Sélection de Paramètres pour la Discrimination Parole/non Parole d'Emissions Radio Diffusées

Leila Zouari et Gérard Chollet

CNRS-LTCI
École Nat. Sup. des Télécommunications de Paris,
Département Traitement du Signal et des Images,
46 rue Barrault, 75314 Paris, France.
Tél : int+ 33 1 45 81 71 44, Fax : int+ 33 1 45 81 37 94
zouari,chollet@enst.fr

Résumé En reconnaissance automatique de la parole grand vocabulaire d'émissions radio diffusées, une étape crutiale de segmentation en parole/non parole est nécessaire. Or souvent, les segments de parole sont mélangés avec d'autres sons tels que la musique. Par conséquent, dans cet article, on se propose de trouver une paramétrisation adéquate aussi bien pour la parole que pour un mélange de parole+musique afin de bien les discriminer. On s'intéressera en particulier aux paramètres MFCC (Mel Frequency Cepstral Coefficients), LFCC (Linear Frequency Cepstral Coefficients) et à leurs combinaisons, et on évaluera trois types de combinaison, par fusion des paramètres, par fusion des scores et par fusion des décisions.
Nos expériences montrent que les coefficients MFCC sont plus performants en détection de la parole, que les paramètres LFCC le sont en reconnaissance de la musique+parole et que leur combinaison constitue un bon compromis lorsque des signaux de parole et de parole + musique sont tous les deux présents.
Mots clés Segmentation audio, Parole, Musique, MFCC, LFCC.

1 Introduction

Souvent dans les émissions radio, plusieurs types de signaux sont présents : parole, musique, parole + musique, jingles, silence, etc. Suivant le type d'application, différentes segmentations sont envisageables :

- segmentation musique/non-musique pour le traitements de la musique (classification par genre ou par instrument, etc),
- séparation parole/fond musical des segments de parole + musique pour le mixage audio ou la séparation des sources, etc,
- segmentation parole/non-parole pour la transcription orthographique et éventuellement la recherche d'information.

Ce travail s'insère dans le cadre du développement d'un système de transcription automatique des émissions radio. De ce fait, on se propose dans un premier temps de réaliser une segmentation du flux audio dans le but d'extraire les parties contenant de la parole ou de la parole mélangée avec la musique. La construction d'un tel système suscite le choix d'une

paramétrisation adéquate aussi bien pour la parole que pour la parole et la musique.
Les coefficients MFCC ont prouvé leur efficacité en traitement automatique de la parole. Leur succès provient, entre autres, de l'utilisation de l'échelle MEL qui favorise les basses fréquences. Dernièrement, Logan [5], a montré que les MFCC peuvent aussi représenter la musique sans pour autant se prononcer sur leur optimalité. L'échelle MEL n'étant pas optimale pour la musique puisqu'il peut y avoir autant d'information en basses fréquences qu'en hautes fréquences. Par conséquent, dans cet article, on s' est proposé de trouver la meilleure paramétrisation de la musique mélangée avec la parole. Les paramètres retenus seront combinés avec les coefficients MFCC pour mieux discriminer la parole et les mélanges de parole et de musique. Trois techniques de fusion ont été évaluées : fusion des paramètres, fusion des scores et fusion des décisions.
Après avoir présenté notre système de segmentation et décrit notre corpus de données, nous détaillerons les expériences de discrimination de la parole et de la parole+musique et nous en tirerons les conclusions.

2 Système de segmentation

Classiquement, la segmentation du flux audio est réalisée en deux temps. Dans un premier temps, on extrait du signal les paramètres jugés pertinents. Ces derniers doivent caractériser au mieux les classes à discriminer. Puis un processus de segmentation/classification permet d'affecter chaque partie du signal à une classe.

2.1 Paramétrisation

Notre paramétrisation est basée sur les coefficients cepstraux. Le signal audio est extrait de la séquence vidéo, échantionné à 16khz, puis les coefficients MFCC (MEL Frequency Cepstral Coefficients) et LFCC (Linear Frequency Cepstral Coefficients) sont calculés à partir d'un banc de 24 filtres. Ces filtres, de type MEL (échelle logarithmique) ou linéaires, sont appliqués toutes les 10 ms sur une fenêtre glissante de durée 20ms. Aux coefficients statiques (12 MFCC ou LFCC + énergie) nous rajoutons les dérivées première et seconde, ce qui permet d'obtenir des vecteurs de paramètres de dimension 39.

2.2 Classification

Les systèmes de segmentation de l'état de l'art font appel à des techniques tels que les modèles de Markov cachés (HMM) [4], les Modèles de Mélange Gaussien (GMM) [1,3], les k-plus proches voisins (KNN) [6], les réseaux de neurones, et plus récemment les Machines à Vectors de Supports (SVM) [2]. Comme ces travaux sont menés dans un objectif de reconnaissance de la parole, une approche de type GMM est adoptée. Ainsi, chaque classe est modélisée par un Modèle de Mélange Gaussien (GMM).
La classification, réalisée toutes les 10ms, est basée sur la règle suivante : soient N classes $C_1, C_2, .., C_N$ et le vecteur de test O. Le vecteur O est assigné à la classe la plus vraisemblable càd celle pour laquelle la vraisemblance $P(O/C_i)$ est maximale.
Quatre classes ont été utilisées : parole, musique, parole + musique et autres. Un nombre de composants de 256 gaussiennes par GMM est choisi empiriquement.

