Fusion de paramètres pour une classification automatique parole/musique robuste

Séparation parole/musique dans les fichiers audio

Julien Pinquier — Jean-Luc Rouas — Régine André-Obrecht

Laboratoire IRIT, UMR 5505 CNRS UPS INP
118, route de Narbonne
F-31062 Toulouse Cedex 4
{Julien.Pinquier, Jean-Luc.Rouas, Regine.Andre-Obrecht}@irit.fr

RÉSUMÉ. *Dans cet article, une nouvelle approche relative à l'indexation de la bande sonore de documents audiovisuels est proposée, son but est de détecter les composantes parole et musique. Trois nouveaux paramètres sont extraits : la modulation de l'entropie, la durée des segments (issue d'une segmentation automatique) et le nombre de ces segments par seconde. Les informations issues de ces trois paramètres sont ensuite fusionnées avec celle issue de la modulation de l'énergie à 4 Hz. Une première expérience, effectuée sur un corpus de parole lue et de diverses sortes de musique, permet de montrer l'intérêt de chacun des paramètres par sa distribution. Ensuite, un deuxième corpus est utilisé afin de vérifier la robustesse des paramètres et du système de fusion proposé. Cette expérience, réalisée sur un corpus radiophonique, donne un taux d'identification correcte supérieur à 90 %.*

ABSTRACT. *This paper deals with a novel approach to speech/music segmentation. Three original features, entropy modulation, stationary segment duration and number of segments are extracted. They are merged with the classical 4Hz modulation energy. The relevance of these features is studied in a first experiment based on a development corpus composed of collected samples of speech and music. Another corpus is employed to verify the robustness of the algorithm. This experiment is made on radio corpus and shows performances reaching a correct identification rate of 90 %.*

MOTS-CLÉS : *classification, fusion, documents sonores, paramètres acoustiques, segmentation, distribution, durée, entropie, énergie.*

KEYWORDS: *classification, merging, audio documents, acoustic parameters, segmentation, distribution, duration, entropy, energy.*

1. Introduction

A l'heure actuelle, les méthodes d'indexation en audio et vidéo sont principalement manuelles : un opérateur humain doit lire, écouter et/ou regarder le document numérique de façon à sélectionner les informations recherchées. Cette tâche d'indexation doit être automatisée car le volume de données s'accroît énormément et le traitement de plusieurs requêtes devient extrêmement fastidieux.

Le document audio ou la bande sonore d'un document audiovisuel est très complexe puisqu'il résulte d'un mixage entre plusieurs sources sonores. Si l'on se réfère à la norme MPEG7, indexer un document sonore signifie rechercher les composantes primaires (parole, musique), identifier des sons clés (applaudissements, effets spéciaux...), détecter et identifier les locuteurs, trouver des mots-clés ou des jingles, ou rechercher des thèmes. La discrimination entre parole et musique est alors une étape obligée du processus.

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) [SAU 96], [ZHA 98] tandis que la variation de la magnitude spectrale (le « Flux » spectral) permet de détecter les continuités harmoniques [SCH 97]. D'autre part, dans la communauté du traitement automatique de la parole, les paramètres cepstraux sont privilégiés pour extraire les zones de parole [GAU 99] et [FOO 00].

Trois approches sont communément utilisées pour la classification : les modèles de mélanges de lois gaussiennes ([SCH 97] et [WOL 99]), les k plus proches voisins [CAR 99] et les modèles de Markov cachés ([ZHA 98] et [KIM 96]). Un état de l'art en indexation audio [CAR 00] permet d'avoir de plus amples informations sur les paramètres existants pour la classification de documents sonores.

Dans une étude précédente [PIN 02], nous avons utilisé un système d'indexation basé sur une modélisation différenciée, où il n'était plus question de chercher à discriminer la parole de la musique, mais à les caractériser au mieux de façon indépendante afin de faire une séparation de type classe/non-classe (c'est-à-dire parole/non-parole et musique/non-musique). L'approche était mise en œuvre à partir de modèles de mélanges de lois gaussiennes (MMG) en utilisant des paramètres spectraux (fréquentiels) pour la classification musique/non-musique et cepstraux (MFCC) pour la classification parole/non-parole. Cette approche est concluante en terme de résultats (environ 90 %) mais nécessite un changement (ou une réestimation) de nos modèles (parole, non-parole, musique et non-musique) lorsque nous utilisons un corpus de test différent de celui de l'apprentissage.

Dans cet article, une nouvelle méthode de classification est présentée. Celle-ci consiste toujours à détecter de manière disjointe les segments de parole et de musique.

L'originalité de ce travail est l'utilisation de paramètres inhabituels : la modulation de l'entropie, la durée des segments (issue d'une segmentation automatique) et le nombre de ces segments par seconde. Ces paramètres se révèlent tout aussi performants que ceux classiquement utilisés dans la littérature pour la classification parole/autres ou musique/autres ([SCH 97]).

Le choix de ces paramètres a été fait de manière à détecter ces deux composantes aussi bien dans des zones simples (zones contenant seulement de la parole ou de la musique) que dans des zones critiques (zones contenant à la fois de la parole, de la musique et/ou du bruit). Ces paramètres sont fusionnés avec la modulation de l'énergie à 4 Hz. L'intérêt de cette méthode est qu'une fois les quelques seuils appris, aucun nouvel apprentissage n'est nécessaire pour traiter un document d'un nouveau type et enregistré dans de nouvelles conditions : l'estimation de nos paramètres est faite une fois pour toutes.

Cet article est divisé en trois parties. Une première section permet de décrire le système global de classification et les paramètres. Une première expérience, effectuée sur un corpus de parole lue et de toutes sortes de musique, permet de montrer la pertinence du choix des paramètres par leur distribution et fait l'objet de la deuxième section. Au cours du dernier paragraphe, un deuxième corpus (radiophonique) est employé afin de vérifier la robustesse des paramètres et du système de fusion proposé dans des conditions très diverses (reportages, informations, chansons, interviews...).

2. Le système global et ses paramètres

Le système de fusion d'informations proposé est basé sur l'extraction de quatre paramètres :

– la modulation de l'énergie à 4 Hz,

– la modulation de l'entropie,

– le nombre de segments par seconde,

– la durée de ces segments.

Notre système (figure 1) se décompose en deux systèmes de classification correspondant aux deux détections disjointes de la parole et de la musique. Nous les appellerons respectivement système parole/non-parole et système 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 fusionnant les scores (vraisemblances) issus de la modélisation de chacun des paramètres.

