Détection de la parole et de la musique dans les documents sonores : fusion de deux approches

Julien PINQUIER, Jean-Luc ROUAS, Julie MAUCLAIR, Régine ANDRÉ-OBRECHT

Équipe SAMOVA, IRIT, UMR 5505 CNRS INP UPS
118, route de Narbonne, 31062 Toulouse cedex 04, FRANCE
{pinquier, rouas, mauclair, obrecht}@irit.fr

Résumé – Dans cet article, une segmentation de la bande sonore est effectuée en détectant les composantes parole et musique. Cette segmentation résulte de la fusion de deux approches de classification. La première, classique, est basée sur une analyse spectrale et des Modèles de Mélanges de Gaussiennes (MMG). La seconde, originale, utilise des paramètres "simples" et robustes : la modulation de l'énergie à quatre hertz, la modulation de l'entropie, la durée des segments (issus d'une segmentation automatique) et le nombre de ces segments par seconde. Notre système global se décompose en deux sous-systèmes de classification (Parole/NonParole et Musique/NonMusique). Il atteint respectivement 94 % d'accuracy pour la parole et 90 % pour la musique sachant qu'une décision est prise sur chaque seconde du signal. Il apparaît très intéressant d'améliorer un système classique, basé sur une analyse spectrale et des MMG, par des paramètres "simples" et robustes.

Abstract – In this paper, we present and merge two speech / music classification approaches we have developed. The first one is a differentiated modeling approach based on spectral analysis, which is implemented with GMM. The other one is based on three original features: entropy modulation, stationary segment duration and number of segments. They are merged with the classical 4 Hertz modulation energy. Our classification system is a fusion of the two approaches. It is divided in two classifications (Speech/NonSpeech and Music/NonMusic) and provides 94 % of accuracy for speech detection and 90 % for music detection. Decision is taken on each second of signal. Beside the spectral information and GMM, classically used in speech / music discrimination, simple parameters bring complementary and efficient information.

1 Introduction

Le document sonore est un document très difficile à indexer, car l'extraction de l'information élémentaire se heurte à l'extrême diversité des sources acoustiques. Il peut être intéressant de rechercher des "bruits" ou des sons sémantiquement significatifs (applaudissements, effets spéciaux...), de repérer les passages musicaux pour les isoler et les identifier, de détecter les locuteurs équivalents à des tours de parole dans un dialogue. Enfin la transcription du discours ou la recherche de mots clés (mots isolés, groupes de mots...) fournissent une information importante sur le contenu du message verbal, et permettent l'accès à la recherche d'informations telle qu'elle est pratiquée dans des documents textuels. Ces prétraitements peuvent être assimilés à la recherche de niveaux de description plus ou moins élémentaires, la plus élémentaire étant la décomposition de la bande sonore en ses composantes de base "parole, musique, bruit".

Plusieurs méthodes de discrimination Parole / Musique ont été décrites dans la littérature. Elles peuvent se classer en deux groupes selon les paramètres discriminants utilisés. 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. 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) [1], [2] tandis que la variation de la magnitude spectrale (le "Flux" spectral) permet de détecter les discontinuités harmoniques [3]. D'autre part, dans la communauté du traitement automatique de la parole, les paramètres cepstraux sont privilégiés pour extraire la composante parole ([4], [5]).

De par la nature même des signaux de parole et de musique, leur indexation ne peut découler de l'utilisation d'outils communs. Le système de classification, que nous avons développé, résulte de la fusion de deux sous-systèmes. Le premier, noté système I, utilise la modélisation différenciée basée sur une analyse cepstrale et spectrale [6]. Le second, appelé système II, est issu de l'extraction de paramètres "originaux" [7] : la modulation de l'entropie, le nombre de segments (issus d'une segmentation automatique), la durée de ces segments et la modélisation de l'énergie à quatre hertz. Dans chaque sous-système la décision élémentaire repose sur une approche bayésienne où les lois sont des gaussiennes ou des mélanges de lois gaussiennes.

Cet article est divisé en trois parties. Une première section permet de décrire le système global de classification et chacun des paramètres utilisés. Ensuite, les méthodes de fusion d'information étudiées sont présentées. Au cours du dernier paragraphe, un corpus radiophonique est employé afin de valider chacun des systèmes (I et II) et de vérifier la robustesse de notre système de fusion dans des conditions très diverses.

2 Le système de classification

Notre système (figure 1) se décompose en deux systèmes de classification correspondants aux deux détections disjointes de la parole et de la musique. Ainsi, les passages contenant de la parole, de la musique mais aussi les deux simultanément sont

détectés. Chacun des systèmes emprunte des paramètres issus du système I et du système II et leur modélisation statistique respective.

