MajecSTIC 2009
Avignon, France, du 16 au 18 novembre 2009

Identification Musicale à l'aide de Technologies Vocales

Hugo Mauchrétien, Georges Linarès, Corinne Fredouille et Tania Jiménez.

CERI-LIA Avignon - BP1228 - 84911 Avignon Cedex 9 - France

Contact : hugo.mauchretien@etd.uapv.fr

Résumé
L'identification musicale est un processus d'appariement d'un extrait de musique à un morceau de musique connu. Les applications d'un tel système sont multiples, comme la protection des droits d'auteurs, ou plus simplement, permettre à un utilisateur d'identifier le morceau de musique qu'il écoute. Du fait des intérêts que peut présenter une telle application, plusieurs approches ont déjà été étudiées, le plus généralement basée sur des méthodes de reconnaissance des formes. Nous proposons d'utiliser les techniques de traitement de la parole, efficaces dans des environnements difficiles. Notre approche est basée sur les mixtures de gaussiennes (GMM), et les modèles de Markov cachés (HMM) qui sont des concepts très usités dans les domaines de traitement de la parole. Nous avons appliqué à la musique une méthode de segmentation du type regroupement en locuteurs. Les résultats de nos travaux sont convainquants puisque notre système est résistant aux bruits et à la compression du signal. Avec 25 dB de bruit, nous obtenons 100% d'identification correcte en une seconde de signal. En encodant nos morceaux en MP3 à 56 kbits, nous obtenons un taux d'identification de 100% avec trois secondes de signal.

Abstract
Music identification is a matching process of song snippet to a song known by the system. There are various use cases of a such system, as protection of copyright, or simply, to allow the user to identify the piece of music he listens, and to retrieve information like the artist name or the album title. Because of the interest of such application, several approaches have already been studied, most commonly based on methods of pattern recognition. We propose to use speech processing techniques, effective in difficult environments. Our approach is based on Gaussian mixtures (GMM) and hidden Markov models (HMM), which are widely used concepts in the speech processing domain. We applied to the music a speakers type segmentation method. Our experiments demonstrate an identification accuracy of 100% with 25 dB of additive noise and with low bitrate encoded MP3, sugesting our system is resistant to noise and signal compression.

Mots-clés : identification musicale, récupération d'information en fonction du contenu, acoustique, indexation, parole.

Keywords: Music identification, content-based information retrieval, acoustic, indexation, speech.

---

1. Introduction

L'intérêt pour des solutions d'identification musicale est relativement récent, et les scénarii d'utilisation de tels systèmes sont multiples. Pour les distributeurs de contenu multimédia, il peut se révéler financièrement intéressant de savoir quelle radio diffuse tel ou tel contenu. Il en va de même pour les droits vidéos par rapport à des sites comme Youtube. D'un point de vue de l'utilisateur final, il peut être très pratique de pouvoir identifier un morceau de musique avec un extrait de moins de dix secondes, avec un téléphone portable ou encore un ordinateur. Le système d'identification que nous proposons doit permettre d'identifier aussi vite que possible un morceau de musique diffusé dans des conditions potentiellement variables, cette variabilité pouvant introduire du bruit lié à la transmission du signal, ou encore à l'environnement d'acquisition ou de diffusion. La robustesse aux variabilités acoustiques est

un problème clef du traitement automatique de la parole. Nous proposons d'évaluer une approche basée sur le traitement automatique de la parole pour l'identification musicale, et d'evaluer sa robustesse a differents types de distorsions acoustiques.
Cet article est structuré comme suit : la deuxième *section* présente l'état de l'art, la troisième *section* présente le système proposé, avec l'extraction des paramètres, la segmentation, l'apprentissage et l'identification. Les expériences réalisées et leurs résultats sont présentés dans la *section* quatre.

