Une nouvelle approche fondée sur les ondelettes pour la discrimination parole/musique

E. Didiot, I. Illina, O. Mella, D. Fohr, J.-P. Haton

LORIA-CNRS & INRIA Lorraine
BP 239, 54506 Vandoeuvre-les-Nancy, France
Mél : {didiot,illina,mella,fohr,jph}@loria.fr

ABSTRACT

The problem of Speech/Music discrimination is a challenging research problem which significantly impacts Automatic Speech Recognition (ASR) performance. This paper proposes new features for the Speech/Music discrimination task. We use a decomposition of the audio signal based on wavelets which allows a good analysis of non stationary signals like speech or music. We compute different energy types in each frequency band obtained from wavelet decomposition. We use two Class/Non-Class classifiers : one for speech/non speech, one for music/non music. On a broadcast corpus, using the proposed wavelet approach, we obtained a significant improvement (35%) compared to MFCC parameters.

1. INTRODUCTION

La discrimination entre la parole et la musique consiste à segmenter le flux audio en zones acoustiquement homogènes comme la parole, la musique ou la parole sur fond musical. C'est un problème important pour l'indexation de documents audio et pour la transcription automatique de programmes radiophoniques. Dans le cadre de la transcription, la séparation entre parole et musique permet d'éliminer les segments ne contenant que de la musique et donc diminuer le nombre d'erreurs de reconnaissance.
La discrimination parole/musique nécessite deux étapes : la paramétrisation puis la classification du signal audio. L'étape de paramétrisation consiste à extraire du signal des caractéristiques discriminantes entre la parole et la musique. De nombreux paramétrages ont été proposés dans l'état de l'art du domaine. Ils peuvent être classés en paramétrages fréquentiels, temporels, mixtes et dans le domaine cepstral.
Les paramètres fréquentiels sont issus de la DSP (Densité Spectrale de Puissance). La DSP d'un signal est issue de la transformée de Fourier de la fonction d'autocorrélation. Les paramètres fréquentiels permettent de capter le contenu fréquentiel d'un signal a un moment donné (formants, harmoniques, fréquence fondamentale, etc.). Le centroïde spectral, le flux spectral et le *spectral rolloff point* [12, 13] sont les plus utilisés.
Les paramètres temporels sont calculés directement à partir du signal audio. Ils permettent de capter des éléments relatifs à la variation de l'onde acoustique, comme par exemple, la périodicité ou les variations de l'onde sonore au cours du temps. Parmi ces paramètres, citons la mesure de rythmicité (*Pulse metric*), l'énergie et le taux de passage par zéro (*Zero Crossing Rate*) [12, 13, 14].
Les paramètres mixtes proviennent à la fois d'analyses fréquentielle et temporelle du signal. La modulation d'énergie à 4Hz [13], la modulation basse fréquence de l'amplitude du signal dans différentes bandes de fréquence de Karneback [6] ou encore le pourcentage de trames de faible énergie [12] en sont des exemples.
Les Coefficients Cepstraux (MFCC) permettent une représentation compacte du spectre du signal en prenant en compte, grâce à l'échelle Mel, des informations perceptives. Ces paramètres, très utilisés en reconnaissance de la parole, sont utilisés non seulement en discrimination parole/musique [1, 7, 10] mais aussi en classification de genres musicaux [15].
Cet article présente une nouvelle approche de discrimination parole/musique fondée sur l'utilisation de la décomposition en ondelettes du signal. A notre connaissance, une telle approche n'a jamais été utilisée pour cette tâche. Notre motivation pour l'utilisation des ondelettes est qu'elles mesurent les variations temporelles des composantes spectrales et permettent donc d'extraire les caractéristiques temporelles et fréquentielles du signal. De plus, la décomposition multi-bandes, qu'effectue la transformée en ondelettes dyadique, se rapproche fortement de ce que fait l'oreille humaine [4], c'est à dire qu'elle ressemble à une échelle logarithmique. Par rapport aux coefficients MFCC, les ondelettes ont une résolution temps-fréquence différente, permettent une représentation plus compacte du signal, possèdent un ensemble plus riche de fonctions de base et sont plus robustes à la non stationnarité du signal et à ses distorsions.
Dans notre travail, nous avons étudié différentes décompositions en ondelettes du signal, en extrayant des paramètres fondés sur l'énergie. Pour valider l'approche proposée, nous avons effectué des expériences sur un corpus d'émissions radiophoniques. Nous avons comparé la paramètrisation proposée au paramètrage MFCC.
Les sections 2 et 3 introduisent nos nouveaux paramètres et décrivent notre système de classification parole/musique. La section 4 présente les corpus utilisés tandis que les résultats de nos expériences sont détaillés dans la section 5. Enfin, la section 6 présente nos conclusions.