La décision finale est prise en fusionnant les scores (vraisemblances). Pour la détection de parole, il s'agit des coefficients cepstraux, de la modulation de l'énergie à quatre hertz et de la modulation de l'entropie. Pour la détection de musique, nous avons utilisé des coefficients spectraux et les deux paramètres issus de la segmentation.

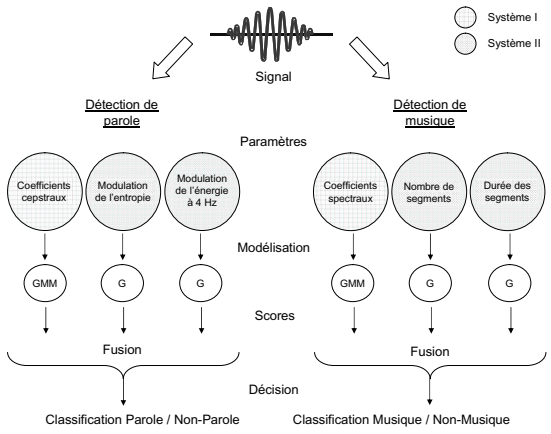

FIG. 1: le système global de classification.

2.1 Description du système I

Ce système, présenté dans [6], utilise la modélisation différenciée. Il se décompose en deux sous-systèmes. Le premier (Parole/NonParole) est basé sur une analyse cepstrale. Pour chaque trame (16 ms) d'analyse, 18 paramètres sont extraits (8 MFCC, l'énergie et leurs dérivées respectives). Ces paramètres sont ensuite normalisés par soustraction cepstrale. Le second (Musique/NonMusique) utilise une analyse spectrale afin d'extraire 29 paramètres (28 coefficients spectraux et l'énergie).

Ces deux sous-systèmes utilisent des GMM pour modéliser chacune de leurs classes (Parole, NonParole, Musique et NonMusique). Ces classes ont été apprises en utilisant deux algorithmes successifs. Le premier, un algorithme de quantification vectorielle (VQ) basé sur l'algorithme de Lloyd [8], permet d'initialiser les modèles. Le second est un algorithme permettant de réestimer les modèles (EM) [9]. Le nombre de lois gaussiennes utilisées a été fixé à 128.

2.2 Paramètres du système II

Une rapide description des paramètres utilisés par notre système de classification est faite dans cette section. Le lecteur désireux d'informations plus complètes pourra se référer à [7].

2.2.1 Modulation d'énergie à 4 Hz

Le signal de parole possède un pic caractéristique de modulation en énergie autour de la fréquence syllabique 4 Hz [10]. Nous avons calculé la modulation (variation) de l'énergie autour de 4 Hz. Ce paramètre a des valeurs plus élevées pour des segments de parole que pour les segments musicaux.

2.2.2 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. Le calcul de l'entropie est une technique de quantification de cet ordonnancement. Nous avons utilisé un paramètre basé sur la variation d'une estimation de la valeur de l'entropie sur des fenêtres de 1 seconde. Ce paramètre prend des valeurs plus élevées pour les segments de parole que pour les segments musicaux.

2.2.3 Paramètres de segmentation

La segmentation est issue de l'algorithme de "Divergence Forward-Backward" (DFB) [11] 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.

$$\begin{cases} y_n = \sum a_i y_{n-i} + e_n \\ var(e_n) = \sigma_n^2 \end{cases}$$

où (y_n) est le signal de parole et (e_n) est un bruit blanc gaussien.

La méthode consiste à détecter les changements des paramètres des modèles autorégressifs au travers des erreurs de prédiction calculées sur deux fenêtres d'analyse successives.
La statistisque est donnée par : $W_n = \sum_{k=1}^{n} w_k$,

$$w_k = \frac{1}{2}\left\{2\frac{e_k^0 e_k^1}{\sigma_1^2} - [1+\frac{\sigma_0^2}{\sigma_1^2}]\frac{e_k^0}{e_0^2} + [1-\frac{\sigma_0^2}{\sigma_1^2}]\right\}$$

et l'erreur de prédiction à l'instant k :

$$e_k^i = y_k - \sum_{j=1}^{p} a_j^i y_{k-j}, i = 0, 1.$$

Cette méthode a été comparée à de nombreuses autres méthodes de segmentation [12]. Elle a déjà fourni des résultats intéressants pour la reconnaissance automatique de la parole : des expériences ont montré que la durée des segments est porteuse d'une information pertinente [13].

Elle permet d'atteindre, notamment pour la parole, une segmentation subphonétique où 3 types de segments se distinguent :

- les segments quasi-stationnaires qui correspondent à la partie stable des phonèmes lorsqu'elle existe,
- les segments transitoires,
- les segments courts (environ 20ms).

