Fusion de paramètres en classification Parole/Musique

Julie Mauclair[1], *Julien Pinquier*[2]

[1]Laboratoire d'Informatique de l'Université du Maine - CNRS FRE 2730
Avenue René Laennec, 72085 Le Mans cedex 09, FRANCE
Tél : ++33 (0)2 43 83 38 43 - Fax : +33 (0)2 43 83 38 68
Mél : mauclair@lium.univ-lemans.fr - http ://www-lium.univ-lemans.fr

[2]Institut de Recherche en Informatique de Toulouse - CNRS UMR 5505
118, route de Narbonne, 31062 Toulouse cedex 04, FRANCE
Tél : ++33 (0)5 61 55 88 35 - Fax : ++33 (0)5 61 55 62 58
Mél : pinquier@irit.fr - http ://www.irit.fr/recherches/SAMOVA/

ABSTRACT

This work addresses the soundtrack indexing of multimedia documents. We present a speech/music classification system based on three original features : entropy modulation, stationary segment duration and number of segments. They were merged by basic score maximisation with the classical 4 Hertz modulation energy. We validate this fusion approach with the use of the probability theory and the evidence theory. The system is tested on radio corpora. Systems are simple, robust and could be improved on every corpus without training or adaptation.

1. INTRODUCTION

Les méthodes d'indexation actuelles en audio (et en vidéo) nécessitent encore pour la plupart un travail manuel : un opérateur humain doit sélectionner les informations désirées en lisant, écoutant et/ou visualisant le document. Cette tâche est peu compatible avec le volume croissant des données à recenser et doit donc être automatisée. En amont d'un travail de recherche de locuteurs, de langues, de mots-clés, la discrimination Parole/Musique devient un enjeu essentiel.

Plusieurs méthodes de discrimination Parole/Musique ont été décrites dans la littérature. Elles peuvent se classer en deux groupes. D'une part, dans la communauté des spécialistes en musique, l'accent porte sur des paramètres permettant de séparer au mieux la musique du reste (non-musique). Par exemple, le taux de passage par zéro (Zero Crossing Rate) et le centroïde spectral sont utilisés pour séparer le bruit des parties voisées (donc harmoniques) [13], [15] tandis que la variation de la magnitude spectrale (le "Flux" spectral) permet de détecter les continuités harmoniques [14]. D'autre part, dans la communauté du traitement automatique de la parole, les paramètres cepstraux sont privilégiés pour extraire les zones de parole [7].

Trois approches sont communément utilisées pour la classification : les Modèles de Mélanges de lois Gaussiennes, les k plus proches voisins [4] et les Modèles de Markov Cachés. Un état de l'art en indexation audio [5] permet d'avoir de plus amples informations sur les paramètres existants pour la classification de documents sonores.

L'IRIT dispose d'un système d'indexation de la parole et de la musique sur la bande audio [12] basé sur une fusion basique (maximisation de scores). Ce système sert de référence à cette étude pour valider notre approche de la fusion en classification Parole/Musique à l'aide de la théorie des probabilités et de la théorie de l'évidence.

Cet article est composé de trois parties : une présentation globale du système de classification, une description des méthodes de fusions et enfin, les résultats obtenus sur un corpus radiophonique.

2. SYSTÈME DE CLASSIFICATION

Le système se décompose (Figure 1) en deux sous-systèmes de classification correspondant aux deux détections disjointes de la parole et de la musique.

- Pour le sous-système de détection de la parole, nous avons utilisé la modulation de l'entropie et la modulation de l'énergie à 4 Hz.
- Pour le sous-système de détection de la musique, une segmentation automatique du signal nous a permis d'obtenir le nombre de segments par unité de temps et la durée de ces segments.

Ainsi, nous avons deux classifications en parallèle : une classification Parole/Non-Parole et une classification Musique/Non-Musique. Ainsi, les passages contenant de la parole, de la musique mais aussi simultanément de la parole et de la musique sont détectés. La décision est prise en maximisant les scores (vraisemblances) issus de la modélisation de chacun des paramètres.

