante du dictionnaire apparait associée à un seul événement. Le deuxième événement est décrit par deux composantes car il se compose de deux motifs temps-fréquence distincts. L'activation de ces composantes nous indique alors la présence des deux événements au cours du temps.

Les β-divergences Le choix de la divergence D constitue un des aspects les plus importants dans l'application de la NMF car il conditionne la qualité et les propriétés de la factorisation. A la différence des distances, les divergences garantissent uniquement la propriété de séparation. En traitement de l'audio, la famille des β-divergences est de loin la plus populaire [Cichocki et Amari, 2010]. En particulier, leur utilisation se résume bien souvent à trois cas d'intérêt :

— la distance euclidienne pour $\beta = 2$;

— la divergence de Kullback-Leibler (KL) généralisée pour $\beta = 1$ se définissant comme $d_\beta(a, b) = a \log(\frac{a}{b}) + b - a$ [Kullback et Leibler, 1951] ;

— la divergence d'Itakura-Saito (IS) pour $\beta = 0$ se définissant comme $d_\beta(a, b) = \frac{a}{b} - \log(\frac{a}{b}) - 1$ [Itakura, 1968].

Le choix de la β-divergence influe notamment sur les propriétés d'échelles de la divergence, en particulier la divergence IS (β=0) garantit une invariance à l'échelle $d(\lambda a, \lambda b) = d(a, b) \forall \lambda \in \mathbb{R}$. Cette propriété permet notamment d'accorder autant d'importance dans la décomposition aux points temps-fréquence de faible et forte énergie. Cela peut s'avérer utile pour la prise en compte d'événements distants ou de faible intensité dans les scènes sonores.

Algorithmes NMF L'étape d'estimation des paramètres de la NMF peut s'écrire comme la minimisation d'une fonction f dépendant d'un ensemble de paramètres θ où pour le problème NMF dans sa forme la plus simple on a :

$$\min_{\theta} f(\theta) = \min_{\mathbf{W},\mathbf{H}\geq 0} D(\mathbf{V}\|\mathbf{WH}). \tag{4.3}$$

Le décomposition NMF a initialement été estimée par l'usage de règles de mises à jour multiplicatives (MU), de l'anglais *multiplicative update rules* [Lee et Seung, 1999]. Ces méthodes s'appuient sur une écriture du gradient de la fonction de coût comme une différence entre deux termes positifs $\nabla_\theta f(\theta) = \nabla^+_\theta - \nabla^-_\theta$. Les paramètres sont alors mis à jour par la règle multiplicative suivante :

$$\theta \leftarrow \theta \times \frac{\nabla^-_\theta}{\nabla^+_\theta} \tag{4.4}$$

Cette mise à jour garantit uniquement un mouvement de θ dans le sens opposé à celui du gradient de f mais ne garantit pas la décroissance du coût dans le cas général. Toutefois, la décroissance de la fonction de coût a été démontrée pour l'utilisation des β-divergences pour $\beta \in [0, 2]$ [Févotte et Idier, 2011] lors de l'utilisation des règles MU suivantes :

$$\mathbf{W} \leftarrow \mathbf{W} \otimes \frac{((\mathbf{WH})^{(\beta-2)} \otimes \mathbf{V})\mathbf{H}^T}{(\mathbf{WH})^{(\beta-1)}\mathbf{H}^T}, \tag{4.5}$$

$$\mathbf{H} \leftarrow \mathbf{H} \otimes \frac{\mathbf{W}^T((\mathbf{WH})^{(\beta-2)} \otimes \mathbf{V}}{\mathbf{W}^T(\mathbf{WH})^{(\beta-1)}}, \tag{4.6}$$

où $\otimes$ désigne l'opération de produit matriciel terme à terme. Durant ces dernières années, bien d'autres algorithmes ont été proposés pour estimer la NMF. On trouve des approches par descente de gradient projeté, par moindres-carrées alternés, par méthodes de Newton ou par descente de coordonnées. Cependant, les MU ont l'avantage de fournir des algorithmes relativement simples, composés uniquement de produits de matrices. De plus, si d'autres algorithmes offrent une convergence plus rapide, Serizel et al. [2016c] ont montré que l'application des algorithmes MU avec des processeurs graphiques (GPU) pouvait accélérer significativement les temps de calcul rendant les règles MU particulièrement attractives. C'est en grande partie pour cette raison que nous nous limiterons à l'utilisation des règles MU pour l'estimation des différentes variantes NMF considérées dans ce travail de thèse.

4.2.3 Différences avec les travaux similaires de l'état de l'art

Nous revenons ici sur quelques points clés concernant les différences entre notre utilisation des méthodes NMF et celles de l'état de l'art introduites dans le chapitre 2. A la date de nos premiers travaux publiés sur l'apprentissage de descripteurs par factorisation de matrices [Bisot et al., 2016a], la plupart des méthodes utilisaient les factorisations de matrices différemment à la fois pour la classification de scènes sonores et la détection d'événements. Pour la classification de scènes, une première approche consistait à se servir de techniques de factorisation de matrices telles que la NMF ou la NMF convolutive pour apprendre un dictionnaire différent par représentation temps-fréquence dans la base de données [Cauchi, 2011; Benetos et al., 2012]. Les différentes scènes sonores étaient alors discriminées en définissant un critère défini comme la distance entre les dictionnaires issus de chaque exemple. Dans ce cas, les dictionnaires n'étaient pas appris sur la base entière mais individuellement sur chaque exemple et étaient utilisés comme descripteurs à la place des matrices d'activations.

Pour les applications à la détection d'événements, les premières approches se servant de factorisations de matrices les utilisaient également comme classifieur [Benetos et al., 2016; Mesaros et al., 2015, 2016b; Bui et al., 2016]. Une approche fréquente était par exemple d'apprendre des éléments de dictionnaire correspondant à chaque étiquette à partir d'enregistrements contenant uniquement des événements isolés. Ensuite, les mélanges étaient projetés sur le dictionnaire et la matrice d'activations était simplement post-traitée et considérée comme contenant directement les activations des différents événements. Cette stratégie est pertinente dès lors que chaque élément de dictionnaire correspond à une seule catégorie d'événements, ainsi une activation suffisamment marquée de ce vecteur de base indique sa présence dans l'enregistrement.

Toutes les approches basées sur les factorisations de matrices que nous utilisons ou proposons diffèrent des stratégies décrites ci-dessus et ont toujours deux points communs : un seul dictionnaire représentant l'ensemble de la base est appris sur les données d'apprentissage et les projections sur ce dictionnaire seront toujours utilisées comme descripteurs pour l'apprentissage d'un classifieur.

4.3 Système de classification de scènes sonores

4.3.1 Systèmes de classification de scènes proposés

Nous commençons par présenter les différents éléments qui composent nos systèmes de classification de scènes utilisant l'apprentissage de descripteurs par factorisation de matrices. Nous introduisons un cadre simple allant de la mise en forme des représentations de bas niveau jusqu'à la classification des activations. La stratégie de construction de la matrice de données que nous proposons est illustrée figure 4.3 et est relativement indépendante des techniques de factorisation de matrices ou des classifieurs choisis. Sa simplicité et son efficacité nous conduiront à la réutiliser durant toute la suite du manuscrit.

Pré-traitements des spectrogrammes La première étape, tout comme dans le chapitre précédent, est d'extraire la représentation temps-fréquence pour chaque exemple de la base de données. Les spectrogrammes seront notés $\mathbf{S} \in \mathbb{R}_+^{F \times T}$, où F est le nombre de bandes de fréquences et T le nombre de trames temporelles. Afin de réduire la dimension des spectrogrammes, tout en gardant une forme de contexte temporel, nous leur appliquons une succession de traitements avant l'étape de décomposition. Pour cela, chaque spectrogramme $\mathbf{S} \in \mathbb{R}_+^{F \times T}$ est découpé en m tranches longues de q trames (ou fenêtres) temporelles, avec $m = \frac{T}{q}$. Ces tranches de l'image contiennent la signature acoustique des événements présents à différents instants de la scène. L'image temps-fréquence $\mathbf{S}$ est alors représentée par une série d'images plus courtes (les tranches) telle que $\mathbf{S} = [\mathbf{S}_1, ..., \mathbf{S}_m]$, où $\mathbf{S}_i \in \mathbb{R}_+^{F \times q}$ est la tranche de q trames commençant $q \times (i-1)$ trames après le début de la scène.

Afin de réduire la dimension, chacune des m tranches est moyennée au cours du temps. Chaque enregistrement de la base est alors représenté par un ensemble de m vecteurs $[\mathbf{s}_1, ..., \mathbf{s}_m]$, où $\mathbf{s}_i \in \mathbb{R}_+^F$ est le vecteur moyen de la tranche $\mathbf{S}_i$, correspondant au contenu fréquentiel moyen à différents instants de la scène. D'autres statistiques peuvent être appliquées afin d'améliorer la modélisation de l'évolution temporelle de l'information fréquentielle. La moyenne s'étant montrée suffisante pour obtenir des performances satisfaisantes, nous ne considérerons pas l'usage de statistiques supplémentaires afin de ne pas détourner l'attention des factorisations de matrices. Après avoir extrait l'ensemble de $[\mathbf{s}_1, ..., \mathbf{s}_m]$ m vecteurs pour chacun des N exemples de la base d'apprentissage, nous les concaténons horizontalement pour former la matrice $\mathbf{V} \in \mathbb{R}_+^{F \times mN}$. Dans la suite du manuscrit nous ferons référence à cette approche à construction par TM (tranches moyennées) de la matrice de données.

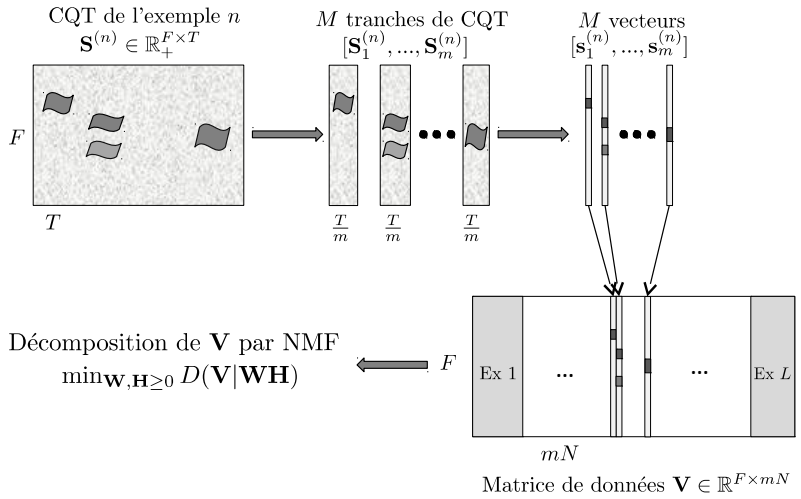


FIGURE 4.3 – Construction de la matrice de données $\mathbf{V}$ à partir des CQT pour la classification de scènes.

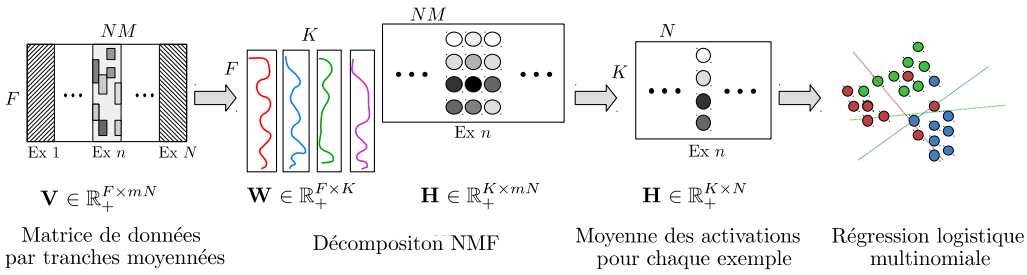


FIGURE 4.4 – Les principales étapes, allant de la matrice de données jusqu'au classifieur, des systèmes de classification de scènes par factorisation de matrices.

Classifier les activations Une fois la matrice de données $\mathbf{V}$ construite, nous la décomposons par factorisation de matrices de sorte à apprendre le dictionnaire d'événements de base. Puis, la matrice $\mathbf{V}$ est projetée sur le dictionnaire afin d'obtenir la matrice d'activations $\mathbf{H} \in \mathbb{R}^{K \times mN}$ utilisées comme descripteurs. Étant donné que nous devons prendre une seule décision pour chaque enregistrement, nous moyennons les m vecteurs correspondant aux projections de chacun des m vecteurs $\mathbf{s}_i$ extraits du même exemple audio de la base. Pour résumer, chaque exemple de la base est représenté par la moyenne des projections sur le dictionnaire de ses différentes tranches temps-fréquence moyennées. Nous illustrons les principales étapes de l'apprentissage de nos systèmes de classification de scènes sur la figure 4.4.

Dans ce chapitre nous choisissons la régression logistique multinomiale comme classifieur. Le première raison de ce choix est la simplicité du modèle, permettant de maintenir l'attention sur la qualité des représentations apprises par factorisation de matrices lors de leur comparaison. Contrairement aux SVM à noyaux, la régression logistique possède moins de paramètres à régler et pose moins de problèmes de complexité lorsque le nombre d'exemples augmente. Ce classifieur a aussi l'avantage de fournir en sortie pour chaque exemple sa probabilité d'appartenance à chaque catégorie. Enfin, les modèles de factorisation supervisée de matrices et par réseaux de neurones que nous présenterons par la suite utilisent également une forme équivalente à la régression logistique

pour l'étape de décision.

4.3.2 Quelques variantes de factorisation de matrices

Nous présentons brièvement dans cette section les formulations des variantes de factorisations de matrices appliquées pour l'apprentissage de descripteurs. Elles correspondent toutes à des extensions de la PCA et de la NMF.

Factorisation de matrices avec contraintes de parcimonie La parcimonie est un concept presque incontournable et désiré dans un grand nombre d'applications des factorisations de matrices. L'objectif est d'obtenir des décompositions plus robustes et interprétables [Plumbley et al., 2010]. Pour la classification de scènes, nous visons à apprendre un dictionnaire d'événements de base capable de représenter les différentes catégories de scènes sonores. Dans ce cas, pour avoir une décomposition plus interprétable, chaque exemple de la base de données devrait être expliqué uniquement par les événements de base les plus pertinents pour caractériser la scène en question. Ainsi, nous cherchons à obtenir des matrices d'activations parcimonieuses, limitant le nombre de composantes actives pour chaque exemple. Il existe de nombreuses manières d'ajouter des contraintes ou de réguler la parcimonie des activations pour la PCA et la NMF. Cependant dans le cas général, les factorisations de matrices avec parcimonie s'expriment comme le problème d'optimisation suivant :

$$\min_{\mathbf{W},\mathbf{H}} D_\beta(\mathbf{V}\|\mathbf{WH}) + \lambda \sum_{k=1}^{K} \|\mathbf{h}_{k:}\|_1 \text{ s.t. } \|\mathbf{w}_k\|_2 = 1; \tag{4.7}$$

où $\mathbf{h}_{k:}$ représente la ligne de $\mathbf{H}$ et $\mathbf{w}_k$ la colonne de $\mathbf{W}$ indexée par k, $1 \leq k \leq K$.

PCA parcimonieuse (SPCA) De multiples approches ont été proposées pour étendre la PCA aux décompositions parcimonieuses. Certaines prennent en compte les contraintes d'orthogonalité des vecteurs de base [Zou et al., 2006], d'autres proposent une approche probabiliste [Guan et Dy, 2009] ou encore en étendent l'approche par déflation de la PCA [Mackey, 2009]. Nous choisissons la formulation telle que définie dans les travaux de Mairal et al. [2009a] qui présentent la SPCA comme un problème d'apprentissage de dictionnaires plus général tout en proposant des algorithmes efficaces pour estimer la factorisation. Dans ce cadre, les matrices $\mathbf{W}$ et $\mathbf{H}$ sont solutions du problème donnée par l'équation (4.7), où D_β est choisie comme la distance euclidienne ($\beta = 2$).

NMF parcimonieuse (SNMF) De même, de multiples approches ont été proposées pour ajouter des contraintes de parcimonie au modèle NMF. La plus intuitive est d'ajouter simplement un terme régulant la norme ℓ_1 de $\mathbf{H}$ dans le problème NMF de l'équation (4.2). Il a été montré que les algorithmes par règles multiplicatives étaient beaucoup plus stables et fournissaient de meilleures décompositions lorsque la norme des vecteurs de base est également contrainte dans le problème SNMF [Eggert et Körner, 2004; Le Roux et al., 2015b]. Dans ce cas, la SNMF s'exprime comme le problème d'optimisation défini à l'équation (4.7) en y ajoutant les contraintes de positivité sur les deux matrices. L'algorithme par règles multiplicatives et les avantages de cette formulation sont discutés dans [Le Roux et al., 2015b].

Décomposition de matrices à noyaux Les familles de décompositions de matrices à noyaux telles que la PCA à noyaux [Schölkopf et al., 1998] (KPCA) et la NMF à noyaux [Zhang et al., 2006] (KNMF) cherchent à décomposer les données dans un espace transformé de plus grande dimension (ce qui permet de traiter plus efficacement les données qui ne sont pas linéairement

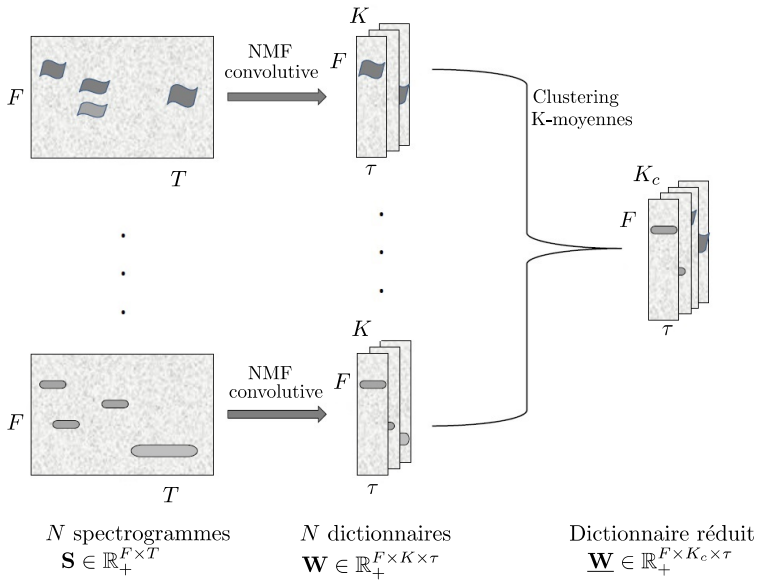


FIGURE 4.5 – Construction du dictionnaire par K-moyennes lors de l'utilisation de la NMF convolutive.

séparables). Étant donné une fonction Φ de l'espace de départ des descripteurs vers l'espace transformé, la factorisation approxime les données $\Phi(\mathbf{V})$ dans cet espace : $\Phi(\mathbf{V}) \approx \mathbf{W}_\Phi \mathbf{H}$, où $\mathbf{W}_\Phi$ est le dictionnaire de vecteurs de base dans l'espace transformé. Bien souvent, les vecteurs de base dans $\mathbf{W}_\Phi$ sont définis comme des combinaisons linéaires des données dans l'espace transformé. Selon le noyau, on peut ne pas avoir accès directement aux données dans l'espace transformé, mais nous connaissons l'expression du produit scalaire dans ce nouvel espace. Ainsi, on peut obtenir la matrice d'activations $\mathbf{H}$ sans connaître explicitement $\Phi(\mathbf{V})$ et $\mathbf{W}_\Phi$ en ayant accès à l'expression de $\mathbf{W}_\Phi^T \Phi(\mathbf{V})$. Les extensions à noyaux de la PCA et la NMF sont basées sur ce même principe et sont décrites avec plus de détails dans [Schölkopf et al., 1998] et [Zhang et al., 2006] respectivement.

NMF convolutive La NMF convolutive [O'Grady et Pearlmutter, 2006; Smaragdis, 2004] est une extension, introduite pour la séparation de sources, intégrant du contexte temporel dans la formulation de base de la NMF. L'objectif de la NMF convolutive est d'apprendre des éléments de dictionnaires représentés en 2 dimensions. Si l'on décompose des représentations temps-fréquence, les éléments du dictionnaire sont des tranches de spectrogrammes représentant des groupes de trames successives. Ainsi, la NMF convolutive nous permet de décomposer les spectrogrammes en différentes tranches de base contenant une représentation temps-fréquence des événements apparaissant au cours de la scène. Si on prend un spectrogramme $\mathbf{S} \in \mathbb{R}_+^{F \times T}$, avec F bandes de fréquences et T trames temporelles, la NMF convolutive vise à décomposer $\mathbf{S}$ en K tranches longues de τ trames. On cherche alors à approximer $\mathbf{S}$ de la manière suivante :

$$\mathbf{S} \approx \sum_{t=0}^{\tau-1} \mathbf{W}_t \overset{t\rightarrow}{\mathbf{H}}, \tag{4.8}$$

où $\mathbf{W}_t \in \mathbb{R}_+^{F \times K}$ et la $k^{\text{éme}}$ colonne de $\mathbf{W}_t$ correspond à la trame t du vecteur de base 2D indexé par k, $1 \leq k \leq K$. Appliquer l'opération $t \rightarrow$ à $\mathbf{H}$ déplace ses colonnes t indices sur la droite en mettant les t premières à 0. Comme la NMF convolutive est particulièrement appropriée à la

décomposition d'images temps-fréquence, nous ne l'appliquerons pas directement à la matrice de données $\mathbf{V}$ présentée précédemment. A la place, nous apprenons un dictionnaire en trois dimensions noté $\mathbf{W}_i \in \mathbb{R}_+^{F \times K \times \tau}$ pour chaque exemple audio i dans l'ensemble d'apprentissage. Les dictionnaires $\mathbf{W}_i$ sont ensuite rassemblés pour former un dictionnaire global $\hat{\mathbf{W}} = [\mathbf{W}_1, ..., \mathbf{W}_N]$. Dans le but de diminuer la taille et la redondance du dictionnaire, nous le réduisons en appliquant un regroupement par K-moyennes sur $\hat{\mathbf{W}}$ donnant un dictionnaire réduit $\underline{\mathbf{W}}$, contenant les K_c centres des *clusters* définis par les K-moyennes. Ce procédé est résumé dans la figure 4.5. Le vecteur de descripteurs appris pour un exemple est alors obtenu en décomposant son spectrogramme sur $\underline{\mathbf{W}}$ et en gardant la moyenne à travers le temps des projections sur le dictionnaire (la moyenne des lignes de $\mathbf{H}$).

4.4 Expériences sur la classification de scènes

4.4.1 Objectifs et bases de données

L'objectif de cette première évaluation des variantes de factorisation de matrices est de comparer leur efficacité pour la tâche de classification de scènes sonores. Nous analysons notamment l'impact de quelques paramètres clés de certaines représentations tels que le choix de la divergence, de la contrainte de parcimonie et la taille des dictionnaires. De plus, nous comparons nos systèmes par NMF et PCA aux approches par extraction de descripteurs présentées dans le chapitre précédent. Les systèmes que nous présentons fournissent des avancées de performance par rapport aux approches par descripteurs mais ils nous servirons également de systèmes de référence crédibles pour la suite de notre travail.

Nous reprenons dans cette section les trois bases de données de classification de scènes utilisées pour l'évaluation des descripteurs d'images dans le chapitre 3 : les bases LITIS, DCASE 2016 et DCASE 2017.

4.4.2 Représentation temps-fréquence

Nous gardons la CQT telle que présentée et discutée dans le chapitre précédent comme représentation temps-fréquence pour la classification de scènes. Les expériences sur les descripteurs d'images ainsi que des expériences préliminaires sur les factorisations de matrices ont confirmé l'intérêt de la CQT pour cette tâche. Les CQT sont extraites en utilisant la toolbox YAAFE [Mathieu et al., 2010]. La première étape est toujours de normaliser le signal audio de sorte à ce que ses valeurs restent dans l'intervalle $[-1, 1]$. Pour la base de données du LITIS, les CQT sont extraites en utilisant 12 bandes par octave de 5 à 11kHz pour donner 134 bandes de fréquences en utilisant des fenêtres de 60 ms sans recouvrement. Pour les bases du DCASE 2016 et 2017, les enregistrements étant de meilleure qualité, l'intervalle de fréquences est choisi de 5 à 22 kHz, ce qui donne 146 bandes de fréquences.

Pour la construction de la matrice de données $\mathbf{V}$ présentée section 4.3.1 et figure 4.3, nous choisissons des tranches de spectrogramme de 2 secondes sans recouvrement pour les bases du LITIS et DCASE 2016 ce qui donne $M = 15$ tranches par exemple. Pour la base du DCASE 2017, du fait de la longueur inférieure des exemples (10 secondes), nous prenons des tranches de 1 seconde sans recouvrement donnant $M = 10$ tranches. La taille des tranches a été réglée sur des expériences préliminaires en utilisant la NMF simple. Nous avons remarqué que prendre des tranches plus courtes n'améliorait pas notablement les résultats, ainsi les valeurs choisies sont un compromis entre performance et complexité.

Enfin nous appliquons une compression de la dynamique des amplitudes en logarithme sur les représentations temps-fréquence pour la PCA et ses variantes. Pour la NMF et ses variantes,

nous appliquons une compression en racine carrée, ces méthodes n'étant définies que pour des données à coefficients positifs. D'autres compressions telles que $log(1 + x)$ assurant la positivité ont été testées sans amélioration notable des performances. Une fois compressée, nous normalisons la matrice de données de sorte à ce que chaque dimension soit de variance unitaire pour toutes les méthodes et de moyenne nulle pour la PCA et ses variantes. Les mêmes compressions des représentations temps-fréquence seront appliquées pour les systèmes des chapitres suivants indépendamment des techniques utilisées.

4.4.3 Classification et métriques

Avant l'étape de classification, les descripteurs appris par factorisation de matrices sont également normalisés de sorte à ce que chaque dimension soit de variance unitaire et de moyenne nulle. Nous rappelons que le classifieur est une régression logistique multinationale apprise avec un algorithme L-BFGS en utilisant l'implémentation de la régression logistique de *scikit-learn* [Pedregosa et al., 2011]. Comme justifié dans le chapitre précédent, nous choisissons le score F1 comme métrique principale afin d'évaluer les méthodes comparées. Nous utiliserons le taux de reconnaissance lors de la comparaison aux méthodes de l'état de l'art.

4.4.4 Résultats

Factorisation de matrices simples

Nous commençons par présenter sur la figure 4.6, les résultats obtenus en utilisant la PCA et la NMF dans leur formulation originale. Pour la PCA, nous présentons des résultats seulement pour $K = 128$ car le nombre de composantes est limité à la dimension de la représentation d'entrée. De plus, nous avons observé expérimentalement que choisir un nombre de composantes inférieur ne fait que dégrader les résultats. Pour la NMF, les décompositions obtenues sont susceptibles de changer d'une initialisation à l'autre car le problème d'optimisation de la NMF n'a pas de solution unique. Cette sensibilité à l'initialisation introduit une part d'aléatoire dans la décomposition. Afin de limiter l'impact de cet effet lors de l'apprentissage du dictionnaire, nous réalisons la NMF sur 10 initialisations aléatoires différentes. Ensuite, nous gardons le dictionnaire correspondant à la décomposition ayant obtenu le critère d'attache aux données le plus faible. Enfin, les descripteurs sont obtenus en projetant les données sur le dictionnaire appris à partir de la base d'apprentissage. Nous utilisons l'implémentation de la NMF par règles multiplicatives ainsi que son équivalent pour processeurs GPU [Serizel et al., 2016c].

Le premier élément remarquable sur la figure 4.6 est que pour les trois bases de données, l'apprentissage de descripteurs par NMF est plus efficace qu'en utilisant une simple PCA. Il s'agit d'une première confirmation que la contrainte de positivité de la décomposition joue un rôle important dans la qualité des représentations apprises. Toutefois, on remarque dans le même temps que les performances de la NMF continuent de croître avec la taille du dictionnaire, où pour les trois bases, les meilleurs scores F1 sont atteints pour $K = 1024$. Nous n'avons pas noté d'améliorations en doublant encore la taille des dictionnaires ($K = 2048$), alors que les calculs deviennent significativement plus coûteux en temps. Ainsi, dans la suite de ce chapitre et du manuscrit nous ne présenterons pas de résultats au delà de $K = 1024$, qui s'est révélé être la meilleure taille de dictionnaire pour la classification de scènes pour la plupart des bases et des méthodes non-supervisées. Dans cette configuration, nous nous éloignons fortement de l'utilisation des factorisations de matrices pour la réduction de dimension. En effet, au lieu de chercher à réduire la taille de la représentation d'entrée, la NMF apprend des descripteurs plus riche par la décomposition de l'ensemble des scènes en une somme d'un grand nombre d'événements élémentaires.

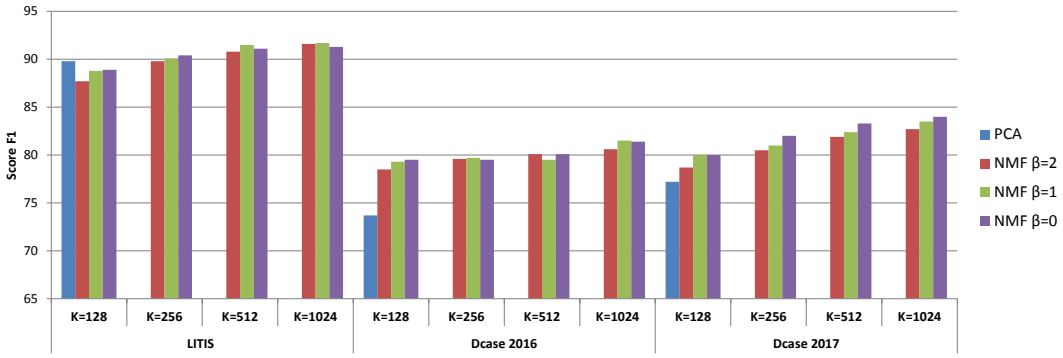


FIGURE 4.6 – Score F1 pour la PCA et la NMF simple sur les trois bases en fonction de la taille du dictionnaire.

Cette première expérience nous permet également de montrer que, pour les trois bases de données, même un système d'apprentissage de descripteurs basé sur une NMF simple et un classifieur linéaire permet d'améliorer significativement les scores par rapport aux systèmes de référence. De plus, même si ce n'est pas le cas pour la base du LITIS, le système NMF permet déjà d'obtenir des performances supérieures aux approches par descripteurs d'images présentées au chapitre 3. La présentation de ces résultats permet également d'observer l'influence du choix de la divergence sur les performances. Comme nous l'avons mentionné dans la section 4.2, il est commun en audio de choisir des coûts d'attache aux données autres que la distance euclidienne, en particulier les divergences KL et IS. Dans cette première application de la NMF à la classification de scènes on peut également remarquer que la distance euclidienne a tendance à fournir des scores F1 légèrement inférieurs aux autres divergences. En revanche, il n'y en a aucune qui domine clairement à travers tous les résultats. Cela a pour conséquence pour nous de considérer le choix de la divergence comme un paramètre supplémentaire à régler en fonction de la base de données, de la variante NMF et de la méthode classification.

Influence des contraintes de parcimonie

Nous nous intéressons maintenant à l'influence de l'ajout de contraintes de parcimonie aux factorisations de matrices par PCA et NMF. Nous présentons sur la figure 4.7 l'évolution du score F1 en fonction de λ_1, le paramètre contrôlant l'influence de la pénalité de norme ℓ_1 sur la matrice d'activations. Pour la SNMF, les résultats sont présentés en utilisant la distance euclidienne ($\beta = 2$) comme coût d'attache aux données. Dans la plupart des cas présentés, la distance Euclidienne permet d'obtenir de légèrement meilleures performances. Le procédé d'apprentissage des descripteurs reste le même que pour la NMF présenté ci-dessus avec la contrainte de parcimonie en plus.

La première tendance que l'on peut observer en regardant les courbes présentées sur la figure 4.7 est que l'ajout de la parcimonie permet de notablement améliorer les performances dans la plupart des configurations testées. Les seules configurations où la parcimonie n'est parfois pas bénéfique sont les cas où les dictionnaires NMF sont de plus petite taille ($K = 128$). En effet, l'ajout de la parcimonie ℓ_1 a pour conséquence de limiter le nombre de composantes actives lors de la projection sur le dictionnaire. Cela permet à un instant donné de ne garder que les composantes (ou événements élémentaires) représentant au mieux le contenu fréquentiel de l'exemple. Limiter le nombre de composantes actives par exemple est d'autant plus important lorsque les dictionnaires sont appris sur l'ensemble de la base de données. Ainsi, ils représentent toute la variabilité des différents événements contenus dans les scènes sonores de chaque base. Limiter la quantité

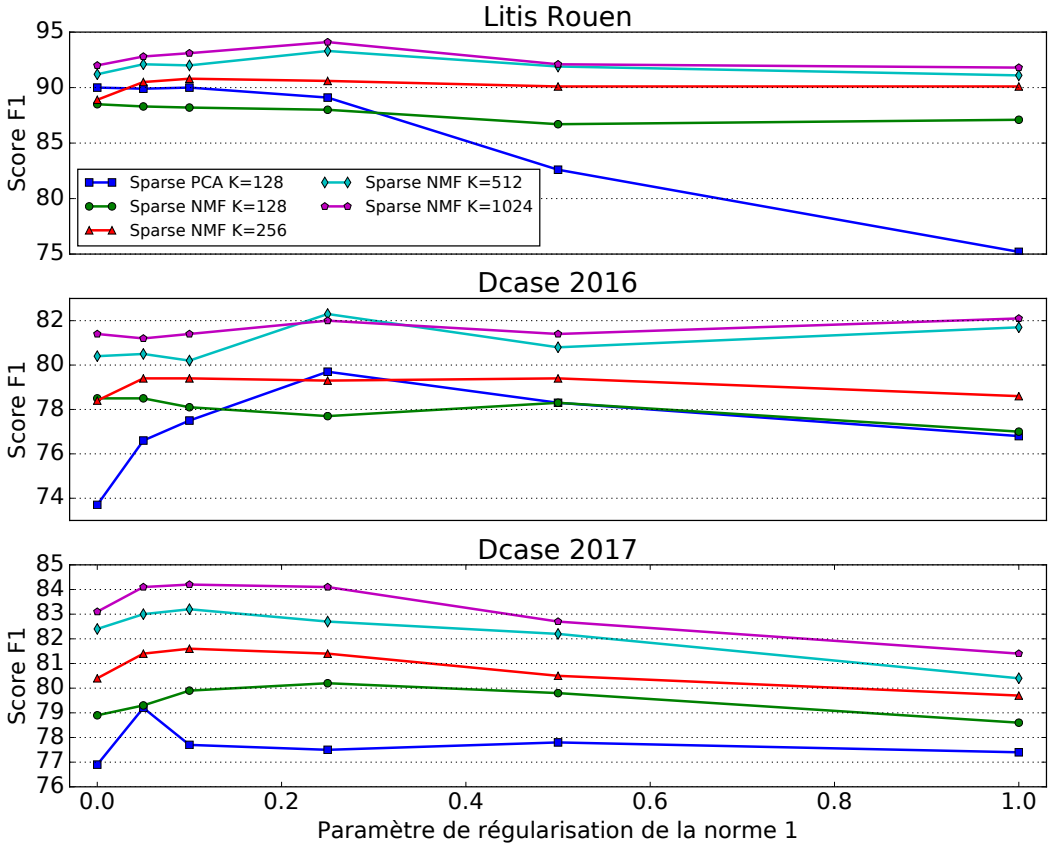


FIGURE 4.7 – Scores F1 pour SPCA et SNMF sur les trois bases de données. Les courbes pour chaque taille de dictionnaire testée donnent le score F1 en fonction de la valeur du paramètre de régularisation de la norme ℓ_1.

d'événements actifs par scène aide à l'interprétabilité de la décomposition à la fois pour l'Homme et pour la machine. On peut également noter, sans surprise, que pour chaque taille de dictionnaire il existe un seuil à partir duquel augmenter la valeur du paramètre de régularisation détériore les résultats. En revanche, ces résultats ne permettent pas de conclure définitivement quant à une valeur optimale du paramètre de régularisation associé à chaque base de données. Celui-ci pouvant également dépendre des autres étapes des systèmes de classification de scènes, nous le considérons comme un paramètre à régler au cas par cas.

Décomposition à noyaux

Nous présentons brièvement quelques résultats obtenus avec l'extension à noyaux de la PCA présentée à la section 4.3.2. Nous utilisons un noyau gaussien dont le paramètre σ est réglé sur un ensemble de développement en parcourant l'ensemble $\{1, 10, 100, 200\}$. Les scores F1 obtenus avec KPCA sur les trois bases de classification de scènes sont indiqués dans le tableau 4.1. Bien que nous ayons mentionné une version à noyaux de la NMF, la complexité de ces méthodes nous empêche de fournir des résultats pour des grandes tailles de dictionnaires. Des résultats avec KNMF sur de petits dictionnaires ont tout de même été présentés dans nos travaux [Bisot et al., 2016a], mais les performances obtenues restent limitées. La présence de la matrice de Gram $\Phi(\mathbf{V})^T\Phi(\mathbf{V}) \in \mathbb{R}^{N\times N}$ dans les règles de mise à jour de l'algorithme rend la KNMF bien plus complexe que la NMF lorsque $N >> F$. En revanche, comme la décomposition KPCA se base

sur des décompositions en valeurs singulières, le problème de complexité est moindre et nous permet d'obtenir des résultats en un temps de calcul raisonnable.

L'utilisation de la formulation à noyaux de la PCA ne limite pas la taille du dictionnaire, tant que $K < N$, ce qui nous permet de présenter des résultats sur une variante de la PCA avec un nombre de composantes plus élevé. Sur les trois bases de données, les scores F1 obtenus par KPCA sont supérieurs à ceux obtenus avec la PCA simple. Tout comme c'est le cas pour certains classifieurs, il semblerait que dans ce cas, l'utilisation de noyaux dans la décomposition permet d'améliorer les qualités des descripteurs appris. Cependant, les performances de la KPCA restent en dessous de celles obtenues par NMF pour les deux bases du DCASE. L'utilisation de noyaux ne permet pas de rattraper l'amélioration de la qualité des décomposions qu'offre la contrainte de positivité dans la NMF. Les bonnes performances de KPCA sur la base du LITIS peuvent s'expliquer par la construction de la base de données. En effet, certains exemples provenant du même lieu peuvent se retrouver à la fois dans l'ensemble d'apprentissage et de test. Les méthodes à noyaux permettent de profiter du biais introduit par la présence d'exemples très similaires dans les deux ensembles.

	LITIS Rouen		**DCASE 2016**		**DCASE 2017**	
	K=512	K=1024	K=512	K=1024	K=512	K=1024
PCA (K=128)	89.8		73.7		76.9	
Kernel PCA	94.3	**95.6**	**79.7**	**79.5**	**81.7**	**82.0**

TABLEAU 4.1 – Score F1 pour KPCA sur différentes tailles de dictionnaires K et sur les 3 bases de classification de scènes.

NMF convolutive

Nous terminons la comparaison de l'évaluation des méthodes d'apprentissage de dictionnaires par la NMF convolutive. Nous rappelons que pour la NMF convolutive, contrairement aux autres méthodes présentées précédemment, nous n'appliquons pas de découpage ni d'intégration temporelle aux représentations temps-fréquence. Dans ce cas le schéma d'apprentissage de descripteurs utilisé suit celui décrit à la section 4.3.2 et sur la figure 4.5. Afin de rendre la comparaison à la NMF simple plus juste, nous l'évaluons également avec le même procédé de décomposition des spectrogrammes que la NMF convolutive. Ainsi de manière analogue, nous décomposons chaque spectrogramme individuellement avec la NMF afin de construire un dictionnaire final en regroupant par K-moyennes tous les sous-dictionnaires appris. Nous ferons référence à cette stratégie d'apprentissage comme *NMF-km*. Pour la NMF convolutive, nous décomposons les CQT avec 80 vecteurs de bases 2D longs de $\tau = 4$ trames pour la base du LITIS et avec 40 vecteurs de base 2D longs de $\tau = 8$ trames pour les bases DCASE. La concaténation de tous les dictionnaires obtenus est réduite par K-moyenne en gardant les K_c centres pour former le dictionnaire final. Nous présentons les scores F1 pour la NMF-Km et la NMF convolutive en fonction du nombre de centres K_c dans le tableau 4.2

Nous nous sommes intéressés à la NMF convolutive dans l'objectif d'évaluer si l'introduction d'une décomposition incluant du contexte temporel dans les vecteurs de base permettait d'améliorer la qualité des représentations apprises. Les résultats présentés dans le tableau 4.2 semblent confirmer dans une certaine mesure les avantages de la NMF convolutive. En effet, sur les trois bases, la NMF convolutive permet d'obtenir de meilleures performances que la NMF simple évaluée en utilisant le même schéma de décomposition par regroupement.

	LITIS Rouen			DCASE 2016			DCASE 2017		
K_c	256	512	1024	256	512	1024	256	512	1024
NMF-km	90.1	92.2	**93.7**	76.1	79.6	79.9	75.7	78.9	82.6
NMF convolutive	90.5	92.6	**94.5**	77.7	80.8	**82.5**	81.0	82.5	**83.4**

TABLEAU 4.2 – Scores F1 pour la NMF convolutive et NMF-km pour différentes tailles de dictionnaires K_c.

Cette expérience nous amène également à discuter du choix de l'architecture pour nos systèmes d'apprentissage de représentations. En effet, contrairement aux autres méthodes, nous utilisons la NMF convolutive et la NMF-km pour décomposer les représentations temps-fréquence telles qu'elles ont été extraites. Au contraire, la stratégie d'intégration temporelle par utilisée pour évaluer les autres variantes NMF altère l'information temporelle des représentations temps-fréquence en moyennant les tranches par blocs. On pourrait supposer que cette dernière approche limite possiblement la qualité des représentations. Le fait de moyenner les tranches sur 2 secondes entraîne nécessairement une perte d'information par rapport aux spectrogrammes complets. Toutefois, les résultats obtenus avec SNMF en utilisant la construction par tranches moyennées (TM) semblent cependant indiquer le contraire. En effet, les performances de la SNMF sont équivalentes ou supérieures à celles obtenues avec la NMF convolutive. Comme nous l'avons vu pour les descripteurs d'images, les scènes sonores se caractérisent par la distribution de l'information temps-fréquence sur le long terme et non localement. Cela explique que nous ne gagnons pas nécessairement à décomposer les représentions temps-fréquence en entier, même en introduisant du contexte temporel dans la décomposition avec la NMF convolutive. De plus la construction par TM des données permet de grandement réduire la taille totale des données à décomposer, ce qui entraine une forte réduction du temps de calcul nécessaire pour apprendre et projeter sur le dictionnaire. Ainsi, dans tous les cas étudiés, la construction par TM offre le meilleur compromis entre temps de calcul et performance.

Comparaison aux approches par extraction de descripteurs

Maintenant que nous avons évalué les différentes méthodes d'apprentissage de descripteurs par factorisation de matrices, nous pouvons comparer leurs performances aux approches par extraction de descripteurs présentées dans le chapitre précédent. Pour les trois 3 bases de données, les méthodes définissant l'état de l'art reposent toutes sur des techniques apprentissage de représentation supervisées que nous ne présenterons volontairement pas dans ce chapitre. Nous excluons également pour l'instant de cette comparaison les méthodes par extraction de descripteurs utilisant des réseaux de neurones profonds étant donné que nous utilisons simplement un classifieur linéaire. Des comparaisons plus complètes des méthodes de l'état de l'art seront proposées dans les chapitres suivants. A ce stade, nous souhaitons discuter de l'hypothèse que les techniques proposées d'apprentissage de descripteurs non-supervisées peuvent permettre d'apprendre de meilleures représentations que celles obtenues à partir de schémas d'ingénierie de descripteurs élaborés pour d'autres tâches. Les performances pour les meilleures variantes de factorisation de matrices proposées sont présentées dans le tableau 4.3 et sont comparées aux meilleures approches par extraction de descripteurs ainsi qu'aux systèmes de référence sur chaque base de données.

