La classification textuelle : un outil privilégié pour la classification de documents audio

Louis Rompré, Ismaïl Biskri, François Meunier

LAMIA - DMI - Université du Québec à Trois-Rivières (Qc), Canada

Abstract

In this article, we present a generalisation of the numerical classification applied to sound documents. The approach presented is a combination of concepts related to natural language processing and frequency analysis. The raw data of the audio documents are converted into strings which allow for numerical analysis. Thus a treatment similar to the one dedicated to the classification of textual documents can be applied. The text analysis performed in collaboration with the audio analysis makes the realisation of a system able to classify heterogeneous documents possible.

Résumé

Dans cet article, nous présentons une généralisation de la classification numérique appliquée à l'audio. L'approche présentée est une combinaison de concepts liés au traitement des langues naturelles et à l'analyse fréquentielle. Les données brutes des documents audio sont transposées en chaîne de caractères alphanumériques propice à l'analyse numérique. Ainsi un traitement analogue à celui voué à la classification de documents textuels peut être appliqué. L'analyse de texte couplée à l'analyse audio donne lieu à la réalisation d'un système capable de classifier de larges corpus hétérogènes.

Mots-clés : classification, n-grammes, WAVE, MIR, TALN.

1. Introduction

Initialement composés quasi totalement de données textuelles, les réseaux informatiques regorgent désormais de données multimédias de toutes sortes. Parmi ces documents, les fichiers sonores sont des plus convoités. Les documents recherchés sont généralement désordonnés et répartis. Une recherche dans cet environnement chaotique peut facilement devenir une tâche fastidieuse qui demande un temps considérable. La mise en place d'une structure favorise la consultation des documents. L'augmentation constante du volume de données et la nature évolutive de l'information requiert l'automatisation de certains processus qui mènent à la création d'une structure de données efficace.

La classification crée une hiérarchie qui améliore la recherche de documents. La création de classes de similarités réduit la complexité de l'environnement ce qui favorise l'obtention des documents souhaités. Les bienfaits de la classification vont au-delà des besoins en matière de recherche documentaire. Elle génère une vue d'ensemble qui favorise la connaissance de l'environnement ciblé. Cet article présente une approche générale qui tend à automatiser la classification de documents audio.

L'approche préconisée repose sur des concepts liés au traitement des langues naturelles (TALN). La classification est réalisée à partir d'une évaluation statistique de séquences discriminantes extraites des signaux audio. L'accent est porté sur la transformation des données brutes en caractères descriptifs sujets à l'analyse numérique.

L'intérêt pour un outil de classification automatisé de données sonores s'étend à plusieurs domaines d'application tels l'amélioration des moteurs de recherche, la création d'outils d'indexation de documents audiovisuels et la mise en place de chaînes de traitements multi formats.

La catégorisation d'un document est réalisée à partir d'une évaluation statistique de ses données. La figure 2.1 illustre ce mécanisme.

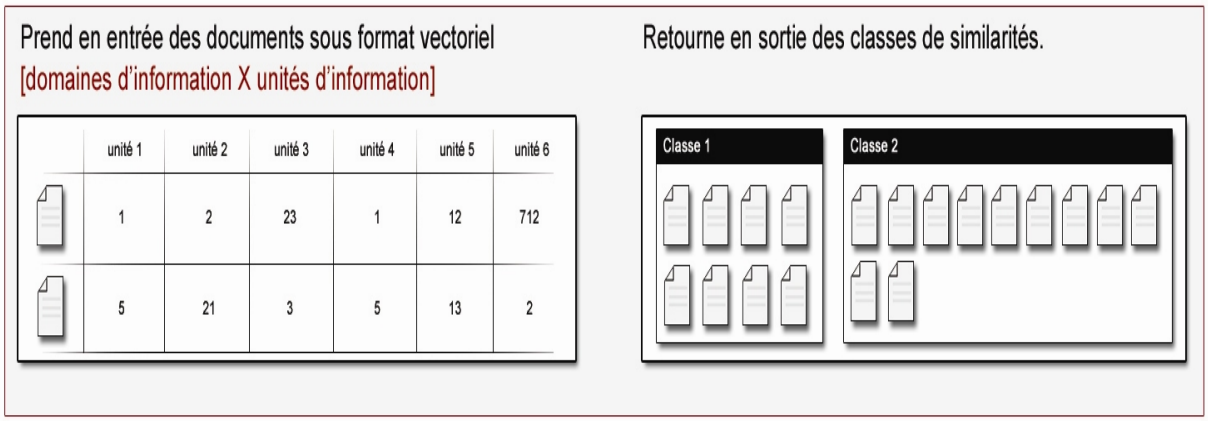

Figure 2.1 : Processus de classification illustré.

Explicitement le processus de catégorisation consiste à (Biskri et al., 2006) :

i. Déterminer une unité d'information à considérer ;

ii. Prendre en entrée des documents ou des portions de document ;

iii. Comptabiliser les unités d'informations ;

iv. Créer un vecteur de cooccurrence pour chacun des documents ;

v. Ordonner les vecteurs de manière à retourner en sortie des classes de similarités.

Les documents considérés constituent le domaine d'information tandis que l'union des différentes unités répertoriées forme l'ensemble des unités d'information.