Leur longueur varie entre 20 et 100 ms pour la parole. Pour la musique, un segment correspond à la tenue d'une note ; il peut être beaucoup plus long.

• 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) [14]. Au niveau de la segmentation, cela se traduit par de nombreux changements. La musique "instrumentale", étant plus tonale (ou harmonique), ne présente pas de telles variations.

Le nombre de segments par unité de temps (ici la seconde) est donc plus important pour la parole que pour la musique.

• Durée des segments

Comme le paramètre précédent, la durée des segments est issue de la même segmentation automatique (DFB). Elle correspond au calcul de la durée moyenne des segments les plus caractéristiques sur chaque seconde de signal. Les segments sont généralement plus longs pour la musique que pour la parole.

3 La fusion

La fusion d'information, qu'il s'agisse de fusion de paramètres ou de fusion de scores, est actuellement largement étudiée [15],[16]... Elle consiste à mettre à profit le maximum d'information sur les données afin de réduire les faiblesses de certaines grâce à d'autres. Dans l'étude présentée ici, la fusion des scores de vraisemblances est privilégiée.

3.1 Fusion par maximisation

Un premier travail a consisté à évaluer la fusion par maximisation. En effet, chacun des paramètres pour la détection de parole (resp. musique), modélisés soit par des mélanges de lois gaussiennes (système I) soit par des lois gaussiennes (système II), nous fournit un score de vraisemblance pour la Parole (resp. Musique) et la NonParole (resp. NonMusique) pour chaque seconde de signal étudié. Afin de prendre une décision, le paramètre possédant le score le plus important est privilégié.

3.2 Approche probabiliste

La théorie des probabilités permet de modéliser quantitativement l'incertitude. Intuitivement, la probabilité d'un évènement est une mesure normalisée (entre 0 et 1) de la vraisemblance de cet évènement. L'un des inconvénients majeurs de cette technique réside dans l'exigence de la connaissance parfaite des probabilités (notamment a priori). Enfin, la notion d'ignorance sur un fait n'est pas prise en compte. Pour pallier ce problème, on s'appuie sur des indices de confiance relatifs à l'expert, la classe ou l'observation.

L'expert nous apporte ses capacités à discriminer la Classe de la NonClasse avec un taux de confiance α_e :

$$\alpha_e = 1 - (P(NC > seuil) + P(C < seuil)).$$ [1]

L'indice de confiance de classe $\beta^e_.$ est en quelque sorte l'expérience que l'on a du modèle expert :

$$\beta^e_C = \frac{P(y=C)}{P(y=C)+P(y=NC)}$$
$$et\ \beta^e_{NC} = \frac{P(y=NC)}{P(y=C)+P(y=NC)}.$$

1. C={Parole,Musique} la Classe, NC={NonParole,NonMusique} la NonClasse et y l'observation

Les indices de classe (β^e_C, β^e_{NC}) permettent de définir une fonction de coût.

La confiance en l'observation, notée $\gamma^e_.$ est naturellement la vraisemblance de cette observation courante :

$$\gamma^e_C = P(y|C)\ et\ \gamma^e_{NC} = P(y|NC).$$

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

$$s^*_e(y) = min\left\{\left\{(1-\beta_{NC}) * \frac{\gamma_C}{P(y)}\right\}, \left\{(1-\beta_C) * \frac{\gamma_{NC}}{P(y)}\right\}\right\}.$$

La décision par fusion est l'argument qui maximise :

$$\alpha_e * s^*_e(y).$$

4 Expériences et évaluation

4.1 Corpus

Une base de données a été réalisée à partir d'enregistrements de RFI [2](Radio France Internationale).Cette base de données, contient de longues périodes de parole, de musique, ainsi que des zones de chevauchement pouvant contenir de la 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 : de nombreux instruments sont représentés. Il y a également des parties de voix chantée. Le corpus est multi-locuteur et multilingue.

Cette base a permis l'apprentissage du système I et l'évaluation du système total. Nous avons utilisé environ 8 heures pour l'apprentissage et 1 heure 30 minutes pour les tests. Les paramètres des lois et les seuils utilisés dans le système II n'ont pas été réappris, ils sont issus d'expériences précédentes [7].

4.2 Résultats

Nous avons testé séparément chaque paramètre. Les deux fonctions discriminantes, basées sur la modulation de l'énergie à 4 Hz et la modulation de l'entropie, montrent que ces différents paramètres fournissent des taux d'identification correcte similaires (autour de 84 %) pour de la classification Parole/NonParole (Tableau 1) : le taux est calculé par rapport un étiquetage manuel Parole/NonParole ramené à la seconde. La fonction discriminante basée sur le nombre de segments issus de l'algorithme de divergence donne un taux supérieur à 86 % pour de la détection de musique (Tableau 2). L'approche bayésienne, avec le paramètre de durée des segments et la loi Gaussienne inverse fournit un taux d'identification correcte légèrement plus bas (76 % pour de la Musique/NonMusique).