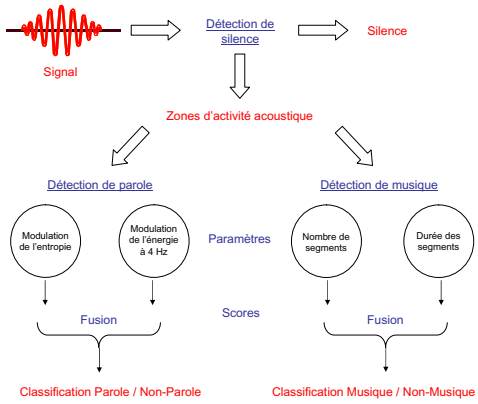

FIG. 1 - Le système de classification Parole/Musique.

2.1. Sous-système Parole/Non-Parole

- Modulation de l'énergie à 4 Hz
  Le signal de parole possède un pic caractéristique de modulation en énergie autour de la fréquence sylla-

bique 4 Hz. [8]. Ce paramètre a des valeurs plus élevées pour les segments de parole que pour les segments musicaux.

- Modulation de l'entropie

  Des observations menées sur le signal et sur le spectrogramme associé font apparaître une structure plus "ordonnée" du signal de musique par rapport au signal de parole. La modulation de l'entropie est plus élevée pour la parole que pour la musique.

2.2. Sous-système Musique/Non-Musique

La segmentation est issue de l'algorithme de "Divergence Forward-Backward" (DFB) [2] qui est basé sur une étude statistique du signal dans le domaine temporel. L'hypothèse de départ est que le signal de parole est décrit par une suite de zones quasi-stationnaires. Chacune est alors caractérisée par un modèle statistique, le modèle autorégressif gaussien. La méthode consiste à détecter les changements dans les paramètres autorégressifs.

FIG. 2a - Segmentation sur 1 seconde de parole.

FIG. 2b - Segmentation sur 1 seconde de musique.

- Nombre de segments

  Ce paramètre est extrait de l'algorithme DFB. Il correspond au nombre de segments présents durant chaque seconde de signal. Les signaux de parole présentent une alternance de périodes de transition (voisées/non-voisées) et de périodes de relative stabilité (les voyelles en général) [3]. Le nombre de segments par unité de temps (ici la seconde) est donc plus important pour la parole (Figure 2a) que pour la musique (Figure 2b).

- Durée des segments

  Les segments sont généralement plus long pour la musique (Figure 2b) que pour la parole (Figure 2a). Nous la modélisons avec une loi Gaussienne inverse. La pdf est donnée par :

$$p(g) = \sqrt{\frac{\lambda}{2\pi g^3}} * e^{\frac{-\lambda(g-\mu)^2}{2\mu^2 g}}, g \geq 0$$

  avec μ = valeur moyenne de g et $\frac{\mu^3}{\lambda}$ variance de g.

2.3. Système de référence global

Pour chaque sous-système, on estime un modèle statistique. La décision est prise en maximisant les scores (vraisemblances) issus de la modélisation de chacun des paramètres. Par la suite, nous validerons cette approche par l'utilisation de nouveaux moyens de fusion : la théorie des probabilités et la théorie de l'évidence.

3. FUSION DE DONNÉES

La fusion de données a depuis peu suscité un intérêt certain dans la communauté scientifique [6] et commence à apparaître dans le domaine du traitement de la parole [11]. Elle consiste à mettre à profit le maximum d'informations sur les données afin de réduire les faiblesses de certaines à l'aide des autres. La théorie des probabilités et la théorie de l'évidence nous apportent des solutions quant à la combinaison d'informations provenant de sources hétéroclites.

3.1. Théorie des probabilités

Les méthodes les plus utilisées pour la fusion de données ont tout d'abord été envisagées sous l'approche bayésienne. Ici, la mise à jour des informations (modélisées par des distributions de probabilités) se fait à l'aide du théorème de Bayes. Ce théorème permet d'estimer la probabilité de l'occurrence d'un évènement futur en observant l'occurrence d'évènements similaires dans le passé. La prise de décision est ensuite réalisée, le plus souvent par le critère du maximum *a posteriori*. L'un des inconvénients majeurs de cette technique réside dans l'exigence de la connaissance parfaite des probabilités, et plus particulièrement de la probabilité *a priori*.