En simplifiant, on peut dire que des documents sont jugés similaires lorsqu'ils sont constitués des mêmes unités d'information à des fréquences presque identiques. Par conséquent, le choix de l'unité d'information a un impact direct sur les résultats de la classification. La richesse du son et la qualité variable de la numérisation impliquent une unité d'information flexible pouvant s'adapter aux variations.

Jusqu'à présent, la recherche dédiée à la classification automatisée de documents textuels est considérablement en avance par rapport à celle consacrée aux documents audio. L'idée d'élargir la portée de ces travaux devient donc une solution intéressante.

Les n-grammes comme unité d'information

La notion de n-gramme suscite un intérêt grandissant depuis le milieu des années 90, principalement dans le domaine du traitement automatisé des langues naturelles. Des recherches sur l'identification de la langue (Greffenstette, 1995) et sur le calcul de similarités (Damashek, 1995) ont démontré que même si les n-grammes ne représentent pas une entité concrète tel que le mot pour le domaine du texte, ils n'entraînent pas de perte d'information.

L'unité d'information est généralement liée au domaine auquel elle est attachée. Cette singularité engendre l'utilisation de plusieurs variantes afin de couvrir un environnement complexe. Or, les progrès réalisés dans le domaine de la compression numérique entraînent l'apparition de nouveaux formats de fichiers. Ce contexte est peu favorable à l'utilisation d'une unité d'information peu flexible.

Un n-gramme est une séquence de *n* caractères descriptifs consécutifs. Le caractère qui le compose correspond à l'élément atomique le plus descriptif contenu dans le document traité. Extraire les n-grammes d'un document consiste à glisser une fenêtre de taille *n* sur ses données. Le déplacement est effectué un caractère à la fois. Ce modèle de découpage est traditionnellement associé au traitement automatisé des langues naturelles. Cependant, il possède des propriétés intéressantes pour la recherche de similarités musicales :

i. Opère sur des données dont la nature peut varier (Dunning, 1994) ;

ii. Tolère un certain ratio de déformation (Miller et al, 1999).

Ces propriétés permettent de tenir compte de la diversité des sons et de la fiabilité variable des enregistrements audio. Les modèles de classification basés sur la notion de n-gramme recherchent les séquences discriminantes au sein du contenu des documents à classifier. Ce concept statistique axé sur l'occurrence de combinaisons particulières peut s'étendre au-delà du domaine du texte. Il représente une avenue prometteuse pour la recherche de similarités musicales. Pour intégrer cette approche il suffit de transposer les données sonores en chaînes de caractères alphanumériques.

L'intérêt de l'inclusion de la notion de n-gramme au processus de classification de documents audio réside dans la portabilité de ce mode de découpage. La possibilité d'appliquer une chaîne de traitements unique à des documents de formats différents laisse entrevoir l'émergence d'outils de classification capables de suivre l'évolution résultant de l'intérêt croissant pour les documents multimédias.

2. Revue de la littérature et état de l'art

Le son est généré par des variations de pressions diffusées dans l'air. Ces vibrations forment un signal caractérisé par une amplitude, une fréquence et une forme. Il est donc, par définition, un signal analogique. Un son peut être pur (figure 2.1) ou complexe (figure 2.2). Un son pur est périodique, c'est-à-dire, qu'il possède une forme sinusoïdale qui se répète selon un intervalle fixe nommé période. La vitesse de répétition de l'onde est appelée fréquence et est mesurée en hertz (Hz).

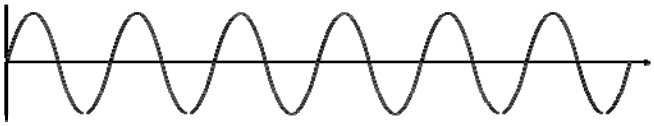

Figure 2.1 : Son pur composé d'une seule sinusoïde.

La majorité des sources sonores produisent des sons dont les vibrations sont de formes complexes composées de plusieurs sinusoïdes de fréquences et d'amplitudes différentes (figure 2.2).

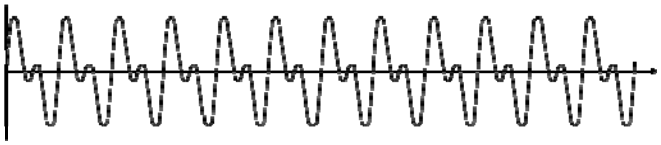

Figure 2.2 : Son complexe composé de deux sinusoïdes.

La hauteur d'un son correspond à la fréquence du signal. Une fréquence rapide génère un son aigu (haut), tandis qu'une fréquence lente génère un son grave (bas). L'oreille humaine perçoit généralement les sons dont la fréquence est comprise entre 20 et 20 000 hertz. Cet intervalle décroît avec l'âge. La différence de hauteur entre deux notes se nomme intervalle. Un intervalle peut être ascendant (deuxième son plus aigu que le premier) ou descendant (deuxième son plus grave que le premier). On nomme octave l'intervalle qui sépare deux sons dont les fréquences fondamentales ont un rapport de fréquence égal à 2 (880 hertz, 440 hertz, 220 hertz, 110 hertz, 55 hertz etc.). Ces sons sont perçus par l'oreille de façon comparable et sont identifiés par la même note.