2. Etat de l'art

2.1. L'identification musicale

Les travaux déjà réalisés dans le domaine de l'identification musicale peuvent être classés en deux groupes : les approches basées sur des empreintes audio et des tables de hachage [16] [8], et les approches plus récentes venant du domaine de la reconnaissance de la parole et du locuteur.
Les approches basées sur les empreintes audio ou ADN audio [13] établissent une empreinte à partir d'un échantillon de musique et interrogent le système d'identification, qui compare l'empreinte à celles qu'il connaît, puis renvoit le résultat ayant la métrique la plus faible par rapport à l'échantillon reçu. Les autres approches, venant généralement des techniques de reconnaissance automatique de la parole, se basent sur l'extraction de paramètres de type « Mel-Frequency Cepstra Coefficients » (MFCC) [14] [17], « Perceptual Linear Predictive » (PLP) [3] [5], et des Chroma [2]. Le tempo d'un morceau de musique est aussi pris en compte dans certains travaux [2].
Ces approchent ont leurs forces et leurs faiblesse, mais certaines sont plus robustes face à une variation de tempo [2], d'autres supportent mieux le bruit [1] ou la compression du signal [12] [17] [16] [1]. Le choix des paramètres dépend donc du contexte d'utilisation de l'application.

2.2. La segmentation

De nombreux algorithmes de segmentation ont été utilisés en traitement automatique de la parole. Celui que l'on retrouve le plus souvent est l'algorithme de Viterbi [15] [9]. Dans le domaine de la segmentation en locuteurs, une méthode appelée « E-HMM » (Evolutive HMM) a récemment été developpée [10] [11]. Nous avons utilisé cette méthode pour réaliser la segmentation des morceaux de musique.

3. Présentation du système proposé

En reconnaissance de la parole, les formes récurrentes sont toutes issues d'un ensemble reduit d'unités élémentaires (phonèmes), chaque mot de la langue étant modélisé, au niveau acoustique, comme une séquence d'unités de base. L'analogie morceau de musique-mot (ou morceau-phrase) supposerait que la musique soit, de la même facon, composée d'une succession de sons élémentaires, ce qui semble peut réaliste au vu de la diversité des voix, des intruments etde leur conjugaison. On s'est donc dirigés vers des techniques non-supervisées, dans lesquelles on ne dispose pas de connaissance *a priori* sur les formes à modéliser ni sur la structure temporelle des séquences. L'approche qu'on propose consiste à modéliser les sons élémentaires d'un morceau de musique par des GMM, la structure temporelle étant codée dans un HMM. Une des principales difficulté de ce type de tâche est lié à l'estimation des modèles, qui repose sur des hypothèses de segmentation et de regroupement des semgents. Cette estimation conjointe des formes acoustiques et de la structure globale de la session se retrouve dans les tâches de segmentation et regroupement en locuteur, dans lesquelles ont doit simultanément estimer les modèles de locuteur et la strucure globale de la session. Nous proposons d'utiliser l'algorithme E-HMM développé au LIA pour cette tâche [10] [11].
Le système connait donc un ensemble de morceaux dont il dispose d'une version de réference. Pour chacun de ces morceaux, nous estimons un HMM par le méthode E-HMM.
En phase d'identification, un graphe de reconnaissance est construit par parallélisation de ces HMMs. La reconnaissance du morceaux est réalisée par recherche du chemin de probabilité maximale dans le graphe sachant la séquence d'observations. L'identification est donc un processus synchrone qui doit évaluer la probablitié qu'une observation (un segment de musique) soit émise par un des modèles référencés dans un dictionnaire estimé a priori. On utilise l'algorithme de Viterbi [15] qui permet d'estimer, à chaque trame, la probabilité à posteriori de chacune des hypothèses de reconnaissance.