2.1. *Le système global*

Un prétraitement commun aux deux sous-systèmes consiste à détecter le silence afin de ne traiter que les zones d'activité acoustique. Cette détection se fait sur la

base de calculs d'énergie par rapport à un seuil. Cette méthode est classiquement utilisée dans la littérature [ZHA 98]. Dans notre système, les résultats sont donnés pour chaque seconde. Donc une classification « silence » signifiera que le silence est majoritairement représenté durant la seconde de test, c'est-à-dire au moins durant 0,5 s.

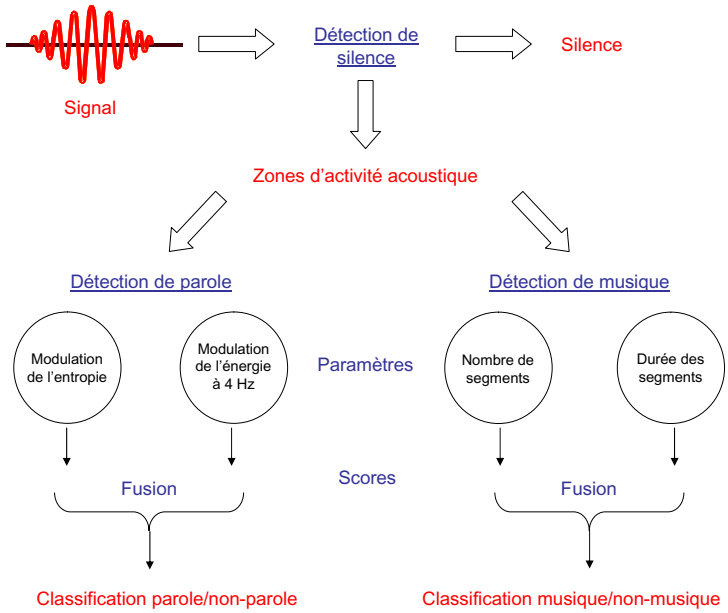

Figure 1. *Le système global de fusion de paramètres*

Le système parole/non-parole repose sur l'extraction de la modulation de l'énergie à 4 Hertz et la modulation de l'entropie tandis que le système musique/non-musique utilise les deux paramètres issus d'une segmentation automatique. Dans le système parole/non-parole, l'information primaire traitée est issue de l'analyse d'une trame de 16 ms, alors que dans le système musique/non-musique, l'information est obtenue à partir du traitement de segments de taille variable : la notion de trame n'existe pas. Dans chaque cas, une décision est prise sur l'analyse complète d'une fenêtre d'une seconde.

Chaque « classifieur », correspondant à un paramètre, est défini à partir d'un modèle statistique : une loi inverse gaussienne pour la durée des segments et une loi gaussienne pour les autres paramètres. Le choix (la réponse) du « classifieur » est donné avec un certain indice de confiance (score de vraisemblance). Ensuite, nous avons fusionné les scores obtenus. Dans ce but, il nous a fallu rechercher une règle pouvant tirer le meilleur parti de chacun de ces paramètres. Nous avons opté pour une classification hiérarchique en deux parties (figure 1).

– Dans la première partie, consacrée à la détection de parole, nous avons fusionné les deux paramètres de modulation de l'énergie à 4 Hz et de modulation de l'entropie par maximisation des scores de vraisemblance. L'indice de confiance le plus important détermine le choix (parole ou non-parole).

– Dans la seconde partie, consacrée à la détection de musique, les deux paramètres de segmentation (le nombre de segments par seconde et la durée moyenne de ces segments par seconde) ont été fusionnés. La méthode de fusion est la même que précédemment : par maximisation des scores.

2.2. *Modulation de l'énergie à 4 Hz*

Le signal de parole possède un pic caractéristique de modulation en énergie autour de la fréquence syllabique 4 Hz [HOU 85]. Pour extraire et modéliser cette propriété, la procédure suivante est appliquée.

1) Le signal est découpé en trames de 16 ms sans recouvrement.

2) Pour chaque trame, 40 coefficients spectraux sont extraits suivant l'échelle Mel et correspondent à l'énergie des 40 bandes de fréquence.

3) Pour chaque bande, cette énergie est filtrée grâce à un filtre à Réponse Impulsionnelle Finie (RIF) passe-bande de fréquence centrale 4 Hz (figure 2).

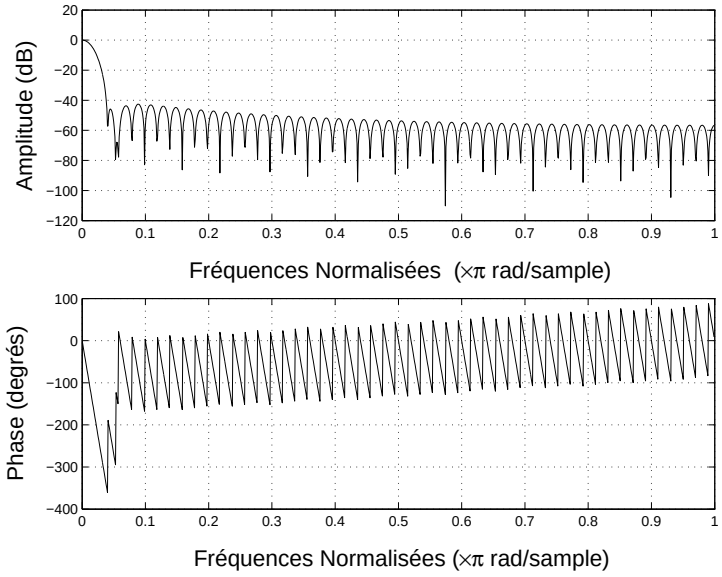

Figure 2. *Réponse en fréquence et en phase du fi ltre RIF*

4) La somme des énergies filtrées est effectuée sur l'ensemble des canaux et est normalisée par l'énergie moyenne.

5) La modulation est obtenue en calculant la variance de l'énergie filtrée en décibels, sur une seconde de signal.

La parole possède une modulation de l'énergie plus forte que la musique (figure 3).