Pour palier le problème de l'ignorance (traduite par une égalité de probabilités), on peut s'appuyer sur des indices de confiance, en se basant sur deux informations : l'expert et la classe [10]. Notons α l'indice de confiance de l'expert dans son propos et β, l'indice de confiance de classe qui est en quelque sorte l'expérience que l'on a du modèle expert. Si on a un seuil séparant la Classe C de la Non-Classe NC, l'expert e sait les discriminer avec un taux de confiance α_e :

$$\begin{aligned} \alpha_e &= 1 - taux\ d'erreur \\ &= 1 - (Pr(NC < seuil) + Pr(C > seuil)) \end{aligned} \quad (1)$$

Ici, les experts sont les quatre paramètres du système de référence (modulation de l'entropie, modulation de l'énergie à 4 Hz, nombre de segments et durée des segments). Pour chaque expert, il existe une matrice de confusion Classe/Non-Classe (Parole/Non-Parole ou Musique/Non-Musique). β_{eC} et β_{eNC} s'expriment alors par :

$$\beta_{eC} = \frac{Pr(y = C|C)}{Pr(y = C|C) + Pr(y = NC|C)}$$

$$\beta_{eNC} = \frac{Pr(y = NC|NC)}{Pr(y = C|NC) + Pr(y = NC|NC)}$$

Où y est l'observation extraite toutes les secondes.

La stratégie bayésienne nous donne la fonction de décision pour chaque expert :

$$\begin{aligned} s_e^*(y) = min \left\{ \left\{ (1 - \beta_{eNC}) * \frac{Pr(y|C)}{P(y)} \right\}, \right. \\ \left. \left\{ (1 - \beta_{eC}) * \frac{Pr(y|NC)}{P(y)} \right\} \right\}. \end{aligned} \quad (2)$$

Au final, on prend la décision avec l'expert e qui maximise la formule :

$$\alpha_e * (1 - s_e^*(y))$$

3.2. *Théorie de l'évidence*

Comme nous l'avons vu précédemment, il existe de nombreuses solutions pour l'association multi-sources, mais chacune de ces méthodes a ses faiblesses. La plupart traite l'imprécision, mais des notions telles que l'incertitude ou encore la fiabilité sont ignorées. La théorie de l'évidence nous permet, comme nous allons le voir, de modéliser et d'utiliser des données incertaines [9] et [1].

Soit un ensemble de N classes θ. Ici, $\theta = \{P, M, PM, B\}$ avec P pour Parole, M pour Musique et B pour Bruit.
A partir de cet ensemble, on en défini un autre :

$$2^\theta = \{A | A \subseteq \theta\}$$

$$2^\theta = \{\emptyset, \{P\}, \{M\}, \{PM\}, \{B\}, \{P \cup M\}, ..., \theta\}.$$

Cet ensemble sert de référentiel de définition pour l'ensemble des grandeurs utilisées par la théorie de l'évidence pour évaluer la véracité d'une proposition. Soit une information provenant de n'importe quelle source (capteur, agent, expert...) traduisant par exemple une opinion sur l'état d'un système. Cette information porte sur les éléments de 2^θ, c'est à dire non seulement parmi les hypothèses singletons, mais aussi parmi les disjonctions de celles-ci. L'opinion sur le système est alors caractérisée par des degrés de croyance dans les différentes hypothèses. Ces degrés de croyance peuvent être définis par une fonction de croyance notée m_θ. La fonction m_θ est alors définie par :

$$m_\theta : 2^\theta \rightarrow [0, 1]$$

et vérifie les propriétés suivantes :

1. $m_\theta(\emptyset) = 0$
2. $\sum_{A \subseteq \theta} m_\theta(A) = 1$

La modélisation issue de cette fonction est appelée jeu de masses. Une telle modélisation consiste à répartir toute notre connaissance disponible sur l'ensemble 2^θ. $m_\theta(A)$ représente la partie du degré de croyance placée exactement sur la proposition A. Les différents jeux de masses (un par expert) sont ensuite fusionnés grâce à la loi de Dempster-Shafer [9] pour construire un jeu de masse final unique et ainsi accéder à une information plus fiable.

Nous avons ainsi quatre experts (paramètres de notre système de classification) qui nous apportent quatre jeux de masses (m_1,...,m_4).
Pour $m_e(\theta)$, qui représente la masse associée à l'ignorance, nous devons considérer la part d'erreur de l'expert. Pour les autres hypothèses, chaque expert fournit un jeu de masses qui provient des probabilités *a priori*.

Voyons, par exemple, le résultat de cette modélisation sur l'expert 1 (modulation de l'énergie à 4 Hz) :
$m_1(y \in \{P \cup PM\}) = m_1(P \cup PM) = Pr(y|Parole)$
$m_1(M \cup B) = Pr(y|Non\text{-}Parole)$
et $m_1(\theta) = Pr(NC < seuil) + Pr(C > seuil)$.