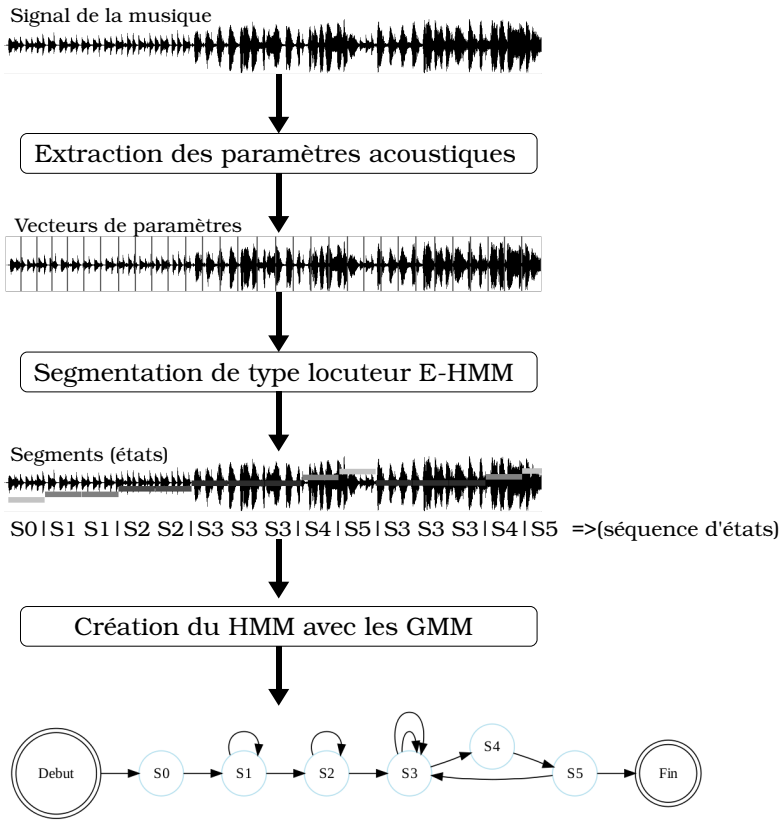

FIG. 1 – *Phases de création d'un modèle.*

Voici les différentes phases nécessaires à la construction de notre système (voir FIG.1) : la paramétrisation des fichiers, la segmentation et enfin l'apprentissage.

3.1. Extraction des paramètres, modélisation acoustique

La première étape est l'extraction de paramètres depuis les fichiers de musique. La paramétrisation consiste à analyser le signal d'un morceau de musique. Nous avons utilisé deux méthodes : PLP [3] [5] et MFCC [14] [17]. On applique au signal un filtre passe-bas de 8 kHz et on élimine sa composante continue. Une fenêtre de temps de 30 millisecondes avec un recouvrement de 10 ms défile sur le signal. Pour chaque unité acoustique contenue dans la fenêtre, la méthode PLP réalise une approximation de la densité spectrale à résolution variable (échelle de Bark) et calcule des coefficients tels que l'énergie pour une bande de fréquence donnée. Les coefficients cepstraux (MFCC), majoritairement utilisés en reconnaissance de la parole, permettent d'obtenir une représentation du signal sur l'échelle de Mel, proche de la perception humaine. Pour ces deux méthodes d'extraction des paramètres, nous avons utilisé les douze premiers coefficients et l'énergie, pour constituer des vecteurs à 13 dimensions, stockés dans des matrices de type HTK. Nous avons également réalisé des expériences en utilisant les dérivées premières et secondes de ses coefficients, pour obtenir des vecteurs de 39 dimensions.

3.2. La segmentation

Le but de cette étape de segmentation est de regrouper les segments homogènes pour en extraire des états (GMM). Lors de ce processus de segmentation, il nous faut atteindre un niveau de granularité assez fin pour avoir suffisament d'états pour modéliser correctement un morceau de musique. Une fois nos fichiers paramétrisés, nous les avons segmentés en utilisant l'outil de segmentation locuteur du CERILIA. Cet outil de segmentation fonctionne sur le principe d'un HMM évolutif [10] [11]. L'originalité

principale de cette méthode porte sur l'exploitation des informations (nouveaux segments détectés et segmentation provisoire) dès leur disponibilité en intégrant dans un même processus la détection de ruptures et la classification. Contrairement à une segmentation classique, toutes les informations disponibles sont exploitées à chaque étape et remises en cause à l'étape suivante.
Les états de ce modèle représentent les segments du morceau de musique ; les transitions entre ces états modélisent les changements de segments. Les segments, i.e. les états du HMM, sont ajoutés un à un à chaque itération du processus de segmentation.
La méthode se décompose en trois étapes précédées d'une initialisation du processus :
Initialisation :
– Le HMM contient un seul état représentant l'ensemble des segments du morceau. La segmentation contient un seul segment couvrant tout le morceau.