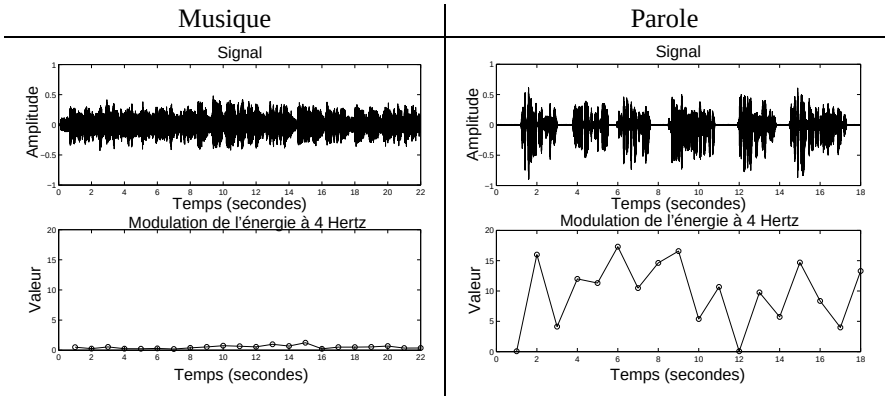

Figure 3. *Modulation de l'énergie à 4 Hz pour la parole (6 phrases de parole lue) et la musique (extrait de Mozart)*

2.3. *Modulation de l'entropie*

Des observations menées sur le signal ainsi que sur le spectrogramme font apparaître une structure plus « ordonnée » du signal de musique par rapport au signal de parole. Pour mesurer ce « désordre », nous avons testé un paramètre basé sur l'entropie du signal [MOD 89] :

$$H = \sum_{i=1}^{k} -p_i log_2 p_i,$$

avec p_i = probabilité de l'événement i et k = nombre d'issues.

Une procédure similaire à celle employée pour le paramètre de modulation de l'énergie à 4 Hz est appliquée.

1) Le signal est découpé en trames de 16 ms sans recouvrement.

2) L'entropie est estimée pour chaque trame grâce à l'estimateur non biaisé décrit ci-après.

Le calcul de l'entropie se fait en deux étapes. Tout d'abord, un histogramme est calculé puis l'entropie est ensuite estimée.

On pose : N = le nombre d'échantillons contenus dans la fenêtre considérée et S les valeurs d'amplitude prises par le signal : $S = \{S_1...S_n...S_N\}$

– Calcul de l'histogramme

La borne minimale de l'histogramme est définie par : $min_h = min(S) - \frac{\Delta}{2}$;
La borne maximale est définie par : $max_h = max(S) + \frac{\Delta}{2}$;
avec $\Delta = \frac{max(S)-min(S)}{N-1}$.

Le nombre de paliers de l'histogramme est défini par arrondi supérieur de la racine carré du nombre d'échantillons $N_h \approx \sqrt{N}$.

– Estimation de l'entropie

Une fois l'histogramme obtenu, on a les probabilités d'apparition des différentes valeurs de l'amplitude (on notera h_i l'effectif de la classe i, $i = 1...N_h$). En considérant que les échantillons sont indépendants ($\widehat{H} = \Sigma_{n=1}^{N} \widehat{H_n}$), on effectue le calcul de l'estimateur biaisé :

$$\widehat{H}_{biased} = \frac{\Sigma_i(-h_i log(h_i))}{N} + log(N) + log(\frac{max_h - min_h}{N_h});$$

Le biais est alors : $nbias = -\frac{N_h-1}{2N}$.
On obtient l'estimateur non biaisé en enlevant le biais : $\widehat{H}_{unbiased} = \widehat{H}_{biased} - nbias$.

3) La modulation est obtenue en calculant la variance de l'entropie sur une seconde de signal. Puisque nous avons choisi de calculer l'entropie sur des trames de 16 ms, nous obtenons 62 valeurs de l'entropie $\widehat{H}$ par seconde. On pose $\Psi = \{\widehat{H_1}...\widehat{H_{62}}\}$, alors la modulation d'entropie est définie par : $modulation_H = var(\Psi)$.

La modulation de l'entropie est plus élevée pour la parole que pour la musique (figure 4).

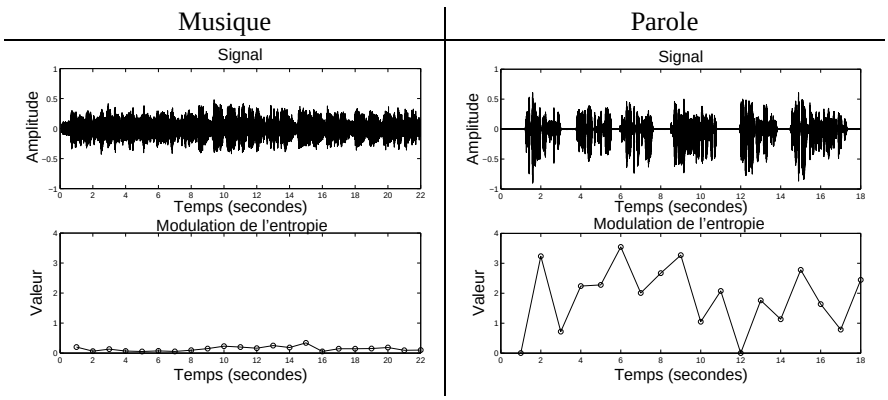

Figure 4. *Modulation de l'entropie pour la parole et la musique*

2.4. *Paramètres de segmentation*

2.4.1. *Segmentation automatique*

La segmentation est issue de l'algorithme de « Divergence Forward-Backward » (DFB) [AND 88] qui est basé sur une étude statistique du signal dans le domaine temporel. En faisant l'hypothèse que le signal de parole est décrit par une suite de zones quasi stationnaires, chacune est 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 de modèles autorégressifs au travers des erreurs de prédiction calculées sur deux fenêtres d'analyse (figure 5). La distance entre ces deux modèles est obtenue à partir de l'entropie mutuelle des deux lois conditionnelles correspondantes.

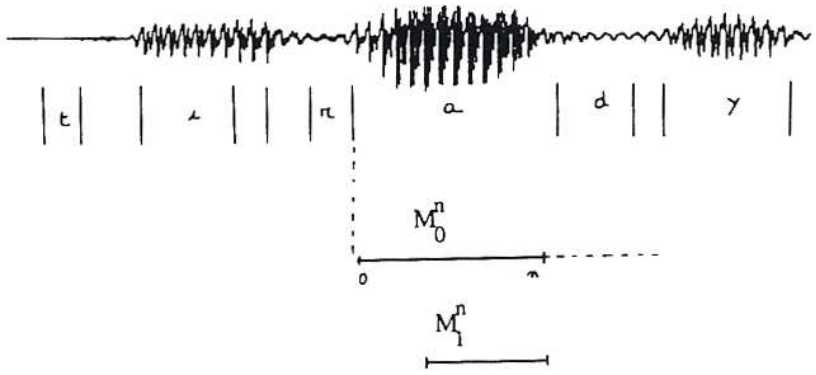

Figure 5. *Localisation des fenêtres d'estimation des modèles M_0^n et M_1^n à l'instant n ; l'instant "0" correspond à la dernière frontière validée. La phrase prononcée est : « il se garantira du... »*

La statistique est définie comme une somme cumulée : $W_n = \sum_{k=1}^{n} w_k$

avec w_k l'entropie mutuelle dans le cas gaussien :

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

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 [AND 93]. 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 [AND 97].