Le choix du signal acoustique comme support à l'analyse s'impose par sa relation intrinsèque avec les données des documents audio. Cependant, puisque le son est un phénomène continu donc composé d'une infinité de valeurs, il est impératif de considérer uniquement une portion de l'information. La conversion des données brutes en données textuelles représentatives permet de répondre à cette condition.

La sémantique vectorielle est un modèle algébrique qui représente un document par l'entremise de vecteurs dans un espace linéaire multidimensionnel (G. Salton et al., 1975). L'objectif est de formaliser un document à l'aide d'un minimum de données. Ainsi, seules les caractéristiques fondamentales sont conservées. Pour le domaine du texte, les vecteurs sont communément composés de mots clés. Ces termes sont des entités intelligibles. Pour le domaine de l'audio, cette notion n'est pas aussi clairement définie. Cette abstraction nécessite l'intervention de système d'analyse musical pour identifier l'unité à considérer.

Nous nous somme intéressé aux systèmes de recherche de similarité musicale (*Music Information Retrieval*) dont l'utilisation ne nécessite pas d'expertise dans ce domaine. Ce choix est motivé par le désir de rendre l'information accessible à un grand nombre d'utilisateurs. Les systèmes retenus sont conçus pour assister l'identification et la localisation des similarités. Les caractéristiques observées correspondent aux mots clés.

Plusieurs auteurs suggèrent les n-grammes comme unité d'information à comparer (Downie, 1999, 2003) (Patel et Mundur, 2005). Seul le caractère descriptif qui les compose varie. Downie est l'un des premiers à proposer l'utilisation d'approches traditionnellement associées au traitement automatisé des langues naturelles. Les documents audio sont représentés à l'aide de chaînes alphanumériques structurées de manière comparable au mot. Les « mots musicaux » sont bâtis à partir du contenu même des documents. Ils ont le mandat de décrire la musique de la même manière et aussi précisément que le font les mots dans le domaine textuel. L'hypothèse soutenue par Downie est qu'une simple représentation, basée sur les intervalles d'une mélodie, contient assez d'information pour réaliser des opérations de catégorisation. Il existe une certaine équivalence entre un n-gramme composé d'intervalles dans le domaine de la musique et un n-gramme composé de lettres dans le domaine du texte. Ces entités informent sur la nature des données. La technique de recherche de similarités proposée suscite un intérêt parce qu'elle propose l'utilisation d'une technique existante dont l'efficacité a déjà été démontrée pour le domaine du texte. Cependant, le projet développé par Downie est destiné uniquement aux fichiers sonores de format MIDI. La particularité de ce format est qu'il contient une description détaillée de la mélodie. Ainsi, un fichier MIDI fournit explicitement le ton, l'intensité, et la durée des notes.

Patel et Mundur se sont intéressés au traitement de documents audio de format WAVE. Ce choix est justifié par le fait que ce format est l'un des plus utilisés. Les données musicales sont transposées en une représentation textuelle composée d'un alphabet réduit à deux

caractères D et U où D équivaut à un intervalle descendant (*downward*) et U un intervalle ascendant (*upward*). La répétition de la même fréquence est interprétée comme un intervalle descendant. Les intervalles sont identifiés suite à l'application d'une transformée de Fourier. Le projet de Patel et Mundur introduit l'analyse fréquentielle. Cependant, dans un contexte où le son est généralement polyphonique (plusieurs sons émis simultanément), il est difficile d'extraire précisément les intervalles.

3. Architecture du système proposé

L'approche préconisée lors de la conception de notre système de classification repose sur l'extraction de deux propriétés complémentaires du signal acoustique :

i. La forme de l'onde ;

ii. Le contenu fréquentiel.

La forme de l'onde réfère au domaine spatial tandis que le contenu fréquentiel réfère au domaine spectral. Ces deux aspects permettent de générer une vue capable de représenter des documents monophoniques mais également polyphoniques. Le domaine d'information sujet à l'analyse se compose de documents WAVE dont la nature du contenu peut varier. Nous avons opté pour une architecture modulaire afin de pouvoir interchanger les différentes composantes du système. Notre approche est semi-automatique puisque l'utilisateur est sollicité lors de la paramétrisation du système. La figure 4.1 illustre cette architecture.