Etape 1 : création d'un nouveau segment
– Des données sont sélectionnées pour initialiser un nouveau segment. La segmentation provisoire et le HMM sont modifiés pour prendre en considération le nouvel état.

Etape 2 : adaptation des segments et segmentation intermédiaire
– Une étape itérative adapte les modèles du HMM et propose une segmentation provisoire.
  Cette étape s'arrête quand le maximum de vraisemblance au niveau de la segmentation provisoire est atteint.

Etape 3 : critère d'arrêt du processus
– Le processus s'arrête quand tous les segments ont été détectés, sinon il reprend à l'étape 1. Dans le modèle de segmentation proposé, la segmentation provisoire est remise en cause à chaque étape du processus.
  Cette opération s'effectue à deux niveaux :
  – Lors de la phase d'adaptation des segments (étape 2), un segment peut être divisé en plusieurs sous-segments.
  – L'arrêt du processus (étape 3) a pour conséquence de supprimer le dernier segment (état) ajouté.

Ce processus de segmentation permet d'obtenir les états (GMM) propres à un morceau de musique. Il permet également d'avoir une séquence unique d'états pour représenter un morceau de musique dans un modèle HMM. Les états ne sont pas partagés entre les morceaux de musique du modèle. Chaque état est composé de seize mixtures de gaussiennes.

3.3. Apprentissage

Cette partie du travail consiste à assembler un modèle permettant de représenter des connaissances extraites des morceaux de musique. Nous avons donc utilisé un outil d'entraînement de GMM/HMM pour créer notre modèle. Cet outil est basé sur l'algorithme EM (*Expectation-Maximization* [4]). Le modèle HMM créé contient des « tags » qui représentent chacun un morceau de musique par sa séquence d'états, déterminée précedemment lors du processus de segmentation. Chaque état du HMM (FIG.2) correspond à une densité de probabilité d'un évènement par rapport à un état et il y a une probabilité de transition d'un état à l'autre. Ceci contraint l'ordre temporel dans lequel les états doivent être observés.
Notre ensemble d'apprentissage est consitué d'une discographie d'un artiste unique. Nous avons utilisé cent-soixante fichiers MP3 compressés à 128 Kb/s, en stereo, à une frequence d'échantillonage de 44100 Hz. Nous avons appris des morceaux de musique, en les représentant par des séquences d'états (GMM) uniques. Chaque morceau de musique du modèle (HMM) est donc comparable à un mot dans le domaine de la reconnaissance automatique de la parole. Le fait de travailler avec la discographie d'un seul artiste rend a priori la tâche d'identification plus difficile qu'avec plusieurs artistes, puisque les morceaux de musique se ressemblent davantage.

3.4. L'identification

Lors du processus d'identification d'un extrait audio, le système parcours simultanément les *tags* du HMM et trouve le *tags* ayant la meilleure vraisemblance par rapport à l'extrait observé. L'identification est une estimation de la vraisemblance d'une séquence d'observations sachant un modèle. On recherche le chemin le plus probable dans la chaîne de Markov à l'aide de l'algorithme de Viterbi qui réalise un alignement dynamique. Cet algorithme teste la vraisemblance d'un vecteur d'observation sachant un état et calcule un coût de transition sous forme d'une probabilité.