Avec le système I, les résultats sont de 90 % de classification correcte (unité : la seconde) pour la détection de parole (coefficients cepstraux) et de 87 % pour la musique (coefficients spectraux).

2. dans le cadre du projet RAIVES (projet CNRS)

TAB. 1: Classification Parole/NonParole

Paramètres	Taux d'identification correcte
(1) Coeff. cepstraux (système I)	90.9 %
(2) Modulation de l'energie à 4Hz	87.3 %
(3) Modulation de l'entropie	87.5 %
(1) + (2) + (3) fusion	**94 %**

TAB. 2: Classification Musique/NonMusique

Paramètres	Taux d'identification correcte
(1) Coeff. spectraux (système I)	87 %
(2) Nombre de segments	86.4 %
(3) Durée des segments	78.1 %
(1) + (2) + (3) fusion	**90 %**

Les résultats fournis par les différents paramètres du système I et II ont été ensuite fusionnés par maximisation des scores de vraisemblance (cf. paragraphe 3.1) et par la théorie des probabilités (cf. paragraphe 3.2). Ceci permet d'augmenter le taux de classification correcte pour arriver respectivement à 94 % d'accuracy pour la parole et 90 % pour la musique. Les 2 méthodes de fusion donnent des résultats similaires.

5 Discussion

Un système de classification parole/musique basé sur la fusion de deux approches a été présenté. L'approche par modélisation différenciée utilise des GMM, une analyse cepstrale pour la parole et une analyse spectrale pour la musique. En parallèle, quatre paramètres (la modulation de l'énergie à quatre hertz, la modulation de l'entropie et deux paramètres issus d'une segmentation) sont employés afin d'exploiter au mieux les propriétés du signal. Chacun des paramètres donne des résultats satisfaisants. La fusion des deux approches permet d'augmenter de le taux de reconnaissance d'environ trois points pour atteindre 94 % pour la détection de parole et 90 % pour la détection de musique.

Il apparaît intéressant d'améliorer un "lourd" (8 heures d'étiquettage manuel nécessaires à l'apprentissage) système de classification, basé sur une analyse spectrale et des MMG, grâce à des paramètres "simples" tels ceux présentés ici. Ces paramètres sont robustes et indépendants de la tâche car leurs seuils de classification ont été appris sur une base personnelle (différente du corpus RFI utilisé ici).

Références

[1] 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*. Nov. 1998, vol. 3527, pp. 398–409, SPIE.

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

[3] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *International Conference on Audio, Speech and Signal Processing*, Munich, Germany, Apr. 1997, pp. 1331–1334, IEEE.

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

[5] J. Foote, "Automatic audio segmentation using a measure of audio novelty," in *IEEE International Conference on Multimedia and Expo*, New-York, USA, 2000, pp. 452–455, IEEE.

[6] J. Pinquier, C. Sénac, and R. André-Obrecht, "Indexation de la bande sonore : recherche des composantes parole et musique," in *Congrès de Reconnaissance des Formes et Intelligence Artificielle*, Angers, France, Jan. 2002, pp. 163–170.

[7] J. Pinquier, Jean-Luc Rouas, and R. André-Obrecht, "Robust speech / music classification in audio documents," in *International Conference on Spoken Language Processing*, Denver, USA, Sept. 2002, vol. 3, pp. 2005–2008.

[8] J. Rissanen, "An universal prior for integers and estimation by minimum description length," *The Annals of Statistics*, vol. 11, pp. 416–431, Nov. 1982.

[9] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. 39 (Series B), pp. 1–38, 1977.

[10] 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*, vol. 77, no. 3, pp. 1069–1077, 1985.

[11] R. André-Obrecht, "A new statistical approach for automatic speech segmentation," *IEEE Transactions on Audio, Speech, and Signal Processing*, vol. 36, no. 1, Jan. 1988.

[12] R. André-Obrecht, "Segmentation et parole ?," M.S. thesis, IRISA, 1993.

[13] R. André-Obrecht and B. Jacob, "Direct identification vs. correlated models to process acoustic and articulatory informations in automatic speech recognition," in *International Conference on Audio, Speech and Signal Processing*, Munich, Germany, 1997, pp. 989–992, IEEE.

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

[15] D. Dubois and H. Prade, "La problématique scientifique du traitement de l'information," *Revue I3 (Information Intéraction Intelligence*, vol. 1, no. 2, July 2002.

[16] M. Detyniecki and R. R. Yager, "A context dependent method for ordering fuzzy numbers using probabilities," *International Journal of Uncertainty, Fuzzines and Knowledge-based systems*, vol. 8, 2000.