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 20 ms).

Leur longueur varie entre 20 et 100 ms pour la parole (figure 6).

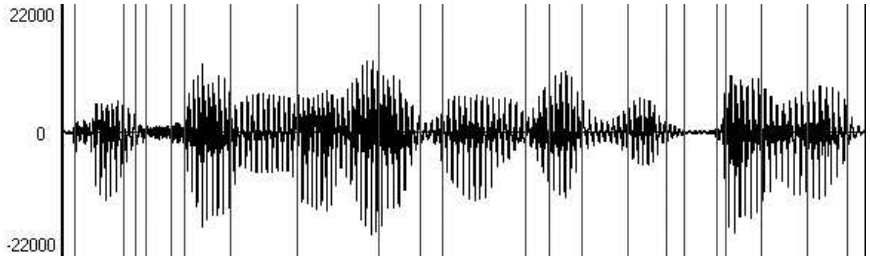

Figure 6. *Résultat de la segmentation sur environ 1 seconde de parole. La phrase prononcée est : « Confi rmez le rendez-vous par écrit »*

Pour la musique, un segment correspond à la tenue d'une note ; il peut être beaucoup plus long (figure 7).

Figure 7. *Résultat de la segmentation sur environ 1 seconde de musique de l'extrait de Mozart déjà utilisé*

2.4.2. *Paramètres*

– 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) [CAL 89]. Au niveau de la segmentation, cela se traduit par de nombreux changements. La musique, 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 (23 segments dans notre exemple de la figure 6) que pour la musique (10 segments dans notre exemple de la figure 7).

– Durée des segments

Comme le paramètre précédent, la durée des segments est issue de la même segmentation automatique (DFB). Afin de limiter la corrélation de ces deux paramètres de segmentation, la durée moyenne des segments sur une seconde n'est pas calculée sur tous les segments des fenêtres mais seulement sur les 7 plus longs. Le nombre de segments caractéristiques a été fixé expérimentalement. Les segments sont généralement plus longs pour la musique (180 ms dans notre exemple de la figure 7) que pour la parole (80 ms dans notre exemple de la figure 6).

2.5. *Les échelles de temps du système*

De manière à bien comprendre le fonctionnement de notre système global et la fonctionnalité de chaque paramètre, il convient de préciser la signification de certains termes qui seront utilisés tout au long de cet article afin de ne pas les confondre.

– La **trame** est l'unité correspondant au découpage du signal (ici, toutes les 16 ms) en vue de calculer les paramètres de modulation de l'énergie à 4 Hertz et de modulation de l'entropie.

– Le **segment** est l'unité représentative de la segmentation par l'algorithme DFB. Il est de taille variable et les paramètres du sous-système de classification musique/non-musique l'utilisent.

– La **fenêtre** de décision est quant à elle représentative de la prise de décision. Dans notre système, celle-ci est de longueur fixe (1 seconde) quel que soit le paramètre utilisé.

La figure 8 représente ces trois échelles de temps sur un même signal.

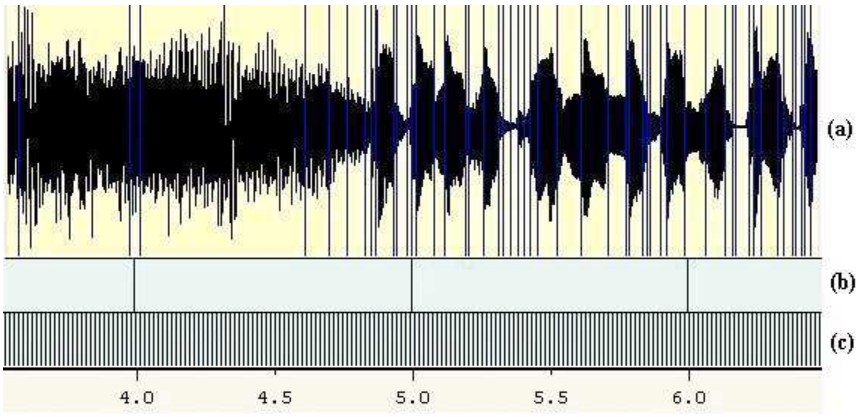

Figure 8. *Représentation des échelles de temps de notre système : les segments de taille variable en (a),les fenêtres de décision (une par seconde) en (b) et les trames d'analyse (une toutes les 16 ms) en (c)*

3. Étude des distributions des paramètres

Pour évaluer la pertinence de ces paramètres, nous avons utilisé d'une part un corpus de parole lue contenant 5 langues européennes, le corpus MULTEXT [CAM 98]. Ce corpus est lu par 10 locuteurs par langue (5 hommes et 5 femmes). Les enregistrements sont de bonne qualité, le taux d'échantillonnage est de 20 kHz. D'autre part, nous avons créé un corpus d'extraits musicaux contenant différentes sortes de musique, allant du rock au classique avec un taux d'échantillonnage de 16 kHz.

La durée totale pour chaque corpus (parole et musique) est d'environ 35 minutes (soit plus de 2000 segments d'une seconde).

3.1. *Modulation de l'énergie à 4 Hz*

L'histogramme obtenu pour le paramètre de modulation de l'énergie à 4 Hz pour la parole et la musique est montré sur la figure 9.

La parole et la musique sont clairement dissociées. L'intersection des deux histogrammes (modulation de l'énergie = 2,5) peut être utilisée comme seuil, dans le cadre d'une approche bayésienne.