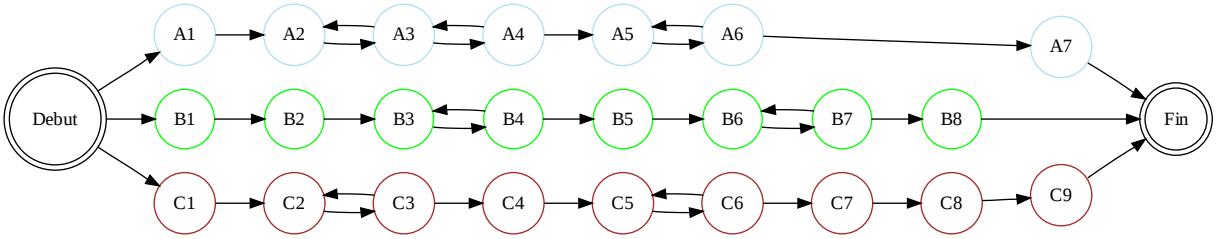

FIG. 2 – *Schéma d'un HMM contenant trois morceaux de musique [A,B,C].*

4. Les expériences

4.1. Protocole expérimental

Pour tester la validité de notre approche et évaluer les différentes méthodes d'extraction de paramètres, nous avons créé plusieurs modèles différents. Tous nos fichiers MP3 ont été mis au format 16 bit, 16 kHz en mono pour ensuite en extraire des paramètres. Nous avons utilisé les paramétrisations MFCC et PLP avec 13 et 39 coefficients pour apprendre quatre modèles différents. Chaque morceau de musique du modèle (HMM) est donc comparable à un mot dans le domaine de la reconnaissance automatique de la parole. Ces modèles ont ensuite été évalués par rapport à plusieurs types de contraintes : la compression/dégradation du signal sonore et le bruit. Les ensembles de tests pour évaluer la robustesse de nos modèles ont été créés à partir des 160 morceaux de musique encodés en MP3. Nous avons travaillé avec des morceaux de musique coupées à trente secondes, pour des questions de rapidité des expériences.
De plus, il ne nous a pas paru intéressant de créer un système qui ne reconnaît pas un morceau de musique en moins de trente secondes.

4.1.1. Opérations de compression et dégradation du signal

Les fichiers MP3 ont été réencodés en MP3 56 Kbps. Nous avons fait du *streaming* avec nos MP3 originaux en 32 et 64 Kbps pour imiter les conditions d'émission d'une radio Internet. Nous avons conservé le taux d'échantillonage original des morceaux, à savoir 44,1 kHz et nous avons réalisé le *streaming* en mono. Nous avons également encodé nos fichiers en GSM (*Global System for Mobile communications*), qui est un codec conçu pour compresser la voix sur les appareils de téléphonie mobiles. Les fichiers ont également été encodés avec le codec G711, fréquemment utilisé en VoIP.

4.1.2. Opérations d'ajout de bruits au signal

Pour simuler des conditions réelles, nous avons choisi quatre niveaux de rapports signal/bruit différents : 3, 6, 9 et 25 dB. Les bruits ajoutés viennent de quatre types d'environnements : des bruits de rues, des bruits de lieux administratifs, des bruits de véhicules et des bruits extérieurs non urbains, comme un parc pour enfant. Pour chaque type d'environnement bruité, nous avons sélectionné en moyenne 39 fichiers que nous avons mixés aléatoirement avec nos 160 morceaux de musique, aux quatre niveaux de rapports signal sur bruit choisis.

4.2. Résultats des expériences

Les résultats sont présentés en deux parties : la première présente les résultats d'expériences sur la compression et la dégradation du signal, la deuxième présente les résultats d'expériences portant sur l'ajout de bruit au signal. Les modèles créés avec une paramétrisation à 13 coefficients donnent des résultats moins bons que ceux créés avec une paramétrisation utilisant les dérivées premières et secondes de ces 13 coefficients. Par exemple, pour des MP3 encodés à 56 kbits, le modèle utilisant le type de paramétrisation PLP a 13 coefficients donne des taux d'identification de plus de 10 points inférieurs aux taux obtenus avec celui en 39 coefficients. On constate donc que le fait de prendre en compte les dérivées premières et secondes des paramètres confère à notre système une meilleure résistance aux dégradations du signal.