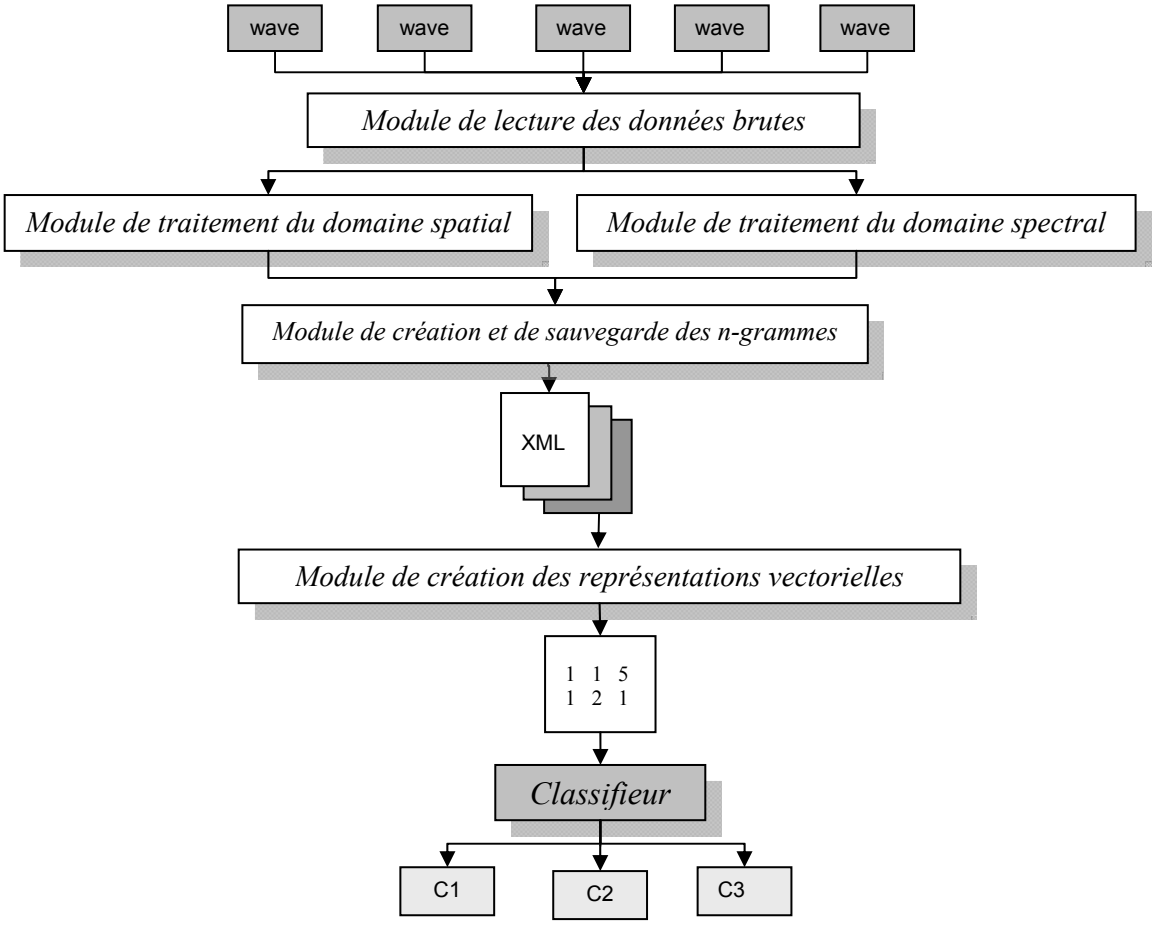

Figure 3.1 : Architecture du système proposé.

Les documents sont lus par un premier module. Les données brutes sont extraites puis transférées au module de traitement spatial ou spectral selon le mode d'analyse effectué. Les données sont apprêtées afin d'être uniformisées. Les caractéristiques spatiales ou spectrales sont prélevées et transformées en caractères descriptifs de manière à créer une chaîne de caractères alphanumériques représentative des documents. Le reste du processus est identique à celui présenté par J. G. Meunier et Biskri pour le traitement du texte (J. G. Meunier et Biskri, 2002). La liste de caractères est dans un premier temps découpée en n-grammes puis l'inventaire est sauvegardé dans un fichier pour être réutilisé. Un fichier est créé par document soumis à la classification. Lorsque tous les documents sont traités, le module de création des représentations vectorielles est sollicité. Les fichiers inventaires sont décortiqués afin qu'une représentation vectorielle uniforme soit créée pour chacun des documents WAVE. Les vecteurs sont alors présentés au classifieur de sorte que les classes de similarités soient déduites et présentées à l'utilisateur.

3.1. Lecture des données

La première étape consiste à faire la lecture des données brutes. La lecture est réalisée à l'aide d'une librairie externe. Par conséquent, l'analyse de différents formats audio requiert uniquement l'ajout des librairies utiles à l'interprétation de ces formats et n'a aucun impact sur le restant de la chaîne de traitements. Cette flexibilité est nécessaire en raison de la nature évolutive de la représentation des données. De manière à couvrir un large éventail de données, nous avons décidé de travailler avec l'un des formats le plus répandu soit le format WAVE.

3.2. Préparation des données

Les prétraitements permettent de normaliser et de simplifier les données brutes pour faciliter la classification. Les paramètres sont :

i. La taille des n-grammes : La dimension influence la probabilité d'apparition des n-grammes. Une taille importante engendre un nombre élevé de n-grammes distincts et par conséquent une distribution faible.

ii. La taille du lexique : Un certain contrôle peut être exercé sur la taille du lexique. Le domaine spatial considère la forme de l'onde sonore. Les points d'échantillonnage permettent de reconstituer cette forme. La réduction de leur quantification offre donc la possibilité de limiter la taille de l'alphabet. Les échantillons codés sur 16 bits sont convertis sur 8 bits. Ce prétraitement permet de limiter l'alphabet à 256 éléments plutôt qu'à 65 535. Un même signal peut être quantifié de manière différente. Même si la qualité de la numérisation est affectée par la réduction de la quantification de chacun des échantillons, le signal peut tout de même être reproduit et reconnu. Pour le domaine spectral, ce sont les fréquences qui sont examinées. L'alphabet est déterminé par le nombre de fréquences considérées.