Le faible nombre d'ensembles de validation croisée (seulement 4) fournis par les auteurs des bases DCASE limite l'estimation des écarts type des résultats obtenus par la moyenne des taux de reconnaissance sur les différents ensembles. Afin de donner une idée de la significativité des ré-

Descripteurs	Classifieur	**LITIS Rouen**	**DCASE 2016**	**DCASE 2017**
Système de référence		91.7*	72.5	74.8
HOG + L-SPD	SVM	94.0	77.7	80.3
MFCC + RQA	SVM	86.0*	67.1	-
KPCA	LR	**96.0**	80.2	82.2
SNMF	LR	94.6	**82.7**	**84.4**
NMF convolutive	LR	94.8	**82.5**	**83.7**

TABLEAU 4.3 – Tableau résumé des meilleurs taux de reconnaissance des techniques d'apprentissage de descripteurs comparés à d'autres approches d'extraction de descripteurs de l'état de l'art pour les trois bases de données. (*) Seul la précision a été donnée pour ces systèmes, les résultats sont inclus à titre indicatif.

sultats, nous calculons pour chaque taux de reconnaissance l'intervalle de confiance à 95% d'une proportion. Le nombre de tirage correspond alors au nombre d'exemple testés au total et le nombre de succès est donné par le taux de reconnaissance. Dans la suite de ce travail, les valeurs en caractères gras affichées dans les tableaux seront toutes celles comprises dans l'intervalle de confiance du meilleur taux de reconnaissance affichés pour chaque base de données. Ces intervalles sont en général de l'ordre de plus ou moins 1 point de taux de reconnaissance pour les bases du DCASE 2017 et LITIS Rouen et de plus ou moins 2 points pour la base DCASE 2016.

La comparaison des techniques de factorisation de matrices pour l'apprentissage de descripteurs que nous avons menée nous a permis d'exhiber 3 variantes différentes de la PCA et la NMF qui offrent des améliorations par rapport aux approches par descripteurs d'images présentées au chapitre 3. Avec la KPCA pour la base du LITIS et la SNMF pour les bases DCASE, nous obtenons une augmentation de 2 points sur le taux de reconnaissance par rapport à la combinaison HOG+L-SPD. Au lieu de fixer des techniques d'extraction du contenu à partir des représentations temps-fréquence, les méthodes proposées dans ce chapitre nous permettent d'apprendre automatiquement l'information facilitant la description des catégories de scènes sonores. Par différentes approches, nous confirmons que les techniques d'apprentissage de dictionnaire sont particulièrement adaptées à la classification de scènes sonores. La projection sur les dictionnaires d'événements de base appris de manière non-supervisée constitue des descripteurs pertinents permettant d'atteindre de bonnes performances même avec un classifieur simple.

La construction de la matrice de données par TM apparait comme une stratégie efficace pour réduire de la taille des données tout en restant performante. Nous serons amenés à la reprendre comme premier bloc des différents systèmes d'apprentissage supervisés présentés dans les chapitres suivants. Nous avons conduit cette étude comparative en nous restreignant à un cadre simple pour la construction des données et la classification afin de faciliter la comparaison des méthodes. S'affranchir de telles limitations laisse des perspectives d'amélioration en ouvrant la possibilité d'utiliser des classifieurs plus complexes ou d'inclure d'autres formes d'intégration et de modélisation temporelle. Parmi les approches comparées, la SNMF fournit les meilleures performances sur les 3 bases de données. Cette dernière possède également l'avantage d'être la moins coûteuse en temps de calcul comparée à la NMF convolutive ou la KPCA, surtout lorsque le nombre de données augmente. La simplicité et l'efficacité de cette approche pour la tâche nous amènera à la considérer comme point de départ pour construire des méthodes plus complexes d'apprentissage supervisé.

Pour finir, nous présentons une matrice de confusion pour la base du DCASE 2017 dans l'ob-

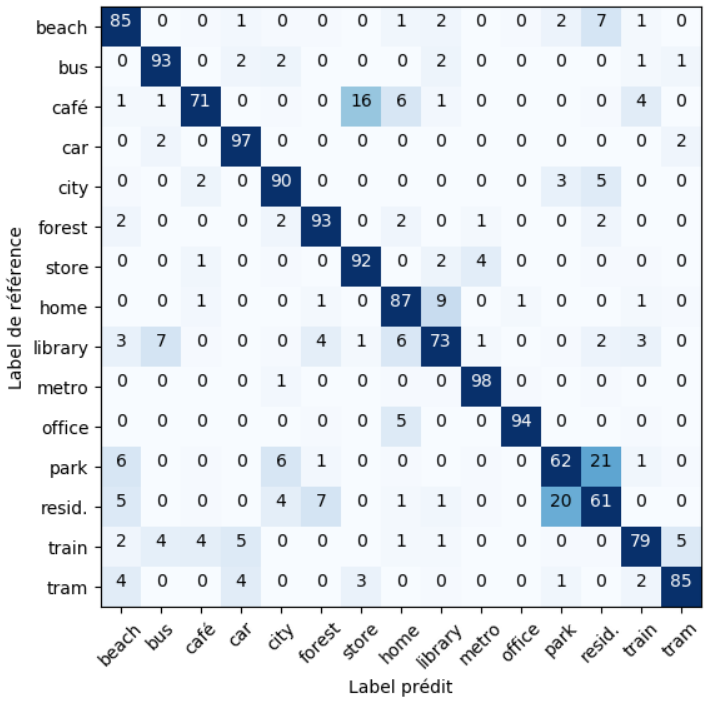


FIGURE 4.8 – Matrice de confusion normalisée pour SNMF appliquée à la base du DCASE 2017.

jectif de rendre compte des difficultés restant à traiter par les systèmes de classification de scènes. La matrice de confusion est obtenue avec le meilleur système SNMF puis elle est normalisée après avoir additionné les confusions pour les 4 ensembles de test. On peut remarquer sur la figure 4.8 que la majorité des confusions se produisent entre des classes correspondant à des environnements acoustiques très similaires. En effet les deux classes les plus confondues sont *residential area* et *park* étant deux environnements plutôt calmes et contenant un grand nombre d'événements caractéristiques communs : bruit de fond de voiture, enfants qui jouent, des oiseaux etc... La même interprétation s'applique également aux autres confusions importantes telles que *store* et *café* ou encore *home* et *library*. Dans ces cas, l'apprentissage non-supervisé ne nous permet pas de dégager clairement des événements élémentaires assez spécifiques pour différencier de telles paires d'environnements. Des techniques augmentant le pouvoir discriminant du dictionnaire deviennent nécessaires afin d'être capable de différencier certaines catégories de scènes sonores aux propriétés similaires.

4.5 Nos premiers systèmes de détection d'événements par NMF non-supervisée

Après la classification de scènes, nous proposons également un cadre simple et efficace pour l'application de techniques d'apprentissage non-supervisé de descripteurs à la détection d'événements avec recouvrement. Contrairement à la classification de scènes, où des décisions sont prises sur plusieurs secondes d'enregistrement, la détection d'événements implique souvent la réalisation de prédictions à une échelle temporelle beaucoup plus fine (allant jusqu'à quelques dizaines de millisecondes). Cela implique l'utilisation d'approches et de métriques différentes, capables de prédire les événements présents à chaque instant d'un enregistrement. De plus, nous nous intéressons plus particulièrement à la détection d'événements avec recouvrement, c'est-à-dire lorsque

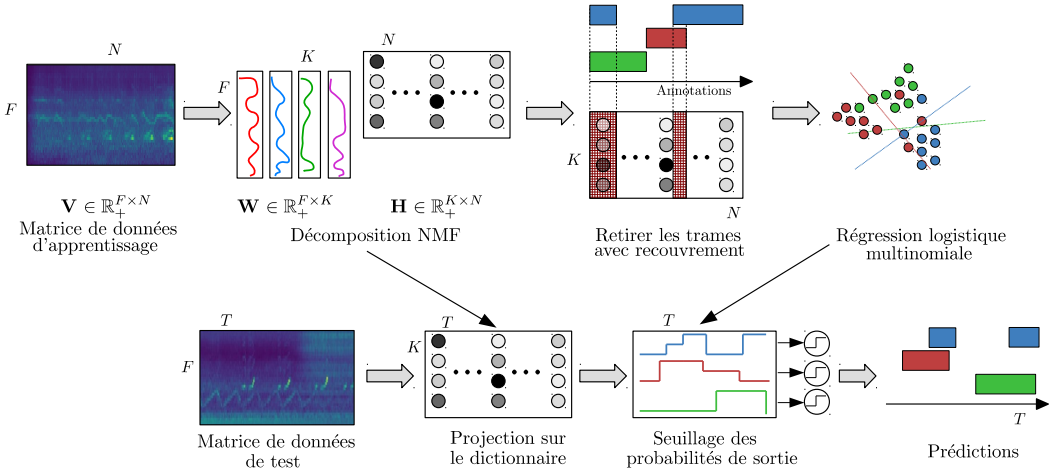


FIGURE 4.9 – Les principales étapes de nos premiers systèmes de détection d'événements avec recouvrement.

plusieurs événements peuvent être présents au même instant dans l'annotation.

Les approches par apprentissage de représentations sont depuis quelques années les approches dominantes en détection d'événements. S'inscrivant dans cette tendance, les méthodes par NMF se sont montrées particulièrement adaptées pour la tâche [Mesaros et al., 2015; Benetos et al., 2016], en particulier de par la capacité de la NMF à apprendre et séparer des représentations caractérisant différent objets se recouvrant. Comme mentionné section 4.3.2, la plupart des méthodes de détection d'événements utilisant la NMF s'en servent comme classifieur. Elles considèrent la matrice d'activations post-traitée comme contenant les occurrences des différents événements à détecter. Á la place, nous proposons un système simple de détection d'événements par NMF en nous en servant uniquement comme technique d'apprentissage de descripteurs. C'est-à-dire que les activations seront utilisées pour apprendre un classifieur qui a pour rôle d'effectuer les prédictions sur les événements présents à chaque instant. L'architecture de détection d'événements que nous proposons dans ce chapitre sera utilisée à plusieurs reprises dans la suite du manuscrit. Les principales étapes du système de détection d'événements que nous proposons sont illustrées sur la figure 4.9 et décrites ci-dessous.

Apprentissage de représentations La matrice de données $\mathbf{V}$ est construite en concaténant les représentations temps-fréquence de tous les enregistrements de la base d'apprentissage. Contrairement à la classification de scènes, nous n'appliquons pas d'intégration temporelle directement sur les représentations bas-niveau car nous devons être capables d'identifier des objets sonores à une échelle temporelle plus courte. Les techniques de factorisation de matrices sont alors appliquées afin d'apprendre un dictionnaire en décomposant la matrice de données $\mathbf{V}$ tel que décrit section 4.2. Les descripteurs sont obtenus en projetant la matrice $\mathbf{V}$ sur le dictionnaire $\mathbf{W}$.

Classification La détection d'événements avec recouvrement est un problème de classification multi-labels. Bien que des techniques existent pour directement traiter les problèmes de classification multi-labels, l'approche la plus courante est de transformer le problème en autant de problèmes de classification à deux classes *un contre tous* qu'il y a d'étiquettes. Nous proposons une stratégie légèrement différente de sorte à traiter la tâche comme un problème multi-classes. Pour

cela, après avoir appris les descripteurs, nous enlevons de la matrice **H** toutes les trames contenant plus d'un événement dans l'annotation. Par ailleurs, nous ajoutons une étiquette *bruit de fond* sur les trames ne contenant pas d'événements, nous avons alors un problème de classification multi-classes. Nous utilisons également une régression logistique multinomiale comme classifieur, en tirant parti de sa capacité à fournir directement les probabilités d'appartenance d'un exemple à chaque classe. En effet, cela nous permet lors de l'étape de prédiction sur les données de tests, de seuiller les probabilités de sortie de manière à rendre possible la prédiction de plus d'un événement par trame. L'approche décrite s'est montrée plus efficace que l'approche *un contre tous* en fournissant de meilleures performances tout en réduisant légèrement le temps de calcul par l'apprentissage d'un seul classifieur.

4.6 Expériences sur la détection d'événements avec recouvrement

4.6.1 Objectifs et bases de données

Dans cette section, nous présentons nos premières expériences sur la détection d'événements sonores avec recouvrement. Cela nous permet d'introduire les bases de données et les métriques pour les chapitres suivants. L'objectif est avant tout de valider l'intérêt de l'apprentissage de descripteurs par NMF sur lesquel nous nous appuierons dans la suite du manuscrit. Nous rappelons qu'une des principales difficultés de la détection d'événements est que les systèmes doivent être capables de caractériser et de détecter des événements se recouvrant parfois dans le temps. Ainsi l'objectif est de confirmer que l'apprentissage de descripteurs par NMF permet de traiter certaines difficultés propres à la tâche. Nous présentons brièvement l'ensemble des bases utilisées dans ce chapitre. La liste des catégories et les statistiques sur le taux de recouvrement sont détaillées en Annexe A.

TUT SED 2016 : Il s'agit de la base de données de détection d'événements avec recouvrement correspondant à la tâche 3 du challenge DCASE 2016 [Mesaros et al., 2016b]. C'est une des seules bases publiques de détection d'événements avec recouvrement en conditions réelles, c'est-à-dire annotées à partir d'enregistrements faits dans différentes scènes sonores. La base de données contient 12 enregistrements de 3 à 5 minutes dans deux environnements différents : *Rue calme* et *au domicile*. Les deux environnements contiennent respectivement une liste de 7 et 11 étiquettes correspondant aux différents événements se produisant dans les enregistrements. La base contient 4 ensembles de validation croisée contenant chacun de 3 à 4 enregistrements.

TUT SED synth 2016 : La base contient 100 mélanges synthétiques d'environ 5 minutes créés à partir d'événements isolés collectés en ligne [Cakir et al., 2017]. La base contient 16 catégories d'événements différentes et les mélanges peuvent contenir des taux de recouvrement allant jusqu'à 5 événements simultanés à un instant donné. Les mélanges sont répartis en 3 ensembles apprentissage-test-validation 60%-20%-20%.

Contrairement aux bases de classification de scènes, assez peu de travaux ont été publiés sur les deux bases de données de détection d'événements que nous considérons pour l'évaluation. Nous pouvons donc nous permettre d'inclure directement les performances des meilleurs systèmes de l'état de l'art. La plupart d'entre eux sont basés sur des réseaux de neurones profonds, en utilisant soit des perceptrons multi-couches (MLP), des réseaux convolutifs (CNN), récurrents (RNN) ou la combinaison de ces deux derniers (CRNN).

4.6.2 Métriques

La détection d'événements avec recouvrement est un problème multi-labels, il nécessite l'utilisation de métriques spécifiques. Nous reprenons les deux métriques les plus largement utilisées

par la communauté, le taux d'erreur (ER pour *Error Rate* en anglais) et le score F1. Ces deux métriques ont initialement été introduites pour certaines tâches de traitement de la parole possédant des problématiques similaires. Elles sont ensuite devenues la norme en détection d'événements sonores [Mesaros et al., 2016a], en particulier dû à leur utilisation dans de nombreuses tâches des challenges DCASE. Le taux d'erreur est calculé comme la somme des insertions, suppressions et substitutions divisée par le nombre total d'événements dans une fenêtre d'une seconde. Le score F1 est la moyenne harmonique entre la précision et le rappel par segments de 1 seconde. Le calcul de ces est présenté plus en détail dans l'annexe B. Bien que les méthodes soient principalement comparées en calculant les métriques sur des segments de 1 seconde, il est également intéressant de réduire la taille des fenêtres de mesure pour évaluer la précision temporelle des modèles. Pour cela nous reportons parfois les métriques par trame que nous notons ER_{tr} et $F1_{tr}$. Ces grandeurs sont obtenues en évaluant les prédictions à la même résolution que les fenêtres d'analyse des représentations temps-fréquence.

Pour les deux bases de données, les métriques sont calculées en une seule évaluation sur tous les ensembles de test, et non moyennées à travers ceux-ci. C'est-à-dire qu'une fois les prédictions réalisées pour chaque enregistrement, l'ER et le score F1 sont calculés en une seule fois sur l'ensemble des trames d'évaluation.

4.6.3 Protocole expérimental

Représentation temps-fréquence Au cours de toutes nos expériences sur la détection d'événements nous utiliserons toujours des spectres Mel comme représentation temps-fréquence. Ils nous ont permis d'obtenir des performances similaires avec moins de bandes de fréquences lorsque nous les avons comparés aux CQT ou aux TFCT au cours d'expériences préliminaires. Les spectres Mel sont extraits après avoir normalisé le signal dans l'intervalle $[-1, 1]$. Nous choisissons des fenêtres de 40ms avec 50% de recouvrement et 40 bandes de fréquence pour la base réelle et 80 bandes pour la base synthétique. Il s'agit de réglages largement utilisés dans le domaine [Cakir et al., 2015b, 2017; Mesaros et al., 2016b].

Réglage de la NMF Les descripteurs sont appris avec la SNMF telle qu'introduite pour la classification de scènes. Nous considérons, pour la détection d'événements, l'utilisation des distances euclidiennes ($\beta = 2$) et de la divergence KL ($\beta = 1$). Après une validation préliminaire, le paramètre de régularisation pour la parcimonie ℓ_1 est réglé à $\lambda_1 = 0.5$ pour la base réelle et à $\lambda_1 = 0$ pour la base synthétique. Les résultats sont présentés pour des tailles de dictionnaire de $K = 16$ pour la base réelle et de $K = 32$ pour la base synthétique. Augmenter la taille des dictionnaires dans les deux cas ne permet pas d'améliorer les performances. Tous ces paramètres sont réglés sur un ensemble de développement associé à chaque base de données.

En particulier pour la base TUT SED synth, le score F1 et l'ER par trame sont considérés comme deux des métriques principales pour la comparaison des travaux état de l'art. Calculer ces métriques à une résolution temporelle aussi basse avantage fortement les systèmes capables de prendre en compte le voisinage temporel de la trame donnée dans leur estimation. Or, nos systèmes effectuent la classification trame par trame et ne profitent donc d'aucune prise en compte du contexte temporel ni pour la décomposition ni pour la classification. Afin de simuler simplement la prise en compte d'une forme de modélisation temporelle, nous proposons le système *NMF contextuelle*. Pour cela, les projections NMF à un instant donné sont remplacées par leur moyenne et leur écart type dans les 20 trames voisines centrées sur la trame courante. Cela revient à faire de la détection d'événements par classification de fenêtres glissantes, avec un décalage d'une trame. Des expériences préliminaires nous ont amené à retenir cette solution et non la NMF convolutive

car les systèmes NMF contextuelle se sont montrés à la fois moins coûteux à apprendre et plus performants.

Classifieurs

De même que pour la classification de scènes, le classifieur utilisé est une régression logistique multinomiale. Afin de s'affranchir de l'aspect multi-labels de la tâche lors de l'apprentissage, nous apprenons le classifieur uniquement sur les trames ne contenant pas de recouvrement. Pour la prédiction, les probabilités de sortie sont seuillées à 0.35 pour l'environnement *residential area* de la base réelle et à 0.3 pour l'environnement *home* ainsi que pour l'ensemble de la base synthétique. Enfin, nous post-traitons simplement les prédictions par un filtre médian long de 7 trames afin de retirer les points isolés temporellement dans nos prédictions.

4.6.4 Résultats

TUT SED 2016

Les résultats sur la base TUT SED 2016 sont présentés dans le tableau 4.4. Ils sont comparés aux systèmes ayant obtenus les meilleurs résultats sur la base de développement du challenge DCASE 2016 auquel est associé la base de données. Les systèmes sont évalués et appris indépendamment sur les deux environnements acoustiques différents présents dans la base. Les résultats sont présentés sous forme de métriques calculées comme la moyenne sur les deux environnements (colonne de droite du tableau).

On peut noter que les approches de l'état de l'art par réseaux de neurones profonds apprises sur les représentations temps-fréquence [Vu et Wang, 2016; Zöhrer et Pernkopf, 2016] atteignent de bien meilleures performances que les systèmes basés sur l'extraction de descripteurs MFCC [Elizalde et al., 2016; Mesaros et al., 2016b]. Nos systèmes NMF partent de représentations bas niveau très similaires à celles des autres méthodes de l'état de l'art mais utilisent un classifieur beaucoup plus simple (une régression logistique). Les deux systèmes que nous évaluons, pour les deux divergences différentes, se placent parmi les meilleurs taux d'erreur sur la base de développement mais avec des scores F1 plus faibles que les meilleurs systèmes. Ainsi, en ajoutant simplement une étape d'apprentissage de descripteurs non-supervisé par NMF on peut obtenir des performances compétitives avec des systèmes par réseaux de neurones profonds plus complexes.

Ces résultats sont tout de même à prendre avec précaution pour deux raisons. La première est que les résultats sur la base de test du challenge (non publique) ont fortement modifié le classement des méthodes par rapport aux résultats sur la base développement (présentés ici). La deuxième réside dans les limitations apportées par la taille et la nature de la base de données. Cette base étant annotée à la main à partir d'enregistrements en conditions réelles, elle ne contient que peu d'enregistrements (20 de 5 minutes au total). De plus, enregistrer en conditions réelles implique que les probabilités d'apparition des différents événements d'intérêt dans la scène sonore sont réalistes. Cela entraîne souvent un fort déséquilibre dans la représentation de chaque étiquette. Ainsi certains événements sont majoritaires tels que *car passing by* dans l'environnement *residential area* par rapport à d'autres plus rares tels que *object impact*. Comme les classifieurs disposent de peu d'information pour apprendre à reconnaître ces événements plus rares, ils ont tendance à uniquement prédire les événements les plus fréquents. Ce phénomène explique en partie les performances très faibles de tous les systèmes sur l'environnement *home* car il contient un nombre important d'événements différents peu représentés. Enfin, on peut également noter que la NMF avec la divergence de KL obtient des performances légèrement supérieures à celles obtenues avec la distance euclidienne. L'intuition étant que la divergence KL permet de limiter l'influence négative du bruit de fond et de la variation d'intensité des événements comme cela a été montré pour

			Home		Res. A.		Moyenne	
	Méthode	Features	ER	F1	ER	F1	ER	F1
[Mesaros et al., 2016b]	GMM	MFCC	-	-	-	-	91	23.7
[Elizalde et al., 2016]	Forêts	MFCC	-	-	-	-	76	38.5
[Zöhrer et Pernkopf, 2016]	GRNN	Mel	-	-	-	-	**73**	47.6
[Vu et Wang, 2016]	RNN	Mel	-	-	-	-	81.5	**49.8**
	NMF $\beta = 2$	Mel	87	29	60	56	**73.5**	42.5
	NMF $\beta = 1$	Mel	86	30	89	58	**72.5**	44

TABLEAU 4.4 – Taux d'erreur et scores F1 sur les deux environnements de la base TUT SED 2016 pour le système NMF proposé et pour les meilleurs systèmes de l'état de l'art.

	Méthode	ER_{tr}	$F1_{tr}$	ER_{1sec}	$F1_{1sec}$
[Mesaros et al., 2016b]	GMM	0.78	40.5	0.72	45.3
[Cakir et al., 2015b]	DNN	0.68	49.2	1.1	50.2
[Cakir et al., 2017]	CNN	0.56	59.8	0.78	59.9
[Cakir et al., 2017]	RNN	0.6	52.8	0.64	57.1
[Cakir et al., 2017]	CRNN	**0.48**	**66.4**	**0.47**	**68.7**
	NMF $\beta = 2$	0.71	40.8	0.74	46.5
	NMF $\beta = 1$	0.72	40.8	0.74	47.0
	Contexte NMF $\beta = 2$	0.64	49.5	0.68	52.5
	Contexte NMF $\beta = 1$	0.62	51.8	0.67	53.6

TABLEAU 4.5 – Taux d'erreur et scores F1 sur les deux environnements de la base TUT SED synth 2016 pour le système NMF proposé comparés aux systèmes état de l'art.

d'autres tâches. Ces différences sont toutefois très faibles, elles ne nous permettent pas de tirer des conclusions quant à la supériorité d'une des deux divergences pour la tâche.

TUT SED synth

Les résultats sur la base TUT SED synth sont exposés dans le tableau 4.5. Ils sont comparés aux seuls systèmes publiés à ce jour sur cette base de données [Cakir et al., 2017].

Comparé à la base réelle, on peut remarquer que sur cette base, nos systèmes NMF simples obtiennent en général des performances bien moins élevées que les réseaux plus complexes proposés. Même si dans certains cas les ER sont plus faibles, les scores F1 ont tendance à être bien supérieurs pour les méthodes par réseaux de neurones. Par contre, l'augmentation des projections NMF par leur moyenne et leur variance dans le contexte NMF entraîne une forte amélioration des performances pour nos systèmes. Cela nous permet d'obtenir des résultats plus proches des méthodes prenant en compte le contexte temporel dans l'apprentissage telles que les CNN ou RNN. Avec le contexte NMF, nous arrivons à de meilleures performances que le DNN tout en ayant moins de 2 points de différences avec les RNN sur les métriques par trame. En revanche le système état de l'art combinant CNN et RNN améliore fortement les scores par rapport à l'ensemble des autres systèmes, et ce indépendamment des métriques. La domination beaucoup plus claire des modèles profonds plus avancés sur cette base s'explique en grande partie par sa taille. En effet, la base TUT SED Synth possède 4 fois plus d'enregistrements d'apprentissage, une densité

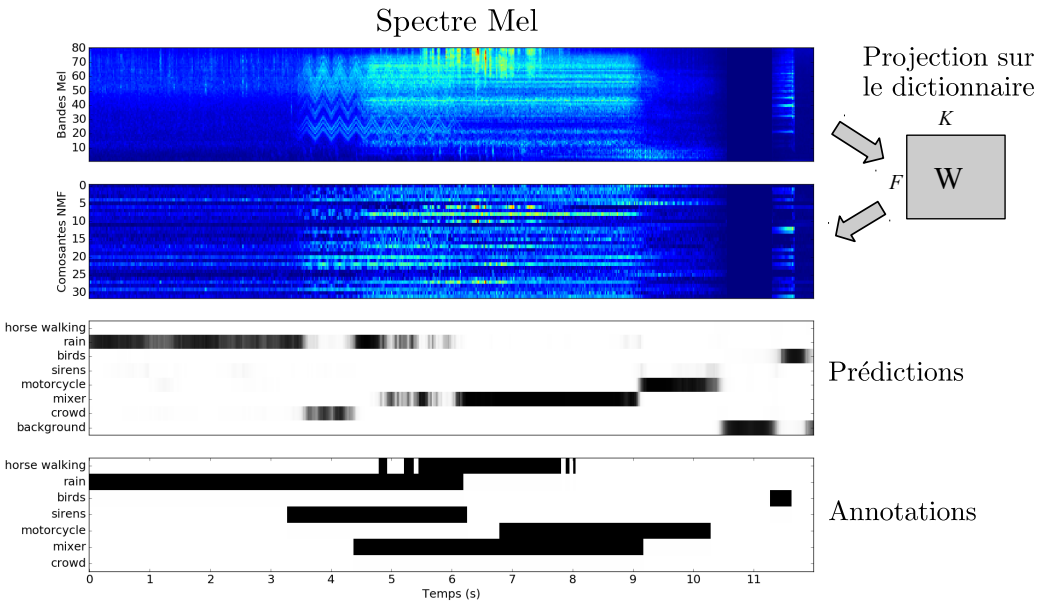


FIGURE 4.10 – Illustrations des différentes étapes pour notre système de détection d'événements par NMF sur la base TUT synth 2016. **De haut en bas** représentation temps-fréquence, projections NMF, séquence d'étiquettes prédite et la vérité terrain.

d'événements plus importante ainsi qu'un meilleur équilibre entre les étiquettes. Ces avantages sont rendu possible par la construction synthétique de la base. De même que pour la base réelle, l'utilisation de la divergence KL pour la NMF amène une légère amélioration des performances, restant cependant toujours en dessous de 2 points de différence sur chaque métrique.

Nous illustrons sur un cas particulier présenté à la figure 4.10, la séquence de prédictions obtenues par nos systèmes de détection d'événements. Nous présentons volontairement un cas contenant des portions avec un fort taux de recouvrement. Cet exemple montre que le modèle NMF proposé arrive rarement à détecter plus d'un événement simultanément. Sur les portions polyphoniques de l'enregistrement, le modèle alterne entre la détection des différents événements se recouvrant. De plus, la classification trame par trame rend les séquences d'étiquettes prédites plus irrégulières. Ces observations confirment la nécessité de modifier la stratégie de classification afin d'améliorer la modélisation de l'évolution temporelle de l'information. Cependant, les systèmes proposés ne sont que la première étape dans la construction de systèmes plus robustes proposés par la suite. Avant d'amorcer le travail sur la recherche de meilleures stratégies de classification au chapitre 6, nous resterons dans un cadre similaire afin d'étudier l'apport d'approches supervisées des modèles SNMF dans la chapitre 5.

4.7 Conclusion

L'utilisation de techniques de factorisation de matrices à partir de spectrogrammes permet de modéliser les scènes sonores comme une combinaison d'éléments de base, représentatifs du contenu fréquentiel des événements constituant la scène. Ces approches, en particulier des variantes de la NMF non-supervisées, peuvent être des outils performants d'apprentissage de repré-

sentations, en se servant des matrices d'activations comme descripteurs. Même avec un cadre très simple, les approches NMF comparées dans ce chapitre atteignent des performances supérieures aux méthodes par ingénierie de descripteurs. Cependant, les systèmes proposés, en particulier pour la détection d'événements sur de plus grosses bases, semblent limités par leur absence de modélisation temporelle. Les approches étudiées permettent de valider une stratégie d'apprentissage de représentations efficace sur laquelle nous nous appuierons dans les chapitres suivants.

Chapitre 5

Apprentissage supervisé de représentations positives

Sommaire

Ce chapitre aborde la question de la factorisation supervisée de matrices pour l'apprentissage de représentations. C'est-à-dire comment, en se servant de l'information sur les étiquettes des données, on peut contraindre ou modifier les décompositions de façon à les adapter à la tâche traitée. Nous proposons une réponse à ce problème en introduisant le modèle TNMF. Le modèle TNMF nous permet d'avoir une extension supervisée des modèles et systèmes NMF introduits dans le chapitre précédent. Les systèmes TNMF sont construits par la modification de certains modèles d'apprentissage de dictionnaires supervisés dans l'objectif de garder les avantages des factorisations positives. En apprenant conjointement le dictionnaire à coefficients positifs et les paramètres du classifieur dans un même problème d'optimisation, les modèles TNMF permettent d'adapter l'apprentissage de la représentation à la tâche traitée. Les travaux de ce chapitre ont également été présentés dans notre article de journal [Bisot et al., 2017b] et ont fait l'objet d'un article de conférence [Bisot et al., 2017a] ainsi que de deux soumissions aux challenges DCASE 2016 et 2017 [Bisot et al., 2016b, 2017c].

5.1 Factorisation supervisée de matrices

5.1.1 Motivations

L'objectif initial des techniques de factorisation de matrices en apprentissage automatique était, la plupart du temps, de réduire la dimension des données ou de faciliter leur interprétation. L'exemple le plus répandu étant le cas de la PCA. La PCA est souvent introduite comme un bloc dans la chaîne de traitement afin de réduire la dimension des observations. Toutefois, elle constitue également un outil facilitant la compréhension et l'interprétation des données, souvent utilisées en statistiques. Bien d'autres approches telles que la NMF, l'ICA (analyse en composantes indépendantes) ou les méthodes de codages parcimonieux sont également très performantes pour des applications de codage, de débruitage ou de séparation de sources. Cette capacité des techniques de factorisation de matrices à faciliter l'interpretabilité des données en font également des outils d'apprentissage de représentations souvent très performants. En effet, pour de multiples applications de classification, ajouter une étape de réduction de dimension dans l'objectif de réduire la complexité des modèles s'accompagne en général d'une amélioration de leurs capacités de généralisation.

L'utilisation de factorisations de matrices pour l'apprentissage de caractéristiques ou pour le *clustering* se fait avant tout dans un cadre non-supervisé. Comme discuté dans le chapitre précédent, les données sont décomposées en résolvant un problème d'optimisation basé sur un critère d'attache aux données, en minimisant une distance ou une divergence entre les données et leur approximation par factorisation de matrices. Nous avons également vu que l'on peut ajouter certaines contraintes ou régularisations au problème afin d'obtenir une décomposition avec des propriétés adaptées à l'application traitée. En revanche, dans beaucoup de cas, en particulier pour des applications de classification, le coût d'attache aux données n'est pas forcément représentatif de la qualité de la représentation apprise. Par exemple, si comme dans notre cas, on utilise des activations (ou projections) sur un dictionnaire comme représentation d'entrée d'un classifieur, le rôle de la décomposition est principalement de faciliter la classification. C'est là qu'interviennent les approches par factorisation supervisée de matrices. Leur objectif est de se servir de l'information sur les étiquettes des données pour contraindre ou modifier les décompositions, de façon à améliorer un critère supervisé cible. Rapprochons nous maintenant de notre cas, c'est-à-dire l'utilisation de la NMF comme méthode d'apprentissage de descripteurs pour la classification de scènes et la détection d'événements. Les approches non-supervisées de la NMF nous ont permis d'apprendre des dictionnaires d'événements de base représentant la signature fréquentielle des événements les plus présents ou les plus représentatifs des scènes dans la base de données. Or, l'objectif final de

nos systèmes est d'être capable de discriminer les différentes catégories de scènes par l'emploi d'un classifieur. Dans ce cas, par l'introduction de modèles NMF supervisés, nous cherchons une décomposition permettant d'apprendre un dictionnaire d'événements élémentaires caractéristiques de certaines catégories, de façon à obtenir des projections facilitant la discrimination des différents étiquettes, des scènes ou des événements.

Nous présentons dans les sections suivantes les principales approches de factorisation de matrices supervisées en portant une attention particulière sur les méthodes appliquées à l'apprentissage de représentations pour la classification. Nous aborderons également les approches plus spécifiques, basées sur des extensions de la NMF, en particulier celles appliquées à l'analyse de sons environnementaux.

5.1.2 Variantes de factorisation supervisée de matrices

Avant de présenter plus en détail certaines variantes de factorisation supervisée de matrices, nous introduisons quelques notations que nous utiliserons tout au long du chapitre. On se donne :

— un ensemble d'apprentissage de N vecteurs de données représenté par la matrice $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_N] \in \mathbb{R}^{F \times N}$,

— un problème de classification à C classes, chaque classe est représentée par son étiquette y dans $\mathcal{Y} = \{1, ..., C\}$,

— un dictionnaire $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_K]$ dans l'ensemble $\mathcal{W} = \{\mathbf{W} \in \mathbb{R}^{F \times K}$ t.q. $\forall j \in \{1, ..., K\}$, $\|\mathbf{w}_j\|_2 = 1\}$,

— des paramètres du classifieur $\mathbf{A}$ pris dans un ensemble $\mathcal{A}$,

— et on note $\mathbf{v}_\Lambda$ le vecteur dans $\mathbb{R}^{|\Lambda|}$ qui contient les entrées de $\mathbf{v}$ indexées par $\Lambda \subseteq \{1, ..., F\}$.

Une des premières et des plus simples approches d'apprentissage de dictionnaires supervisés est d'apprendre un dictionnaire par classe. C'est-à-dire de regrouper les données correspondant à la même classe avant de les décomposer séparément selon un critère non-supervisé [Yang et al., 2010]. L'inconvénient de cette approche est la potentielle redondance dans les composantes entre les dictionnaires correspondant à des classes différentes. Il a été proposé de réduire cette redondance en ajoutant des contraintes d'indépendance des dictionnaires entre différentes classes [Ramirez et al., 2010].

D'autres approches plus complexes apprennent la décomposition et un classifieur linéaire dans un même problème avec l'apprentissage de dictionnaires supervisés (SDL), de l'anglais *supervised dictionary learning*, tel qu'initialement proposé par Mairal et al. [2009b, 2008]. Nous présentons plus en détail le principe des modèles SDL qui sont relativement proches des modèles TDL et TNMF présentés par la suite. Si on se donne ℓ_s une fonction de coût supervisée associée à un classifieur, alors l'approche SDL consiste à résoudre le problème suivant :

$$\min_{\mathbf{A},\mathbf{W},\mathbf{H}} \sum_{i=1}^{N} \ell_s(y_i, \mathbf{h}_i, \mathbf{A}) + \lambda_0 \|\mathbf{v}_i - \mathbf{W}\mathbf{h}_i\|_2^2 + \lambda_1 \|\mathbf{h}_i\|_1 + \frac{\nu}{2}\|\mathbf{A}\|_2^2, \tag{5.1}$$

où $\mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_N]$ est la matrice d'activations, λ_1 le paramètre de régularisation de la parcimonie des activations, et ν le paramètre de régularisation sur la norme ℓ_2 des paramètres du classifieur. Le problème s'exprime comme une somme entre le problème de classification et le coût d'attache aux données, la contribution des deux termes est alors pondérée par un paramètre λ_0. Cette approche a l'avantage de chercher des décompositions qui minimisent à la fois le coût de classification et le terme d'attache aux données. Malgré cet intérêt principal, cette formulation présente quelques

défauts importants. La première limitation provient du nombre important de paramètres du modèle, rendant son apprentissage fastidieux. La deuxième est directement liée à la complexité du problème d'optimisation. En effet le problème donné à l'équation (5.1) n'est pas convexe, ce qui amène souvent, pour ne pas augmenter la difficulté, à se contraindre à l'utilisation de classifieurs linéaires.

La nécessité de connaître l'étiquette de chaque exemple afin d'exprimer le problème d'optimisation dans son entier rend l'usage de SDL complexe en phase de test. La prédiction sur une observation de test implique de chercher $\mathbf{H}$ minimisant le problème équation (5.1) à $\mathbf{W}$ et $\mathbf{A}$ fixé. Cela nécessite, pour inférer l'étiquette d'un exemple de test $\mathbf{v}_i$, de résoudre le problème autant de fois qu'il y a d'étiquettes différentes dans $\mathcal{Y}$. On pourrait imaginer s'affranchir de ce problème en séparant l'inférence en deux étapes durant la phase de test. C'est-à-dire de chercher $\mathbf{H}$ qui minimise uniquement le terme d'attache aux données avant de trouver la valeur de y qui minimise seulement le coût supervisé ℓ_s à $\mathbf{H}$ fixé. Cependant cette solution pose des problèmes d'identifiabilité important entre la représentation latente $\mathbf{H}$ obtenue durant la phase d'apprentissage et la phase test. En effet, les deux matrices d'activations ne sont alors pas obtenues par le même problème d'optimisation, ce qui limite en pratique les capacités de généralisation du modèle.

Parmi les autres approches de factorisation de matrices supervisées on peut trouver des idées similaires à SDL où le coût de classification est remplacé par des contraintes associées au discriminant de Fischer. Des termes de régularisation sont ajoutés sur les activations pour maximiser la similarité intra-classes et la dissimilarité inter-classes [Yang et al.]. On trouve également des variantes discriminatives des décompositions en valeurs singulières [Zhang et Li, 2010; Jiang et al., 2011], des approches réalisant la factorisation dans un espace transformé plus discriminant [Gangeh et al., 2013; Zhang et al., 2013] ou encore des méthodes probabilistes d'apprentissage de dictionnaires supervisés par approches bayésiennes non-paramétriques [Babagholami-Mohamadabadi et al., 2013].

La méthode de factorisation de matrices supervisée sur laquelle nous nous basons dans ce travail est construite à partir du modèle *Task-driven Dictionary Learning* (TDL) [Mairal et al., 2012]. Le modèle TDL a pour objectif d'apprendre un classifieur et un dictionnaire dans un problème d'optimisation à deux niveaux. C'est-à-dire qu'on minimise un coût supervisé vu comme une composition de fonctions entre la projection sur le dictionnaire et le coût de classification. L'intérêt principal du modèle TDL par rapport à SDL est d'être avant tout un modèle d'apprentissage de dictionnaire discriminatif. C'est-à-dire que l'on cherche à apprendre un dictionnaire minimisant uniquement un coût de classification donné qui dépend indirectement du dictionnaire. Le terme d'attache aux données intervient uniquement par l'étape de projection. Le gain en stabilité et en qualité des représentations qu'offre le modèle TDL par rapport à SDL a également été montré expérimentalement [Mairal et al., 2012]. Le modèle TDL est décrit plus en détail dans la section 5.2.

5.1.3 Les variantes supervisées de la NMF

Motivés par le succès de la NMF pour de nombreuses applications, de nombreux travaux cherchent à adapter les techniques de factorisation supervisée de matrices à la NMF. Par exemple, les idées d'apprendre des dictionnaires par classe [Serizel et al., 2016b; Benetos et al., 2012] ou de contraindre les activations par des discriminants de Fisher [Wang et al., 2004; Zafeiriou et al., 2006] se font aussi pour la NMF. On trouve également des approches similaires à SDL en proposant d'apprendre une NMF avec un critère de maximisation de la marge entre les projections correspondant à une même classe [Kumar et al., 2012].

Pour l'analyse de scènes sonores, la majorité des approches par NMF contiennent une forme de supervision. En particulier, les méthodes de détection d'événements par seuillage des activa-

tions NMF impliquent un apprentissage préalable de composantes élémentaires associées à chaque classe [Benetos et al., 2016, 2017; Gemmeke et al., 2013]. Il a également été proposé d'augmenter les données avec une matrice binaire d'étiquettes avant de les décomposer par NMF pour la détection d'événements [Mesaros et al., 2015]. Ce processus associe chaque composante à une ou plusieurs étiquettes, permettant de réaliser l'étape de prédiction des étiquettes simplement en post-traitant la matrice d'activations. Une approche similaire a été utilisée pour la classification de scènes avec l'idée de s'en servir comme méthode d'apprentissage de représentations [Rakotomamonjy, 2017]. Une fois la décomposition NMF augmentée apprise, les activations sont utilisées comme descripteurs pour l'apprentissage d'un classifieur.

Bien que nous nous focalisions sur les méthodes appliquées à des tâches de classification, le critère de classification dans SDL ou TDL peut être modifié pour appliquer ces modèles dans le contexte de débruitage ou de séparation de sources. Dans ce cadre, Sprechmann et al. [2014] proposent une variante positive de TDL modifiée pour le réhaussement de la parole. Au lieu de se servir d'un critère de classification le modèle utilise un critère représentant la séparation des composantes de parole et des composantes de bruit. Avec le même objectif Le Roux et al. [2015a] ont proposé un modèle de NMF *profond* en utilisant la théorie de *dépliement* d'algorithmes itératifs [Hershey et al., 2014]. Le modèle *deep NMF* est construit en dépliant l'algorithme de mise à jour multiplicatif de la NMF résultant en la construction d'un modèle profond nécessitant l'apprentissage de plusieurs dictionnaires.

5.2 Le modèle TNMF

Dans cette section, nous présentons en détail le modèle TNMF, une variante supervisée de la NMF, que nous proposons pour l'analyse de sons environnementaux. Le modèle TNMF peut être vu comme une extension positive des approches TDL. Nous commençons par détailler le principe de ces modèles. Ensuite nous décrivons les modifications apportées au modèle TDL afin de l'adapter aux problèmes d'analyse de sons environnementaux. En particulier, nous proposons un nouvel algorithme qui présente de meilleurs capacités de généralisation.

5.2.1 TDL : apprentissage de dictionnaires adaptés à la tâche

L'objectif principal du modèle TDL, comme bien d'autres approches par factorisation supervisée de matrices, est d'apprendre un dictionnaire permettant d'extraire des représentations adaptées à la tâche traitée. A la différence de la plupart des autres approches, les étapes d'apprentissage de dictionnaires et de classification sont regroupées dans un problème d'optimisation à deux niveaux. Ce problème s'exprime comme l'apprentissage conjoint d'un dictionnaire et de paramètres du classifieur par la minimisation du coût supervisé associé au classifieur.