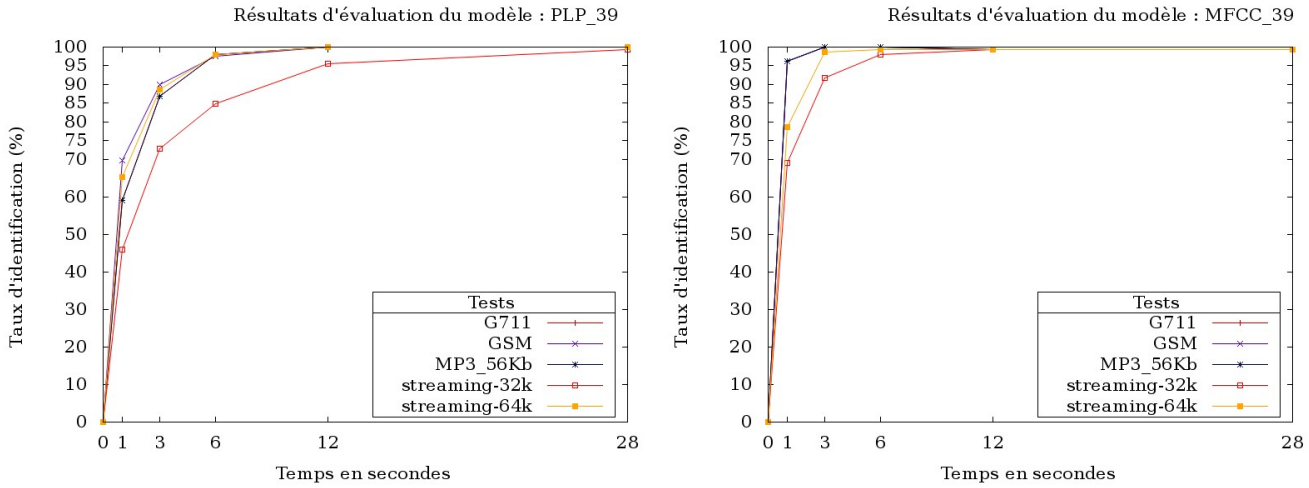

FIG. 3 – *Resultats des expériences sur la dégradation du signal.*

4.2.1. Compression et dégradation du signal

Au niveau de la perception humaine, l'expérience ayant le plus dégradé le signal est celle utilisant le codec GSM. A l'écoute, on distingue des distorsions plus importantes que dans les autres expériences.
Sur le graphique du modèle nommé « MFCC_39 » (voir FIG.3), la courbe qui correspond à l'expérience avec le codec G711 se confond avec celle de l'expérience avec les morceaux encodés en MP3 à 56Kbits. On peut donc dire que ce modèle, qui donne les meilleurs taux d'identification, reconnaît aussi bien le son dégradé par le codec G711 que par un encodage MP3 à 56Kbits. Sur ce même graphique, on voit que pour les deux codecs GSM et G711, ainsi que pour les MP3 encodés à 56 kbits, le taux d'identification ateint 100% au bout de trois secondes. Pour le test de *streaming* à 64 kbits, le meilleur taux d'identification est de 99.37%, ateint au bout de six secondes, dans une intervalle de confiance de +/- 1.2%. Pour le test de *streaming* à 32 kbits, au bout de six secondes, on ateint un taux d'identification de 98.11%, avec une intervalle de confiance de +/- 2.1%.
De manière générale, les paramètres de type PLP donnent un taux d'identification inférieur aux MFCC dans nos expériences de compression et dégradation du signal. Cependant, à propos du modèle PLP avec des paramètres à 39 dimensions, on peut noter qu'il identifie mieux à la première seconde les fichiers compressés avec le codec G711 que le modèle MFCC.
Des deux méthodes d'extraction de paramètres utilisés, à savoir MFCC et PLP, la plus robuste à la compression et à la dégradation du signal semble être la méthode MFCC.

4.2.2. Ajout de bruit au signal

Pour synthétiser nos expériences, nous avons fait la moyenne des résultats obtenus avec les quatre types de bruits d'environnement FIG.4.
Commentaires à propos du modèle MFCC à 39 coefficients (voir FIG.4) :
- On peut voir qu'avec un rapport signal sur bruit de 25 décibels, l'identification ne pose pas de difficultées puisqu'on identifie correctement plus de 99,8% des morceaux de musique (avec un intervalle de confiance de +/- 0.7%), dès la première demi-seconde de signal, et on ateint 100% à une seconde.
- Pour un rapport signal sur bruit fixé 9 dB, l'identification est plus difficile puisque le meilleur taux, qui est de 91,6%, est ateint à une seconde, avec une intervalle de 4.2%.
- En revanche, pour un rapport signal sur bruit à 6 dB, les résultats sont moins bons car le meilleur taux d'identification, qui est de 83,5%, est ateint au bout d'une seconde, avec une intervalle de 5.7%.