iii. La projection des valeurs : La projection consiste à déterminer un seuil de tolérance permettant de regrouper certaines valeurs jugées similaires. Pour le domaine spatial la projection consiste à définir une échelle des valeurs admissibles. Les valeurs réelles sont projetées sur cette échelle. La projection des valeurs est définie à partir de la formule 3.1 où f_i représente la *m* ième valeur, $f_{\min}$ la valeur minimum, $f_{\max}$ la valeur maximum et a une constante représentant la taille de l'intervalle de projection.

$$(f_i - f_{\min} / f_{\max} - f_{\min}) * a \qquad (3.1)$$

Dans le domaine spectral, la projection équivaut à regrouper certaines fréquences du signal.

i. Le lissage : Le lissage est une opération qui élimine certains détails et le bruit que peut contenir un signal (Meunier et al., 2001). Le filtre gaussien 1-D est utilisé. La forme du noyau de ce filtre est donnée par l'équation 3.2, où σ correspond à la dimension du filtre numérique.

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \quad (3.2)$$

Le filtre gaussien est un opérateur de lissage qui a comme effet d'éliminer certaines hautes fréquences.

3.3. Localisation des caractères descriptifs et extraction des n-grammes

Ces opérations permettent de lier l'analyse audio au processus de classification habituellement utilisé pour le domaine textuel. Le découpage en n-grammes est effectué à partir d'une liste de caractères informationnels dont la nature dépend du mode d'analyse effectué.

Les paires composées d'amplitudes et de demi-périodes forment les caractères descriptifs pour le domaine spatial. Pour obtenir ces paires, le signal est parcouru à la recherche des passages par zéro. Le nombre de points d'échantillonnage recensé entre deux passages par zéro est prélevé et considéré comme une demi-période. Ce nombre doit être supérieur à un certain seuil prédéfini sans quoi l'intervalle est ignoré. L'amplitude, quant à elle, équivaut au maximum local pour les valeurs supérieures à zéro ou au minimum local pour les valeurs inférieures à zéro. La figure 3.2 illustre les paires ainsi formées pour un signal quelconque.

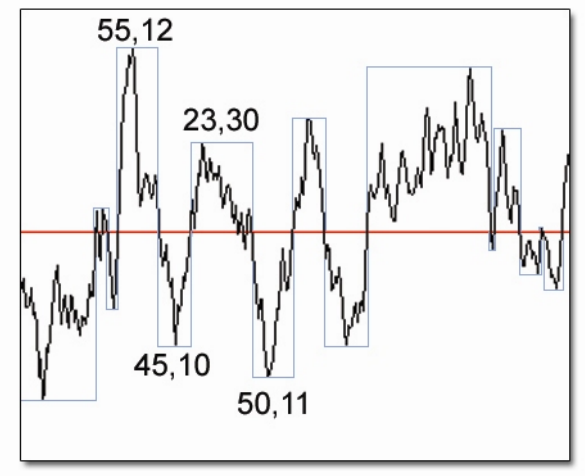

Figure 3.2 : Le caractère descriptif issu de la forme de l'onde.

Les paires sont stockées en mémoire dans une liste en ordre d'apparition. Cette liste est par la suite découpée en n-grammes. Le tableau 3.1 montre le résultat d'un découpage en bi-grammes réalisé à partir de l'équivalence « amplitude – demi-période » du signal de la figure 3.2 où le premier élément du couple représente l'amplitude et le second la demi-période.

Bi-Gramme
(52-12)(45,10), (45,10)(23,30), (23,30)(50,11),…

Tableau 3.1 : N-Grammes composés de paires d'amplitudes et de demi-périodes.

Pour le domaine spectral, les bandes de fréquences auxquelles appartiennent les fréquences dominantes du signal sont définies comme caractère descriptif. Pour obtenir ces valeurs, le signal est dans un premier temps segmenté en tranches de 4 000, 8 000, 16 000 ou 32 000 points d'échantillonnage avec chevauchement (Cavicchi, 2000). La taille du segment influence la précision de l'opération. Une transformée de Fourier est appliquée à chacun des segments. Cette opération mathématique permet d'obtenir le contenu fréquentielle d'un signal complexe. La puissance $P(u)$ est ensuite calculée à l'aide de l'équation (3.3) où u représente la fréquence de rang u.

$$P(u) = \left|F(u)\right|^2 = R^2(u) + I^2(u) \tag{3.3}$$

La puissance donne l'amplitude des fréquences du signal (Gonzalez et Woods, 1992). La courbe résultante est filtrée afin d'isoler les maximums locaux. Cette opération est effectuée à l'aide d'un opérateur de convolution et du filtre de la dérivée première 1-D dont le noyau est donné par :

$$G'(x) = -(\frac{1}{2.50663 * \delta^3}) * x * e^{\frac{-x^2}{2*\delta^2}} \tag{3.4}$$

La variable δ représente la taille du filtre. Les maximums représentent les fréquences dominantes du segment source. Ceux-ci sont triés en ordre décroissant et les k plus grands sont conservés. Les bandes de fréquences auxquelles appartiennent ces fréquences sont définies comme caractère descriptif. Cette association est illustrée à la figure 3.3 où les plus hauts sommets représentent des fréquences dominantes.