3 Expériences et résultats

3.1 Base de données

Nous avons utilisé une base de données de la variété télé "le grand échiquier". L'enregistrement dure trois heures et demi. Il contient de la parole, de la musique, la combinaison des deux (chants, jingles, ..) et des sons divers tels que les rires, les applaudissements, les effets spéciaux. Après avoir étiqueté manuellement cette base, nous l'avons découpée comme suit : 2h30 pour l'apprentissage et le reste pour l'évaluation. Le contenu de ces deux parties est explicité sur la figures Fig.1.

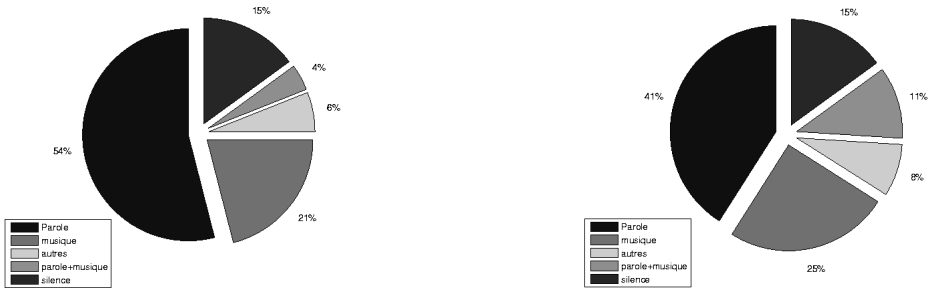

Fig. 1. Contenu des bases d'apprentissage (à gauche) et de test (à droite)

3.2 Mesure de performance

L'évaluation est réalisée trame par trame (toutes les 10ms). Les performances sont mesurées sur la base du rappel (R) et de la précision (P).

$$R = \sum_c T(c|c) / \sum_c T(c) \tag{1}$$

$$P = \sum_c T(c|c) / \sum_c T(c) + T(c|nc) \tag{2}$$

où $T(c|nc)$ est le temps où l'évènement c a été détecté à tort, $T(nc|c)$ le temps où c n'a pas été détecté à tort, $T(c)$ le temps où c est présent et $T(nc)$ le temps où c n'est pas présent.

Les évènements sont la parole, la musique et la parole+musique. Les performances des systèmes seront comparées sur la base de la F-mesure définie par :

$$F - mesure = 2 * R * P/(R + P) \tag{3}$$

Les temps seront mesurés en secondes. Dans les expériences qui suivent, on notera les valeurs de $F - mesure$ pour différentes $marges$, où $marge$ correspond à l'écart toléré (des limites) entre la segmentation automatique et la segmentation manuelle (en millisecondes).

3.3 Systèmes de MFCC/LFCC

Comme nous l'avons précisé dans l'introduction, l'objectif de ce travail et de trouver une paramétrisation adéquate et pour la parole et pour la parole mélangée avec la musique. Pour ce faire, nous avons commencé par développer deux systèmes de segmentation parole/ musique/ parole+musique/ autres en utilisant les paramétrisations MFCC et LFCC.

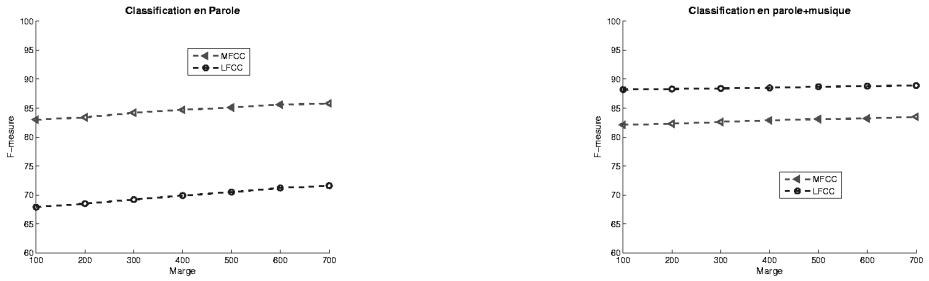

Fig. 2. Performances des paramétrisations MFCC et LFCC

Sur la figure Fig.2 sont reportées les performances du système en classification en parole et en parole + musique. On remarque un écart important entre la classification de la parole en utilisant les MFCC et celle en utilisant les LFCC. Pour la classification en parole + musique, les LFCC sont plus performants que les MFCC. Néanmoins, la différence n'est pas très importante.
Bien qu'intéressantes, ces constatations ne nous permettent pas de trancher entre MFCC et LFCC car dans les données de test, on ne connaît pas a priori les proportions de segments de parole et de parole + musique.

3.4 Combinaison des paramètres

Il s'agit d'une simple concaténation des paramètres MFCC et LFCC.