Nous avonc donc quatre jeux de masses, un par expert que l'on peut fusionner en un jeu de masses global avec la loi de Dempster-Shafer [9] : $m_\theta = m_1 \oplus m_2 \oplus m_3 \oplus m_4$.
Par exemple, si on fusionne l'expert 1 et l'expert 2 (la loi de Dempster-Shafer étant associative et commutative, cet exemple est valable pour la fusion de nos quatre experts en un jeu de masse final), on obtient un jeu de masse intermédiaire :

$$m_{12}(A) = \sum_{A_i \cap B_j = A} m_1(A_i) * m_2(B_j)$$

où A_i et B_j sont des éléments focaux du premier et du deuxième expert respectivement, définis sur 2^θ.
On effectue une renormalisation afin d'avoir $m_\theta(\emptyset) = 0$.

La décision est effectuée sur le maximum de plausibilité pour tenir compte du poids des disjonctions d'hypothèses qui apparaissent dans le jeu de masses final :

$$Pl_\theta(A) = \sum_{B \cap A \neq \emptyset} m_\theta(B)$$

La théorie de l'évidence permet donc de gérer des hypothèses composées et de réduire la part d'*a priori* au sein de la modélisation du problème.

4. EXPÉRIMENTATIONS

4.1. *Corpus*

Le corpus expérimental correspond à une base de données qui a été réalisée à partir d'enregistrements de RFI (Radio France Internationale, projet RAIVES du CNRS). Elle comprend des émissions de radio de tous genres très compressées et dont le taux d'échantillonnage est de 16 kHz. Cette base de données contient de longues périodes de parole, de la musique et/ou du bruit. La parole est enregistrée dans différentes conditions (parole téléphonique, enregistrements en extérieur, bruit de foule et deux locuteurs simultanément). La musique est présente sous diverses formes également, avec de nombreux instruments et des parties de voix chantée. Le corpus est multi-locuteur et multilingue.

4.2. *Résultats*

Le système de classification (cf. 2) propose des résultats en décomposant les parties Parole/Non-Parole et Musique/Non-Musique. Pour évaluer la modélisation du système, nous avons effectué un partage entre les experts Parole/Non-Parole (tableau 1) et les experts Musique/Non-Musique (tableau 2) pour pouvoir comparer avec les résultats précédents.

TAB. 1 - Classification Parole/Non-Parole

Sous-système P/NP	Taux d'identification correcte
Modulation de l'énergie à 4Hz	87.3 %
Modulation de l'entropie	87.5 %
Système de référence (max)	90.5 %
Théorie des probabilités	**90.7 %**
Théorie de l'évidence	**90.9 %**

TAB. 2 - Classification Musique/Non-Musique

Sous-système M/NM	Taux d'identification correcte
Nombre de segments	86.4 %
Durée des segments	78.1 %
Système de référence (max)	89 %
Théorie des probabilités	**84.8 %**
Théorie de l'évidence	**86.9 %**

Le système de référence provient d'un travail précédent [12] où nous avions utilisé une fusion par maximisation des scores de vraisemblance.

Le modèle qui s'appuie sur la théorie des probabilités offre des résultats similaires au système de base. En effet, cette théorie est sous-jacente dans ce dernier et la modélisation ne peut que conforter les résultats obtenus. La pondération des indices de confiance apporte une augmentation pour la discrimination Parole/Non-Parole. Les résultats Musique/Non-Musique peuvent s'expliquer par le fait que la durée des segments n'a pas un taux de confiance d'expert suffisamment élevé (α_4 = 53 %) et que ce paramètre ne rentre donc pas en compte dans la fusion finale. Celle-ci repose uniquement sur la décision du paramètre "nombre de segments".
Pour la théorie de l'évidence, les jeux de masse offrent une amélioration des résultats. Le score obtenu en Musique/Non-Musique s'explique de la même façon que pour la théorie précédente.

5. DISCUSSION

Nous décrivons dans cet article quatre paramètres basés sur différentes propriétés du signal. Tous ces paramètres considérés séparément sont pertinents pour une classification Parole/Musique et les taux d'identification corrects varient entre 76 et 84 %. Ensuite, les méthodes de fusion proposées permettent d'améliorer ces scores pour atteindre plus de 90 %.