2. PARAMÈTRES FONDÉS SUR LES ONDELETTES

L'approche fondée sur les ondelettes est une approche de traitement du signal. Elle a été utilisée avec succès pour une large variété de problèmes, comme par exemple, les problèmes de débruitage ou récemment en reconnaissance automatique de la parole [11].
La transformée en ondelettes discrète (*Discrete Wavelet Transform*, DWT), que nous utilisons, analyse le si-

gnal dans différentes bandes de fréquence à différentes résolutions. Une telle analyse permet de contrôler les variables temps et fréquence du signal. Stéphane Mallat [8] a montré qu'une telle décomposition peut être obtenue par des filtrages passe-bas et passe-haut du signal temporel. Après chaque filtrage le signal est sous-échantillonné d'un facteur deux. Ce processus de décomposition est itéré sur les résultats du filtrage passe-bas jusqu'à l'obtention du nombre de bandes de fréquence désiré. Au final, le signal est décomposé en coefficients d'*approximation* et en coefficients de *détail*. Les coefficients d'approximation correspondent à des moyennes locales du signal. Les coefficients de détail, appelés « coefficients d'ondelettes », représentent les différences entre deux moyennes locales successives, c'est à dire entre deux approximations successives du signal.
Dans notre étude, nous utilisons la transformée en ondelettes dyadique correspondant à un banc de filtre en bandes d'octave. Dans le cas discret, cette transformée en ondelettes peut être définie par l'équation suivante :

$$W(k,j) = \sum_j \sum_k x(n) 2^{\frac{-j}{2}} \Psi(2^{-j}n - k)$$

où $x(n)$ correspond au signal à l'instant n, $\Psi()$ est une fonction temporelle de moyenne nulle, d'énergie finie et à décroissance rapide, appelée « ondelette mère ». La transformée en ondelettes *dyadique* effectue une décomposition du signal en bandes de fréquence non uniformes, ce qui permet d'obtenir une résolution fréquentielle décroissante lorsque les fréquences augmentent. Ainsi, cette décomposition en ondelettes donne une analyse multi-résolution du signal : une haute résolution temporelle et une basse résolution fréquentielle dans les hautes fréquences et inversement dans les basses fréquences. Elle offre ainsi une bonne modélisation du système auditif humain. Dans cet article, nous utilisons uniquement les coefficients d'ondelettes pour paramétrer le signal acoustique. L'utilisation des coefficients d'ondelettes permet de capter les modifications brutale du signal. En effet, les coefficients d'ondelettes prennent une grande valeur lors de tels événements. Dans chaque bande de fréquence, nous calculons sur les coefficients d'ondelettes, différents paramètres d'énergie de résolution temporelle variable [2].
Nous calculons :

– *Logarithme de l'énergie* (**E**) : l'énergie instantanée :

$$f_j = \log_{10}\left(\frac{1}{N_j}\sum_{k=1}^{N_j}(w_k^j)^2\right)$$

où w_k^j dénote le coefficient d'ondelettes à la position k et à l'échelle j, N_j le nombre de coefficients à l'échelle j, et f_j le vecteur de paramètres à l'échelle j.