Ce modèle est donc le meilleur, même si le modèle créé à partir des paramètres PLP est assez proche, puisqu'il est à un ou deux point d'écart maximum. Cependant, pour un rapport signal sur bruit fixé à 3 dB, on remarque que le modèle PLP créé avec 39 coefficients (FIG.4) a un taux d'identification supérieur au modèle MFCC. Pour ce rapport signal sur bruit, il y a 1.28 point d'écart en moyenne entre les deux

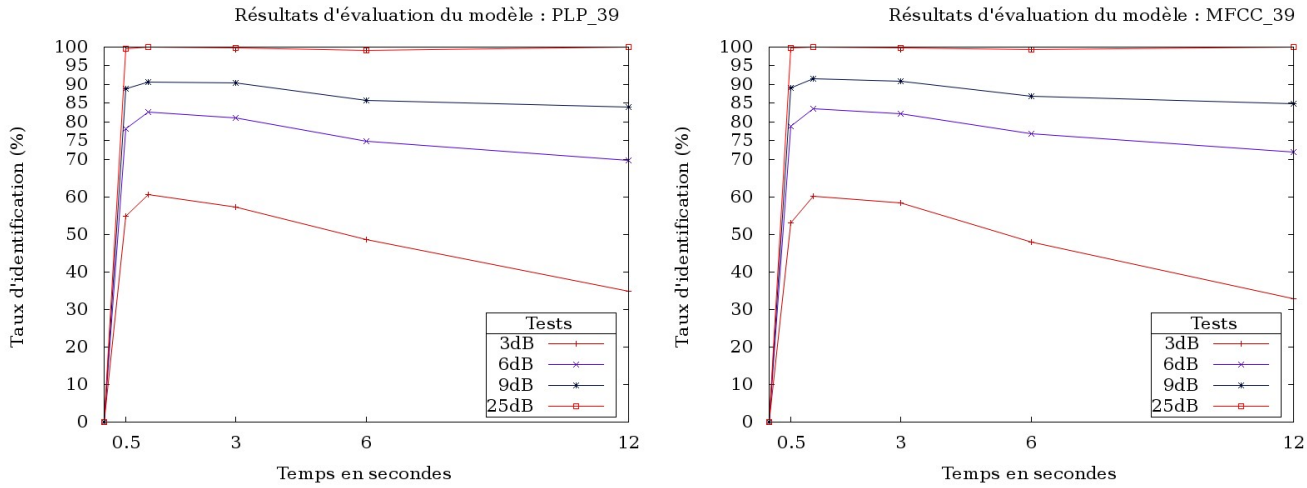

FIG. 4 – *Resultats des expériences réalisées en « bruitant » le signal.*

modèles.
Le fait que les courbes soient décroissantes à partir d'une seconde s'explique de plusieurs façons :
– il y a probablement des problèmes pour modéliser la durée des morceaux dans le HMM. Si la structure temporelle n'est pas bien représentée, cela est source d'erreur lors du décodage des modèles.
– une autre explication est que les bruits ajoutés ne sont pas très influants dans les premières secondes car leur amplitude est trop faible.
– une autre explication, moins probable, est que le HMM est davantage discriminant au début car les premiers états des morceaux sont plus différents que les suivants (par exemple le delai avant le premier son diffère dans chaque morceau).

5. Conclusion, perspectives

En utilisant les techniques venant du traitement automatique de la parole dans notre travail d'identification musicale, nous avons conçu un système résistant aux bruits et à la dégradation du signal. Les résultats sont convaincants, puisqu'ils sont d'un niveau apparement équivalent aux travaux récemment réalisés dans ce domaine [12] [1] [17], même si la démarche expérimentale diffère.
Les travaux dont fait l'objet cet article permettent de faire le point sur les avantages des paramètres de type MFCC par rapport aux PLP dans le domaine de l'identification musicale. Les paramètres MFCC résistent mieux aux bruits, à la compression et à la dégradation du signal. Le fait de prendre en considération la vitesse et l'accélération du signal lors de la paramétrisation permet de créer des modèles plus résistants aux distorsions du signal.
Pour la phase de segmentation, il serait intéressant de tester des segmenteurs fonctionnant sur le principe de l'algorithme de Viterbi [18], ou encore sur la méthode BIC (*Bayesian information criterion*) [6], [7].
Il serait intéressant de rendre le décodage plus rapide en rendant les modèles plus discriminants. Ceci pourrait être réalisé d'une part en apprenant les morceaux dans leur globalité et en réalisant une classification rendant plus discriminants les premiers états des modèles. Et d'autre part, en trouvant une méthode pour prendre en compte la durée dans les modèles, ce qui donnerait au système un gain de précision et de rapidité.