Formulation originale du problème La première étape dans la présentation du modèle TDL est de définir la fonction de projection optimale $\mathbf{h}^\star$ permettant d'obtenir la représentation latente utilisée en entrée du classifieur. Cette fonction se définit comme la solution d'un problème d'optimisation, le problème *elastic-net*. Ainsi $\mathbf{h}^\star(\mathbf{v}, \mathbf{W})$ correspond à la fonction $\mathbf{h}^\star$ évaluée en $\mathbf{v}$ et $\mathbf{W}$. Elle représente la projection de l'exemple $\mathbf{v}$ sur le dictionnaire $\mathbf{W}$ et se définit par :

$$\mathbf{h}^\star(\mathbf{v}, \mathbf{W}) = \min_{\mathbf{h} \in \mathbb{R}^K} \frac{1}{2}\|\mathbf{v} - \mathbf{W}\mathbf{h}\|_2^2 + \lambda_1 \|\mathbf{h}\|_1 + \frac{\lambda_2}{2}\|\mathbf{h}\|_2^2. \tag{5.2}$$

Nous considérons maintenant que chaque vecteur de données $\mathbf{v}$ est associé à une étiquette y dans un ensemble fixe d'étiquettes $\mathcal{Y}$. On définit alors un coût de classification $l_s(y, \mathbf{A}, \mathbf{h}^\star(\mathbf{v}, \mathbf{W}))$, où

$\mathbf{A}$ contient les paramètres du classifieur. Le problème est alors exprimé comme la minimisation conjointe de $\mathbf{W}$ et $\mathbf{A}$ de l'espérance du coût de classification :

$$\min_{\mathbf{W}\in\mathcal{W},\mathbf{A}} f(\mathbf{W},\mathbf{A}) + \frac{\nu}{2}\|\mathbf{A}\|_2^2, \tag{5.3}$$

avec

$$f(\mathbf{W},\mathbf{A}) = \mathbb{E}_{y,\mathbf{V}}[l_s(y,\mathbf{A},\mathbf{h}^\star(\mathbf{v},\mathbf{W}))]. \tag{5.4}$$

Ici, ν est un paramètre de régularisation positif appliqué aux poids des classifieurs pour éviter le sur-apprentissage. Le modèle TDL s'exprime alors comme un problème d'optimisation à deux niveaux de la manière suivante :

$$\begin{cases} \mathbf{h}^\star(\mathbf{v},\mathbf{W}) = \min_{\mathbf{h}\in\mathbb{R}^K} \frac{1}{2}\|\mathbf{v}-\mathbf{W}\mathbf{h}\|_2^2 + \lambda_1\|\mathbf{h}\|_1 + \frac{\lambda_2}{2}\|\mathbf{h}\|_2^2 \\ \min_{\mathbf{W}\in\mathcal{W},\mathbf{A}} \mathbb{E}_{y,\mathbf{v}}[\ell_s(y,\mathbf{A},\mathbf{h}^\star(\mathbf{v},\mathbf{W}))] + \frac{\nu}{2}\|\mathbf{A}\|_2^2. \end{cases} \tag{5.5}$$

Ainsi le dictionnaire est appris en minimisant un critère supervisé défini comme la composition de deux fonctions. La première est la fonction de projection optimale des données sur le dictionnaire dont la solution est donnée en entrée de la deuxième fonction, qui correspond au critère supervisé associé au classifieur. Si on considère par exemple le cas de la régression logistique à deux classes avec $\mathcal{Y} = \{-1,+1\}$, le coût de classification se définit par $\ell_s(y,\mathbf{A},\mathbf{h}^\star(\mathbf{v},\mathbf{W})) = \log(1+e^{-y\mathbf{a}^T\mathbf{h}^\star(\mathbf{v},\mathbf{W})})$ et le problème TDL s'écrit alors :

$$\min_{\mathbf{W}\in\mathcal{W},\mathbf{a}\in\mathbb{R}^K} \mathbb{E}_{y,\mathbf{v}}[\log(1+e^{-y\mathbf{a}^T\mathbf{h}^\star(\mathbf{v},\mathbf{W})})] + \frac{\nu}{2}\|\mathbf{a}\|_2^2 \tag{5.6}$$

Par souci de compacité, nous noterons $\mathbf{h}^\star = \mathbf{h}^\star(\mathbf{v},\mathbf{W})$ dans ce qui suit.

Optimisation du problème : La principale difficulté est que $\mathbf{h}^\star$, solution du problème (5.2), n'est pas dérivable par rapport à $\mathbf{W}$. Les auteurs dans [Mairal et al., 2012] se servent du fait que $\mathbf{h}^\star$ est continue, Lipschitzienne et dérivable presque partout (sauf en les points où elle change de signe) pour donner et démontrer la proposition suivante :

Proposition 5.2.1 *Si l_s est deux fois continue dérivable et selon certaines conditions acceptables sur $\mathcal{Y}$ et l'ensemble des données [Mairal et al., 2012]. La fonction f est dérivable et*

$$\begin{cases} \nabla_{\boldsymbol{A}} f(\boldsymbol{A},\boldsymbol{W}) = \mathbb{E}_{y,\boldsymbol{v}}[\nabla l_s(y,\boldsymbol{W},\boldsymbol{h}^\star)], \\ \nabla_{\boldsymbol{W}} f(\boldsymbol{A},\boldsymbol{W}) = \mathbb{E}_{y,\boldsymbol{v}}[-\boldsymbol{W}\boldsymbol{\beta}^\star\boldsymbol{h}^{\star T} + (\boldsymbol{v}-\boldsymbol{W}\boldsymbol{h}^\star)\boldsymbol{\beta}^{\star T}], \end{cases} \tag{5.7}$$

avec

$$\boldsymbol{\beta}^\star_{\Lambda^c} = 0 \text{ et } \boldsymbol{\beta}^\star_\Lambda = (\boldsymbol{W}_\Lambda^T\boldsymbol{W}_\Lambda + \lambda_2\boldsymbol{I})^{-1}\nabla_{\boldsymbol{h}_\Lambda} l_s(y,\boldsymbol{A},\boldsymbol{h}^\star) \tag{5.8}$$

où Λ est l'ensemble des indices des coefficients non-nuls de $\boldsymbol{h}^\star$.

Cette propriété nous donne la démarche pour obtenir les gradients du coût par rapport au dictionnaire et aux paramètres du classifieur. Ces gradients sont obtenus avec un procédé similaire à la rétro-propagation des gradients pour les réseaux de neurones, c'est-à-dire en s'appuyant sur la règle de dérivation des fonctions composées. Il est proposé par Mairal et al. [2012] de résoudre le problème par descente de gradient projeté stochastique que nous détaillons dans l'algorithme 5.1. Pour cela, on commence par tirer aléatoirement un couple (**v**,y) dans l'ensemble d'apprentissage à chaque itération. Les gradients du coût de classification équation (5.3) sont ensuite calculés en reprenant la proposition 5.2.1. Les étapes de mises à jour du dictionnaire et des poids du classifieur sont effectuées par gradient projeté. C'est-à-dire qu'une fois avoir retiré aux paramètres une quantité proportionnelle à leur gradient, on les projette sur les ensembles sur lesquels ils sont définis.

En pratique ces ensembles sont relativement simples rendant l'étape de projection triviale. En particulier, nous n'avons pas d'étape de projection pour les paramètres du classifieur. Le dictionnaire lui, est défini dans $\mathcal{W}$ et correspond à l'ensemble des matrices aux colonnes de norme ℓ_2 unitaire. Dans ce cas l'étape de projection sur $\mathcal{W}$ revient simplement à normaliser chaque colonne de $\mathbf{W}$ après avoir retiré le gradient correspondant. L'algorithme *least angle regression* (LARS) est utilisé pour résoudre l'équation (5.2) en utilisant la fonction *lasso* de la librairie *spams* [Mairal et al., 2010]. Enfin, l'opération $\Pi_{\mathcal{W}}$ correspond à la projection orthogonale sur l'ensemble $\mathcal{W}$.

Dans sa formulation initiale, une itération dans l'algorithme 5.1 correspond à la mise à jour des paramètres pour le tirage aléatoire d'un couple (**v**,y). Dans ce travail, nous utiliserons également la notion d'époque. Une époque correspond à un passage de l'algorithme sur tous les points d'une permutation aléatoire d'un ensemble d'apprentissage de taille N. Les approches par descente de gradient nécessitent le choix d'un pas du gradient ρ. Il existe de multiples approches pour choisir et adapter la valeur du pas au cours des itérations de manière à accélérer ou garantir la décroissance du critère. Une des plus répandues est connue sous le nom *backtracking line search* en anglais. Ces approches nécessitent souvent l'ajout de multiples opérations intermédiaires. Afin de simplifier et d'accélérer la procédure de choix du pas du gradient, nous suivons l'heuristique proposée par Mairal et al. [2012] pour gérer l'évolution de la valeur de ρ. Il s'agit de sélectionner une valeur initiale du pas ρ et de la faire décroître au fil des TN mises à jour où N représente le nombre d'exemples et T le nombre d'époques. La décroissance du pas est donnée par la règle suivante : $\rho_{tn} = \min(\rho, \rho t_0/((t-1)N+n))$ où t est l'indice d'époque dans $\{1, ..., T\}$, n indique le nombre d'exemples parcourus dans l'époque courante et t_0 est fixé à $TN/10$.

Algorithme 5.1 : Descente de gradient stochastique pour l'apprentissage de dictionnaire supervisé.

prend en entrée
$\lambda_1, \lambda_2, \nu, \mathbf{W} \in \mathcal{W}, \mathbf{A} \in \mathcal{A}, T$(nombre d'époques), N(nombre d'exemples), t_o, ρ
pour $t = 1$ à T **faire**
pour $n = 1$ à N **faire**
Tirage d'un couple $(\mathbf{v}, y)$ dans l'ensemble d'apprentissage
Calculer $\mathbf{h}^\star$ solution de (5.2)
Déterminer Λ (coefficients non-nuls de $\mathbf{h}^\star$)
$\boldsymbol{\beta}^\star_{\Lambda^c} = 0$ et $\boldsymbol{\beta}^\star_{\Lambda} = (\mathbf{W}_\Lambda^T \mathbf{W}_\Lambda + \lambda_2 \mathbf{I})^{-1} \nabla_{\mathbf{h}_\Lambda} l_s(y, \mathbf{A}, \mathbf{h}^\star)$
$\rho_{tn} = \min(\rho, \rho \frac{t_0}{(t-1)N+n})$
$\mathbf{A} \leftarrow \Pi_{\mathcal{A}}[\mathbf{A} - \rho_{tn}(\nabla_{\mathbf{A}} l_s(y_t, \mathbf{A}, \mathbf{h}^\star) + \nu \mathbf{A})]$
$\mathbf{W} \leftarrow \Pi_{\mathcal{W}}[\mathbf{W} - \rho_{tn}(-\mathbf{W}\boldsymbol{\beta}^\star \mathbf{h}^{\star T} + (\mathbf{v} - \mathbf{W}\mathbf{h}^\star)\boldsymbol{\beta}^{\star T})]$
fin pour
fin pour

5.2.2 TNMF

Nous présentons dans cette section le modèle TNMF, la variante supervisée de la NMF que nous introduisons pour la classification de sons environnementaux. Nous discutons des modifications apportées par rapport au modèle TDL dans l'objectif d'appliquer TNMF à la classification de scènes. Son application à la détection d'événements n'est qu'un cas particulier du modèle et des algorithmes que nous introduisons. Nous illustrons sur la figure 5.1 l'architecture générale des modèles TDL et TNMF.

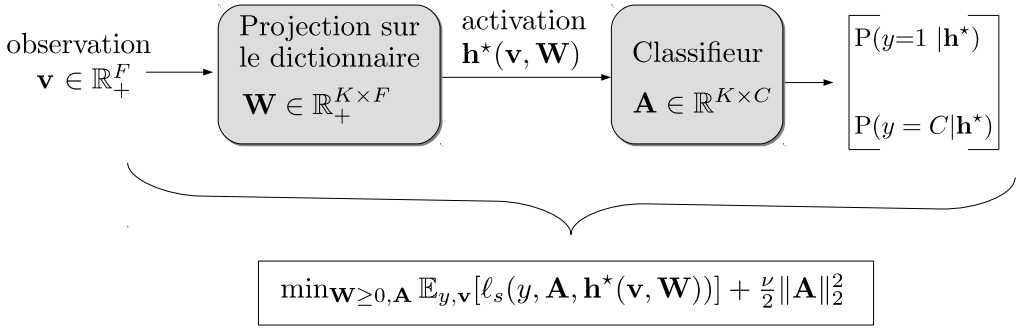


FIGURE 5.1 – Illustration du modèle TNMF dans le cas de la régression logistique multinomiale.

Cas de le régression logistique multinomiale

Dans sa formulation originale, le modèle TDL est présenté dans le cas de la régression logistique bi-classes en utilisant des schémas *un-contre-tous* pour traiter les problèmes de classification multi-classes. La principale conséquence du choix de l'approche *un-contre-tous* est d'avoir à apprendre le modèle autant de fois que d'étiquettes dans la base de données, résultant en l'apprentissage d'un dictionnaire différent par classe. Si ce choix peut avoir des avantages en termes d'interprétabilité des dictionnaires appris, cela ne se traduit pas par de meilleures performances. En effet, à la fois pour la classification de scènes et la détection d'événements, nous avons trouvé expérimentalement que la formulation multi-classes de la régression logistique offrait de meilleures performances que les schémas *un-contre-tous*. De plus, notre motivation dans la construction du modèle TNMF est de proposer une version supervisée de la NMF parcimonieuse (SNMF), ainsi par souci de cohérence nous utiliserons une régression logistique multinationale comme classifieur dans TNMF. Ce choix a également l'avantage d'apprendre un dictionnaire discriminatif représentant l'ensemble des données, dans le même esprit des techniques non-supervisées présentées au chapitre 4.

On se donne un vecteur de données $\mathbf{v}$ de la base d'apprentissage associé à l'étiquette c ainsi que sa projection optimale sur le dictionnaire $\mathbf{h}^\star = \mathbf{h}^\star(\mathbf{v}, \mathbf{W})$. Dans le cas multi-classes, la régression logistique estime la probabilité $\mathrm{P}(y = c|\mathbf{h}^\star)$ pour $c \in \{1, ..., C\}$ qui s'écrit en intégrant la projection optimale sur le dictionnaire :

$$\mathrm{P}(y = c|\mathbf{h}^\star) = \frac{e^{(b_c + \mathbf{a}_c^T \mathbf{h}^\star)}}{\sum_{j=1}^{C} e^{(b_j + \mathbf{a}_j^T \mathbf{h}^\star)}}. \tag{5.9}$$

Les $\mathbf{a}_c \in \mathbb{R}^K$ pour $c \in 1, ..., C$ et les biais b_c sont les poids du classifieur contenus dans $\mathbf{A}$. La fonction de coût supervisée l_s pour un couple $(\mathbf{v}, y)$ utilisée dans le modèle TDL s'exprime alors $l_s(y_c, \mathbf{A}, \mathbf{h}^\star) = -\log(\mathrm{P}(y = c|\mathbf{h}^\star))$. Ainsi, la minimisation de la fonction $f(\mathbf{A}, \mathbf{W})$ du problème TDL équivaut à estimer les paramètres du modèle par maximum de vraisemblance. Comme nous avons un ensemble fixe de données annotées, la fonction de coût du modèle TNMF s'exprime ainsi :

$$\min_{\mathbf{W} \in \mathcal{W}, \mathbf{A}} f(\mathbf{A}, \mathbf{W}) + \frac{\nu}{2}\|\mathbf{A}\|_2^2 = \min_{\mathbf{W} \in \mathcal{W}, \mathbf{A}} \sum_{i=1}^{N} -\log(\mathrm{P}(y_i|\mathbf{h}_i^\star)) + \frac{\nu}{2}\|\mathbf{A}\|_2^2. \tag{5.10}$$

Le changement du classifieur pour la régression logistique multinomiale implique essentiellement un changement dans l'expression des gradients $\nabla_{\mathbf{h}_\Lambda} l_s(y, \mathbf{A}, \mathbf{h}^\star)$ et $\nabla_{\mathbf{A}} l_s(y, \mathbf{A}, \mathbf{h}^\star)$ nécessaires pour

la résolution du problème (cf. proposition 5.2.1). Ici, l_s s'exprimant comme somme et produit de fonctions usuelles, le calcul des gradients ne pose pas de problèmes particuliers.

Version positive de TDL

Nous avons montré les avantages des variantes NMF pour les applications que nous traitons dans le chapitre précédent. Ainsi, nous souhaitons avoir la possibilité d'utiliser le modèle TDL dans sa formulation positive, c'est-à-dire de contraindre les coefficients du dictionnaire et des projections optimales à être positifs. Si une telle formulation a été mentionnée comme possible dans les travaux de Mairal et al. [2012], cela a été confirmé dans un second temps par Sprechmann et al. [2014]. Ces derniers proposent une première variante positive de TDL pour le rehaussement de la parole, changeant ainsi le critère de classification par un critère représentant la séparation de la parole et du bruit. En pratique, les changements à apporter au modèle TDL pour avoir son équivalent positif sont relativement indépendants du coût supervisé. Dans notre cas, l'espace $\mathcal{W}_+$ des dictionnaires est maintenant défini par $\mathcal{W}_+ = \{\mathbf{W} \in \mathbb{R}_+^{F\times K}$ t.q. $\forall j \in \{1, ..., K\}, \|\mathbf{w}_j\|_2 = 1\}$, l'espaces des matrices à coefficients positifs ayant ses colonnes de normes unitaires. Ainsi, lors de l'étape de mise à jour de $\mathbf{W}$ par gradient projeté, l'étape de projection revient à seuiller les coefficients de $\mathbf{W}$ à zéro avant de normaliser ses colonnes en norme ℓ_2.

Avec l'arrivée d'une formulation positive du modèle TDL se pose la question du choix de la divergence pour le coût d'attache aux données. Il a été montré par Sprechmann et al. [2014] que le modèle TDL peut également s'appliquer en utilisant la β-divergence, tant que β reste dans l'intervalle $[1, 2]$ afin d'assurer la convexité de l'étape de projection. Dans ce cas, la fonction de projection optimale $\mathbf{h}^\star$ devient :

$$\mathbf{h}^\star(\mathbf{v}, \mathbf{W}) = \min_{\mathbf{h}\in\mathbb{R}_+^K} D_\beta(\mathbf{v}\|\mathbf{W}\mathbf{h}) + \lambda_1\|\mathbf{h}\|_1 + \frac{\lambda_2}{2}\|\mathbf{h}\|_2^2. \tag{5.11}$$

Les modifications que cela induit sur l'expression des gradients sont rapidement présentées dans les travaux de Sprechmann et al. [2014]. Nous nous servirons de ce résultat pour appliquer une version de TNMF utilisant la divergence de KL généralisée pour la détection d'événements.

Intégration temporelle

Nous rappelons que le système de classification de scènes par NMF introduit dans le chapitre précédent contient une étape d'intégration temporelle par moyenne des activations avant la classification. Cela vient de la construction par tranches moyennées (TM) de la matrice de données, où chaque exemple de la base est divisé dans le temps en plusieurs vecteurs de données. Nous proposons d'introduire la même étape entre la projection sur le dictionnaire et la classification dans le modèle TNMF. Dans le modèle tel que présenté jusqu'ici, chaque projection $\mathbf{h}^\star(\mathbf{v}, \mathbf{W})$ est classifiée individuellement. A la place, nous souhaitons classifier la moyenne des projections correspondant à l'ensemble des vecteurs de données $\mathbf{v}_m^{(n)}$ extraits de l'exemple audio $n \in [1, N]$ avec $\mathbf{V}^{(n)} = [\mathbf{v}_1^{(n)}, ..., \mathbf{v}_M^{(n)}]$. Nous noterons $\mathbf{h}^{(n)} = \frac{1}{M}\sum_{i=1}^M \mathbf{h}^\star(\mathbf{v}_i^{(n)}, \mathbf{W})$ la moyenne des projections des vecteurs de l'exemple d'indice n. Le coût de classification du modèle TNMF s'exprime alors de la manière suivante :

$$f(\mathbf{W}, \mathbf{A}) = \sum_{n=1}^{N} l_s(y, \mathbf{A}, \mathbf{h}^{(n)}). \tag{5.12}$$

Adaptation de l'algorithme

Le modèle TNMF se distingue du modèle TDL par les modifications que nous venons de décrire mais également par la proposition d'un nouvel algorithme. Le modèle TDL à été introduit

avec un algorithme par descente de gradient stochastique pour l'estimation de ses paramètres. Cet algorithme est similaire aux algorithmes stochastiques basés sur la rétro-propagation des gradients pour les réseaux de neurones. En revanche, de tels algorithmes stochastiques sont plus rarement utilisés pour les méthodes d'apprentissage de dictionnaires ou pour l'apprentissage de régressions logistiques, à moins que le nombre de données les rende incontournables. A la place nous proposons un algorithme de mise à jour alternée du dictionnaire et des paramètres du classifieur. Les paramètres du classifieur sont mis à jour à partir de la base de données entière, en utilisant une itération de l'algorithme L-BFGS pour la régression logistique multionomiale. Nous avons observé en pratique qu'effectuer plus d'une itération sur le classifieur diminue le rôle de la mise à jour du dictionnaire ce qui se traduit par un sur-apprentissage plus important.

Ensuite, une fois $\mathbf{A}$ fixé, le dictionnaire est mis à jour par descente de gradient stochastique sur une époque en utilisant le gradient tel que donné par la proposition 5.2.1. En apprenant le classifieur séparément sur une époque, cette approche a l'avantage de proposer une meilleure estimation des paramètres du classifieur avant de mettre à jour le dictionnaire. De plus l'utilisation d'approches du second ordre telles que L-BFGS pour mettre à jour $\mathbf{A}$ fait que le choix du pas du gradient impacte uniquement la mise à jour du dictionnaire. Cela aide à stabiliser l'apprentissage et facilite le réglage des paramètres. Les changements constituant le nouvel algorithme définissant le modèle TNMF sont présentés dans l'algorithme 5.2. Nous validerons cet algorithme expérimentalement dans la section 5.3 car rien ne garantit sa convergence théorique. Toutefois, nous avons observé la décroissance du coût ainsi qu'une augmentation de la capacité de généralisation pour l'essentiel des cas testés pour la classification de scènes et la détection d'événements. La comparaison des courbes du coût et du taux de reconnaissance au cours des itérations pour les algorithmes est donnée à la section 5.3. Il est important de noter que l'algorithme proposé s'étend facilement aux approches par descente de gradient stochastique par mini-batchs. Ayant pour effet d'accélérer la mise à jour du dictionnaire sans forcement ralentir la convergence, nous utiliserons la version par mini-batchs dans certains de nos systèmes. Le code source pour l'algorithme 5.2 est disponible au lien suivant : github.com/rserizel/TGNMF/blob/master/tnmf.py.

Algorithme 5.2 : Descente de gradient stochastique alternée par époque pour l'apprentissage de dictionnaire supervisé.

prend en entrée $\lambda_1, \lambda_2, \nu, \mathbf{W} \in \mathcal{W}, \mathbf{A} \in \mathcal{A}, T$(nombre d'époques)$, t_o, \rho$
  pour $t = 1$ à T **faire**
    $\forall n \in [\![1, N]\!]$ calculer $\hat{\mathbf{h}}^{(n)} = \frac{1}{M} \sum_{m=1}^{M} \mathbf{h}^\star(\mathbf{v}_m^{(n)}, \mathbf{W})$
    $\hat{\mathbf{H}}^\star(\mathbf{V}, \mathbf{W}) = [\hat{\mathbf{h}}^{(1)}, ..., \hat{\mathbf{h}}^{(N)}]$
    Mise à jour de $\mathbf{A}$ par une itération de L-BFGS
    pour $j = 1$ à mN **faire**
      $\rho_{tj} = \min(\rho, \rho \frac{t_0}{(t-1)mN+j})$
      Tirage d'un $\mathbf{v}$ d'étiquette y dans $\mathbf{V}$
      Calculer $\mathbf{h}^\star = \mathbf{h}^\star(\mathbf{v}, \mathbf{W})$
      Calculer Λ (coefficients non-nuls de $\mathbf{h}^\star$)
      $\boldsymbol{\beta}^\star_{\Lambda^c} = 0$ et $\boldsymbol{\beta}^\star_\Lambda = (\mathbf{W}_\Lambda^T \mathbf{W}_\Lambda + \lambda_2 \mathbf{I})^{-1} \nabla_{\mathbf{h}_\Lambda} l_s(y, \mathbf{A}, \mathbf{h}^\star)$
      $\mathbf{W} \leftarrow \Pi_{\mathcal{W}_+}[\mathbf{W} - \rho_{tj}(-\mathbf{W}\boldsymbol{\beta}^\star \mathbf{h}^{\star T} + (\mathbf{v} - \mathbf{W}\mathbf{h}^\star)\boldsymbol{\beta}^{\star T})]$
    fin pour
  fin pour

Autre variante proposée : TNMF avec prise en charge de contraintes de groupe

Bien que nous présentions TNMF dans le cadre de l'analyse de sons environnementaux, nous avons également, à l'occasion d'une collaboration, appliqué le modèle à la reconnaissance de locuteurs [Serizel et al., 2017]. Le modèle proposé combine TNMF avec la NMF par groupes, une autre approche de NMF supervisée proposée pour la reconnaissance de la parole [Serizel et al., 2016b]. Le modèle TNMF est modifié pour y ajouter les termes de régularisation sur la distance intra-classes ou intra-sessions des composantes du dictionnaire qu'implique la NMF par groupes.

De manière générale le modèle TNMF permet aisément d'ajouter des régularisations dépendant uniquement du dictionnaire. Si on se donne une fonction $g(\mathbf{W})$ dépendant uniquement du dictionnaire alors le modèle TNMF s'exprime ainsi :

$$\min_{\mathbf{W}\in\mathcal{W},\mathbf{A}} f(\mathbf{W},\mathbf{A}) + \frac{\nu}{2}\|\mathbf{A}\|_2^2 + \lambda_g g(\mathbf{W}), \tag{5.13}$$

où λ_g est le paramètre contrôlant la régularisation qui dépend du dictionnaire. Dans ce cas, il suffit d'ajouter au gradient par rapport à $\mathbf{W}$ le terme correspondant au gradient de la régularisation pour que l'algorithme TNMF puisse être utilisé de manière identique. Comme nous l'avons proposé pour la reconnaissance de locuteur, cela permet aisément d'incorporer de la connaissance dans le modèle tout en apprenant une représentation adaptée au critère et à la tâche traitée. Par exemple, le modèle peut aisément s'étendre à des cas semi-supervisés, où la fonction $g(\mathbf{W})$ serait un coût d'attache aux données permettant de prendre en compte les données sans étiquettes dans l'apprentissage du dictionnaire.

5.3 Étude expérimentale des algorithmes TNMF

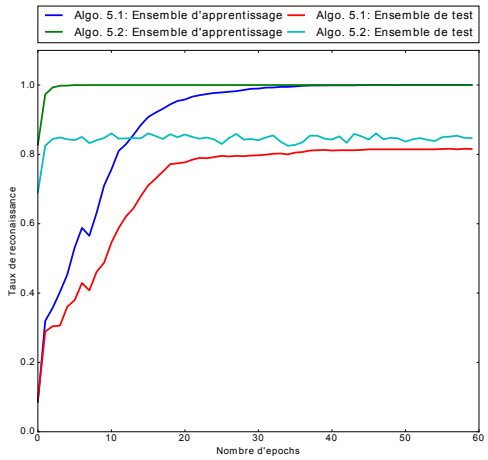
Nous étudions dans cette section, les différences entre les deux algorithmes : l'algorithme original de TDL et la modification proposée pour TNMF. Nous les appliquons au problème de classification de scènes en utilisant la base du DCASE 2017. Nous reprenons exactement les mêmes représentations temps-fréquence et procédures de construction de la matrice de données par TM introduites au chapitre 4. Nous réglons les paramètres du modèle pour les deux algorithmes séparément en cherchant ceux qui maximisent le taux de reconnaissance sur un ensemble de développement correspondant à une sous-partie de l'ensemble d'apprentissage.

Les meilleurs paramètres trouvés sont ($\rho = 0.001$, $\nu = 0.0001$, $\lambda_1 = 0.1$, $\lambda_2 = 0$) pour l'algorithme 5.1 et ($\rho = 0.001$, $\nu = 1$, $\lambda_1 = 0.1$, $\lambda_2 = 0$) pour l'algorithme 5.2. Pour les deux algorithmes l'étape de calcul de la projection optimale se fait en utilisant l'algorithme LARS, pour l'*elastic-net* en utilisant la librairie python *spams* [1] [Mairal et al., 2009a]. Pour l'algorithme 5.2 le classifieur est mis à jour avec l'algorithme L-BFGS pour la régression logistique de la librairie python *scikit-learn* [Pedregosa et al., 2011].

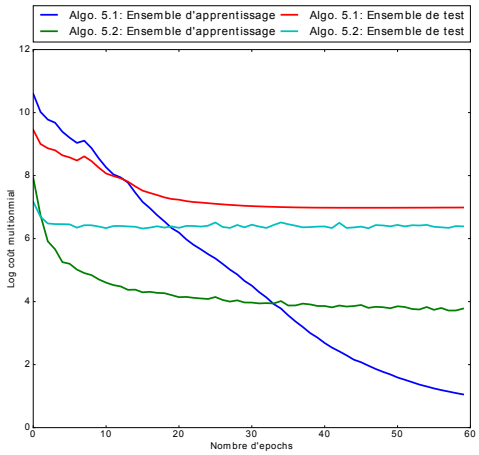
Dans une première expérience nous étudions les évolutions du coût de classification multinomiale de la régression logistique et du taux de reconnaissance au cours des itérations. Pour cela nous initialisons les deux algorithmes avec le même dictionnaire et nous apprenons les modèles sur 60 époques. Nous comparons l'évolution du coût multinomial et du taux de reconnaissance pour les ensembles d'apprentissage et de test afin d'étudier à la fois la capacité d'apprendre et de généraliser de chaque algorithme. Les courbes sont présentées sur la figure 5.2 pour $K = 256$ composantes et sur la figure pour $K = 512$ composantes. Les courbes proviennent des modèles appris sur le premier ensemble apprentissage-test de la base du DCASE 2017.

Dans les deux cas présentés sur la figure 5.2, l'algorithme 5.2 se stabilise en seulement 4 ou 5 époques alors que l'algorithme 5.1 en demande une vingtaine. De plus, on peut voir à la fois

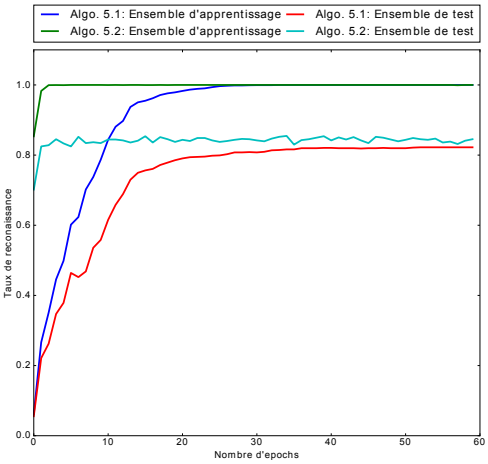
1. http ://spams-devel.gforge.inria.fr/



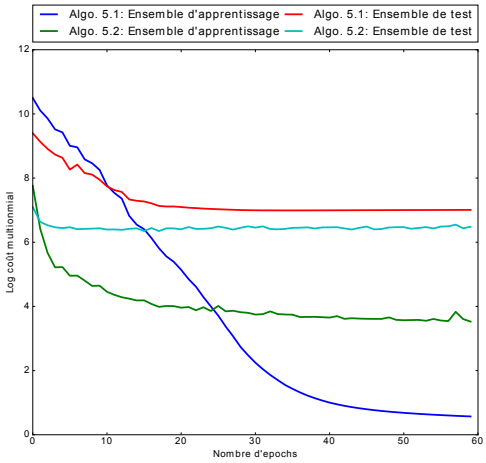
(a) Évolution du taux de reconnaissance pour les deux algorithmes avec $K = 256$ composantes.



(b) Évolution du coût multinomial pour les deux algorithmes avec $K = 256$ composantes.



(c) Évolution du taux de reconnaissance pour les deux algorithmes avec $K = 512$ composantes.



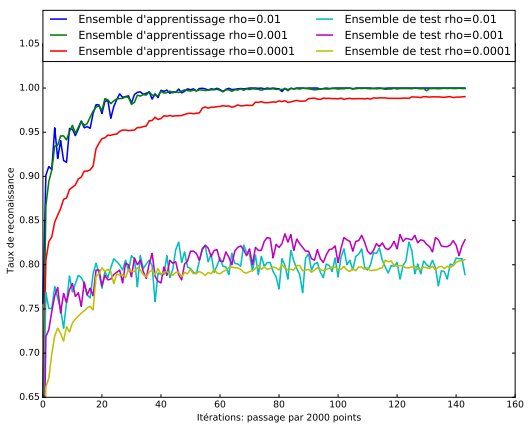
(d) Évolution du coût multinomial pour les deux algorithmes avec $K = 512$ composantes.

FIGURE 5.2 – Comparaison de l'évolution du taux de reconnaissance et du coût multinomial pour les deux algorithmes et pour les ensembles de test et d'apprentissage sur le premier ensemble de la base DCASE 2017.

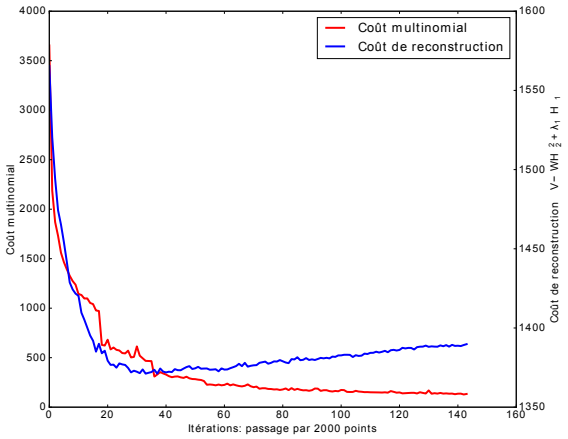
sur les courbes de taux de reconnaissance et de coût de classification, que les performances sur l'ensemble de test sont constamment supérieures pour l'algorithme 5.2, même lorsque l'algorithme semble s'être stabilisé.

Les figures 5.2 (b) et (d) nous montrent en revanche que le coût multinomial décroît vers des valeurs plus basses pour l'algorithme 5.1. Cela est uniquement dû à la contrainte de régularisation des paramètres du classifieur, celle-ci étant plus faible pour l'algorithme 5.1, ce qui diminue forcément la valeur du coût à configuration égale. Par ailleurs, nous avons remarqué qu'à régularisation égale, le coût de classification décroît toujours plus rapidement pour l'algorithme 5.2. Cela est principalement dû à la mise à jour des paramètres du classifieur par une étape de l'algorithme quasi-newton pour la régression logistique, nécessitant en général moins de passages par l'ensemble des données que les algorithmes stochastiques.

Nous proposons également d'observer l'influence du pas du gradient sur l'apprentissage de l'algorithme 5.2 ainsi que le compromis fait par l'algorithme TNMF entre l'erreur de reconstruction et l'erreur de classification durant l'apprentissage du dictionnaire. Nous reprenons exactement



(a) Évolution du taux de reconnaissance en fonction du pas initial du gradient.



(b) Évolution du coût multinomial et du coût de reconstruction au cours des mises à jours.

FIGURE 5.3 – **(a)** Comparaison de l'évolution du taux de reconnaissance en fonction du pas initial du gradient sur le premier ensemble de la base DCASE 2017. **(b)** Évolution du coût multinomial et du coût de reconstruction sur l'ensemble d'apprentissage du premier ensemble de la base DCASE 2017.

le processus décrit ci-dessus, à la différence que les grandeurs sont calculées après chaque passage par 2000 points. La seule différence étant que nous initialisons le dictionnaire de manière aléatoire de façon à observer l'impact du modèle supervisé sur le coût de reconstruction de la NMF. Sur la figure 5.3 (a), on peut noter le compromis nécessaire entre stabilité et capacité de généralisation qu'implique le choix du pas du gradient. Le comportement de l'algorithme lorsque l'on modifie la valeur initiale du pas n'a rien de surprenant. Des valeurs de ρ trop grandes entraînent une instabilité de l'algorithme tandis que choisir un pas trop faible ralentit la décroissance, voire limite la capacité de généralisation.

Enfin, l'objectif des courbes 5.3 (b) est d'étudier l'évolution du coût de reconstruction de la NMF au cours des mises à jours du dictionnaire lors de l'apprentissage de TNMF par l'algorithme 5.2. Les modèles TDL et TNMF apprennent les dictionnaires selon un critère supervisé, ainsi il n'est en aucun cas garanti que le coût de reconstruction de la NMF décroisse toujours au cours des itérations. Comme on peut le remarquer, une fois le classifieur plus proche de la convergence, le modèle a tendance à faire des modifications plus fines sur le dictionnaire de façon à le rendre plus discriminatif, l'éloignant ainsi de sa valeur optimale recherchée pour la NMF non-supervisée.

5.4 Systèmes de classification de scènes

Dans cette section, nous évaluons les approches TNMF proposées sur le problème de classification de scènes. En particulier, nous choisissons de présenter les systèmes TNMF sur les bases de développement des deux campagnes d'évaluation internationales DCASE 2016 et 2017. Notre système soumis à l'édition 2016 du DCASE est uniquement basé sur TNMF, ce qui nous permet dans un second temps de comparer les performances de notre approche aux systèmes état de l'art proposés par les participants aux challenges. Les systèmes TNMF sont construits à partir de la même base que les approches par SNMF proposées au chapitre 4. Les représentations temps-fréquence, les stratégies de construction de la matrice de données et les processus de classification étant similaires, cela nous permet dans un premier temps de comparer la TNMF aux approches par NMF non-supervisées. De plus, comme ces travaux ont été effectués dans le contexte d'un challenge, ils permettent de se comparer à l'état de l'art dans un cadre contrôlé. Bien que nous

n'introduisions pas ces systèmes en détail, les rapports techniques associés ainsi que les résultats sont disponibles sur les sites des challenges. [2]

5.4.1 Ensembles de développement DCASE 2016 et 2017

Présentation des systèmes TNMF

Le système TNMF proposé est simplement une extension supervisée des systèmes SNMF présentés dans le chapitre 4. En effet, nous reprenons la stratégie de construction de la matrice de données par TM, et nous gardons une régression logistique comme classifieur. La différence étant que le classifieur et le dictionnaire sont appris conjointement en minimisant le coût de classification par TNMF en se basant sur l'algorithme 5.2.

Nous introduisons également une approche par fusion tardive de TNMF dans l'objectif d'améliorer la robustesse du système soumis au challenge DCASE 2016. L'objectif est simplement de s'affranchir des aspects aléatoires de TNMF introduits par l'initialisation et par l'algorithme de descente de gradient stochastique.

Nous rappelons que la base du DCASE 2016 se divise en deux ensembles, comme présenté en annexe A. Les modèles proposés seront comparés aux approches non-supervisées sur l'ensemble de développement et aux autres approches de l'état de l'art sur l'ensemble test du challenge.

Représentations temps-fréquence et construction de la matrice de données

Nous reprenons, à quelques paramètres près, la même stratégie d'extraction des représentations temps-fréquence et de construction de la matrice de données présentée au chapitre 4. Pour les deux bases, les CQT sont extraites avec 24 bandes par octave de 5 à 22kHz en utilisant des fenêtres de 30ms sans recouvrement. La construction de la matrice de données par TM se fait en utilisant des tranches de 1 seconde. Nous avons donc 30 tranches par exemple de 30 secondes pour la base DCASE 2016 et 10 tranches par exemple de 10 secondes pour la base DCASE 2017.

Réglages de la SNMF et TNMF

Nous utilisons la formulation de la SNMF présentée au chapitre 4 et dans [Le Roux et al., 2015b]. Nous gardons la décomposition fournissant le meilleur coût d'attache aux données parmi 5 initialisations. Le paramètre de parcimonie est fixé à $\lambda_1 = 0.2$.

Pour l'apprentissage du modèle TNMF, nous utilisons l'algorithme 5.2 ainsi que les outils d'implémentation détaillés section 5.3. Les paramètres ont été réglés sur les ensembles de développement construits à partir de chaque ensemble apprentissage. Les paramètres de régularisation ℓ_1 et ℓ_2 sont trouvés à $\lambda_1 = 0.2$ et $\lambda_2 = 0$, la régularisation du classifieur à $\mu = 0.1$ et le pas du gradient à $\rho = 0.001$. Nous reprenons les dictionnaires obtenus par SNMF afin d'initialiser l'algorithme.

Classification et fusion

Le modèle TNMF a l'avantage de regrouper l'apprentissage du dictionnaire, le calcul de la représentation positive latente (les activations) ainsi que l'apprentissage du classifieur. En revanche, le dictionnaire et les paramètres du classifieur appris peuvent varier d'une occurrence à l'autre de l'algorithme. Cela est dû à sa sensibilité à l'initialisation ainsi qu'à la nature de l'algorithme (descente de gradient stochastique). Afin de se servir de l'aspect aléatoire du modèle TNMF à notre

2. www.cs.tut.fi/sgn/arg/dcase2016/index

	Dcase 2016			**Dcase 2017**		
Référence		72.5			74.8	
	K=256	K=512	K=1024	K=256	K=512	K=1024
SNMF	81.2	82.6	83.1	79.3	83.1	84.4
TNMF	**85.0**	**84.8**	**84.5**	85.0	**86.3**	**85.9**
TNMF fusion		**86.2**			**86.8**	

TABLEAU 5.1 – Taux de reconnaissance pour TNMF et SNMF pour différentes tailles de dictionnaires K.

avantage nous utilisons un schéma de fusion tardive relativement simple se composant des étapes suivantes :

— Apprendre deux dictionnaires différents sur deux initialisations de TNMF avec $K = 256$. Faire de même pour $K = 512$.

— Extraire les 4 représentations en calculant les projections optimales sur chaque dictionnaire.

— Entraîner une régression multinomiale différente sur chaque représentation apprise.

— Moyenner les log-probabilités de sortie des 4 classifieurs sur chaque exemple de la base et prendre la décision finale en choisissant la classe ayant la probabilité finale la plus élevée.

Ce processus permet de rendre notre système de classification de scènes plus robuste. Dans tous les cas testés expérimentalement, la fusion de différentes occurrences du modèle nous permet d'obtenir de meilleurs scores que la moyenne des résultats des modèles individuels.

Résultats sur les ensembles de développement

Nous commençons par présenter les résultats sur les ensembles de développement fournis aux participants des compétitions DCASE 2016 et 2017. Il s'agit des ensembles publics mis à disposition des participants pour développer leurs systèmes tel que décrit en annexe A. Ces ensembles correspondent exactement aux bases de données DCASE 2016 et 2017 utilisées dans les chapitres précédents. Nous rappelons que les résultats sont présentés comme la moyenne des scores sur les 4 ensembles apprentissage-test.

Les performances des différents systèmes comparés sont exposées dans le tableau 5.1. Nous incluons les résultats du système de référence de la base de données à titre indicatif ainsi que les résultats obtenus par extraction de descripteurs d'images. Nous notons "TNMF fusion" le système utilisant le schéma de fusion tardive décrit précédemment.

Cette première évaluation du modèle TNMF permet de confirmer plusieurs de ses avantages par rapport à son équivalent non-supervisé. Le premier est en termes de performance, les systèmes TNMF permettent d'obtenir un gain de taux de reconnaissance de 2 points par rapport à SNMF. Le deuxième aspect remarquable est que TNMF atteint de bien meilleurs taux de reconnaissance pour de plus petits dictionnaires. En effet, les meilleures performances sont atteintes pour un nombre de composantes du dictionnaire de $K = 256$, alors qu'elles sont maximum pour $K = 1024$ avec SNMF. C'est un premier moyen de confirmer la possibilité d'améliorer la qualité des représentations NMF en les apprenant conjointement avec un classifieur. En adaptant le dictionnaire à la tâche traitée, le modèle TNMF apprend les événements de base du dictionnaire permettant de mieux discriminer les scènes sonores. Ainsi, le modèle est moins demandeur en nombre de composantes pour des performances équivalentes comparé aux techniques non-supervisées. En effet, les approches non-supervisées apprennent les événements de base de façon à décrire au mieux les différentes scènes sonores de la base de données car elles sont apprises à partir d'un critère

d'attache aux données. Cependant, on peut supposer que certaines d'entre elles sont communes à un grand nombre de scènes, et par conséquent sont moins utiles pour les discriminer.