Fig. 3. Classification parole/musique/parole+musique/autres

La figure Fig.3 montre que les performances du système issu de la combinaison des paramètres sont entre celles du système MFCC et celles du système LFCC.

3.5 Combinaison des scores

Disposant des systèmes MFCC pour la parole et LFCC pour la parole et la musique, la combinaison des scores est réalisée en leur affectant des poids différents afin de privilégier l'une ou l'autre des paramétrisations. A chaque instant t, si $P(O_{MFCC};t)$ et $P(O_{LFCC};t)$ sont les vraisemblances d'une observation O calculées avec les systèmes MFCC et LFCC, alors son score de fusion peut s'exprimer par : $P(O_{fusion};t) = \lambda P(O_{MFCC};t) \times (1-\lambda) P(O_{LFCC};t)$

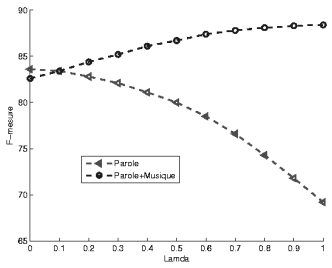

Fig. 4. Performances de la classification de la parole et de la parole + la musique en fonction de λ

Nous avons fait varier le poids λ entre 0 et 1. D'après la figure Fig.4, on peut constater que les performances de détection de la parole se dégradent lorsque λ augmente et le contraire pour la parole + la musique. Mais, vue la forte dégradation des performances de détection de la parole, on est tenté par l'utilisation de valeurs faibles de λ.

3.6 Combinaison des décisions

Il s'agit de fusionner les meilleurs systèmes de segmentation parole/non parole (P/NP) et parole+musique/non parole+musique (MP/NMP) pour en déduire une segmentation en 4 classes : parole (P), parole+musique (MP), musique (M) et autres (A). Les règles de fusion sont explicitées dans le tableau Tab.1.

Tab. 1. Règles de combinaison des décisions

Système MFCC	Fusion	Système LFCC
P → P	P	NMP ← P
MP → P	MP	MP ← MP
A → NP	A	NMP ← A
M → NP	M	MP ← M

On remarque (Fig.5) que la combinaison des systèmes apporte une amélioration par rapport aux coefficients LFCC pour la détection de la parole sans dépasser les performances avec les MFCC ce qui confirme leur supériorité en représentation de la parole. Pour la classification de la parole + la musique les performances de la fusion dépassent celles des

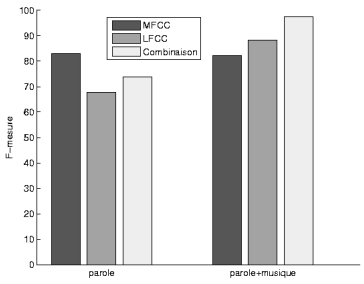

Fig. 5. Histogrammes de la fusion des décisions

deux systèmes de base. Par conséquent, la fusion des décisions pourrait constituer le meilleur compromis dans le cas où la nature des données de test est inconnue.

4 Conclusion

Dans le cadre d'une transcription automatique des émissions radio diffusées, une segmentation audio est réalisée. L'objectif d'une telle segmentation est de caractériser au mieux les segments de parole et de parole+musique qui seront par la suite transcrits. Les paramètres testés sont les coefficients MFCC, LFCC et leurs combinaisons. Les résulats de nos expériences montrent que les coefficients MFCC sont plus adéquats pour la discrimination de la parole, les coefficients LFCC le sont pour la musique+parole et leur combinaison l'est pour un flux audio contenant à la fois la parole et la parole mélangée avec de la musique.

5 Remerciements

Ce travail a été réalisé dans le cadre du pôle de compétitivité Cap-Digital et du réseau d'excellence KSpace. Nous tenons à remercier Youssef Boukhabrine pour sa contribution aux expériences.

Références

1. G. Linares C. Fredouille, D. Matrouf and P. Nocera. Segmentation en Macro-classes Acoustiques d'Emissions Radiophoniques dans le cadre d'ESTER. In *Journées d'Etude sur la Parole JEP*, 2004.
2. G.D. Guo and S.Z. Li. Content-based Audio Classification and Retrieval by Support Vector Machines. In *IEEE transactions on Neural Network*, Janvier 2003.
3. JL. Rouas J. Pinquier and R. A. Obrecht. A Fusion Study in Speech/Music Classification. In *proceedings ICASSP*, 2003.
4. O. Mella J. Razik, D. Fohr and P. Valles. Segmentation Parole/Musique pour la Transcription Rapide. In *Journées d'Etude sur la Parole JEP*, 2004.
5. B. Logan. Mel Frequency Cepstral Coefficients for Music Modelling. In *proceedings of the International Symposium on Music Information Retrieval*, 2000.
6. E. Scheirer and M. Slaney. Construction and Evaluation of a Robust Mainframe Speech/Music Discriminator. In *IEEE International Conference on Audio Speech and Signal Processing*, number 1331-1334, Munich, Germany, 1997.