Merci à Driss Matrouf

Bibliographie

1. E. Batlle, J. Masip, E. Guaus, and P. Cano. Scalability issues in an hmm-based audio fingerprinting. In *Proc. IEEE International Conference on Multimedia and Expo ICME '04*, volume 1, pages 735–738, 30–30 June 2004.
2. D.P.W. Ellis and G.E. Poliner. Identifying 'cover songs' with chroma features and dynamic programming beat tracking. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007*, volume 4, pages IV–1429–IV–1432, 2007.
3. H. Hermansky, B. Hanson, and H. Wakita. Perceptually based linear predictive analysis of speech. In *Proc. IEEE International Conference on ICASSP '85. Acoustics, Speech, and Signal Processing*, volume 10, pages 509–512, Apr 1985.
4. S. Huda, J. Yearwood, and R. Togneri. A constraint-based evolutionary learning approach to the expectation maximization for optimal estimation of the hidden markov model for speech signal modeling. *Systems, Man, and Cybernetics, Part B : Cybernetics, IEEE Transactions on*, 39(1) :182–197, Feb. 2009.
5. J.-C. Junqua, H. Wakita, and H. Hermansky. Evaluation and optimization of perceptually-based asr front-end. 1(1) :39–48, 1993.
6. Hyoung-Gook Kim, D. Ertelt, and T. Sikora. Hybrid speaker-based segmentation system using model-level clustering. volume 1, pages 745–748, 18-23, 2005.
7. M. Kotti, E. Benetos, and C. Kotropoulos. Automatic speaker change detection with the bayesian information criterion using mpeg-7 features and a fusion scheme. pages 4 pp.–, 0-0 2006.
8. Y. Liu, H. S. Yun, and N. S. Kim. Audio fingerprinting based on multiple hashing in dct domain. 16(6) :525–528, June 2009.
9. A. Ljolje and M.D. Riley. Automatic segmentation and labeling of speech. In *Proc. International Conference on Acoustics, Speech, and Signal Processing ICASSP-91*, pages 473–476, 14–17 April 1991.
10. S. Meignier, J.-F. Bonastre, C. Fredouille, and T. Merlin. Evolutive hmm for multi-speaker tracking system. volume 2, pages II1201–II1204 vol.2, 2000.
11. S. Meignier, J.-F. Bonastre, and S. Igounet. E-hmm approach for learning and adapting sound models for speaker indexing. In *Proc. Odyssey Speaker and Language Recognition Workshop*, pages 175–180, 2001.
12. M. Mohri, P. Moreno, and E. Weinstein. Efficient music identification with weighted finite-state transducers. *IEEE Transactions on Audio, Speech, and Language Processing*, PP(99) :1–1, 2009.
13. H. Mayer P. Cano, E. Batlle and H. Neuschmied. Robust sound modeling for song detection in broadcast audio. In *Proc. AES 112th Int. Conv*, pages 1–7, 2002.
14. Sigurdur Sigurdsson, Kaare Br, T Petersen, and Tue Lehn-schiøler. Mel frequency cepstral coefficients : An evaluation of robustness of mp3 encoded music. In *Proceedings of the International Symposium on Music Information Retrieval*, 2006.
15. A. Viterbi. Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *Information Theory, IEEE Transactions on*, 13(2) :260–269, Apr 1967.
16. Avery Wang. The shazam music recognition service. *Commun. ACM*, 49(8) :44–48, 2006.
17. E. Weinstein and P. Moreno. Music identification with weighted finite-state transducers. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2007*, volume 2, pages II–689–II–692, 2007.
18. L. Wilcox, F. Chen, D. Kimber, and V. Balasubramanian. Segmentation of speech using speaker identification. volume i, pages I/161–I/164 vol.1, Apr 1994.