En faisant l'hypothèse que le volume et la diversité des données sont suffisamment significatifs, nous pouvons estimer les probabilités d'erreur :

$$\begin{cases} Pr(musique|parole) = Pr(musique > seuil) = 6,4\%. \\ Pr(parole|musique) = Pr(parole < seuil) = 3,2\%. \end{cases}$$

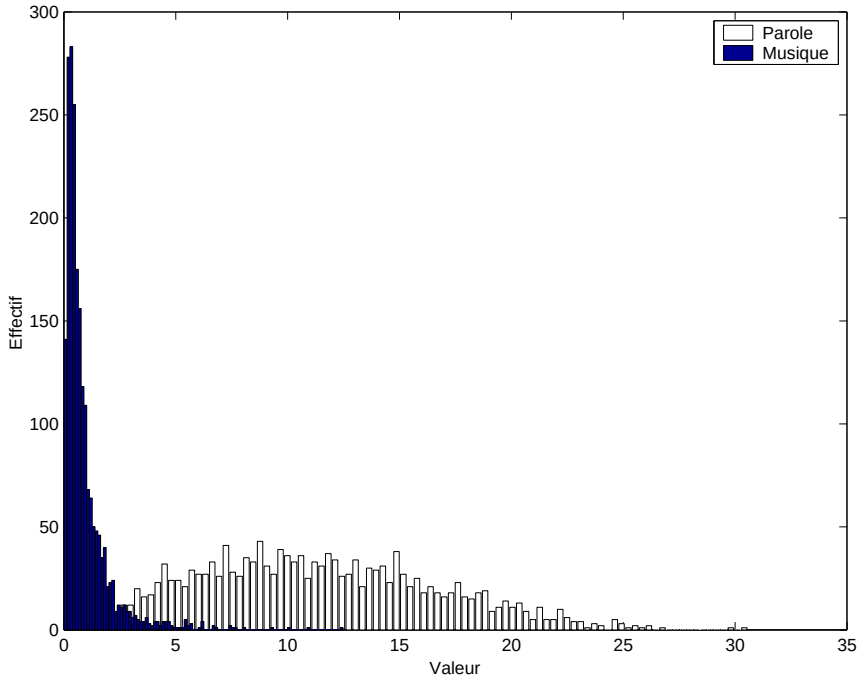

Figure 9. *Distribution de la modulation de l'énergie à 4 Hz par seconde pour la parole (blanc) et pour la musique (noir)*

3.2. *Modulation de l'entropie*

La même expérience a été reconduite avec le paramètre de modulation de l'entropie. Les résultats sont montrés sur la figure 10.

Ce paramètre est également pertinent dans la tâche de séparation parole/musique. Chaque histogramme est clairement séparé, et nous pouvons également déterminer un seuil expérimental (modulation de l'entropie = 0,5).

En faisant la même hypothèse que précédemment sur le volume et la diversité des données, nous pouvons également estimer les probabilités d'erreur :

$$\begin{cases} Pr(musique|parole) = Pr(musique > seuil) = 7,2\%. \\ Pr(parole|musique) = Pr(parole < seuil) = 3,4\%. \end{cases}$$

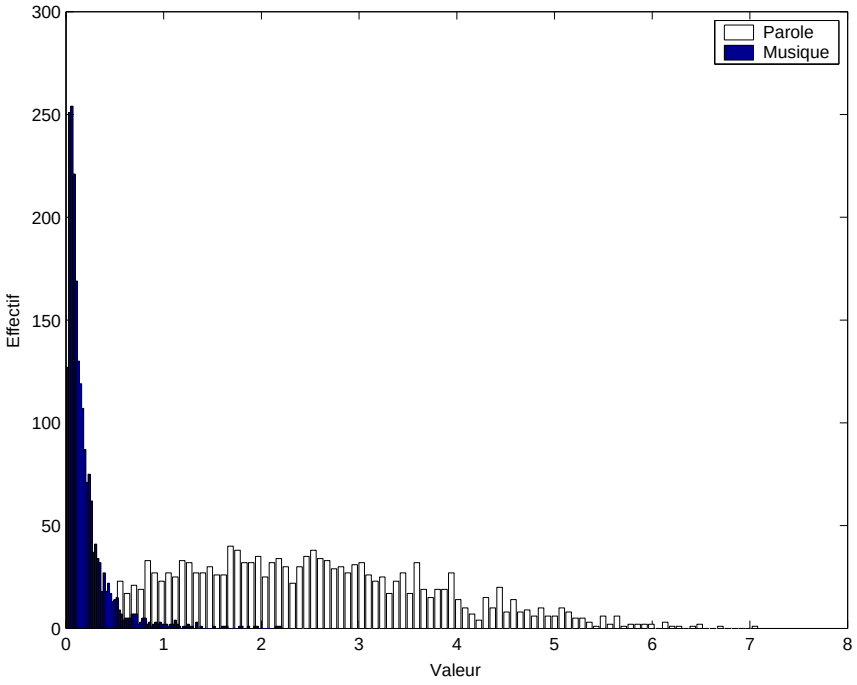

Figure 10. *Distribution de la modulation de l'entropie par seconde pour la parole (blanc) et pour la musique (noir)*

3.3. *Paramètres de segmentation*

3.3.1. *Nombre de segments*

La répartition du nombre de segments obtenus automatiquement est représentée sur la figure 11.

Les deux histogrammes montrent que ce paramètre est pertinent, la parole et la musique peuvent être discriminées au moyen d'un simple seuillage (seuil = 17 segments par seconde). Les probabilités d'erreur sont :

$$\begin{cases} Pr(musique|parole) = Pr(musique > seuil) = 11,6\%. \\ Pr(parole|musique) = Pr(parole < seuil) = 3,6\%. \end{cases}$$

3.3.2. *Durée des segments*

Comme le montre la figure 12, fixer un seuil avec une hypothèse gaussienne dans le cas de la stratégie bayésienne (utilisée pour les paramètres précédents), est impossible.

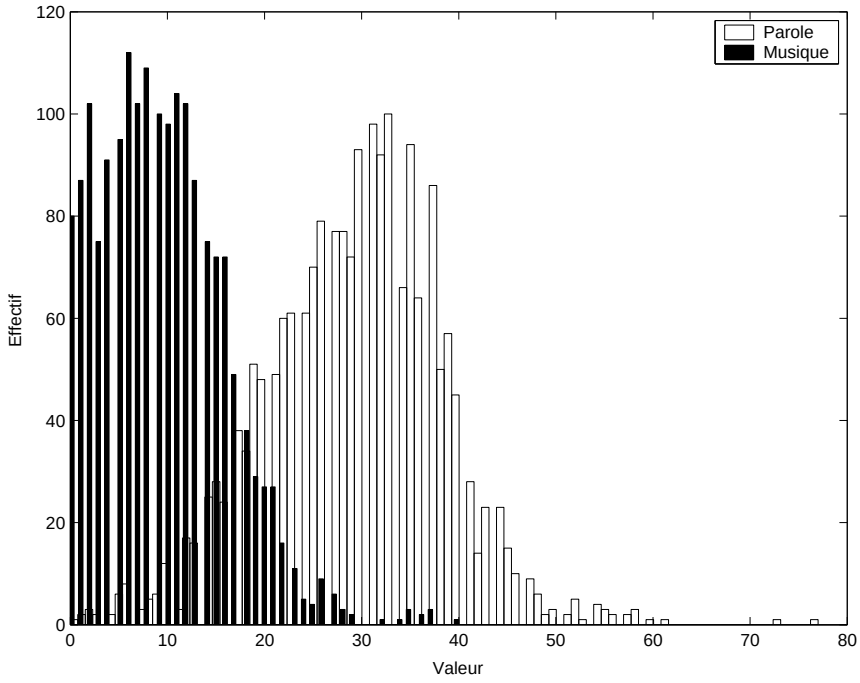

Figure 11. *Distribution du nombre de segments par seconde pour la parole (blanc) et pour la musique(noir)*