Réduire la taille des dictionnaires permet également de réduire le nombre de paramètres du modèle et par conséquent de réduire les temps de calcul durant la phase de test. Bien que l'apprentissage du modèle TNMF soit plus complexe que SNMF, la complexité au test est la même. Ainsi pour des cas d'applications réelles, il peut devenir important d'avoir un système performant avec la taille la plus réduite possible. Enfin, nous obtenons des taux de reconnaissance légèrement supérieures sur les deux bases en utilisant la fusion tardive de différentes occurrences du modèle. Cela confirme que la fusion proposée permet de tirer parti des aspects aléatoires du modèle afin d'obtenir un système plus robuste. Cette fusion reste relativement simple. Nous l'avons proposée telle quelle dans le but de s'assurer un résultat stable pour le système que nous avons soumis au challenge sans trop éloigner l'attention portée sur l'apprentissage de représentations. Pour aller plus loin, il serait éventuellement possible d'améliorer encore les scores en utilisant davantage de modèles dans la fusion ou en incluant d'autres classifieurs.

Résultats sur l'ensemble du challenge DCASE 2016

Après avoir comparé nos méthodes entre elles sur l'ensemble de développement, nous les comparons aux meilleurs systèmes sur l'ensemble de test du challenge DCASE 2016. Ici, notre système correspond à celui de TNMF-fusion appris sur toute la base de développement. Les étiquettes n'étant connues que par les organisateurs, aucune forme de réglage ne peut être fait sur l'ensemble de test, assurant une évaluation équitable des systèmes. Nous incluons, en plus du système de référence, les scores des meilleurs systèmes du challenge. Notre système s'étant classé 3ème sur 49 soumissions, nous ne pouvons présenter l'ensemble des résultats, ils sont disponibles dans le détail sur le site du challenge. Les taux de reconnaissance et les rangs des meilleurs systèmes sont présentés dans le tableau 5.2. Nous mentionnons également la nature des représentations d'entrée et des classifieurs utilisés pour chaque système. Les systèmes classés 2 et 6 ne sont pas inclus dans le tableau car ils correspondent aux systèmes des mêmes auteurs fusionnés pour obtenir le premier système.

Si l'on observe l'ensemble des systèmes soumis par les participants au challenge, on peut s'apercevoir que 16 parmi les 20 premiers systèmes sont basés sur une classification par réseaux de neurones. C'est également le cas des meilleurs systèmes présentés dans le tableau 5.2. Les résultats obtenus avec des réseaux convolutifs seuls (sans stratégie de fusion) sont particulièrement surprenants. En effet, malgré leur domination dans bien d'autres tâches de traitement de l'audio, celle-ci reste à prouver pour la classification de scènes sonores, en particulier pour des bases de petite et moyenne tailles. Les approches par CNN sont réputées nécessiter un nombre plus important d'exemples pour assurer la capacité de généralisation des modèles. Au contraire les approches par NMF ont l'avantage de fournir de bonnes représentations avec moins de données.

On peut également noter que la grande majorité des systèmes utilise des descripteurs audio issus d'autres tâches comme entrée de leur système de classification. Par exemple, les descripteurs MFCC sont utilisés dans 7 des 10 premiers systèmes ainsi que dans le premier et le deuxième système. Ainsi, notre système TNMF se démarque de la plupart des approches performantes du challenge à la fois par la stratégie de classification et par la représentation d'entrée utilisée. En effet, TNMF permet d'apprendre des représentations d'entrée adaptées à la tâche traitée uniquement à partir des représentations temps-fréquence. La qualité des représentations apprises à l'aide de TNMF nous rend compétitifs par rapport aux systèmes par réseaux de neurones complexes en utilisant simplement des régressions logistiques comme classifieurs. Dans ce travail, notre attention s'est focalisée sur l'apprentissage des représentations pour la classification, contrairement aux autres approches qui préfèrent souvent se concentrer sur l'étape de classification. Procéder

	DCASE 2016 Challenge set			
	Entrée	Classifieur	Accuracy	Rang/49
[Mesaros et al., 2016b]	Mel Spectres	GMM	77.2	39
[Valenti et al., 2016]	Mel Spectres	CNN	86.2	7
[Marchi et al., 2016]	Descripteurs audios	Fusion de DNN	86.4	5
[Park et al., 2016]	Coeffs. cepstraux	DNN + GMM	**87.2**	4
[Eghbal-Zadeh et al., 2016]	MFCC + Spectres	CNN + I-vector	**89.7**	1
Nous	**CQT**	**Fusion TNMF**	**87.7**	3

TABLEAU 5.2 – Taux de reconnaissance et rangs des systèmes les mieux classés au challenge DCASE 2016.

ainsi nous permet d'aborder le problème dans l'ordre de la chaîne de classification. Maintenant que ces résultats confirment l'efficacité des variantes NMF proposées pour décrire et discriminer les scènes sonores, nous avons des bases solides pour travailler davantage sur la stratégie de classification dans le chapitre suivant.

5.5 Systèmes de classification d'événements

Nous avons montré dans le chapitre précédent que les techniques d'apprentissage de représentations par NMF permettaient de traiter efficacement certaines difficultés de la détection d'événements sonores. Le pouvoir de séparation de sources de la NMF est particulièrement utile pour les cas avec recouvrement entre événements de classes différentes. On peut ainsi supposer que cette tâche constitue un autre moyen d'appuyer les avantages qu'amène l'extension supervisée de la NMF qui fait l'objet de ce chapitre. De plus, l'application à la détection d'événements nous permet d'étudier le changement du coût d'attache aux données euclidien présent dans la formulation originelle du problème TDL, pour aller vers l'utilisation de divergences plus largement plébiscitées dans les travaux sur la NMF pour l'audio telles que les β-divergences. En particulier, nous souhaitons montrer qu'il est également possible d'apprendre la NMF avec la divergence de KL conjointement avec un classifieur. Nous évaluons nos systèmes sur les deux mêmes bases de données de détection d'événements que dans le chapitre précédent : la base TUT SED 2016 en conditions réelles et la base synthétique TUT SED synth. L'architecture des systèmes est équivalente à celle proposée pour la SNMF non-supervisée appliquée à la détection d'événements dans le chapitre 4. La seule différence étant que l'application de TNMF regroupe l'étape d'apprentissage de représentations et d'apprentissage du classifieur. Les représentations de bas niveau et la stratégie de prédiction des étiquettes sont les mêmes.

5.5.1 En conditions réelles : TUT SED 2016

Nous reprenons l'évaluation entamée chapitre 4 sur la base DCASE 2016 de détection d'événements avec recouvrement en conditions réelles. La base de données est décrite en détail annexe A. Nous gardons les mêmes représentations temps-fréquence en entrée, c'est-à-dire des spectres en bandes Mel avec 40 bandes et des fenêtres de 40ms avec 50% de recouvrement. Nous rappelons également, que le problème étant multi-labels, nous enlevons toutes les trames contenant plus de deux événements dans l'annotation lors de la phase d'apprentissage, tel que décrit au chapitre 4.

			Home		Res. A.		Moyenne	
	Méthode	Features	ER	F1	ER	F1	ER	F1
[Mesaros et al., 2016b]	GMM	MFCC	-	-	-	-	91	23.7
[Elizalde et al., 2016]	Forêts	MFCC	-	-	-	-	76	38.5
[Zöhrer et Pernkopf, 2016]	GRNN	Mel	-	-	-	-	73	47.6
[Vu et Wang, 2016]	RNN	Mel	-	-	-	-	81.5	**49.8**
Nous	NMF $\beta = 2$	Mel	87	29	60	56	73.5	42.5
Nous	NMF $\beta = 1$	Mel	86	30	59	58	72.5	44
Nous	TNMF $\beta = 2$	Mel	85	36	56	64	**70.5**	**50**
Nous	TNMF $\beta = 1$	Mel	86	34	55	64	**70.5**	**49**

TABLEAU 5.3 – Taux d'erreur et scores F1 sur les deux environnements de la base TUT SED 2016 pour le système NMF proposé comparés aux systèmes état de l'art.

Réglages de TNMF

Nous utilisons le modèle TNMF avec l'algorithme 5.2, avec la différence que nous considérons chaque trame comme un exemple à classifier. Ainsi nous n'avons pas l'étape de d'agrégation par moyenne des activations introduite pour la classification de scènes. Les changements qu'impliquent l'utilisation de la divergence de KL sont mentionnés section 5.2 et sont détaillés par Sprechmann et al. [2014]. Nous apprenons le système sur 6 époques de l'algorithme 5.2, avec un pas du gradient à $\rho = 0.001$, la contrainte du classifieur à $\nu = 10$ et les paramètres de parcimonie à $\lambda_1 = 0.5$ et $\lambda_2 = 0$ (identiques à ceux pour SNMF).

Analyse des résultats

Les scores F1 et les taux d'erreur pour les deux environnements de la base sont exposés dans le tableau 5.3. Plus de détails sur le calcul de ces métriques sont donnés dans l'annexe A. Le système TNMF proposé est comparé aux meilleurs systèmes de l'état de l'art sur cette base de données ainsi qu'à son équivalent non-supervisé. Nous rappelons que "NMF $\beta = 2$" désigne la NMF avec le coût euclidien et "NMF $\beta = 1$" celle avec la divergence de KL.

Tout comme nous l'avons vu pour la classification de scènes, l'apprentissage supervisé de la NMF permet d'adapter les représentations au problème traité se traduisant par une amélioration des performances. Les résultats obtenus pour les deux métriques et sur les deux environnements sont supérieurs à ceux obtenus par SNMF, et ce indépendamment du coût d'attache aux données. Alors que l'apprentissage de descripteurs par NMF non-supervisée nous donnait déjà des résultats prometteurs par rapport aux meilleures approches par réseaux de neurones, TNMF nous permet dans tous les cas d'au moins égaler ceux-ci. En effet, nous obtenons avec TNMF les taux d'erreur les plus bas sur la base de données et un score F1 similaire à celui obtenu avec la meilleure méthode par réseaux de neurones récurrents [Vu et Wang, 2016]. Par la même occasion, ces résultats confirment un des avantages des NMF par rapport aux approches par réseaux de neurones : leurs performances sur des petites bases de données. Notre approche est compétitive avec des approches capables de tirer parti du contexte temporel pour prédire les événements en utilisant des variantes de réseaux récurrents [Vu et Wang, 2016; Zöhrer et Pernkopf, 2016]. Cela peut paraître surprenant car notre utilisation de TNMF se fait par l'usage d'un classifieur appris trame par trame, complètement indépendamment du contexte temporel. Nous supposons que nous profitons dans une certaine mesure de la faible taille de la base de données. Pour certaines catégories d'événements

comme *object impact*, seules 15 occurrences sont disponibles dans toute la base, ce qui est loin d'être suffisant pour apprendre un modèle complexe. Les modèles par réseaux récurrents sont relativement complexes et sont réputés nécessiter un nombre important de données. Ainsi pour ce genre de tâches, où récupérer des données annotées est coûteux, le modèle TNMF constitue une alternative efficace aux modèles profonds. Nous verrons dans la section suivante comment notre système se compare aux modèles de l'état de l'art sur une plus grosse base de données synthétique.

Ces résultats nous amènent également à discuter de l'impact du choix du coût d'attache aux données. Nous avions remarqué dans le chapitre précédent que la divergence de KL amenait une légère amélioration des performances pour SNMF. Les scores dans le tableau 5.3 indiquent que ce n'est pas le cas pour sa variante supervisée. Les taux d'erreur sont égaux pour les deux divergences avec un score F1 légèrement supérieur pour $\beta = 2$ ne permettant pas de suffisamment les différencier. Il semblerait que le modèle TNMF apprenne de meilleurs dictionnaires indépendamment du coût d'attache aux données. En revanche les résultats confirment qu'il est possible d'apprendre le modèle TNMF pour la classification avec la divergence de KL, ce qui pourrait s'avérer utile pour des bases ou des tâches plus sensibles à ce choix.

5.5.2 Base synthétique : TUT SED synth

Nous évaluons maintenant nos approches sur la base TUT SED synth dans la continuité des travaux présentés dans le chapitre précédent. Nous gardons les mêmes représentations temps-fréquence en entrée, c'est-à-dire des spectres en bandes Mel avec 80 bandes et des fenêtres de 40ms avec 50% de recouvrement. Tout comme dans la section précédente, nous gardons la même stratégie pour l'apprentissage de dictionnaire et la prise en compte du recouvrement pour la classification.

Réglages de TNMF

Comme un ensemble de développement est clairement établi sur cette base, nous nous en servons comme critère d'arrêt. Pour cela nous apprenons le modèle sur 20 itérations et gardons le couple $(\mathbf{W}, \mathbf{A})$ ayant fourni le meilleur score F1 par trame sur l'ensemble de développement. Le pas du gradient initial est fixé à $\rho = 0.001$ en suivant l'heuristique de décroissance mentionnée à la section 5.2, la contrainte du classifieur à $\nu = 10$ pour $\beta = 2$ et $\nu = 1$ pour $\beta = 1$. Tout comme pour SNMF, nous apprenons des dictionnaires avec $K = 32$ composantes. Les paramètres de parcimonie sont laissés à $\lambda_1 = 0.0$ et $\lambda_2 = 0$, ces valeurs sont identiques à celles pour SNMF mais elles sont également les meilleures valeurs trouvées pour TNMF sur l'ensemble de développement.

Analyse des résultats

Les scores F1 et les taux d'erreur (ER) pour les deux environnements de la base sont présentés dans le tableau 5.4. Comme dans le chapitre précédent, nos systèmes sont comparés aux différentes approches par réseaux de neurones proposées par [Cakir et al., 2017] ainsi qu'aux approches par SNMF non-supervisées.

On peut commencer par noter que sur cette base de données, l'apport de TNMF par rapport à son équivalent SNMF non-supervisée est moins clair, du moins pour les métriques calculées sur les segments de 1 seconde. Dans le même temps, on peut également remarquer un gain notable en performance pour les métriques d'évaluation par trame. Une des principales raisons est le fait que le dictionnaire est appris avec TNMF de façon à minimiser un coût de classification calculé trame par trame. Ainsi, les scores F1 et ER par trame sont plus proches du critère optimisé par le modèle TNMF. De plus, nous pouvons également observer, comme pour la base réelle, qu'il n'y a que très peu de différences en termes de performance entre l'utilisation des deux divergences.

	Méthode	ER_{tr}	$F1_{tr}$	ER_{1sec}	$F1_{1sec}$
[Mesaros et al., 2016b]	GMM	0.78	40.5	0.72	45.3
[Cakir et al., 2015b]	DNN	0.68	49.2	1.1	50.2
[Cakir et al., 2017]	CNN	0.56	59.8	0.78	59.9
[Cakir et al., 2017]	RNN	0.6	52.8	0.64	57.1
[Cakir et al., 2017]	CRNN	**0.48**	**66.4**	**0.47**	**68.7**
Nous	NMF $\beta = 2$	0.71	40.8	0.74	46.5
Nous	NMF $\beta = 1$	0.72	40.8	0.74	47.0
Nous	TNMF $\beta = 2$	0.68	44	0.74	49
Nous	TNMF $\beta = 1$	0.68	44.3	0.73	48.1

TABLEAU 5.4 – Taux d'erreur et scores F1 sur les deux environnements de la base TUT SED synth 2016 pour les systèmes NMF et TNMF proposés comparés aux systèmes état de l'art.

L'apprentissage trame par trame limite naturellement les performances que peuvent atteindre nos modèles. En effet, en l'absence de modélisation ou de prise en compte de l'information temporelle il est plus difficile d'être compétitif avec des modèles de type RNN ou CNN, en particulier sur des plus grosses bases de données. Les conclusions tirées de cette évaluation sont semblables à celles obtenues dans le chapitre 4 : il devient crucial pour cette tâche de s'attaquer à l'étape de classification et de détection. En effet, le modèle TNMF proposé traite uniquement la question de l'apprentissage de la représentation. Si les résultats obtenus sont prometteurs, nos systèmes restent largement handicapés par la simplicité des approches de classification. Nous nous intéresserons de plus près au choix d'un classifieur plus approprié dans le chapitre suivant, en remplaçant la régression logistique par des modèles de réseaux de neurones profonds capables de modéliser l'évolution temporelle des représentations NMF.

5.6 Conclusion

Les techniques d'apprentissage de descripteurs par factorisation de matrices classiques peuvent souvent être améliorées par l'introduction d'extensions supervisées, conditionnant les décompositions à fournir des décompositions adaptées aux tâches traitées. En particulier pour la classification de sons environnementaux, le modèle TNMF introduit permet d'adapter l'apprentissage de dictionnaires de sorte à obtenir des descripteurs facilitant l'étape de classification. En apprenant le dictionnaire NMF et le classifieur conjointement dans un problème d'optimisation à deux niveaux, le modèle TNMF conserve les avantages de la NMF tout en l'adaptant au critère de classification cible. Le modèle TNMF, avec l'algorithme proposé, permet d'améliorer les performances des systèmes de classification de scènes et de détection d'événements. Certains des résultats obtenus sont compétitifs avec des modèles par réseaux de neurones profonds qui constituent l'état-de-l'art dans de nombreuses autres tâches d'apprentissage automatique. Toutefois, les modèles TNMF se basent sur des stratégies de classification relativement simples limitant naturellement leur performance sur des tâches plus exigeantes en termes de modélisation temporelle.

Chapitre 6

Approches par réseaux de neurones profonds

Sommaire

6.1 Motivations

Nous entamons ce dernier chapitre dans l'objectif de s'attaquer au dernier maillon de la chaîne : le classifieur. Le cœur de nos contributions était jusqu'ici l'apprentissage de représentations par NMF. Dans ce contexte, nous avons laissé de côté le choix du classifieur pour rester focalisé sur les approches par apprentissage de descripteurs. Ce choix de méthodologie nous a amené à nous limiter à l'utilisation d'une régression logistique comme modèle de classification, à la fois pour la classification de scènes et la détection d'événements. Nous avons vu dans les chapitres précédents que, malgré l'usage d'un classifieur relativement simple, la qualité des représentations apprises par NMF et TNMF nous a parfois permis d'obtenir des résultats compétitifs avec l'état de l'art. En revanche, nous avons également montré que ce n'est pas tout à fait le cas pour la détection d'événements. Lors de l'utilisation de bases de données de plus grande taille, nos systèmes de détection d'événements par régression logistique souffrent de leur manque de modélisation temporelle. En effet, la comparaison avec des modèles par réseaux de neurones profonds plus complexes montre qu'il reste du chemin à parcourir en termes de performance pour les systèmes NMF que nous avons présenté jusqu'ici. Outre la profondeur des réseaux, c'est également leur capacité à apprendre et à modéliser l'évolution de l'information temporelle avec des couches récurrentes ou convolutives qui permet de traiter certaines difficultés de la détection d'événements.

Arrivée des premiers réseaux pour l'ACSES

Si notre étude des approches par réseaux de neurones s'inscrit dans la suite logique de notre travail, leur étude est devenue incontournable pour les applications traitées, ne serait-ce que pour avoir un système de référence crédible. Comme nous l'avons vu dans le chapitre 2, cela fait maintenant plus de deux ans que les différents acteurs de l'analyse de sons environnementaux s'intéressent à la question des réseaux de neurones profonds pour les différentes tâches du domaine. Une première phase d'exploration a permis de valider l'intérêt des approches par apprentissage profond en proposant différentes avancées en performance avec des réseaux relativement simples. Cela s'est remarqué par exemple avec les premiers travaux d'approches par réseaux de neurones pour la détection d'événements. Des travaux ont proposé d'utiliser de simples perceptrons multicouches (MLP) améliorant les systèmes de référence combinant HMM et NMF [Gencoglu et al., 2014; Cakir et al., 2015b]. Une des raisons du succès de ces modèles dans le cas de la détection d'événements avec recouvrement est leur capacité à directement traiter l'aspect multi-labels par le choix de fonctions de coût et d'activations appropriées dans la dernière couche du réseau [Cakir et al., 2015a]. Les travaux basés sur des MLP pour la détection d'événements traitent un problème de détection temporelle avec des modèles effectuant un apprentissage trame par trame. Dans ce contexte, l'introduction de couches récurrentes dans les réseaux constitue une avancée importante en termes de modélisation et de performances pour le domaine. Depuis l'introduction de couches récurrentes avec les unités LSTM (de l'anglais *long short term memory*) pour la détection d'événements avec recouvrement [Parascandolo et al., 2016], une part de plus en plus importante des travaux sur cette tâche incluent de telles couches dans leurs réseaux.

En revanche, les applications telles que la classification de scènes ou d'événements ont pour objectif d'attribuer ue étiquette à un segment audio et n'impliquent pas de détection temporelle précise. Pour ces tâches de classification, toute l'attention s'est rapidement portée sur les réseaux de neurones convolutifs (CNN). Les CNN devenant état de l'art dans de plus en plus d'applications du traitement du signal dans sa généralité, l'exploration de leur potentiel pour l'analyse de sons environnementaux est inévitable. Ce qui rend l'application des CNN d'autant plus naturelle est l'habitude de représenter les signaux audios par des images correspondant à leurs représentations temps-fréquence, comme discuté chapitre 3. D'autres domaines de recherche ont une avance si-

gnificative quant au développement d'approches par CNN efficaces. Évidemment, la majorité des avancées proviennent de la vision par ordinateur, les CNN étant initialement introduits pour traiter ce genre de problème. Mais d'autres tâches du traitement du signal audio telles que le traitement de la parole ou de la musique ont connu des premières approches par CNN qui ont précédé celles pour l'analyse des sons environnementaux.

De nombreux travaux ont contribué à l'exploration de l'application de CNN à la classification de scènes et d'événements. La majorité d'entre eux s'appuie sur des architectures similaires à celles pour la vision tout en ayant un nombre de couches naturellement limité par la taille des bases de données disponibles [Rakotomamonjy, 2017; Piczak, 2015a; Valenti et al., 2016; Eghbal-Zadeh et al., 2016]. Si les premiers CNN obtiennent des résultats prometteurs, ils restent en compétition avec des approches par factorisation de matrices [Salamon et Bello, 2015b], I-vecteurs [Eghbal-Zadeh et al., 2016] ou encore par simple MLP [Park et al., 2016; Marchi et al., 2016]. Suivant cette phase d'exploration, avec l'augmentation de la taille des bases de données et l'élargissement de la communauté, les CNN définissent maintenant l'état de l'art sur de nombreuses tâches et bases de données du domaine. En effet, les différentes approches proposées depuis suivent de très près et s'inspirent des avancées théoriques, algorithmiques et applicatives des réseaux de neurones profonds. On peut noter par exemple l'utilisation de réseaux récurrents combinés aux réseaux convolutifs [Cakir et al., 2017; Xu et al., 2017c], des réseaux génératifs adversaires (GAN) [Mun et al., 2017] ou des réseaux résiduels [Jee-Weon et al., 2017; Zhao et al., 2017].

Organisation et contributions

Parmi les multiples défis que posent les réseaux de neurones profonds nous nous intéressons plus particulièrement au choix de la représentation d'entrée. Le choix de la représentation d'entrée, ou des descripteurs, est un des facteurs limitant potentiellement la capacité d'apprentissage des réseaux. En effet, aussi complexe que puisse être le réseau, sa capacité à traiter efficacement la tâche donnée est dépendante de la quantité d'information et de la nature de la représentation d'entrée fournie pour son apprentissaget. Cette question se pose tout particulièrement pour le traitement du son car, contrairement à l'image, on ne traite en général pas directement le signal brut. Dans ce cas, les descripteurs audio ou les représentations temps-fréquence choisies ont pour rôle de réduire la dimension tout en donnant une représentation du signal plus facilement interprétable par le réseau. Pour l'analyse de sons environnementaux, l'approche de loin la plus répandue est d'extraire une représentation temps-fréquence (en majorité des spectres Mel). La plupart des travaux se focalisent sur le réseau lui-même, les grandeurs telles que le nombre de bandes de fréquences, la nature des filtres ou la taille des fenêtres d'analyse sont alors simplement traitées comme des paramètres à régler. Il existe néanmoins quelques travaux s'intéressant de plus près au choix de la représentation [Piczak, 2017] ou qui ont proposé de traiter directement le signal brut, soit en apprenant un banc de filtres conjointement au réseau soit en apprenant un réseau convolutif directement sur la forme d'onde [Cakir et al., 2016; Dai et al., 2017].

Dans ce chapitre nous nous intéressons non pas directement à la représentation temps-fréquence mais à l'ajout d'une étape d'apprentissage de représentation par NMF entre le spectrogramme et le réseau. Cette étape entraîne l'utilisation de descripteurs appris automatiquement par factorisation de matrices comme représentation d'entrée des réseaux. Du point de vue des travaux de l'état de l'art, notre approche se situe effectivement dans l'étude du choix de la représentation d'entrée. En revanche du point de vue du déroulement de nos travaux, cela revient également à travailler sur la recherche de stratégies de classification adaptées aux approches par NMF. Une des idées principales de notre approche est de laisser en grande partie le rôle d'apprentissage de représentations intermédiaires aux modèles NMF. En effet, les réseaux de neurones, en particulier les réseaux convolutifs, sont également considérés et utilisés comme techniques d'apprentissage de descrip-

teurs supervisés. De nombreux travaux ont montré qu'il suffit de retirer la couche de classification d'un réseau de neurones profonds pour s'en servir comme extracteur de descripteurs. Si ce réseau a été appris sur un nombre suffisant de données représentatives du problème, il peut être capable de fournir des descripteurs performants même interprétés avec des classifieurs simples (SVM, régression logistique). Le modèle TNMF proposé dans le chapitre précédent s'inscrit dans cette logique, en proposant un apprentissage supervisé de descripteurs obtenant des performances compétitives avec une simple régression logistique. Ainsi, outre l'amélioration du classifieur, l'objectif de ce chapitre est également de discuter et d'observer en quoi l'étape de NMF peut remplacer les couches cachées classiques dans le rôle d'apprentissage de représentations intermédiaires à partir des spectrogrammes.

Nous commencerons par présenter quelques généralités sur le fonctionnement des différentes couches de base, nécessaires à la compréhension de la discussion et des modèles utilisés. Ensuite nous discuterons de la compatibilité entre la NMF et les couches classiques des réseaux de neurones. Dans ce sens, nous verrons en quoi l'étape de NMF peut être vue comme analogue au pré-apprentissage de couches cachées. De plus, nous montrerons que le modèle TNMF est équivalent à un MLP à une couche cachée en précisant le choix des activations et des contraintes appliquées aux paramètres. Nous présenterons ensuite les modèles enchainant NMF et réseaux de neurones profonds que nous proposons d'appliquer à la classification de scènes et à la détection d'événements. Nous montrerons qu'un modèle MLP simple, appris sur les représentations NMF et TNMF, nous permet d'améliorer les performances sur la classification de scènes. Nous reprendrons le même genre d'évaluation pour la détection d'événements, en comparant l'intérêt de prendre la NMF en entrée de modèles MLP et de réseaux récurrents par rapport à l'utilisation de spectres Mel. Ensuite, nous nous inspirerons du modèle récurrent convolutif [Cakir et al., 2017], état de l'art en détection d'événements, en l'adaptant à l'utilisation de représentations NMF en entrée. Nous proposons une architecture similaire avec des couches convolutives filtrant uniquement l'axe du temps des activations NMF. Enfin, nous présenterons une première exploration du modèle DNN-TNMF qui regroupe l'apprentissage du réseau et du dictionnaire NMF en reprenant le cadre de TNMF.

6.2 Quelques notions et notations sur les modèles utilisés

Nous présentons dans cette section des généralités sur les réseaux de neurones utiles au déroulement des discussions qui suivent ainsi qu'à la compréhension des modèles proposés. Les concepts décrits sont relativement indépendants des tâches traitées. Ils sont décrits plus de détails dans l'ouvrage de Goodfellow et al. [2016] et sont résumés dans le cadre de l'analyse de sons environnementaux par McFee [2018]. Nous résumons le fonctionnement de base d'un réseau de neurones profonds ainsi que les principales famille de couches cachées dans la figure 6.1.

6.2.1 Nature et architecture des réseaux étudiés

Nous commençons par présenter quelques notations que nous utiliserons au cours du chapitre.

— Un ensemble d'apprentissage de N vecteurs de données représenté par la matrice $\mathbf{V} = [\mathbf{v}_1, ..., \mathbf{v}_N] \in \mathbb{R}^{F \times N}$.

— Un problème de classification à C classes, l'appartenance d'un exemple à chaque classe est représentée par y_c dans $\{0, 1\}$, indiquant si la classe est active ou non. Pour les problèmes multi-classes standards, seul un des y_c peut être non nul. Pour les problèmes multi-labels, plusieurs y_c peuvent être actifs pour un même exemple.

— Un dictionnaire $\mathbf{W} = [\mathbf{w}_1, ..., \mathbf{w}_K]$ dans l'ensemble $\mathcal{W}_+ = \{\mathbf{W} \in \mathbb{R}_+^{F \times K}$ t.q. $\forall j \in \{1, ..., K\}$, $\|\mathbf{w}_j\|_2 = 1\}$.

— Les paramètres des différentes couches des réseaux seront notés $\mathbf{A}^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$ où (l) représente l'indice de profondeur dans le réseau et d_l représente le nombre d'unités dans la couche l.

— Nous noterons $\mathbf{z}^{(l)}$ à la fois l'entrée de la couche $l+1$ et la sortie de la couche l.

En général, les réseaux de neurones profonds se définissent comme une succession d'opérations non-linéaires appliquées aux observations afin de produire la sortie désirée. Si on prend un vecteur d'entrée $\mathbf{v}$, un réseau à L couches enchaîne L opérations non-linéaires f_l, une par couche, de sorte à produire la sortie $\mathbf{z}^{(L)}$. La succession de ces opérations se résume par une composition de L fonctions appliquées à $\mathbf{v}$ telles que :

$$f(\mathbf{v}) = (f_L \circ f_{L-1} \circ ... \circ f_1)(\mathbf{v}) \tag{6.1}$$

Ainsi, f_l correspond à l'opération effectuée par une couche du réseau sur la sortie de la couche $l-1$.

6.2.2 Le perceptron multi-couches

Nous commençons par le perceptron multi-couches (MLP) [Rosenblatt, 1958], qui est à la base de la plupart des modèles présentés par la suite. En particulier nous définissons les opérations f_l effectuées par les différentes couches du MLP. On se donne une fonction non-linéaire $\sigma : \mathbb{R}^{d_l} \rightarrow \mathbb{R}^{d_l}$ souvent appelée activation. Dans ce cas l'opération effectuée par la couche l correspond à une opération non-linéaire appliquée à une transformation affine de l'entrée de la couche :

$$f_l(\mathbf{z}^{(l-1)}) = \sigma(\mathbf{A}^{(l)}\mathbf{z}^{(l-1)} + \mathbf{b}^{(l)}). \tag{6.2}$$

Les paramètres de la couche l sont alors la matrice de poids $\mathbf{A}^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$ et le vecteur de biais $\mathbf{b}^{(l)} \in \mathbb{R}^{d_l}$. Les couches de MLP standards sont également appelées *fully-connected layers* en anglais faisant référence au fait que chaque neurone de la couche est connecté à toutes les entrées, c'est-à-dire à toutes les dimensions de $\mathbf{z}^{(l-1)}$. Il existe un large choix de fonctions d'activations dans la littérature, telles que les fonctions sigmoïdes ou tangentes hyperboliques. Comme le font la majorité des travaux actuels, nous utiliserons l'activation dénommée "ReLU", de l'anglais *rectified Linear Unit* [Nair et Hinton, 2010; Hinton et al., 2012]. Les activations ReLU effectuent simplement un seuillage de la sortie de la couche à 0, ainsi σ se définit comme $\sigma(x) = \max(0, x)$. L'introduction de ce genre d'activation est considéré comme un des facteurs facilitant l'apprentissage de réseaux plus profonds et plus complexes, contribuant à l'augmentation relativement récente des performances de classification dans de nombreuses tâches.

6.2.3 Fonction objectif

Notre utilisation des réseaux se limite à des applications de classification. Travailler dans ce cadre conditionne le type d'opérations effectuées par la dernière couche du réseau. Ainsi, il est commun de faire référence à cette dernière couche comme étant la couche de classification. Pour les problèmes multi-classes, on utilise la fonction *soft-max* définie comme :

$$\sigma_{sm}(\mathbf{x})_c = \frac{\exp(x_c)}{\sum_{i=1}^{C} \exp x_i} \text{ pour } \mathbf{x} \in \mathbb{R}^C. \tag{6.3}$$

Ainsi, en pratique l'opération effectuée par la dernière couche correspond à :

$$f_L(\mathbf{z}_{L-1}) = \sigma_{sm}(\mathbf{A}^{(L)}\mathbf{z}^{(L-1)} + b^{(L)}). \tag{6.4}$$

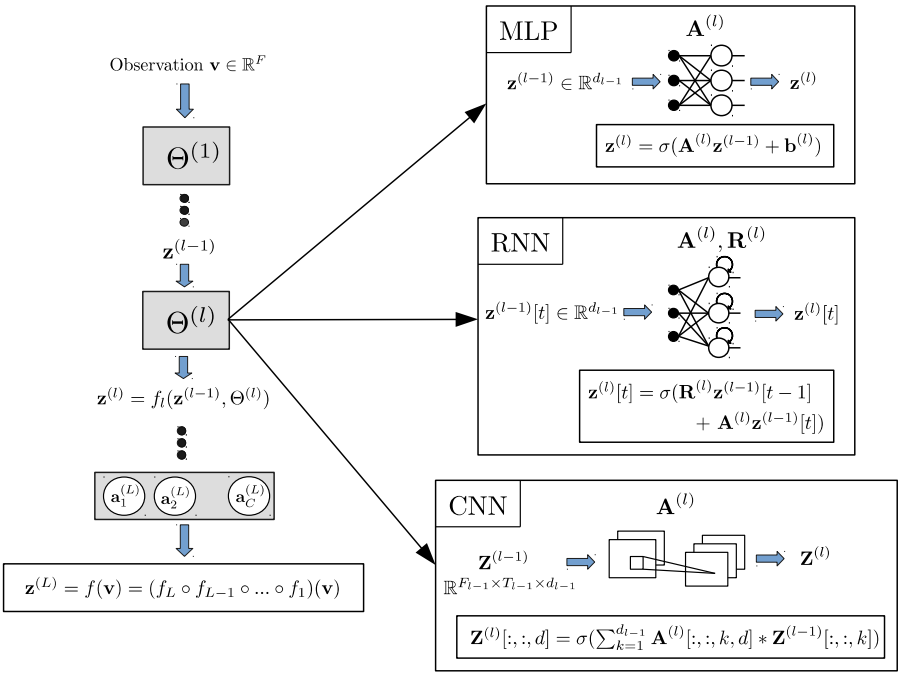


FIGURE 6.1 – Illustration et résumé des notations pour les réseaux de neurones profonds et les différentes couches utilisées.

Dans ce cadre, la fonction de coût standard est la *categorical cross-entropy* représentant l'erreur de classification $\ell_s(y, \hat{y})$ entre la vraie étiquette y et l'étiquette estimée $\hat{y}$:

$$\ell_s(y, \hat{y}) = -\sum_{c=1}^{C} y_c \log(\sigma_{sm}(\mathbf{a}_c^{(L)}\mathbf{z}^{(L-1)} + b_c^{(L)})). \tag{6.5}$$

La fonction *soft-max* a pour effet de normaliser la sortie de la dernière couche de sorte à ce que le vecteur de sortie soit de somme unitaire. Elle s'interprète alors comme donnant la probabilité d'appartenance à chaque classe et elle est utilisée avec le coût *categorical cross-entropy*. La dernière couche du réseau se comporte comme une régression logistique multinomiale. Ainsi, si on garde uniquement la couche de classification, un MLP est équivalent à une régression logistique apprise selon un critère de maximum de vraisemblance.

Pour les problèmes multi-labels, il est commun d'utiliser une activation de type sigmoïde pour la dernière couche :

$$\sigma_{sg}(x) = \frac{1}{1 + \exp(-x)}. \tag{6.6}$$

Dans ce cas la fonction de coût utilisée est la cross-entropie binaire, de l'anglais *binary cross-entropy*, et se définit comme suit :

$$\ell_s(y, \hat{y}) = -\sum_{c=1}^{C} y_c \log(\sigma_{sg}(\mathbf{a}_c^{(L)}\mathbf{z}^{(L-1)} + b_c^{(L)}) + (1 - y_c) \log(1 - \sigma_{sg}(\mathbf{a}_c^{(L)}\mathbf{z}^{(L-1)} + b_c^{(L)})). \tag{6.7}$$

Enfin, si on définit $\Theta = \{\mathbf{A}^{(1)}, ..., \mathbf{A}^{(L)}\}$ comme l'ensemble des paramètres du réseau, le

problème d'optimisation associé à l'apprentissage du réseau devient :

$$\min_{\Theta} \sum_{i=1}^{N} \ell_s(y_i, \hat{y}_i, \Theta), \tag{6.8}$$

où l'on cherche les paramètres Θ minimisant le coût de classification associé au réseau.

6.2.4 Couches convolutives

Les couches convolutives [LeCun et al., 1998] ont été introduites dans l'objectif d'apprendre des représentations intermédiaires à partir d'images tout en répondant à certaines limitations des MLP. En appliquant des opérations de filtrage localement sur les observations, les CNN permettent d'apprendre et modéliser des caractéristiques locales des images. De plus, ces opérations de filtrage appliquée localement entraînent une forte réduction du nombre de paramètres à apprendre dans les modèles tout en permettant de prendre en compte certaines invariances par transformation des images. Pour les CNN, les observations sont dans la majorité des cas représentées par des matrices ou par une série de matrices. Sans perte de généralités, on considère $\mathbf{Z}^{(l)} \in \mathbb{R}^{F_l \times T_l \times d_l}$, la sortie de la couche convolutive l, comme une représentation temps-fréquence avec F_l bandes, T_l trames temporelles et d_l filtres ou canaux de sortie. On définit également $\mathbf{A}^{(l)} \in \mathbb{R}^{I \times J \times d_{l-1} \times d_l}$ les paramètres de la couche l où I et J sont les dimensions de la réponse impulsionnelle des filtres et d_l le nombre de sorties. Ainsi pour le canal d, la sortie de la couche l s'obtient en appliquant la succession des opérations d'activation et de convolution telle que :

$$\mathbf{Z}^{(l)}[:,:,d] = \sigma(\sum_{k=1}^{d_{l-1}} (\mathbf{A}^{(l)}[:,:,k,d] * \mathbf{Z}^{(l-1)}[:,:,k])), \tag{6.9}$$

où $*$ représente le produit de convolution 2D.

6.2.5 Couches récurrentes

Le dernière catégorie incontournable de réseaux de neurones sont les réseaux récurrents [Rumelhart et al., 1988; Elman, 1990]. Ils permettent de traiter des données séquentielles en modélisant les interactions temporelles entre les différents instants de la séquence d'entrée. Dans le cas des RNN, les observations ainsi que les représentations intermédiaires obtenues en sortie des couches récurrentes sont des séquences. Si on considère $\mathbf{z}^{(l)}[t]$ comme étant la sortie de la couche à l'indice temporel t, la sortie de la couche récurrente s'obtient alors par l'opération suivante :

$$\mathbf{z}^{(l)}[t] = \sigma(\mathbf{R}^{(l)}\mathbf{z}^{(l)}[t-1] + \mathbf{A}^{(l)}\mathbf{z}^{(l-1)}[t] + \mathbf{b}^{(l)}). \tag{6.10}$$

Ici, la matrice $\mathbf{R}^{(l)} \in \mathbb{R}^{d_l \times d_l}$ correspond aux poids récurrents qui s'appliquent à la sortie de la couche à l'instant précédent et la matrice $\mathbf{A}^{(l)} \in \mathbb{R}^{d_l \times d_{l-1}}$ correspond aux poids d'entrée s'appliquant au vecteur d'entrée de la couche récurrente.

Les couches récurrentes souffrent, lors de leur apprentissage, du problème bien connu de la disparation des gradients [Pascanu et al., 2013], de l'anglais *vanishing gradient*, empêchant l'apprentissage de dépendances temporelles longues par les modèles RNN classiques. Afin de s'affranchir de cette limitation, il a été proposé de remplacer les unités récurrentes par des unités plus complexes telles que les GRU [Cho et al., 2014], de l'anglais *gated recurrent units* ou les LSTM [Hochreiter et Schmidhuber, 1997], de l'anglais *long short-term memory*. Les unités GRU et LSTM permettent de modéliser efficacement des dépendances temporelles à beaucoup plus long terme par l'ajout de mécanismes de mémoire et d'oubli de l'information passée. Les unités GRU

et LSTM sont des cas particuliers plus avancés des unités récurrentes, nous choisissons de ne pas présenter leur fonctionnement. Toutefois, elles sont décrites et expliquées en détail dans les travaux de Graves et al. [2012] et dans l'article de Olah [2015].

6.3 NMF et MLP

Nous présentons dans cette section nos motivations pour l'utilisation de représentations NMF en entrée de réseaux de neurones profonds. Nous discuterons également de l'équivalence entre le modèle TNMF et un réseau à une couche cachée nous permettant d'introduire le modèle DNN-TNMF.

6.3.1 Améliorer le classifieur

Pour commencer, d'un point de vue purement applicatif, une première motivation pour l'utilisation de descripteurs NMF comme représentation d'entrée des réseaux de neurones provient de leur performance avec des classifieurs relativement simples. Nous avons montré dans les chapitres 4 et 5 que les modèles NMF étudiés fournissaient des représentations adaptées à l'analyse de sons environnementaux, en offrant de bonnes performances avec l'usage d'une régression logistique comme classifieur. Or, comme suggéré section 6.2.3, la dernière couche des réseaux classiques se comporte exactement comme une régression logistique. Dans ce cas, l'apprentissage d'un modèle MLP profond sur des descripteurs NMF correspondrait à l'ajout de couches cachées à la régression logistique dans nos systèmes présentés dans le chapitre 4.

De nombreux travaux du domaine ont suivi cette logique en se servant de MLP uniquement comme classifieur pour interpréter des descripteurs audio classiques. Certains travaux obtiennent des performances compétitives avec les meilleures approches CNN en apprenant simplement des MLP à partir de descripteurs cepstraux combinés à d'autres descripteurs bas-niveau [Marchi et al., 2016; Li et al., 2017; Park et al., 2016]. En revanche, ces approches ont le défaut de limiter les capacités d'apprentissage de représentations intermédiaires des modèles profonds. En effet, de tels descripteurs sont construits de sorte à caractériser certains aspects précis du signal et n'offrent pas les mêmes libertés d'apprentissage aux réseaux que les représentations temps-fréquence. Or, en se servant des activations NMF comme descripteurs, nous donnons au réseau simplement une projection des représentations temps-fréquence sur une nouvelle base définie par le dictionnaire. Cette projection a pour objectif de faciliter l'interpretabilité et la flexibilité de la représentation sans pour autant grandement réduire la quantité d'information qu'elle contient. Il semble alors raisonnable d'émettre l'hypothèse que remplacer les spectrogrammes par des NMF en entrée de MLP rendrait possible l'apprentissage de réseaux potentiellement plus simples et plus performants. Dans ce sens nous définissons le modèle "NMF + MLP", correspondant simplement à l'utilisation de descripteurs NMF en entrée de MLP standards.