– *Logarithme de l'énergie Teager* (**T_E**) : nous utilisons ici l'opérateur TEO (*Teager Energy Operator*) introduit par Kaiser [5]. Cet opérateur permet d'obtenir des paramètres robustes :

$$f_j = \log_{10}\left(\frac{1}{N_j}\sum_{k=1}^{N_j-1}|(w_k^j)^2 - w_{k-1}^j w_{k+1}^j|\right) \quad (1)$$

– *Logarithme de l'énergie hiérarchique* (**H_E**) : l'énergie hiérarchique correspond au calcul de l'énergie au centre de la fenêtre d'analyse en utilisant le même nombre de coefficients quelque soit la bande :

$$f_j = \log_{10}\left(\frac{1}{N_J}\sum_{k=(N_j-N_J)/2}^{(N_j+N_J)/2}(w_k^j)^2\right)$$

où J correspond à la résolution la plus basse. Cette énergie a été utilisée avec succès en paramétrisation pour la reconnaissance automatique de la parole [3].

3. SYSTÈME DE DISCRIMINATION PAROLE/MUSIQUE

3.1. Paramètrisation

Le signal est échantillonné à 16kHz. Après la préaccentuation, nous utilisons une fenêtre de Hamming de 32ms avec un déplacement de 10ms. Nous avons comme paramètres :
Paramètres MFCC de référence : 12 coefficients MFCC avec leur première et seconde dérivées (36 coefficients).
Paramètres fondés sur les ondelettes : ces paramètres sont calculés en utilisant deux familles d'ondelettes : daubechies et coiflet. Les paramètres multi-résolutions sont calculés pour différents niveaux de décompositions, c'est-à-dire différents nombres de bandes (de 5 à 7).

3.2. Description du système

Pour classifier les segments audio nous avons utilisé l'approche « anti-modèles »[9] : les modèles de classe et de non classe sont utilisés en parallèle pendant la classification. Des HMM sont utilisés pour modéliser chaque classe. Deux sous-systèmes sont mis en oeuvre : parole/non parole et musique/non musique.
Après le regroupement des décisions de ces deux sous-systèmes, le signal audio est classé en 3 catégories : parole (P), musique (M) et parole sur fond musical (PM). Pour éviter les segments de très courte durée, nous avons imposé une durée minimale (0.5s) pour chaque segment reconnu. Pour trouver la meilleure séquence de modèles qui décrit le signal, l'algorithme de Viterbi est utilisé.

4. CORPUS

4.1. Corpus d'apprentissage

Nous avons entraîné nos modèles sur deux corpus : *CDs audio* et *programmes radio*. Le corpus *CDs audio* (120 mn) est constitué de morceaux de musique instrumentale et de chansons, extraits de CDs. Le corpus '*programmes radio* (976 mn) contient des programmes de radios françaises : des journaux, des interviews et des programmes musicaux. Ces programmes sont très variés en terme de parole et de musique, de styles d'élocution, de locuteurs, de conditions d'enregistrement (téléphone, studio, interviews, bruits).

4.2. Corpus de test

Nous avons effectué nos expériences sur un corpus composé de 3 parties :

– La partie *News* est composée de trois fichiers d'une heure de des bulletins d'information de radios françaises (« France-Inter »et « Radio France International ») et contient principalement de la parole.
– La partie *Entertainment* est composée de trois

émissions de 20 minutes chacune (interviews et programmes musicaux). Cette partie est considérée comme difficile. En effet, elle comporte beaucoup de segments superposés (parole avec musique ou chanson) et des effets de *fade in-fade out*. Cette partie comprend aussi une alternance de parole de qualité studio et de qualité téléphonique. De plus, certaines interviews sont très bruitées.
– La partie *Scheirer* correspond au corpus de test construit et utilisé par E. Scheirer et M. Slaney [13]. Tous les fichiers audio sont homogènes et ont la même durée de 15 secondes : 20 fichiers de parole studio ou téléphonique et 41 fichiers de musique ou de voix chantée. Les styles musicaux sont plus nombreux (jazz, pop, country, etc.) que dans la partie *Entertainment*. Cette partie ne contient pas de parole sur fond musical.

Au total, ce corpus de test contient 74% de parole, 12% de parole sur musique et 14% de musique.