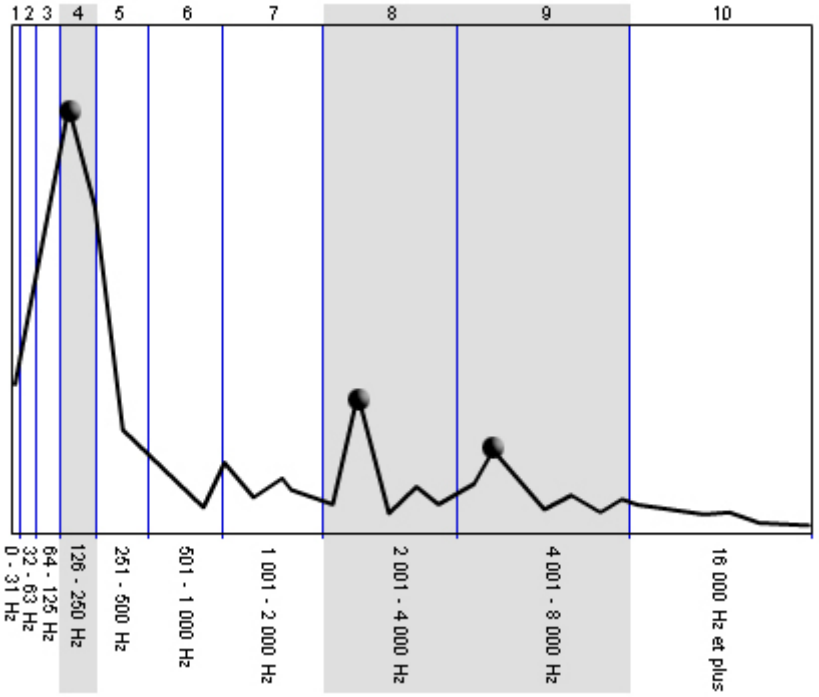

Figure 3.3 : Processus d'extraction des fréquences dominantes.

Les *k* bandes de fréquences dominantes de chacun des segments sont stockées en mémoire dans une liste par ordre d'apparition. Cette liste est par la suite découpée en n-grammes. Le tableau 3.2 donne un exemple de résultat obtenu.

Bi-Gramme
48, 89, 94, 48, 89, 93, 36, 68, …

Tableau 3.2 : N-Grammes composés de bandes de fréquences.

3.4. Création de la représentation vectorielle

Cette opération consiste à adapter les données au format attendu par le classifieur. Une matrice 3 par *n* est créée pour représenter l'ensemble des fichiers où *n* est le nombre total de n-grammes différents comptabilisés. Une rangée peut être interprétée comme le nombre d'apparitions du n-gramme *x* dans le fichier *y*.

Identifiant du fichier	**Identifiant du n-gramme**	**Nombre d'apparitions**
1	1	2
1	2	54
1	3	1
. . .	. . .	. . .

Tableau 4.1 : Représentation vectorielle des documents

Le tableau 4.1 illustre la représentation vectorielle définie pour deux fichiers. Un fichier est représenté par plusieurs rangées successives. Un n-gramme doit avoir le même identifiant pour l'ensemble des fichiers. La matrice est conservée dans un fichier texte afin d'être exploitée par le classifieur.

3.5. Présentation des données au classifieur

La représentation vectorielle des fichiers est présentée au classifeur. Dans le cadre de nos expérimentations, ART a été utilisé. Le choix de ART n'est pas dicté par des raisons de performances particulières. L'objectif n'est pas d'optimiser la classification mais plutôt d'explorer la génération de classe pouvant mener à des interprétations appropriées sur les documents audio. Le choix de ART été effectué en continuité avec les travaux déjà effectués (Biskri et al., 2002).

4. Expérimentations et discussions

Une première expérimentation a été réalisée à partir d'un domaine d'information constituée de 72 fichiers monophoniques contenant chacun une note jouée à différentes octaves. Le tableau 5.1 donne la liste des fichiers utilisés.

do	ré b	ré	mi b	mi	fa	fa #	sol	sol #	la	si b	si
C2	Db2	D2	Eb2	E2	F2	F#2	G2	G#2	A2	Bb2	B2
C3	Db3	D3	Eb3	E3	F3	F#3	G3	G#3	A3	Bb3	B3
C4	Db4	D4	Eb4	E4	F4	F#4	G4	G#4	A4	Bb4	B4
C5	Db5	D5	Eb5	E5	F5	F#5	G5	G#5	A5	Bb5	B5
C6	Db6	D6	Eb6	E6	F6	F#6	G6	G#6	A6	Bb6	B6
C7	Db7	D7	Eb7	E7	F7	F#7	G7	G#7	A7	Bb7	B7

Tableau 4.1 : Domaine d'information de l'expérimentation I.

Dans le premier volet de l'expérimentation, la forme de l'onde a été considérée. La qualité des classifications obtenues varie considérablement selon la valeur des paramètres spécifiés. Lorsque bien calibré, plusieurs classes produites regroupent des notes d'une même octave ou d'octaves adjacentes. Ces associations résultes du fait que les notes d'une même octave ou les notes d'octaves adjacentes possèdent une forme similaire.