La théorie des probabilités permet d'utiliser nos connaissances *a priori* du système pour pratiquer une fusion d'informations grâce aux indices de confiance. Cette théorie apporte des résultats satisfaisants pour la discrimination Parole/Non-Parole ce qui valide la modélisation de ce sous-système. Pour la Musique/Non-Musique, elle obtient des scores inférieurs à ceux du paramètre "nombre de segments" utilisé seul, ce qui semble montrer que la durée des segments n'est pas forcément le meilleur soutien pour celui-ci.

La théorie de l'évidence nous apporte les meilleurs résultats en classification Parole/Non-Parole. Les résultats en Musique/Non-Musique peuvent être améliorés en révisant la façon de calculer les jeux de masses pour les deux paramètres issus de la segmentation. Cette théorie est au premier abord la meilleure solution pour formaliser le système de référence.

Les deux théories semblent s'adapter à un problème de reconnaissance des formes en indexation sonore. La modélisation du problème peut aussi être affinée en utilisant d'autres techniques de calcul et de combinaison des taux de confiance ou des jeux de masses.

Ces techniques de fusion peuvent également être appliquées dans d'autres domaines comme l'identification des langues où l'on peut combiner par exemple les modèles décrivant les différentes sources d'informations discriminantes : les informations acoustiques, phonotactiques et prosodiques.

Beaucoup d'erreurs en classification automatique sont dues à une mauvaise discrimination Parole/Voix Chantée. Nous espérons pouvoir utiliser ces techniques de fusion pour ce domaine de l'indexation audio.

RÉFÉRENCES

[1] H. Altinçay and M. Demirekler. Speaker identification by combining multiple classifiers using dempster-shafer theory of evidence. *Speech Communication*, 2003.

[2] R. André-Obrecht. A new statistical approach for automatic speech segmentation. *IEEE Transactions on Audio, Speech, and Signal Processing*, 36(1), January 1988.

[3] Calliope. *La parole et son traitement automatique*. Masson, Paris, France, 1989.

[4] M. J. Carey, E. J. Parris, and H. Lloyd-Thomas. A comparison of features for speech, music discrimination. In *International Conference on Audio, Speech and Signal Processing*, pages 149–152, Phoenix, USA, March 1999. IEEE.

[5] M. Carré and P. Pierrick. Indexation audio : un état de l'art. *Annales des télécommunications*, 55(9-10) :507–525, 2000.

[6] D. Dubois and H. Prade. La fusion d'informations imprécises. *Traitement du signal*, 11(6), 1994.

[7] J. L. Gauvain, L. Lamel, and G. Adda. Systèmes de processus légers : concepts et exemples. In *International Workshop on Content-Based Multimedia Indexing*, pages 67–73, Toulouse, France, October 1999. GDR-PRC ISIS.

[8] T. Houtgast and J. M. Steeneken. A review of the mtf concept in room acoustics and its use for estimating speech intelligibility in auditoria. *Journal of the Acoustical Society of America*, 77(3) :1069–1077, 1985.

[9] F. Janez. *Fusion d'informations définies sur des référentiels non-exhaustifs différents*. PhD thesis, Université d'Angers, 1996.

[10] P. Leray, H. Zaragoza, and F. d'Alché Buc. Pertinence des mesures de confiance en classification. In *Congrès de Reconnaissance des Formes et Intelligence Artificielle*, Paris, December 2000.

[11] N. Moreau, D. Charlet, and D. Jouvet. Confidence measure and incremental adaptation for the rejection of incorrect data. In *International Conference on Audio, Speech and Signal Processing*, volume 3, pages 1807–1810, Istanbul, Turquie, January 2000. IEEE.

[12] J. Pinquier, J.L. Rouas, and R. André-Obrecht. A fusion study in speech / music classification. In *International Conference on Audio, Speech and Signal Processing*, Hong-Kong, China, April 2003.

[13] J. Saunders. Real-time discrimination of broadcast speech/music. In *International Conference on Audio, Speech and Signal Processing*, pages 993–996, Atlanta, USA, May 1996. IEEE.

[14] E. Scheirer and M. Slaney. Construction and evaluation of a robust multifeature speech/music discriminator. In *International Conference on Audio, Speech and Signal Processing*, pages 1331–1334, Munich, Germany, April 1997. IEEE.

[15] T. Zhang, C. Kuo, and C. J. Hierarchical system for content-based audio classification and retrieval. In *Conference on Multimedia storage and Archiving Systems III*, volume 3527, pages 398–409. SPIE, November 1998.