5. RÉSULTATS EXPÉRIMENTAUX

Pour évaluer notre système, nous avons utilisé 3 scores. Soit n_z^y le nombre de trames reconnues comme z alors qu'elles étaient étiquetées y, et soit nT, le nombre total de trames. Nous calculons :

– Taux de classification correct Global (TG) :

$$100 * (n_{PM}^{PM} + n_M^M + n_P^P)/nT$$

– Taux de classification Musique/Non Musique $(\mathrm{M/NM})$:

$$100 * (n_{PM}^{M} + n_M^{PM} + n_M^M + n_{PM}^{PM} + n_P^P)/nT$$

– Taux de classification Parole/Non Parole $(\mathrm{P/NP})$:

$$100 * (n_{PM}^{P} + n_P^{PM} + n_M^M + n_{PM}^{PM} + n_P^P)/nT$$

Pour le système de référence avec les coefficients MFCC, nous obtenons les résultats suivants : **TG = 82,0%, M/NM = 84,1% et P/NP = 96,2%**. L'intervalle de confiance est de $0,5\%$ à 5% de risque.

5.1. Influence du niveaux de décomposition et du calcul de l'énergie

Nous avons évalué des paramètrisations fondées sur différentes énergies (instantanée (**E**), Teager (**T_E**), hiérarchique (**H_E**)) calculées à partir des coefficients d'ondelettes. Après des expériences préliminaires, nous avons choisi d'utiliser les ondelettes suivantes : daubechies à 4 moments nuls (*db-4*) et coiflet à 2 moments nuls (*coif-1*). Différents niveaux de décomposition (nombre de bandes de fréquence) sont testés : de 5 à 7.

La table 1 montre que les meilleurs résultats en M/NM et P/NP sont obtenus avec l'ondelette *coif-1* : pour P/NP avec l'énergie Teager sur 5 bandes et pour M/NM avec l'énergie hiérarchique sur 7 bandes. Ainsi, en M/NM nous obtenons un gain relatif significatif de 42% (gain absolu de 6,7%) par rapport aux MFCC. En P/NP les résultats obtenus sont comparables aux MFCC. Notons que l'énergie hiérarchique permet une meilleure discrimination en M/NM lorsque le nombre de niveau de décomposition augmente. Le meilleur taux global est obtenu avec l'ondelette *coif-1* et l'énergie Teager sur 7 bandes : nous obtenons une diminution significative du taux d'erreurs de 6,7% comparé aux MFCC, soit un gain absolu de 1,2%. Dans la table 1, la deuxième colonne correspond au nombre de bandes de fréquence utilisé pour la décomposition en ondelettes et donc au nombre de paramètres. Nous observons que nos paramètres offrent une représentation plus compacte du signal que les MFCC, car avec seulement 7 coefficients, les paramètres basés sur les ondelettes sont déjà plus performants que les MFCC avec 36 coefficients. Enfin, pour la discrimination parole/non-parole, nous obtenons avec l'énergie *Teager* des résultats comparable aux MFCC, mais ceci avec moins de coefficients. Cette bonne performance de l'énergie Teager s'explique peut-être par le fait que l'opérateur de Teager permet de prendre en compte la dynamique à très court terme du signal (un coefficient avant et un coefficient après, voir équation (1)).

5.2. Influence des paramètres dynamiques

La durée d'une trame (10ms) n'est pas suffisante à un être humain pour faire la différence entre de la parole et de la musique. E. Scheirer et M. Slaney ont montré que l'utilisation de la variance sur une seconde de leurs paramètres améliorait les résultats en discrimination parole/musique[13]. Ainsi, l'étude de paramètres à plus long terme semble être intéressante.

Pour étudier la dynamique de nos paramètres à moyen terme, nous avons ajouté à nos paramètres statiques leurs dérivées première (Δ) et seconde ($\Delta\Delta$) (voir table 2). Pour étudier la dynamique à plus long terme, la variance dans une fenêtre d'une seconde est également calculée (voir table 3). Nous avons choisi comme paramètres statiques les paramètres qui ont donnés les meilleurs résultats auparavant : *coif-1* avec 7 bandes de fréquence.

La table 2 montre que l'ajout des dérivées a permis d'améliorer la discrimination parole/musique (TG) de 24% (avec Δ) et de 16% (avec $\Delta\Delta$) par rapport aux paramètres MFCC. Toutefois, on constate que l'ajout des dérivées secondes n'a pas apporté d'amélioration par rapport à l'ajout des dérivés premières. En ce qui concerne l'utilisation de la variance dans une fenêtre d'1s, elle permet d'obtenir de meilleurs résultats à la fois pour nos paramètres et pour les paramètres MFCC (table 3). Par exemple, la variance calculée sur les paramètres *énergie*

TAB. 1: Influence du niveau de décomposition et du type d'énergie en utilisant les ondelettes *db-4* et *coif-1*.