Le second volet de l'expérimentation a été dédié à l'analyse fréquentielle. Ce mode d'analyse a permit de réunir un nombre considérable de notes avoisinantes. L'effet de compression généré par le groupement de fréquences a un lien direct avec ce phénomène. La figure 4.1 illustre une portion du spectre d'un son contenant les notes *do* (262 hertz), *mi* (330 hertz) et *la* (440 hertz) de la troisième octave. La fréquence de chacune de ces notes est clairement représentée. Une compression se traduit par le regroupement de ces notes puisqu'elles sont peu distantes l'une de l'autre.

Le système proposé est en mesure de créer des associations pertinentes, particulièrement l'analyse fréquentielle qui génère un nombre intéressant de classes contenant uniquement des notes avoisinantes.

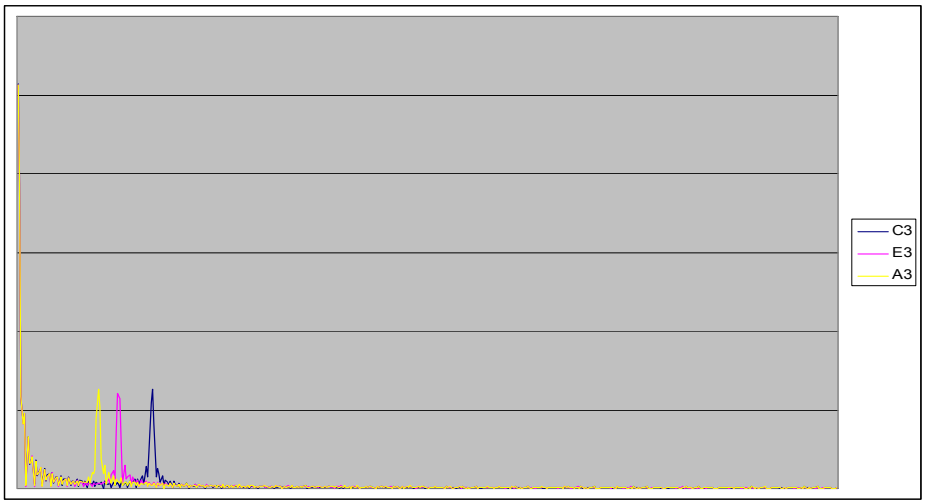

Figure 4.1 : Visualisation du spectre d'un son composé des notes do (C3), mi (E3) et la (A3) de la troisième octave.

Suite à la première expérimentation, le comportement du système lorsque les fichiers sources sont composés de plusieurs notes a été testé. Des fichiers contenant une série de notes d'une même octave ont été utilisés. Afin de réduire les traitements nécessaires à la normalisation des données, la durée des fichiers a été fixée à dix secondes. Les modèles basés sur les caractéristiques spatiales et spectrales sont en mesure d'effectuer une distinction plus ou moins précise des octaves. Bien que cette propriété puisse être désirable à des fins d'analyses, elle peut engendrer des lacunes lorsqu'il s'agit de reconnaissance de mélodies. En effet, la

même mélodie jouée à deux octaves différentes ne pourrait être jugée comme similaire. Cette distinction résulte de l'altération de la forme et du contenu fréquentiel de l'onde attribuable au changement de l'octave. Ce facteur pourrait néanmoins être contourné par une méthode de factorisation permettant de rapporter l'ensemble des notes sur une octave référence.

Classification de musique polyphonique

La musique polyphonique constitue la majorité des documents recherchés sur le web. Par conséquent, il devient primordial qu'un système de classification de fichiers sonores soit en mesure de traiter ce type de documents. Nous avons constitué une banque de données à partir de quatre pièces d'artistes et de genres différents. Le tableau 5.13 présente ces pièces.

Artiste	**Pièce**	**Genre**
Hot Snake	Retrofit	Rock
Elvis	Houng Dog	Rock'n Roll
Calexico	Tres Avisos	Folk
Red Hot Chili Pepper	Higher Ground	Funk

Tableau 4.2 : Pièces de musiques polyphoniques.

La complexité de l'onde sonore augmente le nombre de paires formées de l'amplitude et de la demi-période. Ce facteur génère une signature quasi distincte pour chacun des fichiers. De cette façon plusieurs segments ont été isolés. Malgré la complexité de la forme de l'onde sonore de chacun des extraits, le système a été en mesure de regrouper certains segments. Lors de l'analyse fréquentielle nous avons remarqué que l'ordonnancement et la nature des fichiers influence le classifieur. Un fichier ne peut faire partie de plus d'une classe. La création d'une nouvelle classe dépend du niveau de ressemblance entre le fichier en traitement et ceux existants. L'ordonnancement des fichiers a donc un impact direct sur les résultats produits. Ce phénomène, attribuable à ART, nous force à tenir compte d'un taux d'erreur dans l'évaluation des résultats. L'impact de l'ordonnancement des fichiers sur le processus de classification laisse croire que la nature des fichiers influence le classifieur. Afin de vérifier cette hypothèse, nous avons augmenté le nombre de candidats similaires afin de permettre au classifieur de mieux cibler les caractéristiques recherchées.