6.3.2 NMF, couche MLP et pré-apprentissage

Si on reprend le fonctionnement des couches MLP présentées section 6.2.2, on peut exhiber des ressemblances entre ces couches et le fonctionnement de la NMF comme technique d'apprentissage de descripteurs. Nous rappelons que l'opération effectuée par la première couche cachée sur une observation $\mathbf{v} \in \mathbb{R}_+^F$ se définit dans le cas général par :

$$f_1(\mathbf{v}) = \sigma(\mathbf{A}^{(1)}, \mathbf{v}) \tag{6.11}$$

où σ est une fonction des observations et des paramètres de la couche cachée. Maintenant, considérons que les paramètres de la couche correspondent à une matrice à coefficients positifs $\mathbf{A}^{(1)} =$

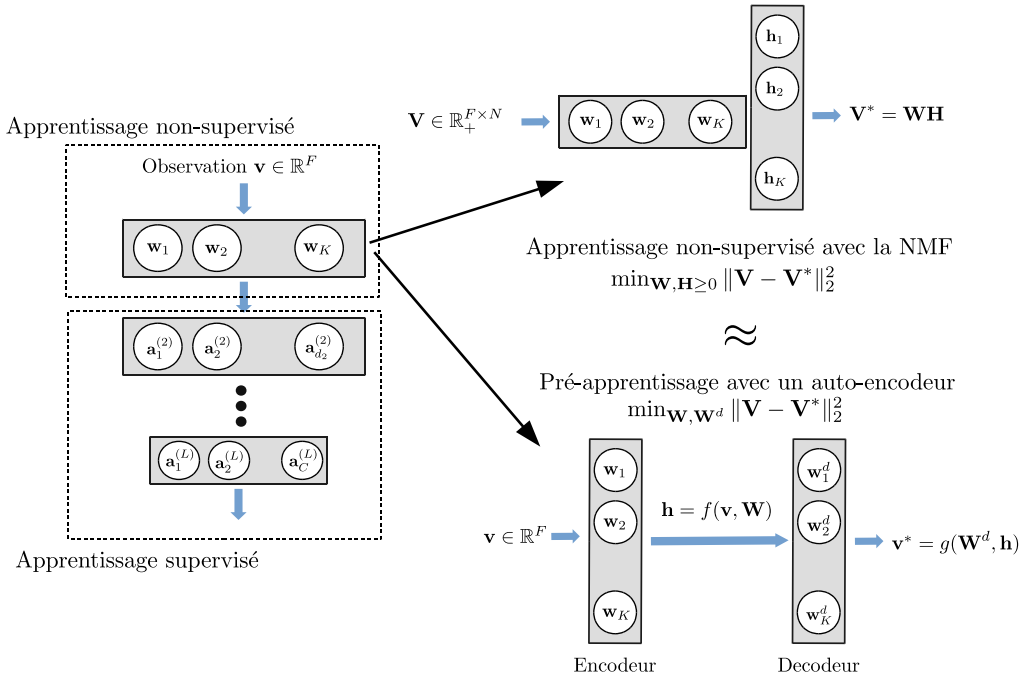


FIGURE 6.2 – Illustration du modèle NMF+MLP et de l'analogie avec le pré-apprentissage de couches.

$\mathbf{W} \in \mathbb{R}^{d_1 \times F}$ et que σ correspond à la fonction de projection positive optimale sur $\mathbf{W}$ définie dans le chapitre précédent :

$$\sigma(\mathbf{w}, \mathbf{v}) = \mathbf{h}^\star(\mathbf{W}, \mathbf{v}) = \min_{\mathbf{h} \in \mathbb{R}_+^{d_1}} D_\beta(\mathbf{v} \| \mathbf{Wh}) + \lambda_1 \|\mathbf{h}\|_1. \tag{6.12}$$

Dans ce cas, l'opération de projection sur le dictionnaire NMF remplace la première couche du réseau en fonctionnant avec le même objectif. Elle permet d'obtenir une représentation intermédiaire par le biais d'une opération non-linéaire fonction des observations et d'une matrice de poids. Cette matrice de poids fait partie des paramètres du modèle et est équivalente ici au dictionnaire. De plus, tout comme la fonction $\mathbf{h}^\star$, les activations de type ReLU fournissent naturellement des représentations positives et parcimonieuses dues à l'opération de seuillage.

En voyant l'étape de NMF comme la première couche du réseau, on peut montrer que la stratégie d'apprentissage du dictionnaire (donc des poids de la première couche) est analogue aux techniques de pré-apprentissage utilisées pour faciliter l'apprentissaget des MLP. Avant l'arrivée d'outils tels que les ReLU ou le *dropout*, il était souvent inévitable d'avoir recours à des stratégies de pré-apprentissage glouton de couches, de l'anglais *Greedy layer-wise pre-training*, pour apprendre des réseaux plus profonds [Hinton et al., 2006; Bengio et al., 2007]. Le principe de base du pré-apprentissage de couches est d'apprendre successivement des modèles de type auto-encodeurs afin de fournir une bonne initialisation aux paramètres des couches avant l'apprentissage supervisé du réseau. L'analogie entre la NMF et le pré-apprentissage de couches cachées est illustrée figure 6.2.

Les auto-encodeurs sont des réseaux de neurones non-supervisés qui s'apprenent à partir d'un critère d'attache aux données. La forme la plus simple d'auto-encodeur se construit à partir de couches MLP standards. La première couche joue le rôle de l'encodeur et se définit par la fonction g_e et ses paramètres $\mathbf{W}_e$. L'encodeur fournit en sortie le "code", donné par la représentation latente $\mathbf{z} = g_e(\mathbf{v}, \mathbf{W}_e)$, souvent de dimension inférieure à celle des observations. De la même manière, la deuxième couche correspond au décodeur, par sa fonction g_d et ses paramètres $\mathbf{W}_d$. Son rôle est de

fournir en sortie, une estimation $\mathbf{v}^\star = g_d(\mathbf{z}, \mathbf{W}_d)$ de l'observation présentée en entrée du décodeur. Les paramètres des auto-encodeurs sont souvent estimés en minimisant une erreur quadratique entre l'observation $\mathbf{v}$ et son estimation $\mathbf{v}$:

$$\min_{\mathbf{W}_e, \mathbf{W}_d} \|\mathbf{V} - \mathbf{V}^\star\|_2^2 = \min_{\mathbf{W}_e, \mathbf{W}_d} \|\mathbf{V} - g_d(g_e(\mathbf{V}, \mathbf{W}_e), \mathbf{W}_d)\|_2^2. \tag{6.13}$$

Lors de la première étape du pré-apprentissage de couches, une fois l'auto-encodeur appris, les paramètres de l'encodeur sont transférés en tant qu'initialisation des paramètres de la première couche du réseau supervisé. Maintenant, si on définit la première couche du réseau comme étant une projection NMF, alors l'apprentissage des paramètres de la première couche (donc du dictionnaire) se fait dans la même logique. En effet, on apprend le dictionnaire en minimisant un coût de reconstruction entre les observations et leurs approximations par NMF. Ainsi dans les deux cas, on apprend les poids de la couche avec un critère non-supervisé dans le but d'augmenter la capacité de généralisation des réseaux et de faciliter leur apprentissage. La NMF nous permet alors de remplacer le rôle de la première couche par une projection sur un dictionnaire à coefficients positifs. Plus spécifiquement pour le traitement de l'audio, remplacer la première couche par cette étape de projection sur le dictionnaire d'événements de base nous permet de garder tous les avantages que possède la NMF pour caractériser les environnements multi-sources tout en restant dans un cadre compatible avec l'apprentissage de réseaux de neurones profonds.

Récemment, quelques travaux ont également cherché à exhiber des liens entre la NMF et les réseaux de neurones profonds ou encore à introduire la notion de profondeur dans des modèles NMF. Par exemple, Smaragdis et Venkataramani [2017] ont montré que, sous certaines conditions, les auto-encodeurs se comportent comme une NMF pour la séparation de sources. Ils proposent d'utiliser des activations de type *soft-plus* ($\log(1 + x)$) à la sortie du décodeur et de l'encodeur afin de garantir la positivité des représentations latentes ainsi que celle de l'estimé des observations. En revanche, le dictionnaire, ou paramètres de l'encodeur, n'est pas contraint à être à coefficients positifs. Les auto-encodeurs positifs sont ensuite appris en minimisant la divergence KL entre les observations et leurs approximations en sortie. Les auteurs montrent que ces modèles se comportent de manière similaire à la NMF pour des applications de séparation de sources. Les types de composantes de base et d'activations obtenues sont alors fortement ressemblantes à celles pouvant être apprises par NMF. Ces modèles ont également été étendus afin de rendre un fonctionnement similaire à la NMF convolutive possible [Venkataramani et al., 2017]. La différence majeure avec nos approches est que les auteurs remplacent la NMF par des réseaux de neurones plus classiques. Au contraire, nous remplaçons simplement la première couche du réseau par une NMF non-supervisée. Dans notre cas, il nous semble important de garder la contrainte de positivité du dictionnaire ainsi que les activations par projection sur ce dictionnaire. L'interprétation comme somme d'événements de base pour l'analyse de scènes sonores qu'offre la NMF est conservée, ce qui ne serait pas le cas en remplaçant l'étape de projection par un simple produit matriciel tel qu'effectué dans Venkataramani et al. [2017].

L'autre catégorie d'approches notable est celle des NMF profondes par dépliement, de l'anglais *Deep unfolding NMF*. Ces méthodes sont basées sur l'idée du *deep unfolding* [Hershey et al., 2014]. Pour des problèmes d'optimisation se résolvant avec des algorithmes itératifs, le *deep unfolding* cherche à construire un réseau à partir des différentes itérations de cet algorithme. Un premier exemple donné par Hershey et al. [2014] est le dépliement de l'algorithme NMF par mises à jour multiplicatives. Dans ce cas un réseau est construit, où le paramètre de chaque couche correspond à un dictionnaire différent. Les observations sont connectées directement à chacune des couches et les activations se font avec l'étape de mises à jour multiplicatives de $\mathbf{H}$ [Wisdom et al., 2016; Le Roux et al., 2015a]. D'autres variantes ont été proposées, notamment une variante construisant une NMF récurrente en dépliement des algorithmes de seuillage itératifs pour le codage parcimonieux comme FISTA, de l'anglais *Fast Iterative Soft-Thresholding* [Wisdom et al., 2017a,b].

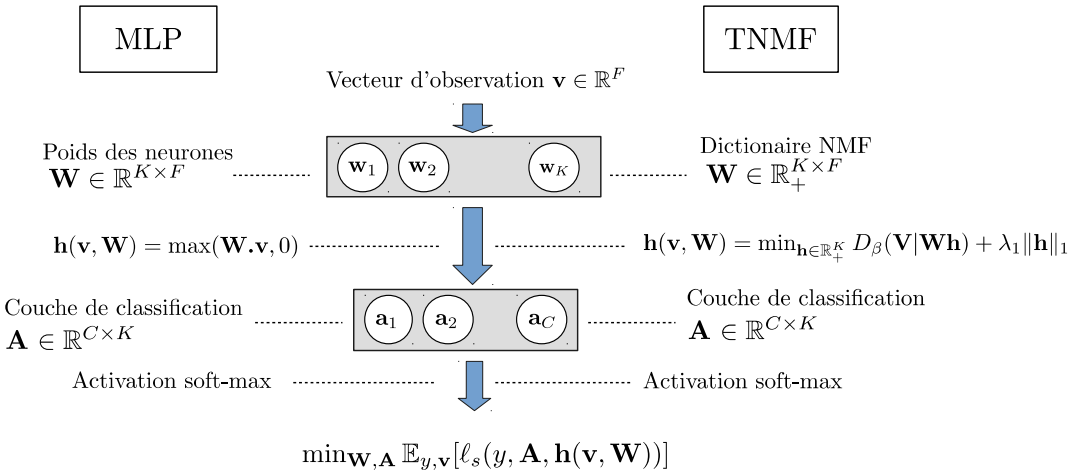


FIGURE 6.3 – Illustration de l'équivalence entre TNMF et un MLP à une couche cachée.

Contrairement aux NMF classiques, les approches deep NMF apprennent autant de dictionnaires que de couches dans le modèle de dépliement. De plus les activations ne sont pas obtenues par la projection sur un dictionnaire mais par la propagation dans les couches du modèle, le rendant ainsi moins facilement interprétable qu'une NMF classique. De plus, ces projections ne sont qu'une estimation des projections NMF optimales car elles sont simplement obtenues par une seule itération des algorithmes MU ou FISTA. Ces modèles, bien que prometteurs, ont uniquement été proposés pour certains cas de séparation de sources et de reconstruction d'image [Wisdom et al., 2017a]. Il reste à savoir si ils peuvent être efficaces dans le rôle d'apprentissage de représentations pour des tâches de classification.

6.3.3 TNMF comme un MLP à une couche cachée

Nous avons discuté de la pertinence de se servir de la NMF non-supervisée comme entrée de MLP. Si on peut apprendre et fixer la première couche comme étant une NMF non-supervisée, il parait naturel d'étudier l'intérêt de réaliser l'apprentissage du dictionnaire par TNMF. Nous montrons ici, en reprenant les notions introduites précédemment, que le modèle TNMF présenté au chapitre 5 est équivalent à un MLP à une couche cachée. Pour commencer nous rappelons que le problème TNMF se définit comme un problème d'optimisation à deux niveaux comme suit :

$$\begin{cases} \mathbf{h}^\star(\mathbf{v}, \mathbf{W}) = \min_{\mathbf{h} \in \mathbb{R}_+^K} D_\beta(\mathbf{V}\|\mathbf{W}\mathbf{h}) + \lambda_1 \|\mathbf{h}\|_1 + \frac{\lambda_2}{2} \|\mathbf{h}\|_2^2 \\ \\ \min_{\mathbf{W} \geq 0, \mathbf{A}} \mathbb{E}_{y,\mathbf{v}}[\ell_s(y, \mathbf{A}, \mathbf{h}^\star(\mathbf{v}, \mathbf{W}))] + \frac{\nu}{2} \|\mathbf{A}\|_2^2 \end{cases} \tag{6.14}$$

Afin de voir TNMF comme un MLP à une couche cachée, il suffit de définir $\mathbf{W}$ (le dictionnaire) comme les poids de la première couche et $\mathbf{A}$ les poids de la couche de classification. Dans ce cas le passage des observations par la couche cachée correspond à la projection $\mathbf{h}^\star(\mathbf{v}, \mathbf{W})$ de $\mathbf{v}$ sur le dictionnaire $\mathbf{W}$, où on considère donc $\mathbf{h}^\star$ comme la fonction d'activation en sortie de la première couche. La représentation intermédiaire $\mathbf{h}^\star(\mathbf{v}, \mathbf{W})$ passe ensuite dans la couche de classification équivalente à la régression logistique multinomiale utilisée dans TNMF. De plus, les paramètres

du modèle TNMF sont estimés et appris en minimisant un coût de classification supervisé équivalent au critère de *categorical cross-entropy* utilisé pour apprendre les réseaux. Ainsi, TNMF se définit comme une composition d'opérations non-linéaires avec une matrice de poids par couche. Le modèle TNMF partage donc avec un MLP à une couche cachée l'architecture, la couche de classification et le critère optimisé. Seuls deux aspects font la différence entre TNMF et les architectures MLP standards. Le premier est que les poids de la première couche sont contraints à être à coefficients positifs. Le second est la particularité de la fonction d'activation. Dans TNMF la fonction d'activation n'est pas un ReLU appliqué au produit scalaire entre les poids et les observations, mais la projection optimale de l'observation sur le dictionnaire défini par les poids des neurones de la première couche. Ce développement est illustré sur la figure 6.3, montrant la représentation de la TNMF comme un MLP.

En plus de NMF+MLP, nous proposons le système TNMF+MLP pour la classification de scènes. Dans une première étape, le modèle TNMF+MLP apprend un dictionnaire NMF discriminatif par le modèle TNMF. Une fois le dictionnaire appris, nous le fixons avant de projeter la matrice de données dessus en utilisant la fonction $\mathbf{h}^\star$ afin d'obtenir la représentation d'entrée du réseau. Nous apprenons ensuite un modèle MLP standard à partir des projections NMF. Nous détaillerons et évaluerons l'intérêt de l'approche TNMF+MLP dans la section 6.4 en la comparant dans un même temps aux modèles TNMF et NMF+MLP. L'objectif étant d'étudier à la fois si apprendre un réseau sur des dictionnaires discriminatifs améliore les performances de TNMF et si l'apprentissage supervisé de NMF permet une meilleure généralisation des réseaux lorsqu'il est utilisé comme première couche fixée de MLP.

6.3.4 DNN-TNMF : Vers un apprentissage conjoint du réseau et du dictionnaire

Les modèles combinant NMF et MLP présentés précédemment se définissent par deux problèmes d'optimisation traités séparément, celui de la NMF et celui de l'apprentissage du réseau. Dans la même idée que le modèle TNMF, on peut naturellement se demander si l'étape de NMF et celle de classification par MLP peuvent se regrouper dans un même problème d'optimisation. Dans ce cas, tous les paramètres de nos systèmes seraient appris conjointement selon le critère supervisé associé à la tâche traitée. De plus, nous avons discuté de l'équivalence entre TNMF et un MLP à une couche cachée. Dans ce contexte on peut s'interroger sur les conséquences d'ajouter des couches cachées au dessus de la couche NMF, afin d'ajouter de la profondeur dans le modèle TNMF. Ainsi, nous présentons dans cette section le modèle DNN-TNMF, extension de TNMF utilisant des réseaux de neurones profonds en guise de remplacement de la régression logistique multinomiale dans le rôle du classifieur. Sans perte de généralité, on se donne un réseau de neurones profonds à L couches ayant pour paramètre $\Theta = \{\mathbf{A}^{(1)}, ..., \mathbf{A}^{(L)}\}$ et ayant une couche MLP standard avec une activation *softmax* comme dernière couche de classification. On définit, tout comme précédemment, la fonction ℓ_s comme étant le coût multinomial ou *categorical cross-entropy*. Si on note $(f_1, ..., f_L)$ les opérations effectuées par chaque couche du réseau alors la propagation d'une observation $\mathbf{v}$ dans le réseau DNN-TNMF se définit par la fonction F comme suit :

$$F(\mathbf{v}, \Theta, \mathbf{W}) = (f_L \circ f_{L-1} \circ ... \circ f_1 \circ \mathbf{h}^\star)(\mathbf{v}). \tag{6.15}$$

Le problème MLP-TNMF se définit alors le problème d'optimisation suivant :

$$\begin{cases} \mathbf{h}^\star(\mathbf{v}, \mathbf{W}) = \min_{\mathbf{h} \in \mathbb{R}_+^K} D_\beta(\mathbf{V}|\mathbf{W}\mathbf{h}) + \lambda_1 \|\mathbf{h}\|_1 + \frac{\lambda_2}{2}\|\mathbf{h}\|_2^2 \\ \\ \min_{\mathbf{W} \geq 0, \Theta} \mathbb{E}_{y,\mathbf{v}}[\ell_s(y, F(\mathbf{v}, \Theta, \mathbf{W}))]. \end{cases} \tag{6.16}$$

Le modèle DNN-TNMF étant un cas particulier de réseaux de neurones, nous proposons dans

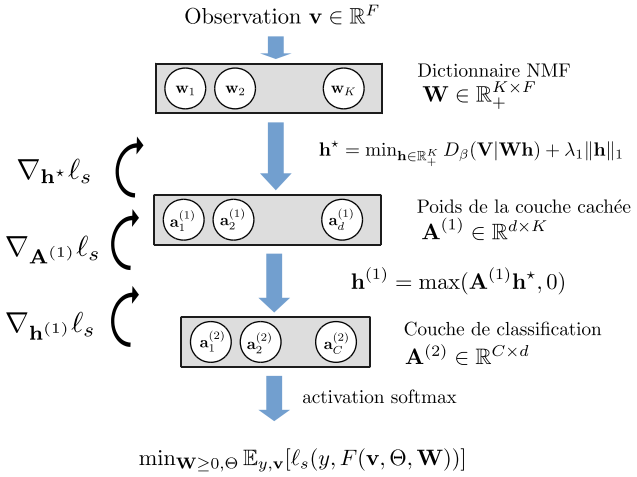


FIGURE 6.4 – Illustration du modèle DNN-TNMF et de la rétro-propagation des gradients.

cette première application du modèle d'utiliser une descente de gradient stochastique (SGD) classique afin d'estimer les paramètres du modèle. L'étape conditionnant notre capacité à apprendre le modèle par SGD est l'obtention de l'expression du gradient du coût de classification ℓ_s par rapport à toutes les variables. La couche NMF étant à l'entrée du réseau, le calcul des gradients de ℓ_s par rapport aux poids $\{\mathbf{A}^{(1)}, ..., \mathbf{A}^{(L)}\}$ ne pose pas de problèmes particuliers et se fait par rétro-propagation standard des gradients dans le réseau. L'algorithme de rétro-propagation des gradients [Rumelhart et al., 1988], de l'anglais *back-propagation*, est à la base de la plupart des algorithmes d'apprentissage de réseaux de neurones profonds. Le seul gradient qui diffère des approches classiques est celui par rapport au dictionnaire $\nabla_{\mathbf{W}} \ell_s$.

Comme DNN-TNMF n'est qu'une extension du modèle TNMF, il suffit simplement de reprendre les résultats présentés chapitre 5 et dans les travaux de Mairal et al. [2012] et Sprechmann et al. [2014]. L'expression de $\nabla_{\mathbf{W}} \ell_s$ est donnée par la proposition 5.2.1. L'interface entre le gradient par rapport au dictionnaire et le réseau se fait par le calcul de $\nabla_{\mathbf{h}^\star} \ell_s$. Cette grandeur correspond au gradient du coût d'un réseau de neurones standard par rapport à son entrée. Si elle n'est pas explicitement calculée dans les algorithmes de rétro-propagation usuels, son calcul est identique à celui du gradient du coût du réseau par rapport à la sortie d'une couche cachée. En effet, la sortie d'une couche est aussi l'entrée de la couche suivante. Or, le calcul du gradient par rapport à la sortie d'une couche est une des étapes nécessaires pour obtenir le gradient par rapport aux poids de cette même couche. Plus d'informations sur les détails des calculs qu'implique l'étape de rétro-propagation des gradients des réseaux standards sont données dans les travaux de Goodfellow et al. [2016] et Rumelhart et al. [1988]. L'architecture du modèle DNN-TNMF est illustrée sur la figure 6.4.

Une fois tous les gradients calculés, l'ensemble des poids (ainsi que le dictionnaire), sont mis à jour par une étape de descente de gradient. Afin de rester dans le cadre de la NMF, nous gardons cependant l'étape de projection du dictionnaire sur l'ensemble $\mathcal{W}_+$ des matrices à coefficients positifs et aux colonnes de norme ℓ_2 unitaire. L'algorithme par descente de gradient stochastique pour DNN-TNMF est donnée dans l'algorithme 6.1. Comme seule particularité de cet algorithme SGD, nous introduisons deux pas du gradient différents ρ_W et ρ_A pour la mise à jour du dictionnaire et des poids des couches standards respectivement. Cette séparation est motivée par le fait que les activations pour la couche NMF et le restant du réseau sont de nature très différentes. L'heuristique de décroissance du pas du gradient au cours des itérations est la même que celle utilisée par Mairal et al. [2012].

Il est important de noter que le modèle DNN-TNMF ne suppose aucune forme particulière

Algorithme 6.1 : Descente de gradient stochastique pour le modèle DNN-TNMF.

prend en entrée $\lambda_1, \lambda_2, \nu, \mathbf{W} \in \mathcal{W}, \Theta, T(\text{nombre d'itérations}), t_o, \rho$
  pour $t = 1$ à T **faire**
    Tirage d'un couple $(\mathbf{v}, y)$ dans l'ensemble d'apprentissage
    Calculer $\mathbf{h}^\star$ solution de (5.2)
    # Rétro-propagation des gradients pour obtenir
    $\nabla_W \ell_s(y, F(\mathbf{v}, \mathbf{W}, \Theta))$ et $\nabla_{\mathbf{A}^{(l)}} \ell_s(y, F(\mathbf{v}, \mathbf{W}, \Theta)) \forall l \in \{1, .., L\}$
    # Mise à jour des poids des couches classiques
    $\mathbf{A}^{(l)} = \mathbf{A}^{(l)} - \rho_A \nabla_{\mathbf{A}^{(l)}} \ell_s(y, F(\mathbf{v}, \mathbf{W}, \Theta))$
    $\rho_W = \min(\rho_W, \rho_W \frac{t_0}{t})$
    # Mise à jour du dictionnaire par gradient projeté
    $\mathbf{W} \leftarrow \Pi_{\mathcal{W}_+}[\mathbf{W} - \rho_W \nabla_{\mathbf{W}} \ell_s(y, F(\mathbf{v}, \mathbf{W}, \Theta))]$
  fin pour

pour les couches du réseau. On peut donc tout à fait le construire avec des couches cachées récurrentes ou convolutives. Cependant, comme première application du modèle, nous l'évaluerons avec des couches MLP classiques pour la classification de scènes dans la section 6.6.

6.4 Approches pour la classification de scènes

Nous évaluons dans cette section nos premiers systèmes par réseaux de neurones profonds utilisant les représentations NMF comme entrée pour la classification de scènes. Nous proposons l'évaluation des deux systèmes NMF + MLP et TNMF + MLP introduits section 6.3. Nous reprenons exactement là où nous nous sommes arrêtés dans le chapitre précédent. C'est-à-dire que nous gardons exactement le même protocole expérimental en changeant uniquement l'étape de classification. De même, nous évaluerons nos méthodes sur les ensembles de développement DCASE 2016 et DCASE 2017 avant de les comparer à d'autres approches de l'état de l'art sur l'ensemble d'évaluation DCASE 2016.

6.4.1 Protocole expérimental

Représentations temps-fréquence Nous gardons les CQT et la construction de la matrice de données $\mathbf{V}$ par TM tels que présentés au chapitre 5. Nous évaluons également le modèle CQT+MLP, utilisant la représentation CQT comme entrée du réseau afin de comparer l'impact du choix de la représentation d'entrée. Ce modèle nous sert de système de référence ayant la même représentation d'entrée que nos approches par NMF.

Réglages de la NMF et de TNMF Nous reprenons également le même protocole pour l'apprentissage des dictionnaires par NMF et TNMF. Ici, nous utilisons TNMF simplement comme un outil d'apprentissage de dictionnaire supervisé. C'est-à-dire qu'une fois le modèle TNMF appris, nous isolons le dictionnaire et projetons les données dessus afin d'obtenir la représentation d'entrée de TNMF+MLP.

Architecture des réseaux Les réseaux de neurones que nous utilisons dans cette section sont des MLP standards. Les couches sont des couches complètement connectées avec des activations ReLU. Nous utilisons également la technique de *drop out* entre chaque couche du modèle avec une probabilité de 0.2 [Hinton et al., 2012]. L'opération de *drop-out* alterne l'annulation aléatoire de

	Dcase 2016			**Dcase2017**		
	Couches	Neurones	λ_1	Couches	Neurones	λ_1
CQT	3	256	-	3	512	-
NMF $K = 256$	2	256	0.2	2	256	0.2
NMF $K = 512$	2	256	0.2	3	256	0.1
NMF $K = 1024$	2	512	0.1	3	512	0.1

TABLEAU 6.1 – Résultats de la recherche d'architectures des réseaux de neurones pour la classification de scènes.

certaines connections du réseau durant son apprentissage. Cette stratégie est réputée pour augmenter la capacité de généralisation des réseaux. L'activation de la dernière couche de classification est un *softmax* et la fonction de coût pour l'apprentissage des paramètres du réseau est la *categorical cross-entropy*. Dans tous les cas, les réseaux sont appris avec une descente de gradient stochastique sur 50 époques avec la librairie *keras* [Chollet, 2015]. Bien qu'il existe de nombreuses alternatives aux approches SGD pour apprendre les réseaux de neurones, nous n'avons pas remarqué de différences notables de performance et de temps de calcul en utilisant d'autres algorithmes. Nous détaillerons le choix du nombre de couches et d'unités par couche dans la section suivante.

Nous rappelons que du fait de la construction de $\mathbf{V}$ par tranches moyennées, nous avons 10 vecteurs de descripteurs par exemple. Pour l'apprentissage du réseau, nous considérons chacun de ces vecteurs comme un exemple séparé, c'est-à-dire que nous n'introduisons pas d'étape d'intégration temporelle avant l'entrée dans le réseau. Pour la prédiction, nous moyennons les log-probabilités de chaque vecteur afin de prendre la décision sur l'exemple entier.

6.4.2 Recherche de paramètres du réseau

Afin de choisir le nombre de couches et le nombre d'unités par couche du réseau, nous effectuons une recherche de paramètres en définissant des ensembles de développement à partir des ensembles d'apprentissage. Nous cherchons le meilleur ensemble de paramètres séparément sur les deux bases de données pour NMF + MLP ainsi que pour CQT+ MLP afin d'adapter la taille du réseau à la représentation d'entrée et à la base. Nous parcourons l'ensemble $\{1, 2, 3, 4\}$ pour le nombre de couches cachées et $\{128, 256, 512, 1024\}$ pour le nombre d'unités par couche ainsi que $\{0, 0.1, 0.2, 0.5\}$ pour la contrainte de parcimonie lors de la projection NMF.

Les meilleurs paramètres trouvés pour les cas étudiés sont présentés dans le tableau 6.1. On peut remarquer que pour les plus petits dictionnaires, l'utilisation de représentations NMF nécessite souvent une couche cachée de moins que pour la CQT afin d'obtenir les meilleures performances. En effet, la NMF jouant déjà un rôle d'apprentissage de représentation à partir des CQT, cela enlève une partie de ce travail aux couches cachées classiques du réseau. Cependant, comme on peut s'y attendre, lorsque nous augmentons la taille des dictionnaires, des couches cachées de plus grande taille sont nécessaires pour être capables de traiter efficacement la plus grande dimension des observations. Par exemple l'architecture des réseaux est la même pour CQT+MLP et NMF+MLP sur la base DCASE 2017 lorsque $K = 1024$.

6.4.3 Résultats sur les ensembles de développement

Les taux de reconnaissance obtenus par les systèmes proposés sur les ensembles de développement sont présentés dans le tableau 6.2. Nous comparons les systèmes NMF+MLP et TNMF+MLP aux autres approches proposées dans le chapitre précédent pour la classification de scènes. Afin

d'expliciter le classifieur utilisé pour chaque système, nous noterons NMF+LR et CQT+LR, ces mêmes représentations classifiées avec la régression logistique.

Ces résultats permettent d'apporter des premières réponses aux questions et aux hypothèses posées section 6.3. La première étant que la NMF supervisée peut fournir de meilleures représentations d'entrée que la CQT, même avec l'utilisation de réseaux de neurones profonds comme classifieur. Sur les deux bases de données, apprendre le MLP avec des descripteurs appris par NMF non-supervisée améliore les performances par rapport à directement apprendre à partir de la CQT. Il apparaît donc que la NMF nous amène à utiliser des réseaux moins profonds, tout en fournissant souvent de meilleurs résultats. Néanmoins, il convient de signaler que si les MLP appris avec la NMF nécessitent une couche en moins, les meilleures performances sont atteintes avec un dictionnaire de taille $K = 1024$ comparé aux 291 bandes de fréquences de la CQT. Ainsi, les réseaux ne possèdent pas forcément moins de paramètres à apprendre. Nous avons un ordre de grandeur de différence entre le nombre de paramètres pour les plus petits réseaux (1.3×10^5 paramètres pour NMF+MLP avec $K = 256$) et celui pour les plus larges (1×10^6 paramètres pour NMF+MLP avec $K = 1024$). Toutefois, même NMF+MLP avec $K = 256$ composantes est suffisant pour dépasser les performances de CQT+MLP.

De plus, on peut également noter les performances similaires entre SNMF+LR et CQT+MLP, montrant que la SNMF non-supervisée peut être tout aussi performante dans le rôle d'apprentissage de représentations intermédiaires qu'un MLP à plusieurs couches cachées. Les meilleurs systèmes NMF+MLP améliorent le taux de reconnaissance d'environ 3 points par rapport à SNMF+LR. Ce résultat confirme l'idée que les descripteurs NMF peuvent être mis en valeur par l'utilisation de meilleurs classifieurs que la régression logistique. Il montre également que si la NMF fournit déjà une représentation latente adaptée à la tâche, ajouter de la profondeur au modèle peut davantage faciliter son interprétation par les classifieurs. Enfin, on remarque qu'apprendre la NMF avec un critère supervisé en se servant de TNMF nous rapproche de la combinaison NMF+MLP. En effet, les différences de taux de reconnaissance entre TNMF et MLP+NMF ne sont de l'ordre que de 1 point. Ainsi, en apprenant la NMF conjointement avec la régression logistique par TNMF, nous obtenons des représentations presque aussi discriminantes que celles apprises par un MLP sur les descripteurs SNMF. En revanche, l'apprentissage de MLP sur les projections des dictionnaires appris par TNMF n'améliore les performances que pour la base DCASE 2016. Toutefois, comme nous l'avons observé avec TNMF dans le chapitre précédent, TNMF+MLP permet d'obtenir de bons résultats avec des dictionnaires de plus petite taille.

6.4.4 Résultats sur l'ensemble du challenge

Nous reprenons dans le tableau 6.3 la comparaison aux méthodes de l'état de l'art sur l'ensemble d'évaluation de la campagne DCASE 2016 entamée au chapitre 5. On peut commencer par remarquer les résultats relativement bons obtenus par CQT+MLP comparés aux systèmes soumis par les participants au challenge. En effet, cela placerait un tel système en cinquième position. Ce résultat montre que nous avons jusqu'ici fait des choix de représentation et de nature des réseaux raisonnables, en plus de fournir une comparaison crédible aux systèmes NMF+MLP. Ensuite on peut noter une hiérarchie similaire entre les méthodes proposées et celles sur les ensembles de développement. En effet, l'approche MLP+NMF améliore les performances par rapport à notre système par fusion de TNMF. Il est également intéressant de noter que MLP+NMF se place au dessus de toutes les méthodes par CNN seuls. Nous montrons ainsi, que pour une base de taille moyenne, l'interprétation des représentations temps-fréquence par des modèles NMF constitue une alternative efficace aux couches convolutives classiques.

Ensuite, sur l'ensemble de données considéré, reprendre les dictionnaires appris par TNMF augmente la capacité de généralisation des MLP appris. Le système TNMF+MLP obtient un taux

	DCASE 2016			DCASE 2017		
Référence		72.5			74.8	
CQT + LR		77.7			82.3	
CQT + MLP		82.8			84.2	
	K=256	K=512	K=1024	K=256	K=512	K=1024
SNMF + LR	81.2	82.6	83.1	79.3	83.1	84.4
TNMF + LR	85.0	84.8	84.5	85.0	86.3	85.9
NMF MLP	**85.6**	**85.7**	**86.2**	84.8	86.3	**87.6**
TNMF MLP	**85.6**	**87.1**	**86.1**	85.8	**87.1**	86.1

TABLEAU 6.2 – Taux de reconnaissance des systèmes NMF et MLP pour différentes tailles de dictionnaires K comparés à SNMF et TNMF.

	Ensemble du challenge DCASE 2016			
	Entrée	Classifieur	Taux de reco.	Rang/49
[Mesaros et al., 2016b]	Mel Spectres	GMM	77.2	39
[Valenti et al., 2016]	Mel Spectres	CNN	86.2	7
[Marchi et al., 2016]	Descripteurs audios	Fusion de MLP	86.4	5
[Park et al., 2016]	Coeffs. cepstraux	MLP + GMM	**87.2**	4
[Eghbal-Zadeh et al., 2016]	MFCC + Spectres	CNN + I-vector	**89.7**	1
Notre soumission	CQT	Fusion TNMF	**87.7**	3
Améliorations récentes	CQT	MLP	86.7	
Améliorations récentes	NMF	MLP	**88.5**	
Améliorations récentes	TNMF	MLP	**90.5**	

TABLEAU 6.3 – Taux de reconnaissance et rangs des systèmes les mieux classés au challenge DCASE 2016.

de reconnaissance de 90.5% se plaçant légèrement au dessus du système proposé par Eghbal-Zadeh et al. [2016] par fusion de CNN et I-Vecteurs. En comparaison aux autres approches proposées pour le DCASE 2016, nos systèmes sont les seuls à considérer l'utilisation d'apprentissage de descripteurs pour extraire les représentations d'entrée de réseaux de neurones. L'idée étant de profiter de l'interprétabilité de la NMF pour extraire des caractéristiques à partir d'environnements multi-sources. Sans utiliser de stratégies d'augmentation ni de fusion de classifieurs, nous avons montré qu'avec la représentation d'entrée appropriée, des modèles profonds relativement simples peuvent être compétitifs avec des systèmes bien plus avancés en termes de taille de modèles ou de nombre d'approches fusionnées.

6.4.5 A propos du challenge DCASE 2017

Nous avons également soumis un système similaire aux approches proposées pour l'édition 2017 de la campagne DCASE 2017. Ce système possède plusieurs différences par rapport au protocole expérimental mis en place dans ce chapitre, à la fois en termes de réglage des CQT, de la construction de $\mathbf{V}$, de la NMF et du choix des réseaux. De plus, par ses améliorations limitées et par soucis de cohérence, nous ne le présenterons pas en détails. Nous en présentons uniquement les grandes lignes. Notre système part du modèle NMF+MLP en y ajoutant une nouvelle branche dans le réseau prenant également en compte la présentation CQT. Nous avons donc un réseau prenant deux représentations différentes en entrée. Ces deux représentations sont passées dans deux branches distinctes du réseau avant d'être regroupées par concaténation pour les dernières couches

du réseau. Ce modèle a été proposé dans l'idée d'apprendre à partir des CQT, l'information discriminative que n'a pas encodé la NMF. Cette approche nous a permis d'obtenir des améliorations d'environ 1 point sur l'ensemble de développement par rapport aux systèmes NMF+MLP. De plus, pour le système final, nous avons également considéré les représentations pour les deux canaux audio. Enfin, le réseau appris est combiné par fusion tardive au modèle TNMF de sorte à construire notre soumission finale.

Le système proposé par notre équipe nous permet de nous classer à la dixième position sur 39 autres équipes. [1] La différence principale avec l'édition 2016 du challenge est le nombre de nouvelles idées, principalement venues de la recherche en apprentissage profond. Si certaines de ces idées ne se sont pas avérées particulièrement efficaces, les premiers systèmes doivent leur succès à la proposition d'approches originales pour la tâche. L'aspect que l'on retrouve chez la plupart des systèmes les mieux classés est l'idée de multiplier le nombre de représentations utilisées en entrée de CNN ou d'augmenter artificiellement la taille de la base de développement. Cela s'est fait par exemple par des augmentations sur le signal d'entrée en changeant la hauteur ou en étirant le signal [Lehner et al., 2017]. Une autre approche est d'augmenter les spectres Mel par de multiples représentations binaurales ou issues de séparation de sources, afin d'augmenter la variété des représentations d'entrée [Y.Hang et Park, 2017]. L'équipe s'étant classée en première position propose de générer de nouveaux exemples en se basant sur des GAN [Mun et al., 2017]. Ces différentes approches ont l'avantage de varier les natures des représentations fournies au réseau, augmentant leur chance d'identifier l'information nécessaire à la discrimination de certaines scènes. De plus, elles permettent d'augmenter artificiellement le nombre d'exemples de la base de données, rendant possible l'apprentissage de réseaux plus complexes et plus profonds en évitant le sur-apprentissage. Ainsi, en comparaison à l'édition 2016, la généralisation de l'utilisation de telles astuces a permis aux participants de s'affranchir de certains inconvénients des modèles CNN. Les techniques mentionnées sont parfaitement compatibles avec nos approches et leur utilisation pourrait également améliorer les performances de nos systèmes. Il sera intéressant d'étudier par la suite comment réaliser des augmentations intelligemment ou de varier les représentations sur lesquelles nous apprenons les NMF afin de rendre nos systèmes plus robustes.

6.5 NMF, CNN et RNN pour la détection d'événements

Dans la section précédente, nous avons évalué nos systèmes par NMF et réseaux de neurones pour la classification de scènes en nous limitant à l'utilisation de MLP. Pour des tâches de détection d'événements les MLP se sont montrés plus fortement limités par rapport à d'autres types de réseaux de neurones profonds. A cause de la nature de la tâche, la modélisation du contexte temporel et la nécessité de prédire des séquences d'observations rendent l'utilisation de réseaux récurrents presque indispensable. Afin de mettre en valeur nos systèmes NMF+MLP, nous proposons d'étudier expérimentalement la compatibilité entre les descripteurs NMF et des couches récurrentes ou convolutives. Pour cela, nous introduisons le système NMF+CRNN reprenant l'idée des systèmes état de l'art en détection d'événements qui enchaînent des couches convolutives et récurrentes (CRNN) [Cakir et al., 2017; Xu et al., 2017c].

1. Le détail des résultats est disponible sur le site du challenge www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-acoustic-scene-classification-results

6.5.1 Protocole expérimental

Base de données et systèmes de référence

Nous proposons d'effectuer l'évaluation sur la base de détection d'événements synthétiques TUT SED 2016 également utilisée dans les chapitres 4 et 5 [Cakir et al., 2017]. Cette base de données possède deux principaux points fort. Elle contient suffisamment d'exemples pour apprendre des modèles plus profonds sans être trop pénalisés par le sur-apprentissage. De plus, la base est associée à des ensembles d'apprentissage, de développement et de test clairement définis, rendant les comparaisons avec l'état de l'art plus justes.

Dans leurs travaux, Cakir et al. [2017] proposent une comparaison rigoureuse de plusieurs familles d'approches par réseaux de neurones appliquées sur la base TUT SED synth. Cette comparaison inclut des modèles MLP, CNN, RNN et CRNN. L'évaluation de ces différentes approches nous permettra de comparer l'apport de la NMF en fonction du type d'architecture utilisé.

Représentation temps-fréquence et NMF

Nous gardons des spectres Mel à 80 bandes comme représentation temps-fréquence avec des fenêtres de 40 ms tels qu'utilisés aux chapitres 4 et 5. Nous gardons également la même stratégie d'apprentissage de descripteurs par NMF non-supervisée. Nous utilisons SNMF pour décomposer l'ensemble de la base d'apprentissage afin d'extraire la représentation d'entrée à partir des spectres Mel. Comme nous avons accès à un ensemble de développement clairement établi, nous sélectionnons le dictionnaire donnant le meilleur score F1 par trame avec une régression logistique simple parmi 10 initialisations aléatoires. Le nombre de composantes dépend du type de réseau utilisé et sera précisé lors de la présentation des choix des paramètres pour les différents modèles.

Réseaux de neurones

Fonction de coût et dernière couche La dernière couche des réseaux sera toujours une couche MLP classique avec autant de neurones que d'étiquettes, c'est-à-dire 16. La particularité du problème de détection d'événements est son aspect multi-labels. Dans ce cas, nous rappelons que l'activation de la dernière couche est une sigmoïde. Ce type d'activation permet d'avoir en sortie du réseau, un vecteur représentant la probabilité de présence de chaque classe dans une trame donnée. Contrairement aux activations *soft-max*, ces probabilités ne sont pas contraintes à être de somme unitaire à travers le vecteur de sortie. Cela autorise la détection simultanée de plusieurs étiquettes, permettant un apprentissage directement multi-labels des réseaux de neurones. Pour ce type de problème la fonction de coût est la cross-entropie binaire, rendant le problème équivalent à l'apprentissage joint d'un détecteur d'activations pour chaque classe.

MLP Mis à part la couche de classification, les MLP utilisés sont équivalents à ceux proposés pour la classification de scènes. Les paramètres du modèle sont alors le nombre de couches cachées et le nombre d'unités par couche. Afin de limiter l'espace des paramètres possibles, nous nous limiterons au cas de couches cachées de même taille. Tel qu'effectué par Cakir et al. [2017], les MLP sont appris trame par trame, sans forme de prise en compte du contexte temporel.

RNN Nous reprenons les architectures RNN suggérées dans Cakir et al. [2017]. Les premières couches du réseau sont des couches récurrentes avec des unités GRU. Ces couches prennent comme entrée une séquence correspondant à une tranche de spectrogramme ou d'activation NMF, selon la représentation choisie. Après être passé par les couches récurrentes, nous obtenons une séquence de même longueur où seule la dimension de la représentation change. Nous ajoutons

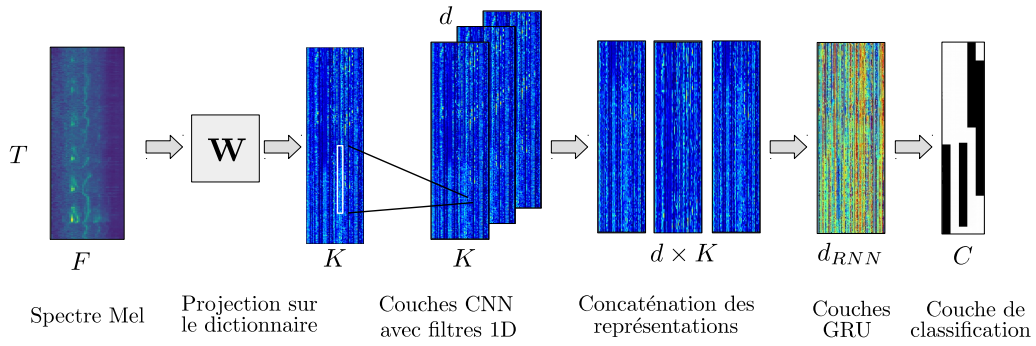


FIGURE 6.5 – Illustration du modèle NMF+CRNN pour la détection d'événements.