Type ond.	Bds	Ener.	M/NM	P/NP	TG
MFCC+Δ+ΔΔ			*84,1*	***96,2***	*82,0*
db-4	5	E	84,2	95,4	81,2
db-4	5	T_E	84,9	95,2	81,4
db-4	5	H_E	84,8	83,3	78,2
db-4	6	E	86,2	93,5	82,2
db-4	6	T_E	86,1	94,6	82,9
db-4	6	H_E	87,7	91,2	81,7
db-4	7	E	82,2	93,7	78,7
db-4	7	T_E	83,0	94,7	80,1
db-4	7	H_E	88,0	89,1	81,6
coif-1	5	E	88,5	95,2	82,0
coif-1	5	T_E	88,6	**96,0**	82,0
coif-1	5	H_E	84,7	82,9	76,8
coif-1	6	E	86,1	93,8	81,1
coif-1	6	T_E	87,5	95,2	82,6
coif-1	6	H_E	90,7	89,0	82,4
coif-1	7	E	88,0	92,4	82,1
coif-1	7	T_E	88,7	93,2	**83,2**
coif-1	7	H_E	**90,8**	85,7	80,1

TAB. 2: Influence des paramètres dynamiques (Δ et $\Delta\Delta$) en utilisant l'ondelette *coif-1* avec 7 bandes.

Param.	Nb	M/NM	P/NP	TG
MFCC+Δ+$\Delta\Delta$	36	*84,1*	*96,2*	*82,0*
E+Δ	14	88,0	95,9	85,9
T_E+Δ	14	**88,3**	**96,2**	**86,3**
H_E+Δ	14	88,6	95,8	84,1
E+Δ+$\Delta\Delta$	21	88,0	95,8	84,7
T_E+Δ+$\Delta\Delta$	21	86,8	96,1	84,9
H_E+Δ+$\Delta\Delta$	21	84,3	95,6	82,2

TAB. 3: Influence des paramètres dynamiques (variance sur 1 seconde) en utilisant l'ondelette *coif-1* avec 7 bandes.

Param.	Nb	M/NM	P/NP	TG
MFCC+Δ+$\Delta\Delta$	36	*84,1*	*96,2*	*82,0*
MFCC+Δ+$\Delta\Delta$(Var1s)	36	*86,4*	*95,9*	*84,2*
E Var 1s	7	90,4	96,0	88,0
T_E Var 1s	7	**90,6**	**96,4**	**88,4**
H_E Var 1s	7	90,3	88,7	84,1

Teager a permis d'améliorer significativement le Taux Global de 35% par rapport aux MFCC. Enfin, notons que nous obtenons ces derniers bons résultats en utilisant seulement 7 coefficients par vecteur de paramètres.

5.3. Expérimentations sur la partie Scheirer

Pour comparer nos résultats avec ceux obtenus par Scheirer et Slaney [13], nous avons utilisé le même corpus de test. Notons que le corpus d'apprentissage diffère de celui de Scheirer et Slaney. Nous comparons notre parametrisation avec la paramètrisation de Scheirer et Slaney suivante : la modulation de l'énergie à 4Hz, la variance du flux spectral et la mesure de rythmicité (*Best 3*). La classification est faite trame par trame, sans contraintes de durée minimum sur les segments. La table 4 montre que nos résultats sont comparables à ceux publiés dans [13].

TAB. 4: Performance par trame (%) sur la partie *Scheirer*.