Une étude statistique [SUA 94] montre que la loi gaussienne inverse (loi de Wald) est une loi de probabilité qui modélise la durée des sons. C'est pourquoi la durée des segments a été modélisée à l'aide d'une loi de Wald paramétrée par μ et λ. Par définition, une variable aléatoire g suit une distribution inverse gaussienne si elle présente une fonction de densité de probabilité (pdf) de la forme [JOH 70] :

$$\begin{cases} p(g) = \sqrt{\frac{\lambda}{2\pi g^3}} * e^{\frac{-\lambda(g-\mu)^2}{2\mu^2 g}}, & \text{si } g \geq 0 \\ p(g) = 0, & \text{sinon} \end{cases}$$

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

La figure 12 décrit la répartition des durées des segments pour la parole et pour la musique. Les lois de Wald, correspondant aux distributions, sont tracées avec les paramètres estimés pour la parole et la musique.

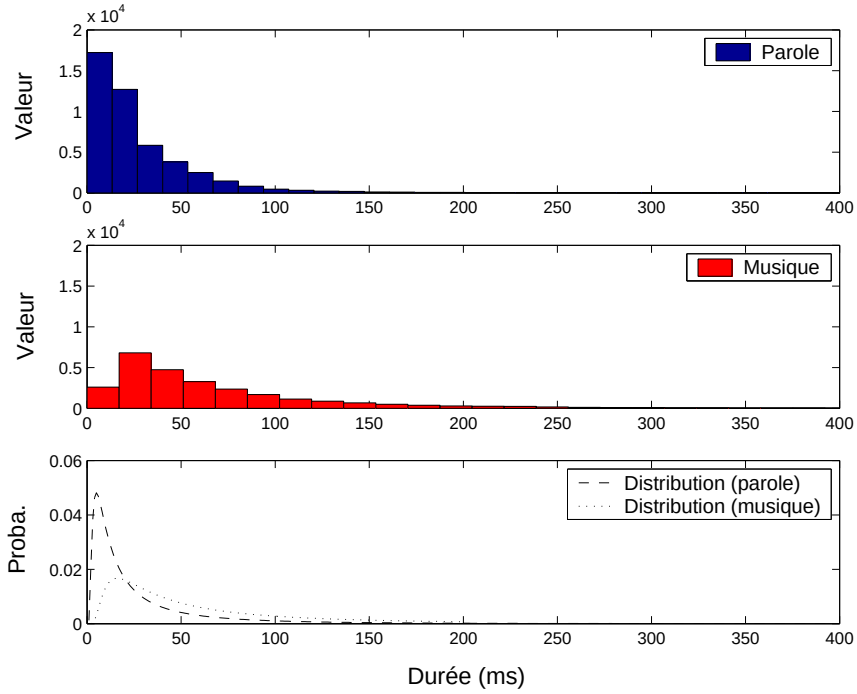

Figure 12. *Répartition des durées des segments pour la parole et la musique ainsi que les lois de Wald correspondantes*

Les paramètres λ et μ du modèle gaussien inverse ont été estimés :

$Parole$	$Musique$
$\lambda = 15,2753$	$\lambda = 50,6069$
$\mu = 30,1865$	$\mu = 74,9350$

Dans le cadre de la stratégie bayésienne, la décision est prise par maximum de vraisemblance en utilisant les deux lois ainsi estimées.

4. Expériences et évaluation

4.1. *Corpus*

Les expériences sont effectuées sur un corpus totalement différent de celui utilisé pour étudier la distribution des valeurs des paramètres et déterminer les seuils, afin d'évaluer la robustesse de nos paramètres et des seuils.

Le corpus expérimental correspond à une base de données qui a été réalisée à partir d'enregistrements de RFI[1] (Radio France Internationale) très compressés et le taux d'échantillonnage est de 16 kHz. 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 multilocuteur et multilingue.

Cette base a aussi permis de faire l'apprentissage des GMM du système de référence (cf. 4.4). Nous avons utilisé environ 8 heures de ce corpus pour faire l'apprentissage et 1 heure 30 minutes pour l'ensemble des tests (sur chacun des paramètres, sur notre système global et sur le système de référence).

4.2. *Etiquetage manuel*

Pour comparer et évaluer les performances de notre système, nous avons étiqueté manuellement une partie du corpus. Cet étiquetage a été effectué sur les deux composantes : parole et musique. Il y a donc deux étiquetages : parole/non-parole et musique/non-musique qui correspondent aux deux sous-systèmes.

Ces étiquetages manuels servent à étiqueter ensuite chacune des fenêtres d'une seconde sur lesquelles seront prises les décisions. Ce deuxième étiquetage est basé sur la classe majoritairement représentée.

La figure 13 présente un exemple d'alignement à la seconde (cf. (b)) obtenue pour une classification manuelle parole/non-parole (p/-) (cf. (a)). Les deux étiquetages à la seconde (parole et musique) ainsi obtenus serviront à la comparaison et l'évaluation des performances de chacun de nos paramètres et de notre système global de fusion.

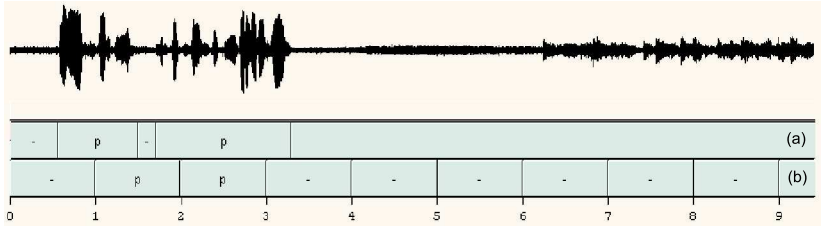

Figure 13. *Exemple d'étiquetage manuel parole/non-parole (p/-) en (a) sur un extrait contenant de la parole, du bruit puis de la musique. La deuxième ligne (b) représente le découpage correspondant (en segments d'une seconde)*

1. Dans le cadre du projet RAIVES (projet CNRS).

4.3. *Evaluation*

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/non-parole (tableau 1) : le taux est calculé par rapport à l'étiquetage manuel parole/non-parole ramené à la seconde (décrit dans le paragraphe précédent).