Classes	Fichiers
1	Elvis - houng dog 1 Elvis - houng dog 2 Elvis - houng dog 3 Copy of Elvis - houng dog 1 Copy of Elvis - houng dog 2 Copy of Elvis - houng dog 3
2	hot snake - retrofit 1 hot snake - retrofit 2 hot snake - retrofit 3 Copy of hot snake - retrofit 1 Copy of hot snake - retrofit 2 Copy of hot snake - retrofit 3
3	Red hot - Higher Ground 1 Copy of Red hot - Higher Ground 1
4	Red hot - Higher Ground 2 Red hot - Higher Ground 3 Copy of Red hot - Higher Ground 2 Copy of Red hot - Higher Ground 3
5	Calexico - Tres Avisos 1 Calexico - Tres Avisos 2 Calexico - Tres Avisos 3 Copy of Calexico - Tres Avisos 1 Copy of Calexico - Tres Avisos 2 Copy of Calexico - Tres Avisos 3

Tableau 4.3 : Répartition des documents lorsque le classifieur est conditionné.

Le tableau 4.3 donne la répartition des classes produites lorsque le classifieur est conditionné. Une classification parfaite est générée. Dans un contexte réel de classification, la duplication des documents ne peut être envisagée. La rétroaction avec l'utilisateur devient donc une option à considérer. Pour produire les résultats escomptés, le classifieur doit être en mesure de bâtir un modèle valable des candidats recherchés.

5. Conclusion

Le processus de classification de document audio peut être automatisé. Pour ce faire, l'attention doit être portée sur les mécanismes d'identification et d'analyse des caractéristiques des documents soumis à la classification. Nous nous sommes intéressé à ces mécanismes dans l'objectif d'automatiser la classification de documents audio.

Nous suggérons une approche hybride qui fusionne des concepts tirés du traitement automatisé des langues naturelles et de l'analyse fréquentielle. La classification est réalisée à partir d'une comparaison statistique de séquences particulières nommées n-gramme et constituées de *n* caractères descriptifs consécutifs. L'avantage de ce modèle de découpage est qu'il assure un niveau d'adaptabilité intéressant en plus de maintenir le lexique à un seuil raisonnable. Ce dernier aspect est un facteur important considérant que le son est un phénomène continu défini par une infinité de valeurs.

Notre approche est innovatrice parce qu'elle utilise une chaîne de traitements flexibles pouvant être utilisée pour différents formats de fichiers. Seule la lecture et la préparation des données diffèrent d'un format à l'autre. La recherche et l'optimisation du caractère descriptif sont les tâches les plus délicates. Dans le cadre de nos travaux, la forme de l'onde et le spectre du signal ont été considérés comme caractères descriptifs puisque ces attributs donnent

plusieurs indications sur la nature et la composition de l'onde sonore. Un mode d'analyse a été dédié à chacune de ces propriétés.

Plusieurs expérimentations ont été effectuées afin de valider le modèle étudié. La nature des résultats obtenus suggère que l'analyse du contenu fréquentiel génère de meilleurs résultats. Le développement futur doit donc accorder une importance plus grande à cet aspect.

L'analyse textuelle couplée à l'analyse audio donne lieu à la réalisation de système capable de classifier de larges corpus hétérogènes.

Références

Biskri I., Rompré L., Laouamer L. & Meunier F. (2006). Classification de documents Multimédias : vers une approche générale. In *Actes du colloque JADT 2006.* Besançon, France.

Grefenstette G. (1995). Comparing Two Language Identification Schemes. In *Actes du colloque JADT 1995.* Rome, Italie.

Damashek M. (1995). Gauging Similarity with n-Grams : Language-Independent Categorization Of Text. *Science*, Vol.267, pages 843-848.

Dunning T. (1994). *Statistical Identification of Language*. Technical Report MCCS 94-273. Computing Research Laboratory, New Mexico State University.

Miller E. L., Shen D., Liu J., Nicholas C. & Chen T. (1999). Techniques for Gigabyte-Scale N-gram Based Information Retrieval on Personal Computers. In *Actes du colloque PDPTA 99.* Las Vegas, États-Unis.

Salton G., Wong A. et Yang C. S. (1975). A Vector Space Model for Automatic Indexing. *Communication of the ACM,* Vol.18, pages 613-620. New York, États-Unis.

Downie J. S. (1999). *Evaluating a simple approach to music information retrieval : conceiving melodic n-grams as text*. PhD Thesis, Faculty of Graduate Studies of the University of Western Ontario.

Patel N. et Mundur P. (2005). *An n-gram based approach to finding the repeating patterns in musical.* Euro/IMSA 2005. Grindelwald, Switzerland.

Meunier J.-G., Biskri I. (2002). SATIM : une plate-forme modulaire pour la construction de chaînes d'analyse de textes assistée par ordinateur. *Colloque international : L'édition électronique en littérature et dictionnairique : évaluation et bilan*. Rouen, France, Juin 2002.

Cavicchi T. J. (2000). *Digital Signal Processing.* John Wiley & Sons, Inc.

Meunier F., Meunier J., Cavayas F. (2001). Synthetic Aperture Radar Image of Agricultural Fields with Surface Network. *Simulation and Spatial Information Retrieval, Optical Engineering*, Vol. 40, Number 10, October 2001, pp. 2319-2330.

Gonzalez R. C. & Woods R. E. (1992). *Digital Image Processing*. Addison-Wesley.

Paulus J. K. & Klapuri A. P. (2003). Conventional and periodic n-grams in the transcription of drum sequences. In *Proceedings of the 2003 International Conference on Multimedia and Expo,* Vol.1.