également quelques couches MLP classiques à la fin du réseau afin d'interpréter les représentations obtenues en sortie des couches récurrentes. De plus, les poids des couches MLP sont partagés avec chaque instant des séquences de sortie des couches récurrentes, c'est-à-dire que la même opération est appliquée à chaque trame de la séquence. Ce type d'architecture permet aux RNN de directement prédire une séquence d'étiquettes associée à la séquence d'observations. Les couches récurrentes ont également la possibilité de traiter les séquences dans les deux sens, par l'utilisation de couches bi-directionnelles [Schuster et Paliwal, 1997]. Les données sont alors lues dans le sens du temps par une moitié des poids et dans le sens contraire par l'autre moitié des poids de la couche bi-directionnelle. Cette stratégie s'est montrée efficace pour le traitement de la parole [Mesnil et al., 2013] ou la détection d'événements sonores [Parascandolo et al., 2016]. Les paramètres à régler dans le cas des RNN sont le nombre de couches RNN, le nombre de poids par couche et le nombre de couches MLP en fin de réseau. Nous contraignons également les couches MLP et RNN à avoir le même nombre d'unités.

CRNN Le modèle CRNN, initialement proposé pour le traitement de la parole [Sainath et al., 2015], se définit par la succession d'une première partie contenant uniquement des couches convolutives et de *pooling* puis d'une deuxième partie contenant uniquement des couches récurrentes. Pour la détection d'événements, le modèle CRNN proposé par Cakir et al. [2017] enchaîne des couches convolutives et des couches de *pooling* sur l'axe des fréquences jusqu'à réduire la dimension de chaque canal à 1 tout en gardant l'axe temporel intact. Ainsi, à la sortie de la dernière couche de *pooling*, la représentation en entrée des couches récurrentes possède une dimension égale au nombre d'unités de la dernière couche convolutive.

La nature des couches convolutives diffère lorsque nous utilisons les activations NMF en entrée du CRNN. L'utilisation habituelle de filtres à 2 dimensions sur les représentations temps-fréquence permet d'extraire des caractéristiques locales à partir de l'image. Cette opération suppose qu'il existe une certaine dépendance entre les pixels voisins de l'image. En effet, pour un spectrogramme, deux points voisins dans l'espace temps-fréquence sont susceptibles d'avoir des valeurs assez proches car la représentation temps-fréquence d'un événement s'étale souvent sur les deux dimensions. En revanche, on ne retrouve pas ce genre de propriétés pour les séquences d'activations NMF. Contrairement aux ondes de Fourier, les composantes des dictionnaires NMF ne sont a priori pas ordonnées. Ainsi rien ne garantit qu'une composante donnée soit particulièrement corrélée avec ses composantes voisines dans le dictionnaire. Ce phénomène empêche l'utilisation de filtres 2D locaux pour traiter les séquences de descripteurs NMF comme une image

	Paramètre explorés	Meilleure valeur		
		MLP	RNN	CRNN
Nbre de couches MLP	$\{2, 3, 4, 5\}$	3	2	0
Nbre de couches convolutives	$\{1, 2, 3, 4\}$	-	-	3
Nbre de couches GRU	$\{1, 2, 3\}$	-	3	2
Nbre de Neurones MLP/RNN	$\{128, 256, 512, 1024\}$	256	512	512
Nbre de filtres CNN	$\{16, 32, 64\}$	-	-	32
Taille des filtres	$\{5, 10, 20\}$	-	-	10
Séquence d'entrée en trames	$\{128, 256, 512, 1024\}$	-	256	512
Nbre de composantes NMF	$\{32, 64, 128\}$	64	128	64

TABLEAU 6.4 – Recherche de paramètres pour les différents types de réseaux évalués sur la détection d'événements.

temps-fréquence. A la place, nous proposons d'utiliser des filtres à une dimension, appliquant la convolution uniquement sur l'axe temporel. Si la dernière couche convolutive contient d filtres, et l'entrée est une tranche d'activations NMF avec K composantes et T trames temporelles, alors la représentation en sortie de cette couche sera un tenseur dans $\mathbb{R}^{T\times K\times d}$. Enfin, en sortie de la dernière couche convolutive, nous aplatissons la dernière dimension de ce tenseur afin de fournir une représentation adéquate à l'entrée des couches récurrentes c'est-à-dire dans $\mathbb{R}^{T\times K.d}$. Une illustration de l'architecture des modèles NMF+CRNN proposés est donnée dans la figure 6.5. Les paramètres à régler pour le modèle NMF+CRNN sont : la taille des filtres, le nombre de couches convolutives, le nombre d'unités par couche convolutive, le nombre de couches récurrentes et le nombre d'unités par couche récurrente.

Apprentissage des réseaux

Généralités Tous les réseaux proposés sont appris avec la librairie *Keras* [Chollet, 2015]. Mis à part pour les couches récurrentes, nous utilisons des activations ReLU en sortie de chaque couche des différents modèles. De manière générale, nous fixons la probabilité de *dropout* à 0.25. Enfin, pour le CRNN nous introduisons une étape de *batch normalisation* entre chaque couche convolutive [Ioffe et Szegedy, 2015]. Il s'agit d'un outil fréquemment utilisé pour accélérer l'apprentissage des réseaux en normalisant de manière adaptative les représentations de sortie des couches pour chaque batch.

Dimension de l'entrée A la fois les RNN et CRNN prennent des séquences d'observation en entrée. Le choix de la longueur de cette séquence devient alors également un paramètre important à régler selon la nature des dépendances temporelles présentes dans les données traitées. En particulier, pour le traitement de la parole il est standard d'utiliser des séquences relativement courtes en entrée de RNN, de l'ordre de quelques dizaines de trames. Il a été montré par Cakir et al. [2017] que pour la détection d'événements, les modèles sont capables de tirer parti de séquences bien plus longues, de l'ordre de quelques secondes. Cette différence s'explique par la longueur moyenne des événements à détecter en fonction de la tâche. Pour un signal de parole, on cherche à identifier des événements souvent très courts. En revanche, de manière générale, le temps entre le début et la fin d'un événement sonore peut être relativement long (une voiture qui passe, un signal d'alarme...).

Taille des batchs et génération de séquences La nécessité de donner des séquences d'observations en entrée des modèles nous oblige à effectuer un découpage des données afin de construire l'ensemble des sous-séquences présentées à l'apprentissage des réseaux. Pour ne pas proposer

constamment les mêmes séquences, nous tirons les instants de début des séquences aléatoirement à chaque époque. Le nombre d'instants de début des séquences est choisi tel que, en moyenne, l'ensemble de la base soit représenté par des séquences se recouvrant à 25%. La taille des batchs est fixée à 64 trames pour les MLP, à 32 séquences pour les RNN et à 16 séquences pour les CRNN.

Algorithme et critères d'arrêt Nous apprenons les réseaux avec l'algorithme ADAM [Kingma et Ba, 2014], un des plus utilisés pour l'apprentissage de réseaux récurrents et convolutifs. Nous gardons les réglages par défaut de l'algorithme ADAM de la librairie *Keras*.

Les réseaux de neurones profonds souffrent souvent du problème de sur-apprentissage. Au cours de l'apprentissage du réseau, il est commun de voir l'erreur de généralisation croître à partir d'un certain nombre d'itérations, tandis que l'erreur d'apprentissage continue de décroître. Afin de traiter ce problème, il est commun de définir des critères d'arrêt, en surveillant l'évolution d'un critère sur un ensemble de développement pendant l'apprnetissage du modèle. Nous choisissons de nous appuyer sur l'évolution du score F1 par trame comme critère d'arrêt. Nous sauvegardons le score ainsi que les poids du modèle à chaque amélioration du score F1 sur l'ensemble de développement. Si le score ne s'est pas amélioré après 40 itérations, nous stoppons l'apprentissage et nous revenons au réseau ayant fourni la meilleure valeur.

6.5.2 Recherche de paramètres

L'ensemble des paramètres considérés ainsi que les meilleures valeurs trouvées sur l'ensemble de développement pour chaque modèle sont donnés dans le tableau 6.4. On peut noter que dans tous les cas la profondeur maximum des réseaux est de 5 couches cachées. Le nombre d'unités par couche MLP et récurrente semble lui aussi dans le même ordre de grandeur pour les trois modèles. Le nombre d'unités convolutives est relativement faible. Ceci est principalement dû à la configuration de notre réseau CRNN et à l'absence de *pooling*. La représentation en sortie des couches convolutives est de dimension proportionnelle au nombre d'unités de la dernière couche et de la dimension d'entrée. Le nombre de filtres est donc limité à 32, et est donc bien inférieur aux 512 filtres du CRNN proposés par Cakir et al. [2017].

6.5.3 Analyse des résultats

Les résultats finaux pour nos systèmes de détection d'événements sur la base TUT SED synth sont présentés dans le tableau 6.5. En plus des réseaux proposés et des systèmes de Cakir et al. [2017], nous incluons les systèmes par NMF avec la régression logistique introduits dans les chapitres précédents. Commençons par observer l'influence de la NMF sur le MLP simple. Les ER et scores F1 obtenus par NMF+MLP sont plus stables à travers les différentes métriques. Mis à part le score F1 par trame, où les résultats sont relativement proches, NMF+MLP affiche de meilleures performances comparé au MLP appris directement sur les spectres Mel. Ces premiers résultats vont dans le sens de nos observations et discussions pour la classification de scènes. Même sans contexte temporel, le système NMF+MLP s'approche des modèles CNN et RNN. Ensuite, si l'on ajoute des couches récurrentes avec NMF+RNN, on remarque également une amélioration des performances par rapport au RNN sur les spectres Mel [Cakir et al., 2017]. Nous avions uniquement discuté des similarités entre la NMF et les couches MLP standards. Toutefois, ce résultat suggère que les descripteurs NMF, en tant que représentation adaptée à la tâche, sont également compatibles avec l'utilisation de couches récurrentes et peuvent amener à l'augmentation des performances des RNN.

	Méthode	ER_{tr}	$F1_{tr}$	ER_{1sec}	$F1_{1sec}$
[Mesaros et al., 2016b]	GMM	0.78	40.5	0.72	45.3
Nous	NMF $\beta = 1$	0.72	40.8	0.74	47.0
Nous	TNMF $\beta = 1$	0.68	44.3	0.73	48.1
[Cakir et al., 2015b]	MLP	0.68	49.2	1.1	50.2
Nous	NMF+MLP	0.66	46.7	0.64	53.0
[Cakir et al., 2017]	RNN	0.6	52.8	0.64	57.1
Nous	NMF + RNN	0.59	59.6	0.59	61.7
[Cakir et al., 2017]	CNN	0.56	59.8	0.78	59.9
[Cakir et al., 2017]	CRNN	**0.48**	**66.4**	**0.47**	**68.7**
Nous	NMF + CRNN	**0.47**	**67.4**	**0.46**	**69.5**

TABLEAU 6.5 – Taux d'erreur et score F1 sur les deux environnements de la base TUT SED synth 2016 pour les systèmes NMF et TNMF proposés comparés aux systèmes de l'état de l'art.

TABLEAU 6.6 – Scores F1 par classe sur la base TUT SED synth 2016.

Durée moy. (s)	Label	MLP NMF	RNN Mel	RNN NMF	CRNN Mel	CRNN NMF
1,2 s	verre qui se brise	69	48	49	54	**80**
1,7 s	arme à feu	18	64	49	73	**75**
2,1 s	miaulement	29	29	30	42	**74**
5,0 s	aboiement	48	51	67	73	**80**
5,9 s	tonnerre	43	46	48	**63**	62
6,1 s	oiseaux	24	41	49	53	**55**
6,4 s	cheveux	16	39	39	**45**	**45**
6,9 s	bébé	52	46	42	**59**	54
7,0 s	moto	45	44	46	47	**52**
7,1 s	bruits de pas	15	34	37	**47**	45
7,3 s	applaudissement	54	57	70	**71**	68
7,8 s	bus	43	55	63	66	**72**
7,9 s	mixeur	36	57	67	**82**	70
8,1 s	supporters	68	64	79	77	**78**
8,2 s	sirènes	49	50	77	66	**81**
8,2 s	pluie	61	59	60	72	**75**
	Moyenne	42	49	55	62	**66**

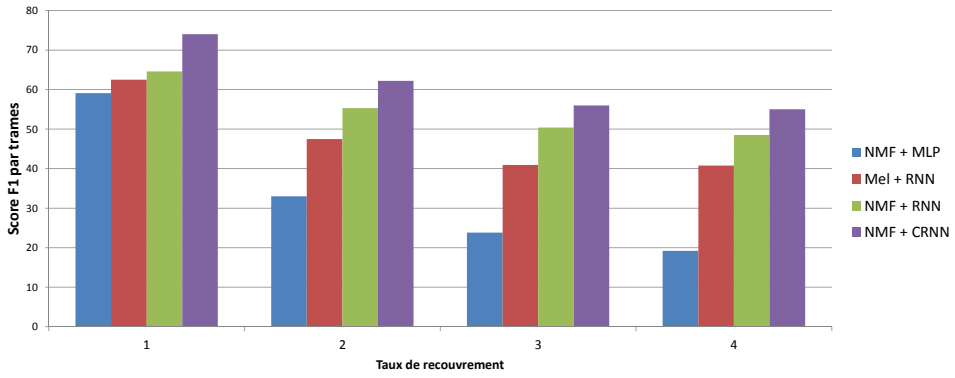


FIGURE 6.6 – Scores F1 en fonction du taux de recouvrement sur la base TUT SED synth.

Le système NMF+RNN est le premier système par NMF modélisant l'évolution temporelle des projections que nous proposons. Cependant, un écart important persiste entre les performances des modèles NMF+RNN et le CRNN sur la représentation d'entrée, montrant que les couches récurrentes ne suffisent pas à elles seules à tirer pleinement parti des représentations NMF. Enfin, nous terminons en observant les performances du modèle NMF+CRNN proposé. Ce dernier améliore les performances d'environ 1 point sur l'ensemble de métriques par rapport au CRNN sur les spectres Mel. Il convient de noter que l'apport de la NMF est moins remarquable dans le cas des CRNN, car le modèle de classification en lui-même répond en grande partie aux attentes d'apprentissage de représentations et de modélisation temporelle de la tâche. En revanche, cela montre bien que le rôle d'apprentissage de représentations uniquement à partir de l'axe des fréquences peut être tout aussi bien réalisé par la NMF. En effet, les couches convolutives dans NMF+CRNN ont uniquement pour rôle de représenter les variations temporelles locales de l'information par l'usage de filtres 1D.

Performances par classe Nous présentons les scores F1 par classe dans la tableau 6.6 pour les modèles RNN+NMF et NMF+CRNN. De manière générale, on peut noter que les différences de score entre les classes sont davantage dépendantes du modèle de réseau utilisé que de la représentation d'entrée. En effet, changer de représentation d'entrée sans changer la nature du réseau, permet d'améliorer les performances de manière relativement homogène à travers l'ensemble des catégories. Cependant, il semble difficile d'isoler des points communs entre les catégories d'événements sur lesquelles les modèles par NMF améliorent les performances. En effet, les plus gros écarts de scores F1 se trouve sur des catégories telles que *sirènes*, *bus* ou *vitre qui se brise*, donc des événements de natures et de structures temps-fréquence très différentes. Ces résultats suggèrent que les modèles par NMF nous permettent d'obtenir des systèmes plus robustes, en traitant la tâche dans son ensemble, sans favoriser la détection d'événements particuliers.

Des différences plus marquées peuvent être observées entre les différentes catégories de réseaux utilisées. Les modèles RNN sont particulièrement performants sur les événements répétitifs étalés sur un temps plus long tels que *sirènes*, *applaudissement* ou *mixeur*. La modélisation des dépendances temporelles à long terme que permettent les unités GRU facilite la détection de tels événements, surtout en comparaison aux modèles MLP. Les couches convolutives du modèle NMF+CRNN semblent être appropriées pour modéliser les évolutions temporelles plus courtes de l'information. En effet, NMF+CRNN permet de fortes augmentations des scores F1 sur des événements plus courts tels que *verre qui se brise*, *arme à feu* ou *aboiement*. Ces événements étant de nature plutôt impulsionnelle, la majorité de l'information les caractérisant se concentre sur quelques trames, ce que permettent de modéliser les filtres 1D sur le temps des couches convo-

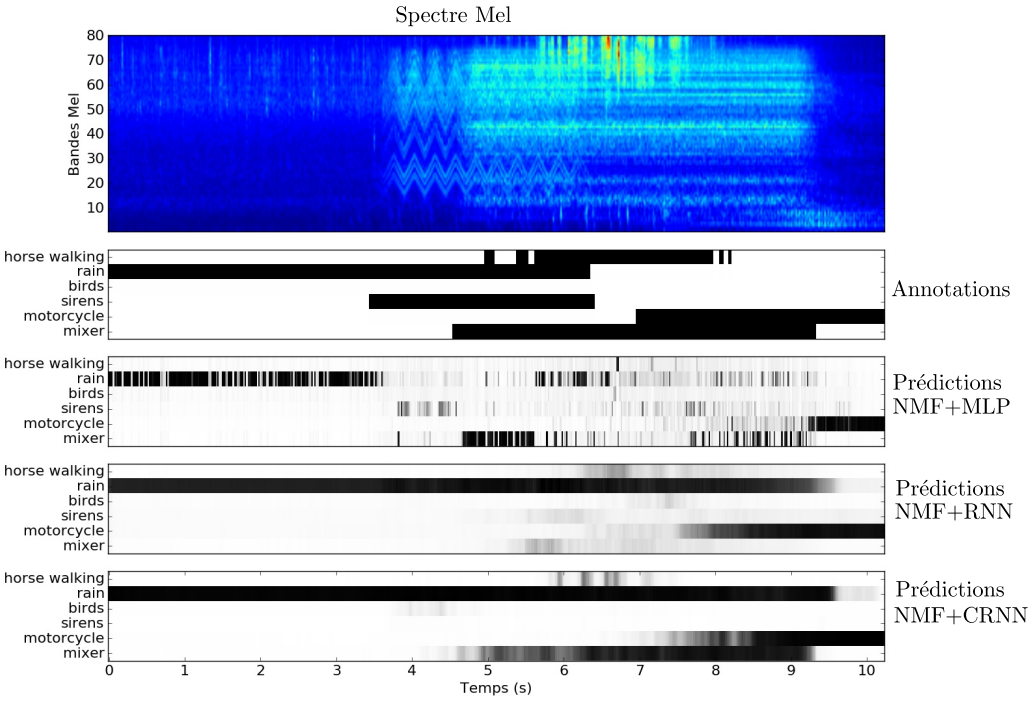


FIGURE 6.7 – Illustration des prédictions obtenues pour différents modèles par NMF+DNN sur une séquence d'observation de la base TUT SED synth.

lutives de NMF+CRNN.

Nous terminons cette analyse des résultats en présentant les performances des différents modèles de détection en fonction du degré de recouvrement sur la figure 6.6. Les scores pour un taux de recouvrement donné sont obtenus en gardant pour l'évaluation seulement les trames contenant ce même taux de recouvrement dans l'annotation. Par exemple, pour un taux de 3, nous retirons toutes les trames contenant plus ou moins de 3 événements dans l'annotation. La figure 6.6 nous permet d'observer le comportement des différents modèles lorsque le taux de recouvrement augmente, c'est-à-dire lorsque plusieurs sources sont présentes simultanément dans l'enregistrement. On remarque que le taux de recouvrement a principalement une influence sur les différences entre NMF+RNN et Mel+RNN. Les deux systèmes ont des performances relativement similaires sur les portions monophoniques des données. En revanche, lorsque le taux de recouvrement augmente, le système NMF+RNN prend l'avantage sur les spectres Mel. Ce résultat va dans le sens de nos principales motivations dans l'utilisation de modèles NMF pour ce genre de tâche. Leur capacité à représenter adéquatement l'information pour des environnements multi-sources en font des outils particulièrement puissants pour la détection en traitement de l'audio en général. Ces résultats montrent également qu'on peut dans un même modèle, tirer parti des avantages de la NMF et de ceux des réseaux de neurones profonds. Nous pouvons également observer le comportement des modèles sur un cas particulier figure 6.7. Comme on peut s'y attendre, le RNN nous permet d'obtenir des prédictions bien plus stables temporellement comparé aux MLP. La prise en compte de dépendances temporelles plus longues par le modèle facilite la détection des événements en un seul bloc, contrairement à la prédiction trame par trame d'autres modèles. Ce cas particulier illustre également le meilleur comportement des CRNN sur les cas avec fort taux de recouvrement. En effet, le modèle NMF+CRNN arrive dans certains cas à détecter la présence de plusieurs

événements simultanément là où les autres modèles échouent.

6.6 Premiers résultats avec DNN-TNMF

Les bonnes performances des modèles TNMF et NMF+MLP nous amènent naturellement vers l'exploration de l'apprentissage conjoint du modèle NMF et du réseau, par le modèle DNN-TNMF introduit section 6.3.4. Nous rappelons que DNN-TNMF revient à remplacer la régression logistique dans TNMF par un réseau de neurones profond. De même, cela revient à apprendre conjointement tous les paramètres des modèles NMF+MLP selon le critère supervisé du MLP. L'objectif de cette section est principalement de montrer que l'apprentissage d'un modèle DNN-TNMF est possible en présentant quelques cas particuliers de classification de scènes.

6.6.1 Protocole

Nous reprenons exactement le même procédé expérimental que celui section 6.4 pour l'évaluation des modèles NMF+MLP sur la base DCASE 2017. Nous gardons les mêmes représentations temps-fréquence ainsi que la même stratégie d'initialisation par NMF non-supervisée. Les modèles DNN-TNMF sont construits avec un réseau de neurones avec une couche cachée, ce qui revient à ajouter une couche MLP classique entre la NMF et la régression logistique dans le modèle TNMF. La couche cachée possède 256 neurones et le pas du gradient pour la mise à jour du dictionnaire est fixé à $\rho_W = 0.005$. Ces paramètres sont réglés sur un des sous-ensembles des ensembles de développement de la base DCASE 2017. Les paramètres du MLP sont mis à jour avec l'algorithme SGD de *Keras* en utilisant les paramètres par défaut, c'est-à-dire un pas de gradient de $\rho_A = 0.01$.

Dans tous les cas, les modèles sont appris par SGD en suivant l'algorithme 6.1. Nous fixons la taille des mini-batchs à 16 et le *dropout* à 0.2. Les gradients du coût par rapport aux activations NMF sont obtenus par le calcul de la matrice Jacobienne de la fonction de coût du réseau grâce à la librairie *Theano*. Cette matrice contient les dérivées partielles de la fonction de coût par rapport aux activations.

6.6.2 Résultats

Observation de la décroissance du coût Dans une première expérience nous souhaitons observer la décroissance du coût du modèle DNN-TNMF en fonction des itérations. Afin d'évaluer la contribution de la mise à jour du dictionnaire dans la décroissance de la fonction du coût, nous étudions 3 cas de figure :

— les paramètres du réseau sont fixés et uniquement le dictionnaire est modifié ;

— le dictionnaire est fixé et les paramètres du réseau sont mis à jour ;

— tous les paramètres sont mis à jour conjointement.

L'évolution de la décroissance du coût logistique multinomial est donnée figure 6.8 pour les trois situations listées ci-dessus. Ces courbes sont obtenues sur le premier ensemble d'apprentissage de la base DCASE 2017. Le coût est calculé pour chaque passage de l'algorithme par 20 mini-batchs. Dans les trois cas, le modèle est initialisé avec le même dictionnaire et les couches MLP sont initialisées avec les mêmes poids afin d'assurer un point de départ strictement équivalent. De plus, la graine aléatoire a été fixée pour l'algorithme SGD de sorte à ce que les données soient lues dans le même ordre à chaque occurrence de l'apprentissage.

On peut observer figure 6.8 la décroissance du coût pour les 3 cas étudiés. Le premier, où le dictionnaire est fixé, n'est pas surprenant compte tenu du fait qu'il correspond à l'apprentissage

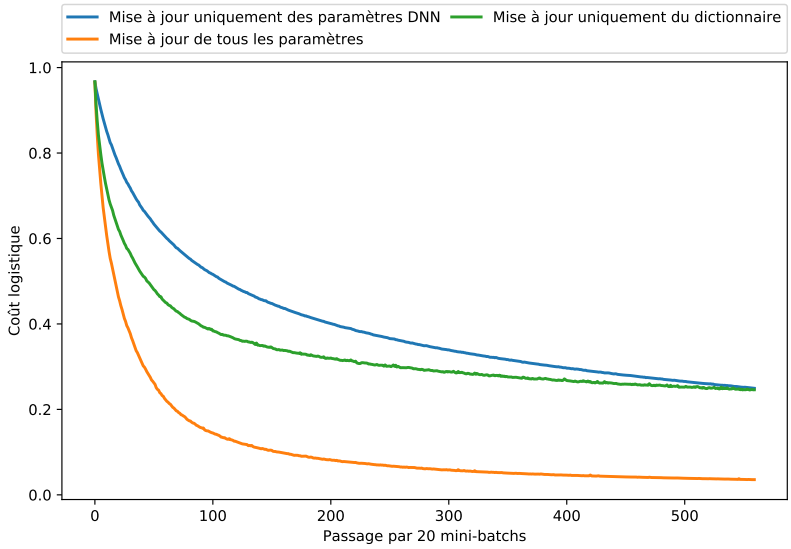


FIGURE 6.8 – Évolution du coût de classification au cours des itérations pour DNN-TNMF sur le premier ensemble de la base DCASE 2017.

	Dcase 2017			
	SNMF	TNMF	NMF+MLP	DNN-TNMF
$K = 256$	79.3	**85.0**	84.3	**85.5**
$K = 512$	83.1	**86.3**	**86.3**	**86.1**

TABLEAU 6.7 – Taux de reconnaissance sur la base DCASE 2017 pour DNN-TNMF et NMF+MLP avec 1 couche cachée sur 2 tailles de dictionnaires.

d'un MLP classique sur une représentation d'entrée fixe. En revanche, le cas le plus intéressant est celui où nous fixons les poids du réseau et autorisons seulement le dictionnaire à être modifié. Dans ce cas, la décroissance observée confirme qu'il est possible de modifier les dictionnaires NMF afin de diminuer le coût supervisé d'un réseau de neurones. La non-convexité des réseaux et la forte non-linéarité qu'introduit la fonction d'activation $\mathbf{h}^\star$ ne permettent pas de garantir la bonne convergence ou la stabilité de l'apprentissage du modèle DNN-TNMF. Ainsi, nous montrons qu'avec des choix raisonnables, il est possible d'apprendre conjointement la NMF et un MLP. Ce travail n'est qu'une première étape vers l'utilisation de modèles plus complexes. La suite logique sera d'observer si nous pouvons obtenir le même genre de comportement du modèle avec l'utilisation de réseaux plus profonds ou avec l'introduction de couches récurrentes ou convolutives.

Première expérience sur la classification de scènes Nous venons de montrer que nous sommes capables d'apprendre le modèle MLP-TNMF. Nous nous demandons maintenant si le système proposé permet uniquement d'arriver aux mêmes solutions plus rapidement que le modèle NMF+MLP ou si il permet en effet d'augmenter la capacité de généralisation des réseaux. Pour cela nous appliquons les modèles DNN-TNMF sur la base DCASE 2017 avec le protocole expérimental détaillé section 6.6.1. Nous nous comparons aux résultats obtenus avec SNMF, TNMF et NMF+MLP en utilisant le même genre de réseaux que ceux choisis pour DNN-TNMF, c'est-à-dire des MLP à une couche cachée.

Les taux de reconnaissance sur l'ensemble de développement de la base DCASE 2017 sont

présentés dans le tableau 6.7. Pour cette première expérience, on peut noter une légère amélioration des performances obtenues avec DNN-TNMF uniquement pour les dictionnaires avec $K = 256$ composantes. En revanche, même pour $K = 512$ il reste avantageux, dans l'état actuel, d'apprendre la NMF et le MLP séparément avec NMF+MLP. Cette différence de comportement en fonction de la taille des dictionnaires est semblable à celle observée pour la comparaison entre TNMF et SNMF chapitre 5. Les modèles de NMF supervisés ont tendance à être d'autant plus avantagés que l'on réduit le nombre de composantes des dictionnaires. En effet, les modèles non-supervisés fournissent des représentations trop générales lorsqu'on impose une faible dimension. Les composantes de base apprises ne permettent pas de représenter des concepts assez fins pour discriminer les scènes. En revanche, en apprenant le dictionnaire selon un critère de classification, ici un MLP, on va chercher uniquement les composantes de base permettant de différencier les catégories. Les MLP permettent de pleinement exploiter le potentiel de ces représentations en étant capable d'interpréter des descripteurs NMF de plus grande dimension (jusqu'à $K = 1024$ dans nos expériences). L'obtention de performances compétitives avec les modèles DNN-TNMF supposerait alors l'augmentation du nombre de composantes des dictionnaires s'accompagnant de la nécessité d'utiliser des modèles construits à partir de réseaux plus complexes. Or, l'augmentation de la taille et de la profondeur des réseaux risque de poser des problèmes de stabilité dans l'apprentissage de DNN-TNMF. Ainsi, il reste du travail à fournir pour des recherches futures pour proposer des algorithmes et des structures de réseaux plus adaptés à la prise en compte de telles représentations d'entrée.

6.7 Conclusion

Les réseaux de neurones profonds dominent maintenant l'analyse de sons environnementaux par leurs performances impressionnantes sur de nombreuses bases de données du domaine. Cependant, la question du choix de la représentation d'entrée pour ces modèles reste peu abordée. Il existe des liens entre certains aspects des réseaux de neurones et les modèles NMF. En proposant et en justifiant l'utilisation de représentations NMF comme entrée de MLP, nous améliorons à la fois le classifieur pour nos systèmes précédents mais aussi la représentation d'entrée pour certains modèles DNN de l'état de l'art. Les performances obtenues par les approches NMF+MLP et NMF+CRNN témoignent de leur intérêt en atteignant les résultats des systèmes de l'état de l'art sur des bases de référence de classification de scènes et de détection d'événements. Enfin, le modèle DNN-TNMF proposé permet d'ouvrir des perspectives quant à l'apprentissage conjoint des réseaux avec les paramètres NMF, regroupant l'apprentissage des différents modèles présents dans nos systèmes en un seul problème.

Chapitre 7

Conclusion

Sommaire

7.1 Bilan de la thèse

Au cours de ce travail de thèse, nous avons établi différentes approches aux problèmes de classification de scènes et de détection d'événements. L'évolution de la nature de nos systèmes au cours de nos travaux de thèse s'inscrit dans un contexte particulier. La rapide évolution de l'intérêt pour le domaine de l'ACSES ainsi que les avancées en apprentissage automatique se sont accompagnées d'une constante transformation des approches plébiscitées par la communauté. Nos travaux suivent et ont contribué aux quelques unes des directions de recherche majeures établies par les différents acteurs de l'ACSES, allant de l'ingénierie de descripteurs jusqu'à l'apprentissage profond. La structure du document traduit l'évolution chronologique de nos approches mais aussi celle des grandes directions prises par la communauté. Nous revenons dans cette section sur quelques points clés des différents systèmes introduits durant cette thèse.

L'analyse de la diversité des approches de l'état de l'art réalisée dans le chapitre 2 nous a permis de dégager plusieurs familles importantes de systèmes pour l'ACSES. En particulier, nos contributions s'inscrivent dans trois de ces courants. Le premier, et le plus ancien, est l'extraction de descripteurs. L'objectif est alors d'identifier des schémas d'extraction de caractéristiques adaptées à la description des catégories classifiées, en s'inspirant d'autres domaines de l'apprentissage automatique. Si elles animaient encore la recherche durant le début de notre thèse, ces approches sont maintenant laissées de côté au profit des méthodes d'apprentissage de représentations. Le développement de ces méthodes, et plus spécifiquement l'apprentissage de descripteurs, a entraîné des avancées de performances importantes sur de nombreux problèmes. Elles ont l'intérêt principal de permettre une modélisation adéquate de certaines propriétés des sons environnementaux tout en laissant la flexibilité au modèle d'adapter la représentation aux données traitées. Aujourd'hui, l'essentiel des méthodes récentes d'ACSES fusionnent les étapes d'apprentissage de représentations et la classification par l'utilisation de réseaux de neurones profonds. Les capacités de modélisation impressionnantes des modèles profonds rendent possible la caractérisation de la grande diversité des événements sonores présents dans notre environnement. Cela se traduit par l'arrivée de systèmes toujours plus robustes, capables de dépasser les performances humaines sur certaines tâches.

Nos premières contributions, introduites au 3, s'inscrivent dans la recherche de descripteurs adaptés à la caractérisation de sons environnementaux. Nous sommes partis de l'idée bien établie de représenter les enregistrements par une image temps-fréquence afin de s'en servir comme point de départ pour des schémas d'extraction de descripteurs. L'objectif de ces descripteurs est d'identifier les événements sonores par leur manifestation sous forme d'objets temps-fréquence dans l'image. En particulier, nous isolons deux descripteurs d'images ayant fait leur preuves en ACSES avant de proposer et de justifier leur complémentarité. La combinaison proposée constitue une alternative performante aux autres approches par descripteurs plus classiques. Cette première approche nous a permis de souligner l'utilité des outils capables de décrire à la fois la répartition et l'évolution du contenu temps-fréquence pour la classification de scènes.

Nous nous sommes ensuite éloigné des méthodes par extraction de descripteurs pour aller vers les approches d'apprentissage de représentations. Nous avons entamé notre étude des techniques d'apprentissage de descripteurs dans le chapitre 4 en nous arrêtant sur les méthodes par factorisation de matrices. Nous nous appuyons principalement sur le modèle NMF qui permet de modéliser une scène sonore comme une somme d'événements élémentaires, s'approchant ainsi de notre compréhension des sons qui nous entourent. Les différentes expériences proposées confirment l'intérêt de se tourner vers des modèles capables d'apprendre par eux-mêmes une représentation adéquate des données. Avec l'apprentissage de représentations interprétables par la NMF non-supervisée, nous avons introduit des systèmes performants malgré l'emploi de stratégies de classification relativement simples. En effet, ces approches offrent des avancées en performance en comparaison

aux systèmes par apprentissage de descripteurs. De plus, elles s'approchent également des performances des modèles profonds sur des bases de données de taille limitée.

Nous sommes entré au chapitre 5 dans le domaine de l'apprentissage supervisé de représentations. Nous nous inspirons de modèles d'apprentissage de dictionnaires supervisés existant dans l'objectif de les adapter au cadre de l'utilisation de la NMF pour la classification de sons environnementaux. Le modèle TNMF introduit permet d'apprendre conjointement le dictionnaire NMF et le classifieur, afin d'adapter la représentation obtenue à la tâche traitée. Le modèle TNMF et son algorithme d'apprentissage sont proposés dans l'objectif d'établir un équivalent supervisé aux systèmes NMF du chapitre précédent. Le modèle TNMF s'est montré particulièrement efficace pour la classification de scènes par l'apprentissage de dictionnaires d'événements élémentaires utiles à la discrimination des catégories. De plus, nos systèmes soumis au challenge DCASE confirment l'intérêt de TNMF pour la tâche par l'obtention de résultats supérieurs aux approches par apprentissage profond. Nous retrouvons également les avantages de TNMF pour la détection d'événements, où nous obtenons des résultats comparables aux systèmes état de l'art sur une base de données en conditions réelles avec recouvrement. Cependant, le modèle se basant sur des stratégies de classification relativement simples, il souffre de son manque de modélisation de l'évolution temporelle de l'information lors de l'apprentissage sur des bases de plus grande taille. Il reste un outil d'apprentissage de représentations performant par la combinaison des approches supervisées avec la NMF et ses avantages pour caractériser les environnements multi-sources.

Nous terminons par la présentation de nos modèles par réseaux de neurones profonds dans le chapitre 6. En s'inscrivant dans la tendance actuelle des approches d'ACSES, nos travaux sur l'apprentissage profond nous amènent à la construction de systèmes de plus en plus performants. Nous tirons parti des enseignements des chapitres précédents afin de proposer l'apprentissage de réseaux profonds sur des représentations NMF, à la différence des approches usuelles partant de spectrogrammes. Il existe des similitudes entre le rôle des couches cachées classiques des réseaux et celui des modèles NMF. Ainsi, nous justifions la pertinence d'entrainer des réseaux à partir de telles représentations d'entrée. Nous gardons les avantages de la NMF pour représenter les sons environnementaux tout en profitant du pouvoir de classification des réseaux profonds. Nos motivations sont ensuite validées par différentes évaluations expérimentales. En effet, à la fois pour la classification de scènes et la détection d'événements, les modèles proposés nous permettent d'atteindre les systèmes état de l'art sur des bases de données de référence. En particulier, nous introduisons un modèle NMF+CRNN enchainant des étapes de représentations NMF, couches convolutives et couches récurrentes pour la détection d'événements polyphonique. Enfin, nous terminons par l'étude de la possibilité d'entrainer conjointement le modèle NMF et des réseaux de neurones profonds. Le modèle DNN-TNMF introduit considère la NMF comme la première couche d'un réseau, dont les paramètres sont entrainés selon les critères supervisés usuels. Nous revenons sur le potentiel d'une telle approche lors du développement de nos perspectives dans la section suivante.

7.2 Perspectives

L'état actuel des meilleurs systèmes d'ACSES rend possible le développement de nouvelles applications grâce à leur capacité à reconnaître les sons environnementaux avec de plus en plus de précision. Néanmoins il reste une demande et une marge de progression importante pour la construction de systèmes plus polyvalents, plus rapides et plus robustes. Dans cette section, nous isolons plusieurs pistes de recherche qui nous semblent prometteuses à explorer pour des travaux futurs. Le choix de ces perspectives est guidé par l'objectif d'améliorer les performances et la robustesse de nos systèmes, tout en restant dans une logique similaire aux approches proposées

dans ce manuscrit.

7.2.1 Rendre nos systèmes plus robustes

Nous avons consacré la dernière partie du chapitre 2 à la présentation des différentes approches employées en classification de sons afin d'améliorer la robustesse des systèmes. Ces techniques constituent rarement le cœur du système mais elles offrent des outils relativement simples pouvant entrainer des gains de performance remarquables. Nous avons laissé de côté certains de ces aspects dans nos travaux afin de nous focaliser principalement sur l'apprentissage de représentations. Nous nous sommes limités à la simple utilisation d'une fusion tardive de classifieurs, et cela uniquement lors de mise en place de nos soumissions aux challenges DCASE. En effet, l'organisation et la popularité de ces challenges ont largement contribué à la généralisation des techniques d'augmentation et de fusion par leur présence récurrente dans les méthodes les mieux classées.

L'exploration de ces différents outils pourrait constituer un moyen simple et efficace d'augmenter les capacités de généralisation de nos systèmes. En particulier pour la détection d'événements, nous nous sommes limités à l'utilisation de spectres Mel avec peu de bandes de fréquences (40 ou 80 selon les cas) et à l'utilisation d'un seul classifieur (allant de la régression logistique au CRNN). Or, les résultats de nombreux travaux récents nous amènent à penser que nous avons beaucoup à gagner à faire varier à la fois les représentations d'entrée et les stratégies de classification. En particulier, les modèles tels que les CRNN sont sensibles au sur-apprentissage, ce qui oblige à mettre en place des critères d'arrêt pertinents durant la phase d'apprentissage. Alors, la simple fusion tardive de différentes occurrences de ces modèles constituerait un moyen simple d'augmenter les capacités de détection de nos systèmes.

De plus, nos approches bénéficiaient du fait d'avoir des représentations temps-fréquence plus variées à leur disposition dans le but de rendre l'apprentissage de représentations plus polyvalent. En effet, nous avons vu dans le chapitre 3 que les différentes constructions des spectrogrammes (STFT, CQT et Mel) retranscrivent différemment le contenu temps-fréquence des sons. Dans ce cas, l'apprentissage d'un dictionnaire NMF d'événements élémentaires à partir de spectrogrammes de nature différentes pourrait augmenter le pouvoir de représentativité et de discrimination des descripteurs NMF. Dans ce sens, il serait également possible de contraindre les activations NMF issues de la décomposition des différentes représentations à être similaires. Cela forcerait les différents dictionnaires à contenir la représentation des mêmes événements élémentaires mais sur des échelles fréquentielles variées.

Enfin, nous nous sommes confrontés durant nos travaux au faible nombre de bases de données publiques de bonne qualité. Celles que nous avons sélectionnées pour nos évaluations présentent des limitations importante de taille (au maximum 25h d'enregistrements) et de diversité (au maximum 16 catégories d'événements). Ainsi, l'arrivée de nouvelles bases de données pousse naturellement à étudier le comportement de nos modèles lorsque le nombre de données augmente, ou lorsque les conditions sont dégradées. En particulier, l'exploration plus poussée des tâches de détection et classification d'événements dans des conditions bruitées pourrait mettre nos approches en valeur. En effet, une des forces de la NMF est sa capacité à modéliser l'addition des différentes sources dans un mélange sonore. Cela ouvre donc la possibilité, lors de l'apprentissage de représentations, de prendre en compte efficacement la différenciation entre les événements d'intérêt et le bruit de fond. Avec la multiplication des applications embarquées de détection d'événements, il devient essentiel de leur donner les moyens de réaliser leur tâche même en présence de conditions sonores délicates.

7.2.2 Données faiblement annotées et problèmes de tagging

Une grande partie des acteurs majeurs d'ACSES se tourne maintenant vers l'étude de modèles appropriés aux traitements de données faiblement annotées. L'intérêt pour ce type de données s'explique principalement par la plus grande facilité de leur collection (voire A) ainsi que par leur relation directe avec les applications d'indexation des méthodes d'ACSES (voire 1). Il nous semble alors naturel d'explorer le potentiel de nos approches en les évaluant et en les adaptant à ce genre de problème. Dans l'objectif d'adapter nos systèmes aux données faiblement annotées, une première approche serait de reprendre les architectures de réseaux de neurones avec mécanisme d'attention proposés dans les travaux récents. En effet, nous avons vu au chapitre 6 que l'apprentissage non-supervisé de descripteurs fournit des représentations d'entrée adaptées à l'apprentissage de réseaux de neurones profonds de différentes natures (MLP, RNN, CRNN). L'introduction de ces mécanismes d'attention dans les réseaux change essentiellement le fonctionnement des couches de classification, ce qui les rend tout à fait compatibles avec les modèles tels que NMF+CRNN. De plus, les capacités de généralisation des modèles pour les tâches de *tagging* sont susceptibles d'être améliorées par un apprentissage non-supervisé adéquat, tel que celui qu'offre les modèles NMF. En effet, l'imprécision des annotations rend l'apprentissage de modèles supervisés performants plus délicat. Ainsi, il pourrait être bénéfique, pour les systèmes DNN de l'état de l'art, de laisser une partie du rôle d'apprentissage de représentations à des modèles non-supervisés. Cette idée suit la logique des modèles NMF+MLP que nous avons pour la classification de scènes. Ensuite, notre modèle TNMF permet, avec quelques modifications, de traiter adéquatement des données semi-supervisées. Dans ce cas, l'apprentissage de dictionnaires peut à fois faciliter la discrimination des données annotées et la modélisation des données sans étiquettes. Enfin, le modèle TNMF permet de mettre en place un apprentissage supervisé de la NMF selon le critère choisi. Une autre piste serait alors d'explorer le potentiel du modèle TNMF sur des données faiblement annotées en reprenant les mécanismes d'attention ou l'apprentissage à instances multiples comme critère supervisé. Un tel modèle garderait l'avantage de la capacité de représentation de la NMF tout en assurant un traitement adapté aux données faiblement annotées.