Param.	M/NM	P/NP
MFCC+Δ+$\Delta\Delta$	*88,9*	*94,6*
MFCC+Δ+$\Delta\Delta$ (Var 1s)	*91,4*	*95,4*
Scheirer (Best 3)	*93,6*	***95,8***
coif-1, 7 bds, T_E (Var 1s)	**94,1**	94,9

6. CONCLUSION

Dans cet article, nous avons proposé de nouveaux paramètres fondés sur la décomposition en ondelettes du signal audio et sur le calcul de différentes énergies. Ces paramètres sont utilisés pour la tâche de discrimination parole/musique. Par rapport aux MFCC, la décomposition en ondelettes fournit une résolution temporelle non uniforme pour les différentes bandes de fréquence. Cette paramétrisation permet également d'obtenir une représentation plus compacte du signal et d'être plus robuste à la non-stationnarité des signaux. La décomposition dyadique que nous effectuons nous fournit une approximation de l'échelle Mel. Les paramètres finaux sont obtenus en calculant différents types d'énergie sur les coefficients d'ondelettes.

Notre nouvelle paramétrisation donne de meilleurs résultats de discrimination parole/musique que les coefficients MFCC avec leurs paramètres dynamiques. Nous avons ainsi un gain relatif significatif de 42% en classification musique/non musique et 6,7% en classification globale. L'utilisation des paramètres dynamiques (dérivées premières et seconde, variance sur 1s) améliore encore plus nos résultats : un gain relatif significatif de 35% en discrimination parole/musique, 3% en P/NP et 40% en M/NM, par rapport à la paramétrisation MFCC. Notons que notre paramétrisation utilise un nombre réduit de coefficients : 7 coefficients par rapport aux 36 coefficients MFCC. Plusieurs perspectives sont envisageables. D'une part, une fusion de différentes paramétrisations nous semble intéressante (MFCC+ondelettes). D'autre part, étant donnée que la variance sur 1s a montré de bons résultats, l'étude d'autres paramètres à long terme peut s'avérer prometteuse.

7. REMERCIEMENTS

Nous remercions Eric Scheirer et Malcolm Slaney pour nous avoir fourni leur corpus de parole et musique.

RÉFÉRENCES

[1] M.J. Carey, E.S. Parris, and H. Lloyd-Thomas. A Comparison Of Features For Speech, Music Discrimination. In *ICASSP-99*, 1999.

[2] M. Deviren. *Systèmes de reconnaissance de la parole revisit´es : R´eseaux Bayesiens dynamiques et nouveaux paradigmes*. PhD thesis, Universit´e Henri Poincar´e, 2004.

[3] R. Gemello, D. Albesano, L. Moisa, and R. De Mori. Integration of Fixed and Multiple Resolution Analysis in a Speech Recognition System. In *ICASSP-01*, 2001.

[4] S. Maes I. Daubechies. A Nonlinear Sqeezing of The Continuous Wavelet Transform based on Auditory Nerve Models. In *Wavelets in Medecine and Biology*, 1996.

[5] J.F. Kaiser. On a Simple Algorithm to Calculate the 'Energy'of a Signal. In *ICASSP-90*, 1990.

[6] S. Karneback. Discrimination between Speech and Music based on a Low Frequency Modulation Feature. In *European Conf. on Speech Comm. and Technology*, 2001.

[7] B. Logan. Mel Frequency Cepstral Coeffi cients for Music Modeling. In *International Symposium on Music Information Retrieval (ISMIR)*, 2000.

[8] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

[9] J. Pinquier. *Indexation sonore : recherche de composantes primaires pour une structuration audiovisuelle*. PhD thesis, Universit´e Paul Sabatier (Toulouse III), 2004.

[10] J. Razik, D. Fohr, O. Mella, and N. Parlangeau-Vall`es. Segmentation Parole/Musique pour la transcription automatique. In *JEP04*, 2004.

[11] R. Sarikaya and J.H.L. Hansen. High Resolution Speech Feature Parameterization for Monophone-based Stressed Speech Recognition. *IEEE Signal Processing Letters*, 2000.

[12] J. Saunders. Real-Time Discrimination of Broadcast Speech/Music. In *ICASSP-96*, 1996.

[13] E. Scheirer and M. Slaney. Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator. In *ICASSP-97*, 1997.

[14] C.C.J. Kuo T. Zhang. Hierarchical System for Content-Based Audio Classifi cation and Retrieval. In *Conference on Multimedia Storage and Archiving Systems III, tome 3527 de SPIE*, 1998.

[15] G. Tzanetakis and P. Cook. Musical Genre Classifi cation of Audio Signals. In *IEEE Transaction on Speech and Audio Processing*, 2004.