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/non-musique).

Les résultats fournis par les différents paramètres ont été ensuite fusionnés (cf. paragraphe 2.1).

– Le premier sous-système, consacré à la détection de parole, résulte de la fusion de la modulation de l'énergie à 4 Hz et de la modulation de l'entropie. Par maximisation des scores de chacun, cette méthode permet d'augmenter le taux d'identification correcte de 3 points pour atteindre 90,5 %.

Paramètres	Performances (Taux d'identification correcte)
Modulation de l'énergie à 4 Hz	87,3 %
Modulation de l'entropie	87,5 %
Fusion (détection de parole)	**90,5 %**

Tableau 1. *Classifi cation parole/non-parole*

– Le second sous-système, consacré à la détection de musique, est issu de la fusion des deux paramètres de segmentation (le nombre de segments par seconde et la durée moyenne de ces segments par seconde). Avec la même méthode de fusion que précédemment (maximisation des scores), le taux d'identification correcte atteint ici 89 % (gain de 2,5 points).

Paramètres	Performances (Taux d'identification correcte)
Nombre de segments	86,4 %
Durée des segments	78,1 %
Fusion (détection de musique)	**89 %**

Tableau 2. *Classifi cation musique/non-musique*

Les résultats des deux sous-systèmes sont ensuite alignés et fusionnés (figure 14).

L'étiquetage résultant possède quatre symboles :

– "P" : correspondant à de la parole et de la non-musique,
– "M" : pour de la musique et de la non-parole,
– "PM" : pour de la parole et de la musique,
– "-" : pour tout le reste (bruit et silence).

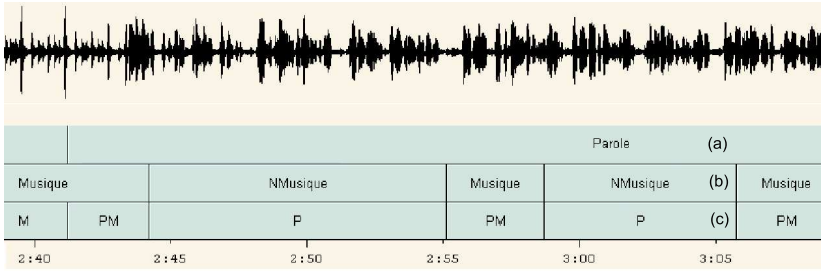

Figure 14. *Exemple de résultats de notre système global pour une classification parole/non-parole (a) et musique/non-musique (b). La fusion est représentée sur la ligne (c) avec "P" pour parole, "M" pour musique et "PM" pour parole et musique*

4.4. *Le système de référence*

4.4.1. *Description du système de référence*

Nous appellerons système de référence le système qui utilise la modélisation différenciée [PIN 02]. Comme notre système global, il se décompose en deux sous-systèmes. Le premier (parole/non-parole) 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/non-musique) 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, non-parole, musique et non-musique). Ces classes ont été apprises en utilisant deux algorithmes successifs. Le premier, un algorithme de quantification vectorielle (VQ) basé sur l'algorithme de Lloyd [RIS 82], permet d'initialiser les modèles. Le second est un algorithme permettant de réestimer les modèles (EM) [DEM 77]. Le nombre de lois gaussiennes utilisées ici a été fixé à 128.

Ce système, dans sa configuration d'origine, est appris sur un corpus composé d'un épisode d'une série télévisée et d'un championnat de patinage artistique. Ses

performances sur ce corpus sont de 93 % d'identification correcte pour la parole/non-parole et 91 % pour la musique/non-musique pour une prise de décision toutes les secondes.

4.4.2. *Comparaison*

Notre système global et le système de référence sont tous deux divisés en deux sous-systèmes parole/non-parole et musique/non-musique. Nous pouvons comparer les sous-systèmes un à un. Nous avons utilisé le système de référence de deux manières : sans nouvel apprentissage et avec un nouvel apprentissage sur le corpus RFI.

– Pour la détection de parole (tableau 3), le résultat du système global (90,5 %) est équivalent à celui obtenu par le système de référence (90,9 %) avec un apprentissage (ou une réestimation) sur le corpus RFI. Cela permet de montrer la pertinence du choix de nos paramètres. Dans les même conditions, c'est-à-dire avec un apprentissage sur des données différentes (système de référence avec sa configuration d'origine), notre système est plus performant que le système de référence (86,1 %). La modélisation statistique de nos paramètres est validée.

Détection de parole	Performances (Taux d'identification correcte)
Système global (fusion)	**90,5 %**
Système de référence	86,1 %
Système de référence avec apprentissage sur RFI	90,9 %

Tableau 3. *Comparaison de notre système global de fusion avec un système de référence pour la détection de parole*

– Pour la détection de musique (tableau 4), le résultat du système global (89 %) est supérieur au système de référence qu'il soit réestimé (87 %) ou non (79,7 %) sur le corpus RFI. Cela nous confirme dans le choix de nos paramètres et dans la fusion utilisée par notre système global.

Détection de musique	Performances (Taux d'identification correcte)
Système global (fusion)	**89 %**
Système de référence	79,7 %
Système de référence avec apprentissage sur RFI	87 %

Tableau 4. *Comparaison de notre système global de fusion avec un système de référence pour la détection de musique*

5. Conclusion

Nous avons présenté dans cet article quatre paramètres basés sur différentes propriétés du signal. Chaque paramètre est pertinent car il permet de faire une discrimination parole/non-parole ou musique/non-musique correcte. En considérant chaque paramètre individuellement, le taux de classification correcte varie d'environ 78 % pour la durée des segments à plus de 87 % pour la modulation de l'entropie.

L'algorithme de classification hiérarchique issu de la fusion entre ces paramètres permet d'améliorer les résultats et d'obtenir environ 90 % de reconnaissance correcte pour chacune des classifications (parole/non-parole et musique/non-musique).

Les expériences décrites dans cet article donnent des résultats meilleurs que ceux obtenus avec l'approche employant des GMM. L'avantage principal de cette nouvelle méthode est qu'elle est utilisable sur tout nouveau corpus sans requérir de nouvel étiquetage : il n'y a plus de phase d'apprentissage et/ou d'adaptation des modèles alors que c'est le cas avec les modèles de mélanges de lois gaussiennes (GMM).

Le corpus de test (radiophonique) est assez difficile car très varié et très compressé : la parole est présente sous diverses conditions, de pure à très bruitée. La qualité des résultats permet de montrer non seulement la robustesse de nos paramètres, mais aussi l'intérêt de notre approche quant à son utilisation sur n'importe quel type de document sonore.