7.2.3 Apprentissage conjoint de représentations et DNN

Nous avons entamé l'étude de l'apprentissage conjoint de l'étape d'apprentissage de représentations par NMF et d'un réseau de neurones profond au chapitre 6. Nous nous sommes arrêtés à l'étude d'un modèle relativement simple (MLP à une couche cachée) appris avec un algorithme par SGD. Les premiers résultats sont prometteurs mais ne justifient par encore l'utilisation de tels modèles sur de plus petites bases de données. Ainsi, un travail conséquent est nécessaire sur l'exploration d'algorithmes d'apprentissage et d'architectures de réseaux appropriées à la particularité du modèle. Toutefois, rien n'empêche en pratique de remplacer les couches MLP par des couches convolutives ou récurrentes, dans l'optique de construire des systèmes de type CRNN-TNMF. Il est cependant possible que ces modèles nécessitent davantage de données d'apprentissage pour commencer à présenter des capacités de généralisation intéressantes. En effet, l'idée d'apprendre toutes les étapes des systèmes de classification audio selon le critère supervisé final devient de plus en plus populaire avec les approches *end-to-end*. Ces méthodes s'appuient sur des réseaux profonds appris directement à partir des signaux audio bruts. Elles sont cependant réputées nécessiter un nombre plus important d'exemples d'apprentissage pour obtenir des performances raisonnables à cause de la grande dimension et du manque d'interprétabilité de la représentation d'entrée. Les résultats obtenus par les premières tentatives en ACSES témoignent de cette limitation. Bien que plus simples, les modèles DNN-TNMF s'inscrivent dans une logique similaire à celle des approches *end-to-end* par le regroupement des différentes étapes de nos systèmes en un seul problème supervisé. Dans ce sens on pourrait également imaginer étendre DNN-TNMF

de sorte à apprendre directement à partir des STFT. L'enchainement des étapes de banc de filtres, de NMF et de classification pourraient être appris conjointement selon le critère de classification désiré. L'application de tels modèles semble envisageable uniquement grâce à la sortie récente de bases de données de plus grande taille. En effet, elles ouvrent la voie à des approches plus complexes, aux capacités de modélisation et de généralisation plus grandes. Il s'agit très certainement d'une étape cruciale dans l'évolution du domaine. Les différents acteurs possèdent maintenant davantage d'outils pour exploiter pleinement le potentiel de l'apprentissage profond pour l'analyse des sons environnementaux.

Annexe A

Bases de données

Sommaire

A.1 Collection et annotation de données pour l'analyse de scènes sonores

Cette annexe présente les bases de données utilisées tout au long du manuscrit. La création de bases de données est un des principaux défis pour de nombreuses tâches de classification. La nature et le contenu des bases choisies pour évaluer les modèles conditionnent notre interprétation des résultats. Ainsi, plusieurs critères sont recherchés dans la constitution d'une bonne base de données. Le premier est sa représentativité des différentes catégories pouvant être rencontrées par la tâche de classification traitée. En effet, les modèles d'apprentissage automatique ne peuvent pas prédire des étiquettes qu'ils n'ont jamais vu. Le deuxième est qu'elle contienne un échantillon suffisant d'exemples pour représenter la variabilité propre à chaque catégorie. Par exemple, pour un événement sonore, cela peut vouloir dire avoir des exemples aux conditions ou au contexte d'enregistrement différents. La vérification de ces propriétés va de paire avec la possibilité de proposer des méthodes d'apprentissage plus performantes mais surtout aux meilleures capacités de généralisation. Outre certains cas particuliers, les bases d'ACSES sont souvent de taille relativement faible en comparaison à d'autres domaines tels que la vision par ordinateur ou le traitement du langage. Afin de palier ce manque de bases de plus grande taille, il convient alors, dans la mesure du possible, de ne pas se limiter à l'évaluation de modèles sur un seul ensemble de données. C'est pourquoi dans ce travail de thèse, nous appliquons les systèmes proposés à plusieurs bases, aux propriétés différentes, afin d'appuyer l'interprétation des performances obtenues.

Les premières bases de données ont souvent été créées en collectant des sons libres de droits sur des sites internet tels que *free-sound.com* ou sur des banques de sons payantes pour le bruitage cinéma. Un des principaux défauts de ce genre de procédé est qu'il compte sur la qualité de l'indexation et des étiquettes ajoutées par les utilisateurs. Ces étiquettes décrivent le contenu de l'enregistrement mais ne caractérisent pas toujours de manière adéquate son contenu en événements sonores. Par exemple, certaines de ces bases sont construites à partir de données audiovisuelles, des étiquettes peuvent alors décrire des objets présents dans l'image n'émettant pas forcement de son distinguable. De plus, certains événements étiquetés peuvent être présents sur seulement une fraction de l'enregistrement, rendant leur identification plus difficile. L'utilisation de ce type de données pour la classification introduit naturellement des imprécisions et des confusions dans la création des étiquettes et nécessite souvent une intervention humaine pour ré-indexer, nettoyer et segmenter les données. Un cas de figure problématique que l'on retrouve fréquemment est par exemple la présence d'une étiquette telle que *klaxon* dans un extrait sonore de plusieurs minutes. Dans ce cas, segmenter l'exemple pour isoler le klaxon permet de grandement faciliter l'apprentissage de modèles de détection ou de classification capables d'identifier cette catégorie d'événements. On parle de données faiblement annotées lorsque l'on sait la présence d'un objet sonore dans le segment audio sans connaître sa position ni sa proportion dans l'audio. Plus récemment, les défis que posent ces stratégies de collection sont compensés par une mise à l'échelle des bases de données, telle qu'effectuée avec l'Audioset. Dans ce cas, la présence d'une quantité très importante d'exemples permet tout de même d'apprendre des systèmes performants malgré les imprécisions introduites par la collection des données. Cependant, la construction de l'Audioset et des sous-bases en découlant ont tout de même nécessité une intervention humaine pour nettoyer et confirmer la pertinence des étiquettes.

La deuxième approche majeure est de réaliser des enregistrements d'environnements sonores spécifiquement dédiés à la construction d'une base de données. Cette stratégie permet en grande partie d'éviter les étapes de nettoyage des données collectées en basculant le coût en temps vers le processus d'enregistrement et non l'annotation. En effet, les enregistrements sont souvent effectués dans un cadre contrôlé, de sorte à ce que nous sachions à l'avance ce que nous allons enregistrer, retirant ainsi l'étape d'annotation. Cependant, cet aspect est beaucoup moins vrai pour

la détection d'événements. La nécessité de connaitre les frontières précises de début et de fin de chaque occurrence d'un événement rend l'ajout d'un processus d'annotation presque inévitable. Il est néanmoins possible de s'en affranchir par la construction de bases synthétiques, en créant des mixtures à partir d'enregistrements isolés. En résumé, les différentes approches à la construction de bases de données s'accompagnent toujours d'un compromis entre de nombreux aspects : le coût, le temps, le réalisme, la taille de la base, le nombre de catégories, la subjectivité des annotations etc.

Dans ce manuscrit nous utiliserons plusieurs bases de données correspondant au différentes sous-tâches de l'analyse de sons environnementaux. Toutes les bases de données sélectionnées sont relativement récentes, elles ont toutes été publiées après le commencement de cette thèse. Ainsi, elles ont largement contribué à la rapide explosion de l'intérêt et du développement de nouvelles méthodes pour l'analyse de sons environnementaux. De plus, elles ont participé à la définition plus rigoureuse des frontières entre les différentes tâches du domaine. En effet, dans les premiers travaux de classification d'événements, la plupart des bases regroupaient des étiquettes correspondant à la fois à des événements brefs isolés tels qu'un klaxon ou à des ambiances sonores décrivant le contexte telles que *dans la rue*. Ces différents niveaux de description du contenu sont maintenant clairement séparés en trois tâches, la classification de scènes, la classification et la détection d'événements.

A.2 Classification de scènes sonores

Nous rappelons que la classification de scènes sonores a pour objectif d'identifier, à partir d'un enregistrement, dans quel type de lieu ou d'environnement le son a été enregistré tel que *dans la rue* ou *métro*. Nous présentons ici trois bases de données que nous utiliserons tout au long de ce travail. Les trois bases sont du même ordre de grandeur et suivent un procédé de collection similaire qui est le suivant :

Une personne équipée d'un microphone se place dans un environnement bien défini comme par exemple un restaurant. Elle lance l'enregistrement tout en notant dans quel type d'environnement elle se trouve. Une fois les enregistrements terminés, ils sont découpés en segments de longueur fixe et l'étiquette de l'environnent dans lequel ils ont été enregistré leur est associé. De plus, comme pour de nombreuses bases de données pour la classification, les auteurs fournissent une séparation en plusieurs ensembles d'apprentissage-test et parfois de développement. Les bases de classification de scènes restent à ce jour de taille limitée. En effet, le processus de collection par enregistrement empêche l'automatisation de la constitution de la base, rendant sa construction couteuse en temps. Le résumé de la configuration des trois bases est donnée dans le tableau A.1, la liste d'étiquettes et leur nombre pour chaque base sont présentés dans le tableau A.2 et leur description plus détaillée est donnée ci-dessous.

A.2.1 LITIS Rouen

Lien de téléchargement : sites.google.com/site/alainrakotomamonjy/home/audio-scene

La base du LITIS Rouen [Rakotomamonjy et Gasso, 2015] reste à ce jour une des plus grandes bases de données publiques de classification de scènes. Elle contient environ 25 heures d'enregistrements pris avec le microphone d'un smartphone Samsung principalement dans des environnements urbains sur le sol Français. Les enregistrements sont disponibles au format "wav" en monophonique avec une fréquence d'échantillonnage de 44.1 khz. Ils sont ensuite découpés en segments de 30 secondes. La base contient 19 étiquettes différentes énumérées dans le tableau A.2. La base est séparée en 20 ensembles apprentissage-test où chaque ensemble d'apprentissage contient un sous-ensemble d'apprentissage et un ensemble de développement de taille égale. Les

Base	**LITIS Rouen**	**Dcase ASC 2016**	**Dcase ASC 2017**
Lieux	France	Finlande	Finlande
Enregistreur	Smartphone Samsung	Soundman OKM II	Soundman OKM II
Nombre d'exemples	3023	1170	4370
Taille des segments	30 secondes	30 secondes	10 secondes
Nombre d'étiquettes	19	15	15
Validation croisée	20 ensembles 80-20%	4 ensembles 75-25%	4 ensembles 75-25%

TABLEAU A.1 – Description des bases de données de classification de scènes sonores

performances des modèles évalués sur cette base seront toujours calculées comme la moyenne des performances sur les 20 ensembles de test. Une des particularités de cette base est que ces ensembles ont été créés aléatoirement. Cela implique par exemple que deux segments de 30 secondes enregistrés dans le même lieu peuvent se retrouver à la fois dans l'ensemble d'apprentissage et dans celui de test. Les modèles ont alors tendance à fournir des performances généralement optimistes.

On peut noter qu'une deuxième version de la base est sortie récemment contenant les mêmes enregistrements mais avec une nouvelle séparation apprentissage-test. Dans celle-ci, les segments issus de la même session d'enregistrement (faits dans le même lieu ou au même moment) sont regroupés dans le même ensemble. Cette deuxième version est sortie bien après la publication de nos travaux sur la première version, d'où le fait que nous présentons des résultats uniquement sur la première version.

A.2.2 Dcase 2016

Lien de téléchargement : www.cs.tut.fi/sgn/arg/dcase2016/task-acoustic-scene-classification

La base DCASE 2016 correspond à l'ensemble de développement de la tâche 1 de l'édition 2016 du challenge DCASE [Mesaros et al., 2016b]. Le processus de collection des données est similaire à celui de la LITIS. Le microphone utilisé est un Soundman OKM II Klassik/studio A3, electret binaural avec un enregistreur Roland Edirol R-09 wave avec une fréquence d'échantillonnage de 44.1 kHz et une résolution de 24 bits. Elle contient 1170 exemples de 30 secondes repartis en 15 catégories différentes énumérées dans le tableau A.2. Les auteurs de la base de données fournissent également 4 ensembles de validation croisée de tailles égales. A la différence de la base du LITIS, tous les segments de 30 secondes ayant été enregistrés dans le même lieu sont regroupés dans le même ensemble. Cette base étant associée à un challenge, elle a l'avantage de comparer les performances d'une multitude de travaux. De plus, la base est associée à un ensemble d'évaluations supplémentaires. Il correspond à l'ensemble utilisé pour comparer les soumissions au challenge. Il contient 26 exemples de 30 secondes par classe.

A.2.3 Dcase 2017

Lien de téléchargement : www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-acoustic-scene-classification

La base du DCASE 2017 est l'extension de la version 2016 pour l'édition 2017 du challenge. Elle se compose du rassemblement entre la version de développement et de l'ensemble d'évaluation de la base 2016. La seule différence étant que les segments ont été réduits à une longueur de 10 secondes. Elle contient donc 4370 segments de 10 secondes répartis en 4 ensembles de validation croisée de mêmes tailles.

TABLEAU A.2 – Liste des catégories et nombre d'exemples par catégorie pour les bases de données de classification de scènes sonores

Étiquettes	**LITIS Rouen**	**DCASE 2016**	**DCASE 2017**
Avion	23	-	-
Bus	192	78	312
Café	120	-	-
Café/restaurant	-	78	312
Voiture	243	78	312
Gare	269	-	-
Aire de jeux pour enfant	145	-	-
Marché extérieur	276	-	-
Métro Paris	139	-	-
Métro Rouen	249	-	-
Billard	155	-	-
Rue calme	90	78	312
Hall étudiant	88	-	-
Restaurant	133	-	-
Rue piétonne	122	-	-
Magasin	203	78	312
Train	164	78	312
TGV	147	-	-
Station de Métro	125	78	312
Rue bruyante/Centre ville	143	78	312
Chemin de Forêt	-	78	312
Plage	-	78	312
Parc	-	78	312
Domicile	-	78	312
Bureau	-	78	312
Bibliothèque	-	78	312
Tram	-	78	312

A.3 Classification d'événements

Lien de téléchargement : serv.cusp.nyu.edu/projects/urbansounddataset/urbansound8k.html

Pour la classification, nous présentons une seule base de données qui est uniquement utilisée dans le chapitre 3. Il s'agit de la base "UrbanSound8k" [Salamon et al., 2014] qui est une collection de segments audio contenant différents événements urbains. En particulier, elle contient 10 étiquettes correspondant aux événements sonores les plus redondants dans les plaintes pour nuisance sonore à la police de New York. Une fois la liste de d'étiquettes choisie, les sons ont été extraits de la banque de sons publique "free-sound". Les enregistrements extraits ont ensuite été découpés de sorte à isoler différentes occurrences des événements afin de former des segments plus courts de 1 à 4 secondes. La base est repartie en 10 ensembles de validation croisée de mêmes tailles regroupant les segments provenant du même enregistrement dans le même ensemble. Cette base a la particularité de regrouper des sons enregistrés avec une large variété de microphones, de qualités et fréquences d'échantillonnage variables.

A.4 Détection d'événements

Nous nous intéressons uniquement à la détection d'événements avec recouvrement. C'est-à-dire qu'à un instant donné, plusieurs événements peuvent être présents simultanément dans l'annotation. Nous avons choisi deux bases différentes, une base annotée à partir d'enregistrements réels et l'autre créée de manière synthétique. Les bases enregistrées dans des conditions réelles nécessitent un effort d'annotation important. En effet, elles demandent à un ou plusieurs annotateurs de placer à la main, les instants de début et de fin de chaque événement présent dans l'enregistrement. Par conséquent, les seules bases en conditions réelles disponibles sont de relativement petite taille. Ainsi, le réalisme de la base se paye par un nombre limité de données d'apprentissage. Ce problème peut être contourné par la construction de bases de détection d'événements synthétiques. Après une étape de collection ou d'enregistrement d'événements isolés préalablement, des mixtures synthétiques sont construites par un processus de génération aléatoire de séquences d'événements avec recouvrement. L'avantage de cette approche est de pouvoir créer un nombre important de séquences et d'avoir un contrôle sur le taux de recouvrement et le niveau de bruit de fond. Bien sûr, les bases synthétiques ont le principal défaut de ne pas être réalistes mais elles permettent l'application de modèles plus demandeurs en quantité de données d'apprentissage.

A.4.1 TUT-SED 2016

Lien de téléchargement : www.cs.tut.fi/sgn/arg/dcase2016/task-sound-event-detection-in-real-life-audio

Il s'agit de la base de développement associée à la tâche 3 de l'édition 2016 du challenge DCASE pour la détection d'événements dans des conditions réelles. La base de données contient 12 enregistrements de 3 à 5 minutes dans deux environnements différents : *Rue calme* et *au domicile*. Les deux environnements contiennent respectivement une liste de 7 et 11 étiquettes correspondant aux différents événements se produisant dans les enregistrements. La vérité terrain contient les instants de début et de fin de l'occurrence de chaque événement dans la liste de étiquettes. Ces frontières ont été placées à la main par un annotateur écoutant les enregistrements a posteriori. La base contient 4 ensembles de validation croisée contenant chacun de 3 à 4 enregistrements. Au total la base contient moins de 2h d'enregistrements annotés. La petite taille de la base, le déséquilibre entre les étiquettes, le recouvrement et la présence de bruit de fond en font une base de données particulièrement exigeante. Cependant c'est une des seules bases publiques de détection d'événements avec recouvrement en conditions réelles, ce qui la rend également particulièrement

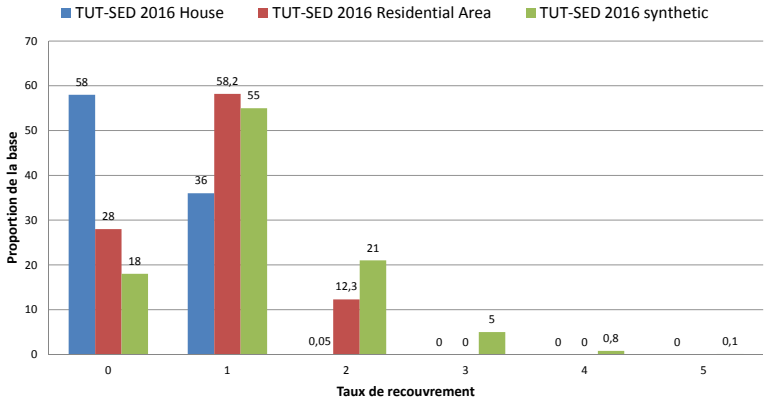


FIGURE A.1 – Pourcentages des données par taux de recouvrement pour les bases de détection d'événements

attrayante. La liste des étiquettes et leur nombre d'occurrences pour chaque environnement sont affichés dans le tableau A.3 et le taux de recouvrement entre étiquettes est présenté dans la figure A.1.

A.4.2 TUT-SED Synthetic 2016

Lien de téléchargement : www.cs.tut.fi/sgn/arg/taslp2017-crnn-sed/tut-sed-synthetic-2016

La base TUT SED synth est composée de mixtures synthétiques d'événements. Les mixtures sont créées à partir de sons isolés répartis en 16 catégories autorisant le recouvrement entre plusieurs événements. Les sons isolés ont été extraits de la banque de sons en ligne *sound-ideas.com*. Un total de 16 catégories contenant suffisamment d'exemples différents a été retenu donnant un total de 976 sons isolés. Les sons isolés extraits ont été découpés de sorte à enlever les périodes

TABLEAU A.3 – Nombre d'occurrences et durée moyenne de chaque catégorie d'événements pour les deux environnements de la base TUT-SED 2016

Residential Area			**Home**		
Étiquettes	Occurences	Durée moyenne	Étiquettes	Occurences	Durée moyenne (s)
Bird singing	130	7.55	water tap running	35	6.0
Car passing by	57	9.16	dishes	94	1.43
Children shouting	23	2.00	cutlery	56	0.74
Object banging	15	0.76	glass jingling	26	0.80
People speaking	40	8.08	cupboard	27	0.65
People Walking	32	5.50	object impact	153	1.01
Wind blowing	22	6.09	object snapping	42	0.45
-	-	-	washing dishes	56	4.44
-	-	-	drawer	22	0.8
-	-	-	object rustling	40	3.28
-	-	-	people walking	24	3.88

TABLEAU A.4 – Liste des catégories et durée totale des événements pour la base TUT-SED synthetic 2016

TUT-SED synthetic 2016			
Étiquettes	Durée totale (s)	Labels	Durée totale (s)
Alarms and Sirens	4405	Footsteps	1173
Baby crying	2007	Glass smash	621
Bird singing	2298	Gun shot	534
Bus	3464	Horse walk	1614
Cat meowing	941	Mixer	4020
Crowd applause	3278	Motorcycle	3691
Crowd cheering	4825	Rain	3975
Dog Barking	716	Thunder	3007

de silence en début, milieu et fin de segment afin de faciliter la précision de l'annotation. La base contient 100 mixtures synthétiques d'environ 5 minutes reparties en 3 ensembles apprentissage-test-validation 60%-20%-20%. Les mixtures peuvent contenir un recouvrement de jusqu'à 5 événements à un instant donné. Ce procédé a l'avantage de permettre d'obtenir facilement un nombre plus important de données annotées par rapport à la base TUT-SED 2016 tout en offrant un meilleur équilibre entre chaque catégorie. La liste des étiquettes et leur durée totale dans la base sont présentées dans le tableau A.4 et le taux de recouvrement entre étiquettes est présenté dans la figure A.1.

Annexe B

Métriques pour la détection d'événements

Sommaire

B.1 Pourquoi des métriques particulières ?

La question du choix de la métrique se pose rarement pour l'évaluation de tâches de classification standards. Pour des problèmes multi-classes, calculer le taux de reconnaissance (ou d'erreur) suffit à donner une bonne première indication des performances des modèles. Comme une seule étiquette est associée à chaque observation, le taux de reconnaissance nous donne simplement le pourcentage de bonnes réponses obtenues par notre système. Dans le cas de la classification de scènes ou d'événements cette observation correspond à une séquence de longueur fixée à la construction de la base de données. Ainsi, la question de la longueur de la fenêtre d'évaluation ne se pose pas.

En revanche, les tâches de détection d'événements avec recouvrement offrent un plus large choix de métriques qui influent sur notre interprétation des résultats. A la différence des tâches de classification, l'axe du temps sur lequel sont placés les instants de début et de fin des événements n'est pas échantillonné. Une grande partie des bases possède des frontières placées à la milliseconde près. Une telle précision temporelle semble excessive et amène donc la question du choix d'une longueur de fenêtre d'évaluation pertinente. C'est là qu'apparait la distinction importante entre les métriques par segments et les métriques par événements. Les métriques par événements s'intéressent uniquement à la bonne détection des instants de début et de fin de chaque événement dans l'annotation. Elles traduisent la capacité du modèle à identifier avec précision, souvent à quelques centaines de millisecondes près, les instants d'occurrence de chaque événement. Ces métriques reviennent à traiter le problème de détection d'événements polyphoniques comme autant de problèmes de segmentation qu'il y a de catégories. Elles sont cependant de moins en moins plébiscitées par la communauté que se tourne principalement vers l'utilisation de métriques par segments.

Les métriques par segments définissent une longueur de fenêtre d'évaluation fixe et évaluent la présence ou l'absence d'un événement dans cette fenêtre indépendamment de la proportion de l'événement dans la fenêtre d'évaluation. C'est-à-dire que si un événement se manifeste uniquement sur quelques dizaines de millisecondes, il sera considéré comme présent même pour des fenêtres d'évaluation d'une seconde. Les métriques par segments accordent moins d'importance à la précision temporelle de la détection. Ainsi, elle permettent de s'affranchir de la subjectivité des annotations en détection d'événements. En effet, les instants de début de certains événements peuvent être flous et sujets à interprétation, en particulier pour les événements se définissant par une augmentation progressive en intensité tels qu'une voiture qui passe. Les métriques par segments vont uniquement traduire la capacité des modèles à lister correctement les événements présents dans une fenêtre donnée. Par exemple, les systèmes soumis aux tâches de détection des challenges DCASE 2016 et 2017 ont été évalués sur des fenêtres de 1 seconde sans recouvrement. Les différentes métriques de détection d'événements sont présentées et discutées en détail dans les travaux de Mesaros et al. [2016a].

B.2 Score F1 et ER par segment

La majorité des métriques, dont celles que nous présentons, se définit à partir de la notion d'épreuves élémentaires. Dans le cas de la détection d'événements, une épreuve correspond à l'identification de la présence ou l'absence d'une catégorie d'événements dans la fenêtre d'évaluation. Il convient alors de rappeler les notions suivantes, permettant de qualifier le résultat de l'épreuve :

— *True positive* (TP) : La classe c est présente à la fois dans la prédiction et dans l'annotation.

— *True negative* (TN) : La classe c n'est présente ni dans la prédiction ni dans l'annotation.

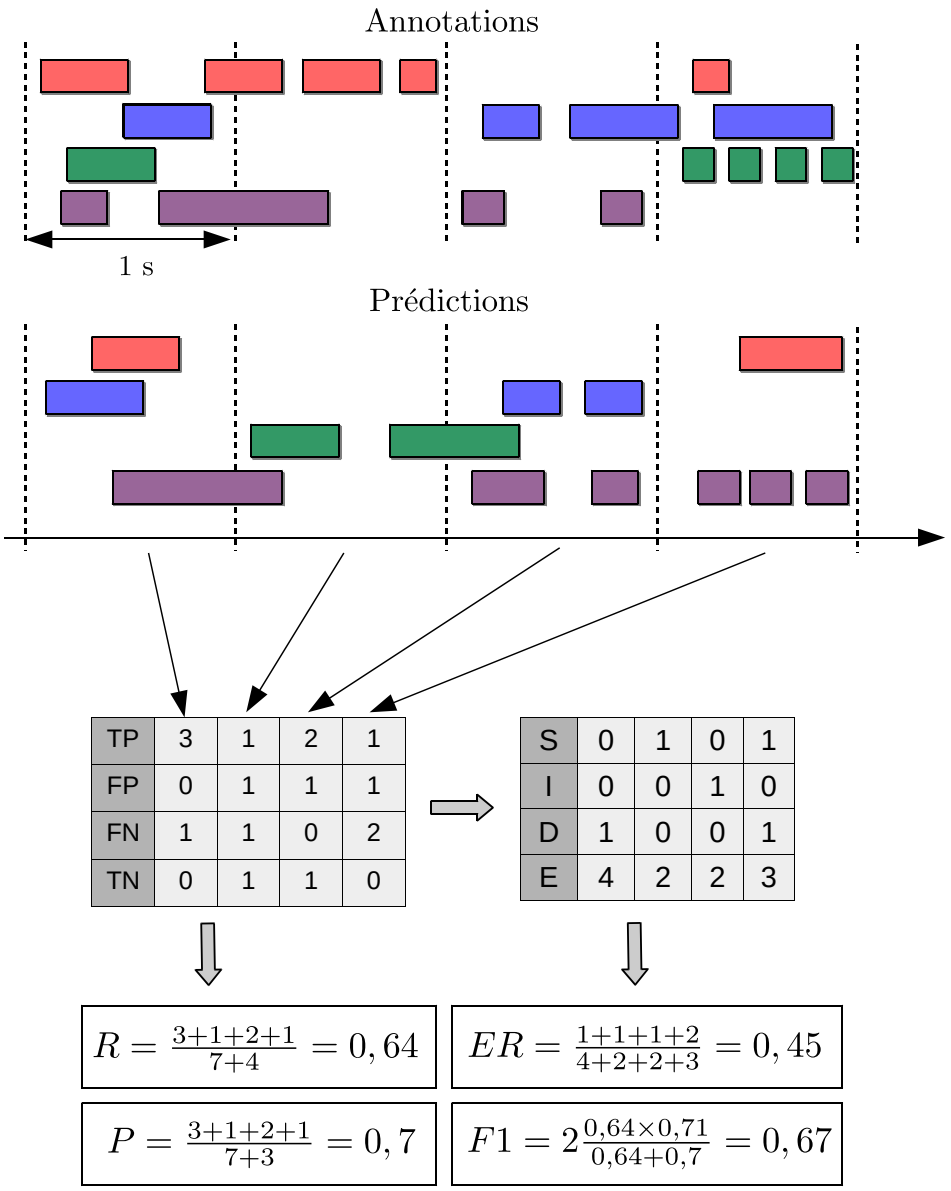


FIGURE B.1 – Exemple du calcul de la précision, du rappel, du score F1 et de l'ER sur quatres fenêtres d'évaluation d'une seconde pour la détection d'événements avec recouvrement.

— *False positive* (FP) : La classe c est présente dans la prédiction mais pas dans l'annotation.

— *False negative* (FN) : La classe c n'est pas présente dans la prédiction mais est présente dans l'annotation.

B.2.1 Scores F1 par segment

Considérons le cas d'un problème de détection d'événements à C classes et N fenêtres d'évaluation. On note FP_n le nombre de faux positifs dans la fenêtre n, cette valeur s'obtient en comptant le nombre de classes ayant donné un faux positif pour cette fenêtre. Il en va de même pour TP_n, TN_n et FN_n. La précision, le rappel et le score F1 par segment s'obtiennent alors en accumulant ces statistiques intermédiaires sur l'ensemble des segments d'évaluation :

$$P = \frac{\sum_{n=1}^{N} TP_n}{\sum_{n=1}^{N} FP_n + TP_n},\ R = \frac{\sum_{n=1}^{N} TP_n}{\sum_{n=1}^{N} FP_n + FN_n} \text{ et } F = \frac{2PR}{P+R}. \tag{B.1}$$

B.2.2 ER par segment

Le taux d'erreur (ER) de l'anglais *error rate*, représente comme son nom l'indique, la proportion d'erreur dans les prédictions du système par rapport à l'annotation. L'ER en détection d'événements a été adapté d'autres métriques similaires telles que le taux d'erreur par mot en reconnaissance de la parole ou le taux d'erreur en détection du locuteur. L'expression de l'ER nécessite l'introduction de trois grandeurs intermédiaires.

— La substitution $S_n = \min(FN_n, FP_n)$: elle traduit l'occurrence simultanée d'un faux positif et d'un faux négatif.

— La suppression (ou *deletion* en anglais) $D_n = \max(0, FN_n - FP_n)$: représente tous les faux négatifs non comptés comme substitution.

— L'insertion $I_n = \max(0, FP_n - FN_n)$: représente tout les faux positifs non comptés comme substitution.

Si on pose E_n le nombre total d'événements présents dans l'annotation pour le segment n, alors l'ER s'obtient comme suit :

$$ER = \frac{\sum_{n=1}^{N} S_n + I_n + D_n}{\sum_{n=1}^{N} E_n}. \tag{B.2}$$

L'obtention du score F1 et de l'ER par segment est illustrée sur un cas particulier figure B.1.

Annexe C

Noyaux de Sinkhorn pour la classification de descripteurs d'images

Sommaire

Cette annexe présente notre étude du choix du noyau pour les SVM lors de la classification de la combinaison des HOG et des SPD. Nous reprenons l'évaluation expérimentale proposée dans nos travaux sur les descripteurs d'images dans le chapitre 3 et dans [Bisot et al., 2015].

C.1 Noyaux de Sinkhorn pour la classification

Nous reprenons brièvement dans cette annexe l'idée d'une construction du noyau du classifieur SVM par la distance de Sinkhorn [Cuturi, 2013], une approximation de l'*Earth Mover's Distance* (EMD). L'EMD est une distance adaptée à la comparaison d'histogrammes et de distributions basée sur la formulation du problème de transport optimal. Si on fixe un coût pour déplacer de l'information d'un bin d'histogramme à un autre, on peut alors définir une distance entre histogrammes comme solution du transport optimal.

Un des avantages de l'EMD est qu'avec le choix de ce coût entre bins d'histogramme, on peut incorporer de la connaissance sur nos données afin de mieux adapter la distance à notre problème. En effet dans notre cas, nous pouvons ajuster le coût de déplacer de l'information d'un couple (intervalle d'amplitude ; fréquence) à un autre. Si on a M intervalles d'amplitude et F bandes de fréquence, alors le descripteur SPD $\mathbf{d} \in \mathbb{R}^{F \times M}$ pour une scène peut s'écrire ainsi :

$$\mathbf{d} = [d_{f_1 a_1}, ..., d_{f_1 a_M}, ..., d_{f_F a_1}, ..., d_{f_F a_M}] \; ; \tag{C.1}$$

où f_i est la bande de fréquence et a_j l'intervalle d'amplitude. Nous utilisons alors la fonction de coût suivante :

$$c(d_{f_k a_l}, d_{f_i a_j}) = |f_k - f_i|^p + |a_l - a_j|^q \; ; \tag{C.2}$$

où les paramètres p et q contrôlent l'influence de l'écart en fréquence et en amplitude des bins de $\mathbf{d}$. Puisque l'information fréquentielle des scènes n'a pas de structure particulière, la fonction de coût reste relativement générale et pénalise simplement l'écart entre bandes de fréquences et amplitudes. Pour finir, on doit calculer la matrice de coût $\mathbf{C}$ contenant le coût de passage entre chaque couple de bins d'histogramme. Lorsque nous utilisons la concaténation des HOG et des SPD, le coût de passage d'un bin des HOG à un bin des SPD est fixé à une valeur arbitrairement grande afin de ne pas autoriser de transfert entre les deux représentations.

Le principal défaut de l'EMD est sa complexité : même les bonnes implémentations ne supportent pas d'histogrammes de dimensions supérieures à quelques centaines de bins. Pour cela nous nous basons sur un travail plus récent sur le transport optimal [Cuturi, 2013] qui permet d'importants gains en temps de calcul par l'ajout d'une contrainte d'entropie au problème. La distance optimale dans ce cas est appelée distance de Sinkhorn et correspond à une borne supérieure de l'EMD. Nous nous reprenons donc cette approximation afin de construire le noyau de Sinkhorn pour la classification. La fonction du noyau de Sinkhorn s'exprime de manière similaire à celle du noyau Gaussien en prenant l'approximation de l'EMD à la place de la distance euclidienne :

$$k(\mathbf{d}, \mathbf{d}^{'}) = e^{-\frac{S(\mathbf{d}, \mathbf{d}^{'})}{\sigma^2}}. \tag{C.3}$$

Ici, $S(\mathbf{d}, \mathbf{d}^{'})$ est la distance de Sinkhorn entre deux vecteurs de descripteurs différents $\mathbf{d}$ et $\mathbf{d}^{'}$.

C.2 Comparaison de l'impact du choix du noyau

Nous effectuons l'évaluation expérimentale de l'impact du choix du noyau uniquement sur la base du LITIS. Cette base était la seule base disponible parmi les quatre bases utilisées dans ce

Noyau Gaussien avec la CQT				
	Précision	Rappel	F1 Score	Accuracy
HOG [Rakotomamonjy et Gasso, 2015]	91.7	-	-	-
HOG	91.2	90.2	90.5	91.2
SPD	90.8	89.2	89.7	90.2
HOG + SPD	93.3	92.5	**92.8**	**93.4**
Noyau de Sinkhorn avec la CQT				
	Précision	Rappel	F1 Score	Accuracy
HOG	91.4	90.3	90.7	91.3
SPD	88.7	86.9	87.4	88.6
HOG + SPD	92.3	90.6	**91.4**	**92.3**

TABLEAU C.1 – Comparaison des descripteurs d'images et noyaux SVMs pour la classification de scènes du LITIS.

chapitre à la période de nos travaux sur le choix du noyau [Bisot et al., 2015] . Les performances des différents descripteurs d'images ainsi que leur combinaison sont présentées pour l'utilisation des deux noyaux sur la base du LITIS dans le Tableau C.1. La première partie du tableau contient les résultats obtenus avec les descripteurs d'images classifiés en utilisant le noyau Gaussien. La méthode de référence sur cette base de données est l'utilisation de HOG proposée par Rakotomamonjy et Gasso [2015] dont les résultats sont uniquement représentés par la précision. Nous pouvons remarquer que les SPD, dans leur formulation originale, ne permettent pas d'améliorer les performances. En revanche en les combinant avec les HOG, nous obtenons un gain en performance sur les 4 métriques. C'est un premier pas dans la confirmation de l'hypothèse qu'ils décrivent des aspects complémentaires des scènes sonores. Nous laissons cet aspect de côté pour le moment, nous étudierons plus en détail par la suite l'intérêt de la combinaison en comparant les performances sur les 4 bases de données.

La construction du noyau de Sinkhorn en utilisant la fonction de coût proposée équation (C.2) nécessite le réglage de plusieurs hyper-paramètres en plus d'être coûteuse en temps de calcul pour estimer la distance. Comme on peut l'observer dans le Tableau C.1, l'utilisation du noyau de Sinkhorn n'apporte aucun gain net en performance, ce quels que soient la métrique ou les descripteurs utilisés. Pour les SPD, le noyau de Sinkhorn dégrade même fortement les performances. Les performances pourraient être améliorées en choisissant un coût différent ou en construisant un noyau différent pour les deux descripteurs en utilisant de l'apprentissage de noyaux multiples. Cependant, le coût en temps de calcul rend l'exploration de différentes solutions difficile. Les premiers résultats n'étant pas prometteurs nous garderons uniquement l'utilisation du noyau Gaussien pour les SVM dans la suite de ce chapitre.

Bibliographie

Abdel-Hamid, O., A.-R. Mohamed, H. Jiang, L. Deng, G. Penn et D. Yu. 2014, «Convolutional neural networks for speech recognition», *IEEE Transactions on audio, speech, and language processing*, vol. 22, n° 10, p. 1533–1545. (page 32)

Adavanne, S., P. Pertilä et T. Virtanen. 2017, «Sound event detection using spatial features and convolutional recurrent neural network», *arXiv preprint arXiv :1706.02291*. (page 26)

Adavanne, S. et T. Virtanen. 2017, «A report on sound event detection with different binaural features», rapport de recherche, DCASE2017 Challenge. (page 26)

Amiriparian, S., M. Freitag, N. Cummins et B. Schuller. 2017, «Sequence to sequence autoencoders for unsupervised representation learning from audio», dans *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*. (page 21, 22)

Aucouturier, J.-J., B. Defreville et F. Pachet. 2007, «The bag-of-frames approach to audio pattern recognition : A sufficient model for urban soundscapes but not for polyphonic music», *The Journal of the Acoustical Society of America*, vol. 122, n° 2, p. 881–891. (page 17)

Babagholami-Mohamadabadi, B., A. Jourabloo, M. Zolfaghari et M. Manzuri-Shalmani. 2013, «Bayesian supervised dictionary learning», dans *Proc. UAI Conference on Application Workshops : Big Data meet Complex Models and Models for Spatial, Temporal and Network Data-Volume 1024*, p. 11–19. (page 72)

Barchiesi, D., D. Giannoulis, D. Stowel et M. D. Plumbley. 2015, «Acoustic scene classification : Classifying environments from the sounds they produce», *IEEE Signal Processing Magazine*, vol. 32, n° 3, p. 16–34. (page 15, 35)

Battaglino, D., L. Lepauloux, L. Pilati et N. Evansi. 2015, «Acoustic context recognition using local binary pattern codebooks», *Porc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. (page 16, 18, 22, 33, 37, 40)

Bello, J. P., C. Mydlarz et J. Salamon. 2018, «Sound analysis in smart cities», dans *Computational Analysis of Sound Scenes and Events*, Springer, p. 373–397. (page 7)

Benetos, E., G. Lafay, M. Lagrange et M. D. Plumbley. 2017, «Polyphonic sound event tracking using linear dynamical systems», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, n° 6, p. 1266–1277. (page 20, 23, 73)

Benetos, E., M. Lagrange et S. Dixon. 2012, «Characterisation of acoustic scenes using a temporally constrained shift-invariant model», dans *Proc. Digital Audio Effects*. (page 20, 22, 23, 49, 72)

Benetos, E., M. Lagrange, M. D. Plumbley et al.. 2016, «Detection of overlapping acoustic events using a temporally-constrained probabilistic model», dans *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6450–6454. (page 20, 50, 62, 73)

Benetos, E., D. Stowell et M. D. Plumbley. 2018, «Approaches to complex sound scene analysis», dans *Computational Analysis of Sound Scenes and Events*, Springer, p. 215–242. (page 23)

Bengio, Y., P. Lamblin, D. Popovici et H. Larochelle. 2007, «Greedy layer-wise training of deep networks», *Proc. Advances in Neural Information Processing Systems (NIPS)*, vol. 19, p. 153. (page 97)

Bisot, V., S. Essid et G. Richard. 2015, «Hog and subband power distribution image features for acoustic scene classification», dans *Proc. European Signal Processing Conference (EUSIPCO)*. (page 30, 37, 40, 136, 137)

Bisot, V., S. Essid et G. Richard. 2017a, «Overlapping sound event detection with supervised non-negative matrix factorization», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 31–35. (page 21, 44, 70)

Bisot, V., R. Serizel, S. Essid et G. Richard. 2016a, «Acoustic scene classification with matrix factorization for unsupervised feature learning», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (page 44, 49, 57)

Bisot, V., R. Serizel, S. Essid et G. Richard. 2016b, «Supervised nonnegative matrix factorization for acoustic scene classification», *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE)*. (page 27, 70)

Bisot, V., R. Serizel, S. Essid et G. Richard. 2017b, «Feature learning with matrix factorization applied to acoustic scene classification», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, n° 6, p. 1216–1229, ISSN 2329-9290. (page 20, 44, 70)

Bisot, V., R. Serizel, S. Essid et G. Richard. 2017c, «Nonnegative feature learning methods for acoustic scene classification», dans *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*. (page 70)

Boersma, P. 2006, «Praat : doing phonetics by computer», *http ://www. praat. org/*. (page 14)

Brown, J. C. 1991, «Calculation of a constant Q spectral transform», *J Acoust Soc Am*, vol. 89, n° 1, p. 425–434. (page 16, 30, 31)

Bugalho, M., J. Portelo, I. Trancoso, T. Pellegrini et A. Abad. 2009, «Detecting audio events for semantic video search.», dans *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, p. 1151–1154. (page 6)

Bui, M.-Q., V.-H. Duong, S. Mathulaprangsan, B.-T. Pham, W.-J. Lee et J.-C. Wang. 2016, «A survey of polyphonic sound event detection based on non-negative matrix factorization», dans *Proc. International Computer Symposium (ICS)*, p. 351–354. (page 20, 50)

Cakir, E., T. Heittola, H. Huttunen et T. Virtanen. 2015a, «Multi-label vs. combined single-label sound event detection with deep neural networks», dans *Proc. European Signal Processing Conference (EUSIPCO)*, p. 2551–2555. (page 24, 90)

Cakir, E., T. Heittola, H. Huttunen et T. Virtanen. 2015b, «Polyphonic sound event detection using multi label deep neural networks», dans *Proc. International Joint Conference on Neural Networks (IJCNN)*, p. 1–7. (page 23, 24, 64, 66, 88, 90, 111)

Cakir, E., E. C. Ozan et T. Virtanen. 2016, «Filterbank learning for deep neural network based polyphonic sound event detection», dans *Proc. International Joint Conference on Neural Networks (IJCNN)*, p. 3399–3406. (page 91)

Cakir, E., G. Parascandolo, T. Heittola, H. Huttunen, T. Virtanen et al.. 2017, «Convolutional recurrent neural networks for polyphonic sound event detection», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, n° 6, p. 1291–1303. (page 25, 33, 63, 64, 66, 87, 88, 91, 92, 106, 107, 108, 109, 110, 111)

Cauchi, B. 2011, *Non-negative matrix factorization applied to auditory scene classification*, mémoire de maîtrise, ATIAM (UPMC / IRCAM / TELECOM ParisTech). (page 20, 49)

Cho, K., B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk et Y. Bengio. 2014, «Learning phrase representations using rnn encoder-decoder for statistical machine translation», *arXiv preprint arXiv :1406.1078*. (page 24, 95)

Chollet, F. 2015, «keras», https://github.com/fchollet/keras. (page 103, 109)

Chu, S., S. Narayanan et C.-C. J. Kuo. 2009, «Environmental sound recognition with time–frequency audio features», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, n° 6, p. 1142–1158. (page 17, 18, 22)

Chu, S., S. Narayanan, C.-C. J. Kuo et M. J. Mataric. 2006, «Where am I ? scene recognition for mobile robots using audio features», dans *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, p. 885–888. (page 7)

Cichocki, A. et S.-i. Amari. 2010, «Families of alpha-beta-and gamma-divergences : Flexible and robust measures of similarities», *Entropy*, vol. 12, n° 6, p. 1532–1568. (page 48)

Clavel, C., T. Ehrette et G. Richard. 2005, «Events detection for an audio-based surveillance system», dans *Proc. IEEE International Conference on Multimedia and Expo (ICME)*, p. 1306–1309. (page 17)

Cuturi, M. 2013, «Sinkhorn distances : Lightspeed computation of optimal transport», dans *Proc. Advances in Neural Information Processing Systems (NIPS)*, p. 2292–2300. (page 37, 136)

Dai, W., C. Dai, S. Qu, J. Li et S. Das. 2017, «Very deep convolutional neural networks for raw waveforms», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 421–425. (page 91)

Dalal, N. et B. Triggs. 2005, «Histograms of oriented gradients for human detection», dans *Proc. Computer Vision and Pattern Recognition (CVPR)*, vol. 1, p. 886–893. (page 34)