6. Bibliographie

[AND 88] ANDRÉ-OBRECHT R., « A New Statistical Approach for Automatic Speech Segmentation », *IEEE Transactions on Audio, Speech, and Signal Processing*, vol. 36, n° 1, janvier 1988, IEEE.

[AND 93] ANDRÉ-OBRECHT R., « Segmentation et parole ? », Master's thesis, IRISA, 1993.

[AND 97] ANDRÉ-OBRECHT R., JACOB B., « Direct Identification vs. Correlated Models to Process Acoustic and Articulatory Informations in Automatic Speech Recognition », *International Conference on Audio, Speech and Signal Processing*, Munich, 1997, IEEE, p. 989-992.

[CAL 89] CALLIOPE, *La parole et son traitement automatique*, Masson, Paris, 1989.

[CAM 98] CAMPIONE E., VÉRONIS J., « A Multilingual prosodic database », *International Conference on Spoken Language Processing*, Sydney, décembre 1998, IEEE, p. 3163-3166.

[CAR 99] CAREY M. J., PARRIS E. J., LLOYD-THOMAS H., « A comparison of features for speech, music discrimination », *International Conference on Audio, Speech and Signal Processing*, Phoenix, mars 1999, IEEE, p. 149-152.

[CAR 00] CARRÉ M., PIERRICK P., « Indexation audio : un état de l'art », *Annales des télécommunications*, vol. 55, n° 9-10, 2000, p. 507-525, Editions Hermès.

[DEM 77] DEMPSTER A. P., LAIRD N. M., RUBIN D. B., « Maximum likelihood from incomplete data via the EM algorithm », *Journal of the Royal Statistical Society*, vol. 39 (Series B), 1977, p. 1-38.

[FOO 00] FOOTE J., « Automatic audio segmentation using a measure of audio novelty », *IEEE International Conference on Multimedia and Expo*, vol. I, New York, 2000, IEEE, p. 452-455.

[FRA 01] FRANZ M., SCOTT MCCARLEY J., WARD T., ZHU W., « Topics styles in IR and TDT : Effect on System Behavior », *EUROSPEECH'2001*, Scandinavia, septembre 2001, ISCA, p. 287-290.

[GAU 99] GAUVAIN J. L., LAMEL L., ADDA G., « Systèmes de processus légers : concepts et exemples », *CBMI'99*, Toulouse, octobre 1999, GDR-PRC ISIS, p. 67-73.

[HOU 85] HOUTGAST T., STEENEKEN J. M., « 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, n° 3, 1985, p. 1069-1077.

[JOH 70] JOHNSON N. L., KOTZ S., *Continuous Univariate Distributions*, Willey interscience publication, New York, 1970.

[KIM 96] KIMBER D., NILCOX L., « Acoustic segmentation for audio browsers », *Proceedings of Interface Conference*, Sydney, Australia, juillet 1996.

[LI 00] LI D., SETHI I., DIMITROVA N., MCGEE T., « Classification of general audio data for content-based retrieval », *Pattern Recognition Letters*, , 2000.

[MOD 89] MODDEMEIJER R., « On Estimation of Entropy and Mutual Information of Continuous Distributions », *Signal Processing*, vol. 16, n° 3, 1989, p. 233–246.

[PIN 02] PINQUIER J., SÉNAC C., ANDRÉ-OBRECHT R., « Indexation de la bande sonore : recherche des composantes Parole et Musique », *RFIA'2002*, Angers, janvier 2002, AFRIF-AFRIA, p. 163-170.

[RIS 82] RISSANEN J., « An universal prior for integers and estimation by minimum description length », *The Annals of Statistics*, vol. 11, 1982, p. 416-431.

[ROS 99] ROSSIGNOL S., RODET X., SOUMAGNE J., COLLETTE J. L., DEPALLE P., « Automatic characterization of musical signals : feature extraction and temporal segmentation », *Journal of New Music Research*, vol. 28, n° 4, décembre 1999, p. 281-295.

[SAU 96] SAUNDERS J., « Real-time discrimination of broadcast Speech/Music », *International Conference on Audio, Speesh and Signal Processing*, Atlanta, mai 1996, IEEE, p. 993-996.

[SCH 97] SCHEIRER E., SLANEY M., « Construction and Evaluation of a Robust Multifeature Speech/Music Discriminator », *International Conference on Audio, Speech and Signal Processing*, Munich, avril 1997, IEEE, p. 1331-1334.

[SUA 94] SUAUDEAU N., « Un modèle probabiliste pour intégrer la dimension temporelle dans un système de reconnaisance automatique de parole », PhD thesis, IRISA, 1994.

[WOL 99] WOLD E., BLUM T., KEISLAR D., WHEATER J., « Classification, search and retrieval of audio », *CRC Handbook of multimedia computing*, CRC Press LLC, 1999.

[ZHA 98] ZHANG T., KUO C., J. C., « Hierarchical System for Content-Based Audio Classification and Retrieval », *Conference on Multimedia storage and Archiving Systems III*, SPIE Vol. 3527, novembre 1998, p. 398-409.

Article reçu le 4 juin 2002
Version révisée le 17 mars 2003
Rédacteur responsable : Isabelle Bloch

Julien Pinquier *a obtenu un diplôme de DEA en Informatique de l'Image et du Langage en 2001 à l'Université Paul Sabatier de Toulouse. Il prépare actuellement une thèse au sein de l'équipe SAMOVA (Structuration Analyse et MOdélisation de la Vidéo et de l'Audio) du laboratoire IRIT (Institut de Recherche en Informatique de Toulouse). Ses travaux portent sur l'indexation sonore, notamment la classification parole/musique/bruit et la recherche de mots-clés ou de sons-clés (jingles).*

Jean-Luc Rouas *a obtenu un diplôme de DEA en Signal, Image et Acoustique en 2001 à l'Université Paul Sabatier de Toulouse. Il prépare actuellement une thèse à l'IRIT au sein de l'équipe SAMOVA. Ses travaux portent sur l'identification automatique des langues et leur caractérisation par la prosodie.*

Régine André-Obrecht *est ancienne élève de l'Ecole Normale Supérieure de Fontenay-aux-Roses (promotion 77). Professeur de l'Université Paul Sabatier, elle dirige l'équipe SAMOVA de l'IRIT. Ses intérêts scientifiques concernent le traitement du signal de parole et la reconnaissance automatique de la parole et de la langue. Les applications visées ont pour cadre l'indexation transmédia des documents audiovisuels.*