Davis, S. et P. Mermelstein. 1980, «Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences», *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, n° 4, p. 357–366. (page 17)

De Cheveigné, A. et H. Kawahara. 2002, «Yin, a fundamental frequency estimator for speech and music», *The Journal of the Acoustical Society of America*, vol. 111, n° 4, p. 1917–1930. (page 14)

Dennis, J., H. D. Tran et E. S. Chng. 2013, «Image feature representation of the subband power distribution for robust sound event classification», *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 21, n° 2, p. 367–377. (page 8, 16, 18, 22, 33, 34, 36)

Dennis, J., H. D. Tran et E. S. Chng. 2014, «Analysis of spectrogram image methods for sound event classification», dans *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*. (page 18, 33, 34)

Dikmen, O. et A. Mesaros. 2013, «Sound event detection using non-negative dictionaries learned from annotated overlapping events», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 1–4. (page 20)

Eggert, J. et E. Körner. 2004, «Sparse coding and nmf», dans *Proc. International Joint Conference on Neural Networks (IJCNN)*, vol. 4, p. 2529–2533. (page 52)

Eghbal-Zadeh, H., B. Lehner, M. Dorfer et G. Widmer. 2016, «CP-JKU submissions for DCASE-2016 : a hybrid approach using binaural i-vectors and deep convolutional neural networks», rapport de recherche, DCASE2016 Challenge. (page 17, 19, 25, 26, 27, 85, 91, 105)

Elizalde, B., A. Kumar, A. Shah, R. Badlani, E. Vincent, B. Raj et I. Lane. 2016, «Experiments on the DCASE challenge 2016 : Acoustic scene classification and sound event detection in real life recording», dans *Proc. of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop*, p. 20–24. (page 65, 66, 86)

Elman, J. L. 1990, «Finding structure in time», *Cognitive science*, vol. 14, n° 2, p. 179–211. (page 95)

Eronen, A. J., V. T. Peltonen, J. T. Tuomi, A. P. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho et J. Huopaniemi. 2006, «Audio-based context recognition», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, n° 1, p. 321–329. (page 15)

Fant, G. 1968, «Analysis and synthesis of speech processes», dans *Manual of phonetics*, édité par B. Malmberg, chap. 8, North-Holland Publishing Company Amsterdam, p. 173–277. (page 31)

Févotte, C., N. Bertin et J.-L. Durrieu. 2009, «Nonnegative matrix factorization with the itakura-saito divergence : With application to music analysis», *Neural computation*, vol. 21, n° 3, p. 793–830. (page 48)

Févotte, C. et J. Idier. 2011, «Algorithms for nonnegative matrix factorization with the β-divergence», *Neural Computation*, vol. 23, n° 9, p. 2421–2456. (page 49)

Fletcher, H. 1940, «Auditory patterns», *Reviews of modern physics*, vol. 12, n° 1, p. 47. (page 16)

Font, F., G. Roma et X. Serra. 2018, «Sound sharing and retrieval», dans *Computational Analysis of Sound Scenes and Events*, Springer, p. 279–301. (page 6)

Gangeh, M. J., A. Ghodsi et M. S. Kamel. 2013, «Kernelized supervised dictionary learning», *IEEE Transactions on Signal Processing*, vol. 61, n° 19, p. 4753–4767. (page 72)

Geiger, J. T., B. Schuller et G. Rigoll. 2013, «Large-scale audio feature extraction and SVM for acoustic scene classification», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. (page 6, 17, 21, 22)

Gemmeke, J. F., D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal et M. Ritter. 2017, «Audio set : An ontology and human-labeled dataset for audio events», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (page 3, 6, 25, 27)

Gemmeke, J. F., L. Vuegene, P. Karsmakers, B. Vanrumste et al.. 2013, «An exemplar-based NMF approach to audio event detection», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 1–4. (page 73)

Gencoglu, O., T. Virtanen et H. Huttunen. 2014, «Recognition of acoustic events using deep neural networks», dans *Proc. European Signal Processing Conference (EUSIPCO)*, p. 506–510. (page 90)

Ghoraani, B. et S. Krishnan. 2011, «Time–frequency matrix feature extraction and classification of environmental audio signals», *IEEE transactions on audio, speech, and language processing*, vol. 19, n° 7, p. 2197–2209. (page 20)

Giannoulis, D., D. Stowell, E. Benetos, M. Rossignol, M. Lagrange et M. D. Plumbley. 2013, «A database and challenge for acoustic scene classification and event detection», dans *Proc. European Signal Processing Conference (EUSIPCO)*. (page 7, 38)

Giordano, B. L. et S. McAdams. 2006, «Material identification of real impact sounds : Effects of size variation in steel, glass, wood, and plexiglass plates», *The Journal of the Acoustical Society of America*, vol. 119, n° 2, p. 1171–1181. (page 14)

Glasberg, B. R. et B. C. Moore. 1990, «Derivation of auditory filter shapes from notched-noise data», *Hearing research*, vol. 47, n° 1, p. 103–138. (page 16)

Goodfellow, I., Y. Bengio et A. Courville. 2016, *Deep learning*, MIT press. (page 92, 101)

Graves, A. et al.. 2012, *Supervised sequence labelling with recurrent neural networks*, vol. 385, Springer. (page 96)

Grey, J. M. et J. A. Moorer. 1977, «Perceptual evaluations of synthesized musical instrument tones», *The Journal of the Acoustical Society of America*, vol. 62, n° 2, p. 454–462. (page 14)

Guan, Y. et J. G. Dy. 2009, «Sparse probabilistic principal component analysis», dans *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, p. 185–192. (page 52)

Guastavino, C. 2018, «Everyday sound categorization», dans *Computational Analysis of Sound Scenes and Events*, Springer, p. 183–213. (page 3)

Gygi, B., G. R. Kidd et C. S. Watson. 2004, «Spectral-temporal factors in the identification of environmental sounds», *The Journal of the Acoustical Society of America*, vol. 115, n° 3, p. 1252–1265. (page 15)

Hayashi, T., S. Watanabe, T. Toda, T. Hori, J. Le Roux et K. Takeda. 2017, «BLSTM-HMM hybrid system combined with sound activity detection network for polyphonic sound event detection», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 766–770. (page 23)

Heittola, T., A. Mesaros, A. Eronen et T. Virtanen. 2010, «Audio context recognition using audio event histograms», dans *Proc. European Signal Processing Conference (EUSIPCO)*, p. 1272–1276. (page 23, 26)

Heittola, T., A. Mesaros, A. Eronen et T. Virtanen. 2013a, «Context-dependent sound event detection», *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2013, n° 1, p. 1. (page 26)

Heittola, T., A. Mesaros, T. Virtanen et A. Eronen. 2011, «Sound event detection in multisource environments using source separation», dans *Proc. CHiME*, p. 36–40. (page 26)

Heittola, T., A. Mesaros, T. Virtanen et M. Gabbouj. 2013b, «Supervised model training for overlapping sound events based on unsupervised source separation.», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 8677–8681. (page 26)

Hershey, J. R., J. Le Roux et F. Weninger. 2014, «Deep unfolding : Model-based inspiration of novel deep architectures», *arXiv preprint arXiv :1409.2574*. (page 73, 98)

Hershey, S., S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold et al.. 2017, «CNN architectures for large-scale audio classification», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 131–135. (page 25, 32)

Hinton, G. E., S. Osindero et Y.-W. Teh. 2006, «A fast learning algorithm for deep belief nets», *Neural computation*, vol. 18, n° 7, p. 1527–1554. (page 97)

Hinton, G. E., N. Srivastava, A. Krizhevsky, I. Sutskever et R. R. Salakhutdinov. 2012, «Improving neural networks by preventing co-adaptation of feature detectors», *arXiv preprint arXiv :1207.0580*. (page 93, 102)

Hochreiter, S. et J. Schmidhuber. 1997, «Long short-term memory», *Neural computation*, vol. 9, n° 8, p. 1735–1780. (page 24, 95)

Imoto, K., Y. Ohishi, H. Uematsu et H. Ohmuro. 2013, «Acoustic scene analysis based on latent acoustic topic and event allocation», dans *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, p. 1–6. (page 26)

Ioffe, S. et C. Szegedy. 2015, «Batch normalization : Accelerating deep network training by reducing internal covariate shift», dans *Proc. International Conference on Machine Learning (ICML)*, p. 448–456. (page 109)

Itakura, F. 1968, «Analysis synthesis telephony based on the maximum likelihood method», dans *International Congress on Acoustics, 1968*, p. 17–20. (page 48)

Jee-Weon, J., H. Hee-Soo, Y. IL-Ho, Y. Sung-Hyun, S. Hye-Jin et Y. Ha-Jin. 2017, «DNN-based audio scene classification for DCASE 2017 : Dual inputfeatures, balancing cost, and stochastic data duplication», rapport de recherche, DCASE2017 Challenge. (page 91)

Jiang, Z., Z. Lin et L. S. Davis. 2011, «Learning a discriminative dictionary for sparse coding via label consistent K-SVD», dans *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 1697–1704. (page 72)

Joder, C., S. Essid et G. Richard. 2009, «Temporal integration for audio classification with application to musical instrument classification», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, n° 1, p. 174–186. (page 21)

Kim, S., S. Narayanan et S. Sundaram. 2009, «Acoustic topic model for audio information retrieval», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 37–40. (page 26)

Kingma, D. et J. Ba. 2014, «Adam : A method for stochastic optimization», *arXiv preprint arXiv :1412.6980*. (page 110)

Komatsu, T., Y. Senda et R. Kondo. 2016a, «Acoustic event detection based on non-negative matrix factorization with mixtures of local dictionaries and activation aggregation», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 2259–2263. (page 20, 21, 22)

Komatsu, T., T. Toizumi, R. Kondo et Y. Senda. 2016b, «Acoustic event detection method using semi-supervised non-negative matrix factorization with a mixture of local dictionaries», rapport de recherche, DCASE2016Challenge. (page 20)

Kong, Q., Y. Xu, W. Wang et M. D. Plumbley. 2017a, «Audio set classification with attention model : A probabilistic perspective», *arXiv preprint arXiv :1711.00927*. (page 28)

Kong, Q., Y. Xu, W. Wang et M. D. Plumbley. 2017b, «A joint detection-classification model for audio tagging of weakly labelled data», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 641–645. (page 27)

Krijnders, J. et G. A. T. Holt. 2013, «A tone-fit feature representation for scene classification», *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*. (page 21)

Krstulović, S. 2018, «Audio event recognition in the smart home», dans *Computational Analysis of Sound Scenes and Events*, Springer, p. 335–371. (page 7)

Kullback, S. et R. A. Leibler. 1951, «On information and sufficiency», *The annals of mathematical statistics*, vol. 22, n° 1, p. 79–86. (page 48)

Kumar, A. et B. Raj. 2016a, «Audio event detection using weakly labeled data», dans *Proc. ACM International Conference on Multimedia*, p. 1038–1047. (page 27)

Kumar, A. et B. Raj. 2016b, «Weakly supervised scalable audio content analysis», dans *IEEE International Conference on Multimedia and Expo (ICME)*, p. 1–6. (page 27)

Kumar, B. V., I. Kotsia et I. Patras. 2012, «Max-margin non-negative matrix factorization», *Image and Vision Computing*, vol. 30, n° 4, p. 279–291. (page 72)

Kunkler-Peck, A. J. et M. Turvey. 2000, «Hearing shape.», *Journal of Experimental psychology : human perception and performance*, vol. 26, n° 1, p. 279. (page 14)

Le Roux, J., J. R. Hershey et F. Weninger. 2015a, «Deep NMF for speech separation», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 66–70. (page 73, 98)

Le Roux, J., F. J. Weninger et J. R. Hershey. 2015b, «Sparse NMF half-baked or well done ?», rapport de recherche, Mitsubishi Electric Research Labs (MERL). (page 52, 82)

LeCun, Y., L. Bottou, Y. Bengio et P. Haffner. 1998, «Gradient-based learning applied to document recognition», *Proceedings of the IEEE*, vol. 86, n° 11, p. 2278–2324. (page 95)

Lee, D. D. et H. S. Seung. 1999, «Learning the parts of objects by non-negative matrix factorization», *Nature*, vol. 401, n° 6755, p. 788–791. (page 7, 20, 47, 49)

Lee, K., D. Lee, S. Lee et Y. Han. 2017, rapport de recherche, DCASE2017 Challenge. (page 27)

Lehner, B., H. Eghbal-Zadeh, M. Dorfer, F. Korzeniowski, K. Koutini et G. Widmer. 2017, «Classifying short acoustic scenes with I-vectors and CNNs : Challenges and optimisations for the 2017 DCASE ASC task», rapport de recherche, DCASE2017 Challenge. (page 27, 106)

Lemaitre, G. et L. M. Heller. 2012, «Auditory perception of material is fragile while action is strikingly robust», *The Journal of the Acoustical Society of America*, vol. 131, nº 2, p. 1337–1348. (page 14)

Lewicki, M. S. 2002, «Efficient coding of natural sounds», *Nature neuroscience*, vol. 5, nº 4, p. 356–363. (page 14)

Li, D., J. Tam et D. Toub. 2013, «Auditory scene classification using machine learning techniques», *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*. (page 22)

Li, J., W. Dai, F. Metze, S. Qu et S. Das. 2017, «A comparison of deep learning methods for environmental sound», *arXiv preprint arXiv :1703.06902*. (page 17, 96)

Lostanlen, V. et J. Andén. 2016, «Binaural scene classification with wavelet scattering», rapport de recherche, DCASE2016 Challenge. (page 18)

Lu, T., G. Wang et F. Su. 2015, «Context-based environmental audio event recognition for scene understanding», *Multimedia Systems*, vol. 21, nº 5, p. 507–524. (page 26)

Mackey, L. W. 2009, «Deflation methods for sparse PCA», dans *Proc. Advances in Neural Information Processing Systems (NIPS)*, p. 1017–1024. (page 52)

Mairal, J., F. Bach et J. Ponce. 2012, «Task-driven dictionary learning», *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, nº 4, p. 791–804. (page 8, 72, 74, 75, 77, 101)

Mairal, J., F. Bach, J. Ponce et G. Sapiro. 2009a, «Online dictionary learning for sparse coding», dans *Proc. International Conference on Machine lLearning (ICML)*, p. 689–696. (page 52, 79)

Mairal, J., F. Bach, J. Ponce et G. Sapiro. 2010, «Online learning for matrix factorization and sparse coding», *The Journal of Machine Learning Research*, vol. 11, p. 19–60. (page 75)

Mairal, J., F. Bach, J. Ponce, G. Sapiro et A. Zisserman. 2008, «Discriminative learned dictionaries for local image analysis», dans *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 1–8. (page 71)

Mairal, J., J. Ponce, G. Sapiro, A. Zisserman et F. Bach. 2009b, «Supervised dictionary learning», dans *Proc. Advances in Neural Information Processing Systems (NIPS)*, p. 1033–1040. (page 71)

Mallat, S. 1999, *A wavelet tour of signal processing*, Academic press. (page 30)

Mallat, S. 2012, «Group invariant scattering», *Communications on Pure and Applied Mathematics*, vol. 65, nº 10, p. 1331–1398. (page 18)

Mallat, S. G. 1989, «Multifrequency channel decompositions of images and wavelet models», *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, nº 12, p. 2091–2110. (page 18)

Marchi, E., D. Tonelli, X. Xu, F. Ringeval, J. Deng, S. Squartini et B. Schuller. 2016, rapport de recherche, DCASE2016 Challenge. (page 17, 85, 91, 96, 105)

Mathieu, B., S. Essid, T. Fillon, J. Prado et G. Richard. 2010, «Yaafe, an easy to use and efficient audio feature extraction software.», dans *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, p. 441–446. (page 38, 54)

McAdams, S., S. Winsberg, S. Donnadieu, G. De Soete et J. Krimphoff. 1995, «Perceptual scaling of synthesized musical timbres : Common dimensions, specificities, and latent subject classes», *Psychological research*, vol. 58, n° 3, p. 177–192. (page 14)

McDermott, J. H. et E. P. Simoncelli. 2011, «Sound texture perception via statistics of the auditory periphery : evidence from sound synthesis», *Neuron*, vol. 71, n° 5, p. 926–940. (page 15)

McFee, B. 2018, «Statistical methods for scene and event classification», dans *Computational Analysis of Sound Scenes and Events*, Springer, p. 103–146. (page 92)

Meddis, R. et L. O'Mard. 1997, «A unitary model of pitch perception», *The Journal of the Acoustical Society of America*, vol. 102, n° 3, p. 1811–1820. (page 14)

Mesaros, A., T. Heittola, O. Dikmen et T. Virtanen. 2015, «Sound event detection in real life recordings using coupled matrix factorization of spectral representations and class activity annotations», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (page 20, 21, 50, 62, 73)

Mesaros, A., T. Heittola, A. Diment, B. Elizalde, A. Shah, E. Vincent, B. Raj et T. Virtanen. 2017a, «Dcase 2017 challenge setup : Tasks, datasets and baseline system», dans *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*. (page 5, 7, 38)

Mesaros, A., T. Heittola, A. Eronen et T. Virtanen. 2010, «Acoustic event detection in real life recordings», dans *Proc. European Signal Processing Conference (EUSIPCO)*, p. 1267–1271. (page 23)

Mesaros, A., T. Heittola et A. Klapuri. 2011, «Latent semantic analysis in sound event detection», dans *Proc. European Signal Processing Conference (EUSIPCO)*, p. 1307–1311. (page 23)

Mesaros, A., T. Heittola et T. Virtanen. 2016a, «Metrics for polyphonic sound event detection», *Applied Sciences*, vol. 6, n° 6, p. 162. (page 64, 132)

Mesaros, A., T. Heittola et T. Virtanen. 2016b, «Tut database for acoustic scene classification and sound event detection», dans *Proc. of European Signal Processing Conference*. (page 5, 7, 19, 38, 50, 63, 64, 65, 66, 85, 86, 88, 105, 111, 126)

Mesaros, A., T. Heittola et T. Virtanen. 2017b, «Assessment of human and machine performance in acoustic scene classification : Dcase 2016 case study», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 1–5. (page 15)

Mesnil, G., X. He, L. Deng et Y. Bengio. 2013, «Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding.», dans *Proc. Annual Conference of the International Speech Communication Association (Interspeech)*, p. 3771–3775. (page 108)

Moore, B. C. J., B. R. Glasberg et T. Baer. 1997, «A model for the prediction of thresholds, loudness, and partial loudness», *Journal of the Audio Engineering Society*, vol. 45, n° 4, p. 224–240. (page 14)

Mun, S., S. Park, D. Han et H. Ko. 2017, «Generative adversarial network based acoustic scene training set augmentation and selection using svm hyper-plane», rapport de recherche, DCASE2017 Challenge. (page 21, 22, 27, 91, 106)

Muroi, T., R. Takashima, T. Takiguchi et Y. Ariki. 2009, «Gradient-based acoustic features for speech recognition», dans *International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, p. 445–448. (page 34)

Mydlarz, C., J. Salamon et J. P. Bello. 2017, «The implementation of low-cost urban acoustic monitoring devices», *Applied Acoustics*, vol. 117, p. 207–218. (page 7)

Nair, V. et G. E. Hinton. 2010, «Rectified linear units improve restricted boltzmann machines», dans *Proc. International Conference on Machine lLearning (ICML)*, p. 807–814. (page 93)

O'Grady, P. D. et B. A. Pearlmutter. 2006, «Convolutive non-negative matrix factorisation with a sparseness constraint», dans *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, p. 427–432. (page 53)

Olah, C. 2015, «Understanding LSTM networks», colah.github.io/posts/2015-08-Understanding-LSTMs/. (page 96)

Olivetti, E. 2013, «The wonders of the normalized compression dissimilarity representation», *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*. (page 22)

Overath, T., S. Kumar, L. Stewart, K. von Kriegstein, R. Cusack, A. Rees et T. D. Griffiths. 2010, «Cortical mechanisms for the segregation and representation of acoustic textures», *Journal of Neuroscience*, vol. 30, n° 6, p. 2070–2076. (page 15)

Oxenham, A. J., J. G. Bernstein et H. Penagos. 2004, «Correct tonotopic representation is necessary for complex pitch perception», *Proc. National Academy of Sciences of the United States of America*, vol. 101, n° 5, p. 1421–1425. (page 14)

Parascandolo, G., H. Huttunen et T. Virtanen. 2016, «Recurrent neural networks for polyphonic sound event detection in real life recordings», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6440–6444. (page 24, 90, 108)

Parizet, E., E. Guyader et V. Nosulenko. 2008, «Analysis of car door closing sound quality», *Applied acoustics*, vol. 69, n° 1, p. 12–22. (page 14)

Park, S., S. Mun, Y. Lee et H. Ko. 2016, «Score fusion of classification systems for acoustic scene classification», rapport de recherche, DCASE2016 Challenge. (page 85, 91, 96, 105)

Pascanu, R., T. Mikolov et Y. Bengio. 2013, «On the difficulty of training recurrent neural networks», dans *Proc. International Conference on Machine Learning (ICML)*, p. 1310–1318. (page 95)

Pauca, V. P., F. Shahnaz, M. W. Berry et R. J. Plemmons. 2004, «Text mining using non-negative matrix factorizations», dans *Proc. SIAM International Conference on Data Mining*, p. 452–456. (page 47)

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al.. 2011, «Scikit-learn : Machine learning in python», *The Journal of Machine Learning Research*, vol. 12, p. 2825–2830. (page 55, 79)

Peeters, G. 2004, «A large set of audio features for sound description (similarity and classification) in the cuidado project», . (page 14, 17, 34)

Peltonen, V., J. Tuomi, A. Klapuri, J. Huopaniemi et T. Sorsa. 2002, «Computational auditory scene recognition», dans *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. (page 17)

Peltonen, V. T., A. J. Eronen, M. P. Parviainen et A. P. Klapuri. 2001, «Recognition of everyday auditory scenes : Potentials, latencies and cues», dans *Proc. Audio Engineering Society Convention (ASE)*. (page 15)

Petetin, Y., C. Laroche et A. Mayoue. 2015, «Deep neural networks for audio scene recognition», dans *Proc. European Signal Processing Conference (EUSIPCO)*. (page 17, 24)

Piczak, K. J. 2015a, «Environmental sound classification with convolutional neural networks», dans *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, p. 1–6. (page 25, 32, 91)

Piczak, K. J. 2015b, «Esc : Dataset for environmental sound classification», dans *Proc. ACM International Conference on Multimedia*, p. 1015–1018. (page 16)

Piczak, K. J. 2017, «The details that matter : Frequency resolution of spectrograms in acoustic scene classification», rapport de recherche, DCASE2017 Challenge. (page 91)

Plinge, A., R. Grzeszick et G. A. Fink. 2014, «A bag-of-features approach to acoustic event detection», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 3704–3708. (page 19)

Plumbley, M. D., T. Blumensath, L. Daudet, R. Gribonval et M. E. Davies. 2010, «Sparse representations in audio and music : from coding to source separation», *Proceedings of the IEEE*, vol. 98, n° 6, p. 995–1005. (page 52)

Rakotomamonjy, A. 2017, «Supervised representation learning for audio scene classification», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, n° 6, p. 1253–1265. (page 20, 21, 33, 34, 73, 91)

Rakotomamonjy, A. et G. Gasso. 2015, «Histogram of gradients of time-frequency representations for audio scene classification», *IEEE Transactions on Audio, Speech and Language Processing*, vol. 23, n° 1, p. 142–153. (page 8, 16, 18, 22, 33, 34, 35, 37, 39, 125, 137)

Ramirez, I., P. Sprechmann et G. Sapiro. 2010, «Classification and clustering via dictionary learning with structured incoherence and shared features», dans *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 3501–3508. (page 71)

Ren, Z., V. Pandit, K. Qian, Z. Yang, Z. Zhang et B. Schuller. 2017, «Deep sequential image features on acoustic scene classification», dans *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE)*. (page 18)

Risset, J.-C. et D. L. Wessel. 1982, «Exploration of timbre by analysis and synthesis», *The psychology of music*, p. 26–58. (page 14)

Roma, G., W. Nogueira et P. Herrera. 2013, «Recurrence quantification analysis features for environmental sound recognition», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. (page 21, 37)

Rosenblatt, F. 1958, «The perceptron : A probabilistic model for information storage and organization in the brain.», *Psychological review*, vol. 65, n° 6, p. 386. (page 93)

Rumelhart, D. E., G. E. Hinton, R. J. Williams et al.. 1988, «Learning representations by back-propagating errors», *Cognitive modeling*, vol. 5, n° 3, p. 1. (page 95, 101)

Sainath, T. N., O. Vinyals, A. Senior et H. Sak. 2015, «Convolutional, long short-term memory, fully connected deep neural networks», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 4580–4584. (page 108)

Salamon, J. et J. P. Bello. 2015a, «Feature learning with deep scattering for urban sound analysis», dans *Proc. European Signal Processing Conference (EUSIPCO)*, p. 724–728. (page 18, 21, 22)

Salamon, J. et J. P. Bello. 2015b, «Unsupervised feature learning for urban sound classification», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (page 38, 91)

Salamon, J. et J. P. Bello. 2017, «Deep convolutional neural networks and data augmentation for environmental sound classification», *IEEE Signal Processing Letters*, vol. 24, n° 3, p. 279–283. (page 27)

Salamon, J., J. P. Bello, A. Farnsworth et S. Kelling. 2017, «Fusing shallow and deep learning for bioacoustic bird species classification», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 141–145. (page 6, 27)

Salamon, J., C. Jacoby et J. P. Bello. 2014, «A dataset and taxonomy for urban sound research», dans *Proc. ACM International Conference on Multimedia*, Orlando, FL, USA. (page 3, 7, 128)

Schlüter, J. et T. Grill. 2015, «Exploring data augmentation for improved singing voice detection with neural networks.», dans *Proc. International Society for Music Information Retrieval Conference (ISMIR)*, p. 121–126. (page 26)

Schölkopf, B., A. Smolar et K.-R. Müller. 1998, «Nonlinear component analysis as a kernel eigenvalue problem», *Neural computation*, vol. 10, n° 5, p. 1299–1319. (page 52, 53)

Schuster, M. et K. K. Paliwal. 1997, «Bidirectional recurrent neural networks», *IEEE Transactions on Signal Processing*, vol. 45, n° 11, p. 2673–2681. (page 108)

Serizel, R., V. Bisot, S. Essid et G. Richard. 2016a, «Machine listening techniques as a complement to video image analysis in forensics», dans *Proc. International Conference on Image Processing (ICIP)*. (page 6)

Serizel, R., V. Bisot, S. Essid et G. Richard. 2017, «Supervised group nonnegative matrix factorisation with similarity constraints and applications to speaker identification», dans *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. (page 79)

Serizel, R., V. Bisot, S. Essid et G. Richard. 2018, «Acoustic features for environmental sound analysis», dans *Computational Analysis of Sound Scenes and Events*, Springer, p. 71–101. (page 7, 34)

Serizel, R., S. Essid et G. Richard. 2016b, «Group non-negative matrix factorisationwith speaker and session similarity constraints for speaker identification», dans *Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. (page 72, 79)

Serizel, R., S. Essid et G. Richard. 2016c, «Mini-batch stochastic approaches for accelerated multiplicative updates in nonnegative matrix factorisation with beta-divergence», dans *Proc. IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, p. 1–6. (page 49, 55)

Shannon, R. V., F.-G. Zeng, V. Kamath, J. Wygonski et M. Ekelid. 1995, «Speech recognition with primarily temporal cues», *Science*, vol. 270, n° 5234, p. 303. (page 15)

Sigtia, S., A. M. Stark, S. Krstulović et M. D. Plumbley. 2016, «Automatic environmental sound recognition : Performance versus computational cost», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 24, n° 11, p. 2096–2107. (page 7)

Smaragdis, P. 2004, «Non-negative matrix factor deconvolution ; extraction of multiple sound sources from monophonic inputs», dans *Independent Component Analysis and Blind Signal Separation*, Springer, p. 494–499. (page 20, 53)

Smaragdis, P. et J. C. Brown. 2003, «Non-negative matrix factorization for polyphonic music transcription», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 177–180. (page 47)

Smaragdis, P. et S. Venkataramani. 2017, «A neural network alternative to non-negative audio models», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 86–90. (page 98)

Sprechmann, P., A. M. Bronstein et G. Sapiro. 2014, «Supervised non-euclidean sparse nmf via bilevel optimization with applications to speech enhancement», dans *Proc. Joint Workshop on Hands-free Speech Communication and Microphone Arrays (HSCMA)*, p. 11–15. (page 73, 77, 86, 101)

Stevens, S. S., J. Volkmann et E. B. Newman. 1937, «A scale for the measurement of the psychological magnitude pitch», *The Journal of the Acoustical Society of America*, vol. 8, n° 3, p. 185–190. (page 16, 30, 31)

Stowell, D. 2018, «Computational bioacoustic scene analysis», dans *Computational Analysis of Sound Scenes and Events*, Springer, p. 303–333. (page 6)

Stowell, D. et D. Clayton. 2015, «Acoustic event detection for multiple overlapping similar sources», dans *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, p. 1–5. (page 23)

Susini, P., S. McAdams, S. Winsberg, I. Perry, S. Vieillard et X. Rodet. 2004, «Characterizing the sound quality of air-conditioning noise», *Applied Acoustics*, vol. 65, n° 8, p. 763–790. (page 14)

Tucker, S. et G. J. Brown. 2003, «Modelling the auditory perception of size, shape and material : Applications to the classification of transient sonar sounds», dans *Proc. Audio Engineering Society Convention (ASE)*. (page 14)

Valenti, M., A. Diment, G. Parascandolo, S. Squartini et T. Virtanen. 2016, «DCASE 2016 acoustic scene classification using convolutional neural networks», rapport de recherche, DCASE2016 Challenge. (page 25, 85, 91, 105)

Venkataramani, S., Y. C. Subakan et P. Smaragdis. 2017, «Neural network alternatives to convolutive audio models for source separation», *arXiv preprint arXiv :1709.07908*. (page 98)

Virtanen, T. 2007, «Monaural sound source separation by nonnegative matrix factorization with temporal continuity and sparseness criteria», *IEEE transactions on audio, speech, and language processing*, vol. 15, n° 3, p. 1066–1074. (page 47)

Virtanen, T., A. T. Cemgil et S. Godsill. 2008, «Bayesian extensions to non-negative matrix factorisation for audio signal modelling», dans *Acoustics, Speech and Signal Processing (ICASSP), 2008 IEEE International Conference on*, IEEE, p. 1825–1828. (page 48)

Virtanen, T., J. F. Gemmeke, B. Raj et P. Smaragdis. 2015, «Compositional models for audio processing : Uncovering the structure of sound mixtures», *IEEE Signal Processing Magazine*, vol. 32, n° 2, p. 125–144. (page 48)

Virtanen, T., M. D. Plumbley et D. Ellis. 2017, «Computational analysis of sound scenes and events», . (page 2, 7)

Vu, T. H. et J.-C. Wang. 2016, «Acoustic scene and event recognition using recurrent neural networks», rapport de recherche, DCASE2016 Challenge. (page 65, 66, 86)

Wang, B. et M. D. Plumbley. 2005, «Musical audio stream separation by non-negative matrix factorization», dans *Proc. UK Digital Music Research Network (DMRN) summer conf.*, p. 23–24. (page 47)

Wang, D. et G. J. Brown. 2006, *Computational auditory scene analysis : Principles, algorithms, and applications*, Wiley-IEEE press. (page 15)

Wang, Y., Y. Jia, C. Hu et M. Turk. 2004, «Fisher non-negative matrix factorization for learning local features», dans *Proc. Asian Conference on Computer Vision*. (page 72)

Wilson, K. W., B. Raj et P. Smaragdis. 2008, «Regularized non-negative matrix factorization with temporal dependencies for speech denoising», dans *Proc. International Speech Communication Association (ISCA)*. (page 47)

Wisdom, S., J. R. Hershey, J. Le Roux et S. Watanabe. 2016, «Deep unfolding for multichannel source separation», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (page 98)

Wisdom, S., T. Powers, J. Pitton et L. Atlas. 2017a, «Building recurrent networks by unfolding iterative thresholding for sequential sparse recovery», dans *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. (page 98, 99)

Wisdom, S., T. Powers, J. Pitton et L. Atlas. 2017b, «Deep recurrent nmf for speech separation by unfolding iterative thresholding», *arXiv preprint arXiv :1709.07124*. (page 98)

Xu, W., X. Liu et Y. Gong. 2003, «Document clustering based on non-negative matrix factorization», dans *Proc. International ACM SIGIR conference on Research and development in informaion retrieval*, p. 267–273. (page 47)

Xu, X., X. Chen et D. Yang. 2017a, «Acoustic scene classification using autoencoder», rapport de recherche, DCASE2017 Challenge. (page 21, 27)

Xu, Y., Q. Kong, Q. Huang, W. Wang et M. D. Plumbley. 2017b, «Attention and localization based on a deep convolutional recurrent model for weakly supervised audio tagging», *arXiv preprint arXiv :1703.06052*. (page 27)

Xu, Y., Q. Kong, W. Wang et M. D. Plumbley. 2017c, «Large-scale weakly supervised audio classification using gated convolutional neural network», *arXiv preprint arXiv :1710.00343*. (page 25, 28, 91, 106)

Yang, M., L. Zhang, X. Feng et D. Zhang. «Fisher discrimination dictionary learning for sparse representation», dans *Proc. IEEE International Conference on Computer Vision (ICCV)*. (page 72)

Yang, M., L. Zhang, J. Yang et D. Zhang. 2010, «Metaface learning for sparse representation based face recognition», dans *Proc. IEEE International Conference on Image Processing (ICIP)*, p. 1601–1604. (page 71)

Yang, W. et S. Krishnan. 2017, «Combining temporal features by local binary pattern for acoustic scene classification», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 25, nº 6, p. 1315–1321. (page 18, 33, 40)

Ye, J., T. Kobayashi, M. Murakawa et T. Higuchi. 2015, «Acoustic scene classification based on sound textures and events», dans *Proc. ACM International Conference on Multimedia*, p. 1291–1294. (page 19)

Y.Hang et J. Park. 2017, «Convolutional neural networks with binaural representations and background subtraction for acoustic scene classification», rapport de recherche, DCASE2017 Challenge. (page 21, 23, 26, 106)

Zafeiriou, S., A. Tefas, I. Buciu et I. Pitas. 2006, «Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification», *IEEE Transactions on Neural Networks*, vol. 17, nº 3, p. 683–695. (page 72)

Zhang, D., Z. Zhou et S. Chen. 2006, «Non-negative matrix factorization on kernels», dans *PRICAI 2006 : Trends in Artificial Intelligence*, Springer, p. 404–412. (page 52, 53)

Zhang, H., Y. Zhang et T. S. Huang. 2013, «Simultaneous discriminative projection and dictionary learning for sparse representation based classification», *Pattern Recognition*, vol. 46, nº 1, p. 346–354. (page 72)

Zhang, Q. et B. Li. 2010, «Discriminative k-svd for dictionary learning in face recognition», dans *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, p. 2691–2698. (page 72)

Zhao, S., T. N. T. Nguyen, W.-S. Gan et J. Douglas L. 2017, «ADSC submission for DCASE 2017 : Acoustic scene classification using deep residual convolutional neural networks», rapport de recherche, DCASE2017 Challenge. (page 91)

Zhou, Q. et Z. Feng. 2017, «Robust sound event detection through noise estimation and source separation using NMF», rapport de recherche, DCASE2017 Challenge. (page 20)

Zöhrer, M. et F. Pernkopf. 2016, «Gated recurrent networks applied to acoustic scene classification», dans *Proc. Workshop on the Detection and Classification of Acoustic Scenes and Events*, p. 115–119. (page 65, 66, 86)

Zou, H., T. Hastie et R. Tibshirani. 2006, «Sparse principal component analysis», *Journal of computational and graphical statistics*, vol. 15, nº 2, p. 265–286. (page 52)

Zue, V. 1985, «Notes on spectrogram reading», *Mass. Inst. Tech. Course*, vol. 6. (page 32)

Zwicker, E., H. Fastl, U. Widmann, K. Kurakata, S. Kuwano et S. Namba. 1991, «Program for calculating loudness according to din 45631 (iso 532b)», *Journal of the Acoustical Society of Japan*, vol. 12, n° 1, p. 39–42. (page 14)

Zwicker, E. et E. Terhardt. 1980, «Analytical expressions for critical-band rate and critical bandwidth as a function of frequency», *The Journal of the Acoustical Society of America*, vol. 68, n° 5, p. 1523–1525. (page 16)

Remerciements

Mes premiers remerciements vont à mes deux directeurs de thèse Slim et Gaël. Vous m'avez fait confiance et mis en confiance dès le début, vous avez été disponibles et à l'écoute jusqu'à la fin. Votre expérience, vos conseils ainsi que votre optimisme et votre bonne humeur à toute épreuve ont été et resterons une source d'inspiration. C'était un privilège et un réel plaisir de travailler avec vous.

Merci aux membres du jury Jimena Royo-Letelier, Annamaria Mesaros, Laurent Daudet, et aux rapporteurs Emmanuel Vincent et Alain Rakotomamonjy d'avoir pris le temps de vous intéresser à mon travail. Vos questions et vos remarques ont permis de nourrir l'échange et de contribuer à améliorer la qualité de la soutenance et du manuscrit. Merci à toutes les super rencontres sources de motivation et d'inspiration en conférence et à l'école d'été. Un merci spécial à la communauté DCASE pour avoir réussi à propulser le sujet sur le devant de la scène et pour nous avoir offert de multiples plateformes pour partager notre travail.

Un merci aux membres permanents de l'équipe AAO (ou ADASP), Umut, Yves, Alexandre, Roland, Bertrand sans oublier Slim et Gaël. J'ai eu l'honneur de vous avoir eu à la fois comme professeurs et comme collègues, autant comme adversaires autour d'un baby que comme camarades autour d'une verre. C'était une chance de travailler dans cette équipe de recherche que vous rendez toute aussi vivante que performante. Merci à Romain et Doğaç, mes deux co-bureaux, super papas chercheurs et amis d'avoir contribué à rendre le travail quotidien plus agréable. Merci pour les collaborations, le support informatique et surtout les moments de partages autant professionnels que personnels. Merci aux doctorants et autres membres de l'équipe S2A, pour l'environnement stimulant pendant et après le travail. Une mention particulière aux doctorants en audio, les chercheurs de tempo Simon et Magdalena et les séparateurs de sources Paul, Simon et Clément, pour la bonne ambiance et les soutenances de thèse inspirantes.

Un merci spécial à Simon Leglaive qui le mérite amplement. En 6 ans on a été camarade de classe, organisateurs de concerts, musiciens live (oui ça compte), collègues, co-bureaux et surtout amis. En plus d'être une machine, t'as jamais hésité à me donner un coup de main, même si il fallait que tu te lèves à 5h30.

Merci à Yoann et Bruno, même sans musique, même avec la distance, le power trio ne meurt jamais, toujours là pour m'aider à préserver ma santé mentale. Et bien sûr merci à tous les autres copains, être bien entouré ça compte. Merci à mes parents pour m'avoir supporté jusque-là, pour m'avoir transmis votre curiosité, pour avoir cru en moi à chaque étape et surtout pour toujours avoir répondu présent quand j'avais besoin de vous. Merci à Juliette et Oscar, les meilleurs frère et sœur qu'on puisse imaginer, qui me rendent toujours de plus en plus fier. Merci au reste de ma famille, d'un côté comme de l'autre de l'Atlantique, qui m'a soutenu jusqu'au bout et s'est toujours intéressé à ce que je faisais.

Enfin, le plus grand des mercis va à Marine, pour ta patience infinie et de rendre mon quotidien toujours plus agréable. Merci de n'avoir jamais douté, de m'avoir supporté, d'avoir rendu les moments difficiles moins pénibles et d'avoir relu avec soin chaque page de ce manuscrit. Par dessus tout, merci pour tout ce bonheur depuis maintenant 8 ans.

Apprentissage de représentations pour l'analyse de scènes sonores

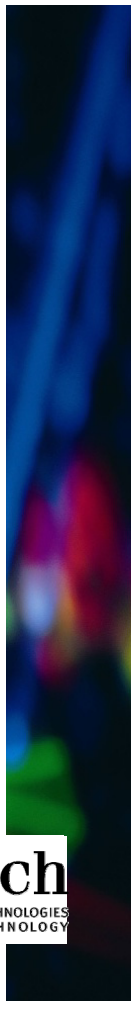
Victor BISOT

Résumé : Ce travail de thèse s'intéresse au problème de l'analyse des sons environnementaux avec pour objectif d'extraire automatiquement de l'information sur le contexte dans lequel un son a été enregistré. Ce domaine de recherche a connu un succès grandissant ces dernières années entraînant une rapide évolution du nombre de travaux et des méthodes employées. Nos travaux explorent et contribuent à plusieurs grandes familles d'approches pour l'analyse de scènes et événements sonores allant de l'ingénierie de descripteurs jusqu'aux réseaux de neurones profonds. Notre travail se focalise sur les techniques d'apprentissage de représentations par factorisation en matrices positives (NMF), qui sont particulièrement adaptées à l'analyse d'environnements multi-sources tels que les scènes sonores. Nous commençons par montrer que les spectrogrammes contiennent suffisamment d'information pour discriminer les scènes sonores en proposant une combinaison de descripteurs d'images extraits à partir des images temps-fréquence. Nous quittons ensuite le monde de l'ingénierie de descripteurs pour aller vers un apprentissage automatique des représentations. Nous entamons cette partie du travail en nous intéressant aux approches non-supervisées, en particulier à l'apprentissage de descripteurs par différentes variantes de la NMF. Plusieurs des approches proposées confirment l'intérêt de l'apprentissage de caractéristiques par NMF en obtenant des performances supérieures aux meilleures approches par extraction de descripteurs. Nous proposons ensuite d'améliorer les représentations apprises en introduisant le modèle TNMF, une variante supervisée de la NMF. Les modèles et algorithmes TNMF proposés se basent sur un apprentissage conjoint du classifieur et du dictionnaire de sorte à minimiser un coût de classification. Dans une dernière partie, nous discutons des liens de compatibilité entre la NMF et certaines approches par réseaux de neurones profonds. Nous proposons et adaptons des architectures de réseaux de neurones à l'utilisation de la NMF. Les modèles introduits nous permettent d'atteindre des performances état de l'art sur des tâches de classification de scènes et de détection d'événements sonores. Enfin nous explorons la possibilité d'entraîner conjointement la NMF et les paramètres du réseau, regroupant ainsi les différentes étapes de nos systèmes en un seul problème d'optimisation.

Mots-clés : Classification de scènes sonores, détection d'événements sonores, apprentissage de représentations, factorisation en matrices positives, apprentissage automatique, réseaux de neurones profonds

Abstract: This thesis work focuses on the computational analysis of environmental sound scenes and events. The objective of such tasks is to automatically extract information about the context in which a sound has been recorded. The interest for this area of research has been rapidly increasing in the last few years leading to a constant growth in the number of works and proposed approaches. We explore and contribute to the main families of approaches to sound scene and event analysis, going from feature engineering to deep learning. Our work is centered at representation learning techniques based on nonnegative matrix factorization, which are particularly suited to analyse multi-source environments such as acoustic scenes. As a first approach, we propose a combination of image processing features with the goal of confirming that spectrograms contain enough information to discriminate sound scenes and events. From there, we leave the world of feature engineering to go towards automatically learning the features. The first step we take in that direction is to study the usefulness of matrix factorization for unsupervised feature learning techniques, especially by relying on variants of NMF. Several of the compared approaches allow us indeed to outperform feature engineering approaches to such tasks. Next, we propose to improve the learned representations by introducing the TNMF model, a supervised variant of NMF. The proposed TNMF models and algorithms are based on jointly learning nonnegative dictionaries and classifiers by minimising a target classification cost. The last part of our work highlights the links and the compatibility between NMF and certain deep neural network systems by proposing and adapting neural network architectures to the use of NMF as an input representation. The proposed models allow us to get state of the art performance on scene classification and overlapping event detection tasks. Finally we explore the possibility of jointly learning NMF and neural networks parameters, grouping the different stages of our systems in one optimisation problem.

Key-words: Acoustic scene classification, acoustic event detection, representation learning, nonnegative matrix factorization, machine learning, deep neural networks



TELECOM ParisTech

EDITE

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY