Caractérisation de la voix chantée dans un contexte d'indexation audio

Hélène Lachambre, Régine André-Obrecht, Julien Pinquier

Institut de Recherche en Informatique de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 9, France
lachambre@irit.fr, obrecht@irit.fr, pinquier@irit.fr

Résumé – Dans un processus de classification ou d'indexation de documents audio, la première étape est souvent la segmentation du signal en composantes primaires : la plupart du temps musique et parole. Très peu de travaux ont cependant été jusqu'alors consacrés à la détection du chant, qu'il soit accompagné ou non.

Nous proposons ici d'utiliser des paramètres simples (vibrato et coefficient harmonique), ainsi qu'une nouvelle segmentation du signal pour caractériser le chant. Nous fusionnons ensuite les résultats avec ceux d'une segmentation classique parole/musique.

Les tests sont réalisés sur un corpus que nous avons composé nous même, de façon à avoir la plus grande diversité possible. Nous testons d'abord notre système sur une tâche d'identification, puis sur une tâche de détection. Dans les deux cas, les résultats sont satisfaisants. Notre système de classification est presque parfait, les seules erreurs sont dues à des styles musicaux anecdotiques. Pour la tâche de détection, nous avons des non-détections, mais très peu de fausses-détections.

Abstract – To extract the content of audio documents, the first step in many methods is to segment the signal in primary components: music and speech. But very few work has been done to detect the singing voice, coupled or not with music.

In this paper, we propose simple parameters (vibrato and harmonic coefficient) and an original segmentation based on a sinusoidal segmentation to characterize the singing voice. The results are then mixed with those from a speech/music decomposition.

We test this classification system on a database composed of various types of sound. We test its performances in classification and in detection. In both cases, the results are good. In our classification system, the only misclassification are due to very rare musical styles. In the detection task, our system misses some of the singing voice segments, but has very few false-alarm.

1 Introduction

Pour indexer un document audiovisuel, la première étape est de déterminer le type d'information présent. Si dans le cas de la bande sonore, de nombreux travaux ont été réalisés pour détecter la musique, la parole, ou encore des sons caractéristiques [1, 2], très peu ont été menés sur le chant [3]. Les caractéristiques du chant se trouvent entre celles de la parole et de la musique ; dans un système parole/musique, le chant est souvent reconnu comme de la musique, mais il est parfois pris pour de la parole !

Notre étude se base sur une segmentation de type parole/musique [4] développée au sein de notre équipe. Celle-ci utilise des paramètres robustes et simples et ne nécessite pas d'apprentissage. Plus précisément, il nécessite le réglage de quelques (4) seuils indépendants du corpus. Nous avons suivi la même idée pour le chant : développer un système sans apprentissage basé sur quelques paramètres.

Nous introduisons une segmentation originale, basée sur une « segmentation sinusoïdale » [5], et y associons deux paramètres simples mais discriminants : le vibrato et le coefficient harmonique [6]. L'analyse de ces deux paramètres, et la fusion avec les informations issues de la segmentation parole/musique, nous permettent de savoir, à chaque instant, lesquelles des trois composantes sont présentes.

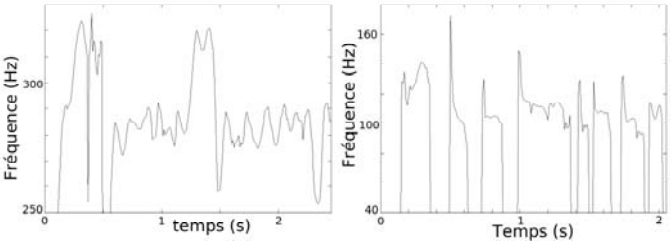
Dans la partie 2, nous décrivons l'état de l'art. Dans la partie 3, nous présentons notre nouvelle segmentation, les paramètres et le processus de décision. Les tests sont résumés dans la partie 4.

2 État de l'art

Nous avons utilisé deux paramètres (le vibrato et le coefficient harmonique) et la « segmentation sinusoïdale ». Nous décrivons ci-dessous ces trois outils.

2.1 Le vibrato

Le vibrato est une oscillation de la fréquence. Le propre du vibrato de la voix est qu'il est toujours présent lorsque nous chantons, mais pas quand nous parlons [7, 8] (voir figure 1). Il est également possible de faire du vibrato avec des instruments (vents et cordes), mais il sera la plupart du temps à une autre fréquence.



(a) Chant (vibrato) (b) Parole (pas de vibrato)

Fig. 1 – Fréquence fondamentale pour des extraits de 2 secondes de chant (a) et de parole (b).

La présence de vibrato est caractérisée par un maximum

entre 4 et 8 Hz dans la DFT de la fréquence fondamentale.

L'inconvénient de ce paramètre est qu'il faudrait, pour de la musique polyphonique, extraire la fréquence fondamentale de chaque instrument (voir section 3.2.2).

2.2 Le coefficient harmonique

Le coefficient harmonique [9] mesure le poids de la plus importante série dans une décomposition en séries harmoniques. Élevé en présence de chant [6], il est calculé ainsi :

– auto-corrélation temporelle R^T :

$$R^T(\tau) = \frac{\sum_{n=0}^{N-\tau-1} [\tilde{s}(n) \cdot \tilde{s}(n+\tau)]}{\sqrt{\sum_{n=0}^{N-\tau-1} \tilde{s}^2(n) \cdot \sum_{n=0}^{N-\tau-1} \tilde{s}^2(n+\tau)}} \quad (1)$$

avec s le signal à analyser, $\tilde{s}$ sa version centrée en zéro et N la taille de la fenêtre d'analyse.

– auto-corrélation fréquentielle R^F :

$$R^F(\omega_\tau) = \frac{\sum_{\omega=0}^{N-\omega_\tau-1} \left[\tilde{S}(\omega) \cdot \tilde{S}(\omega+\omega_\tau)\right]}{\sqrt{\sum_{\omega=0}^{N-\omega_\tau-1} \tilde{S}^2(\omega) \cdot \sum_{\omega=0}^{N-\omega_\tau-1} \tilde{S}^2(\omega+\omega_\tau)}} \quad (2)$$

avec S le module de la transformée de Fourier de s, $\tilde{S}$ sa version centrée en zéro et $\omega_\tau = N/\tau$.

– combinaison des deux auto-corrélations :

$$R(\tau) = \beta \cdot R^T(\tau) + (1-\beta) R^F(\tau) \quad (3)$$

– le coefficient harmonique H_a est alors défini ainsi :

$$H_a = \max_\tau R(\tau) \quad (4)$$

Expérimentalement, [9] trouve $\beta = 0,5$ comme valeur optimale, que nous utiliserons également dans cette étude.

2.3 La segmentation sinusoïdale

Cette segmentation, développée par [5] réalise un suivi automatique des fréquences (voir figure 2). Un segment sinusoïdal est défini par 4 paramètres : les indices de début et fin, les vecteurs des fréquences et de leurs amplitudes.

Le calcul des segments sinusoïdaux se fait ainsi [5] :

– calculer le spectrogramme toutes les 10 ms, avec une fenêtre de Hamming de 20 ms,

– convertir les fréquences en *cent* ($100cent = 1/2ton$) :

$$f_{cent} = 1200 \cdot \log_2 \left(\frac{f_{Hz}}{440 \cdot 2^{\frac{3}{11}-5}} \right) \quad (5)$$

– détecter les maxima du spectrogramme : leurs fréquences f_t^i et leurs amplitudes p_t^i,

– calculer les distances entre les points du spectrogramme :

$$d_{i_1,i_2}(t) = \sqrt{\left(\frac{f_t^{i_1} - f_{t-1}^{i_2}}{C_f} \right)^2 + \left(\frac{p_t^{i_1} - p_{t-1}^{i_2}}{C_p} \right)^2} \quad (6)$$

Deux points $(t, f_t^{i_1})$ et $(t+1, f_{t+1}^{i_2})$ appartiennent au même segment sinusoïdal si $d_{i_1,i_2}(t) < d_{th}$. C_f, C_p et d_{th} sont déterminés expérimentalement : $C_f = 100$ (1 demi-ton), $C_p = 3$ (puissance divisée par 2) et $d_{th} = 5$ (voir [5]).

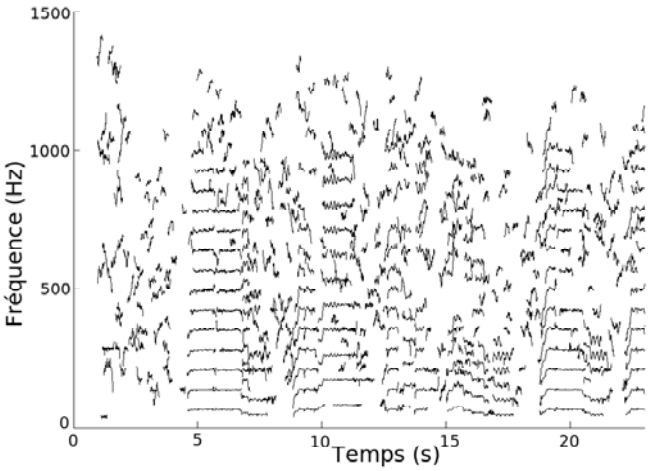


FIG. 2 – Segmentation sinusoïdale d'un extrait de 23 secondes de chant a capella : chaque ligne est un segment.

3 Système Parole/Musique/Chant

Notre système se base sur trois décisions : présence ou non de parole, de musique et de chant. Le travail sur la parole et la musique a été réalisé antérieurement. Le problème du chant nous a amenés à introduire une nouvelle segmentation et une extension de la notion de vibrato.

Dans cette partie, nous décrivons la segmentation et les paramètres. Nous concluons par les règles de décision.

3.1 La segmentation temporelle

L'étude de spectrogrammes nous a amenés à proposer cette nouvelle segmentation, basée sur la segmentation sinusoïdale (voir figure 3) : pendant un son harmonique stable (typiquement une note), la fréquence fondamentale et ses harmoniques commencent et finissent en même temps. Nous analysons donc les corrélations temporelles entre les débuts et les fins des segments sinusoïdaux :

– calculer les segments sinusoïdaux (voir 2.3),

– trouver toutes les extrémités temporelles des segments, en distinguant les débuts des fins,

– placer une limite à l'instant $\boldsymbol{t}$ s'il y a au moins 2 extrémités à $\boldsymbol{t}$ ET 3 débuts ou 3 fins entre $\boldsymbol{t}$ et $\boldsymbol{t+1}$.

Un segment temporel est alors défini par deux limites successives. On en distingue immédiatement deux types :

– les segments longs et stables (durée supérieure à 100 ms),

– les segments courts.

À l'issue de cette segmentation, chaque segment long correspond à une note. Nous analysons ceux-ci, car ils sont discriminants dans l'étude du chant.

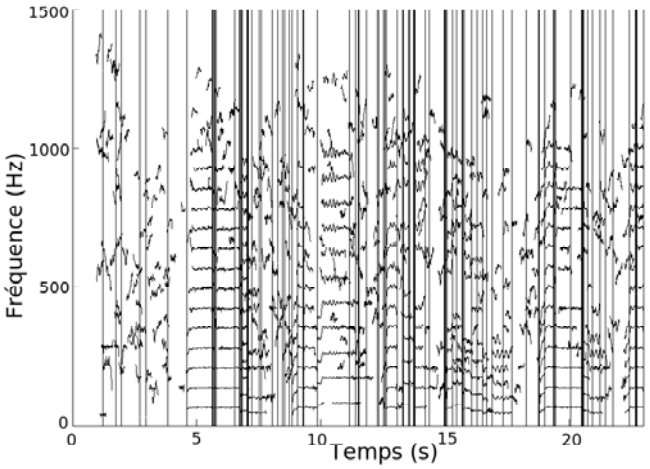


FIG. 3 – Segmentation temporelle du même extrait que la figure 2, les lignes verticales sont les limites des segments.

3.2 La paramétrisation

3.2.1 La musique et la parole

Les paramètres sont ceux de [4] : la modulation de l'énergie à 4 Hz mod_{4Hz}, la modulation de l'entropie mod_H, la durée l et le nombre n de segments stationnaires [10].

Les deux premiers paramètres sont utilisés pour détecter la parole. La modulation de l'énergie à 4 Hz caractérise le fait qu'on prononce en moyenne 4 syllabes par seconde. La modulation de l'entropie fait apparaître le fait que la parole est moins ordonnée que la musique.

Les deux autres paramètres l et n sont utilisés pour détecter la musique. Ils sont issus d'une segmentation du signal en segments stationnaires [10] et sont calculés chaque seconde. n est le nombre de segments par seconde. l est la durée moyenne des 7 plus longs segments de la seconde.

3.2.2 Le chant

Le coefficient harmonique H_a est calculé comme dans la partie 2.2.

Le vibrato est une forte caractéristique du chant. Nous avons étendu cette notion en notant qu'il affecte la fréquence fondamentale, mais aussi ses harmoniques. Ceci nous permet de traiter de la musique polyphonique.

Nous introduisons le paramètre *vibr*, qui mesure, dans un segment temporel, la proportion de segments sinusoïdaux qui ont du vibrato. Les segments sinusoïdaux provenant du chant auront du vibrato mais pas les autres ; ainsi la valeur de *vibr* sera plus élevée en présence de chant.

Comme précisé précédemment, seuls les segments temporels longs sont discriminants. Nous ne calculons *vibr* que pour eux, et attribuons la valeur 0 aux segments courts.

$$vibr = \frac{\sum_{s\in\Gamma} l(s)}{\sum_{s\in\Omega} l(s)} \tag{7}$$

avec :

$l(s)$ la longueur de s,

Γ les segments sinusoïdaux longs (>50 ms) avec du vibrato,

Ω l'ensemble des segments sinusoïdaux longs.

Finalement, H_a et *vibr* sont moyennés sur une seconde, pour être à la même échelle que les autres paramètres.

3.3 Le processus de décision

Nous avons trois décisions à prendre : présence ou absence de parole (P), de musique (M) et de chant (C). Nous complétons les règles données par [4] en y introduisant H_a et *vibr* et en ajoutant une catégorie : le chant.

$$\begin{aligned} P &= (mod_H \geq \lambda_1)\,\&\,(mod_{4Hz} \geq \lambda_2)\,\&\,(H_a \geq \lambda_5) \\ C &= (non(Parole))\,\&\,(vibr \geq \lambda_6) \\ M &= Chant \cup ((n \leq \lambda_3)\,\&\,(l \geq \lambda_4)) \end{aligned} \tag{8}$$

Les seuils λ_1, λ_2, λ_3 et λ_4 sont donnés dans [4] : $\lambda_1 = 0,5$, $\lambda_2 = 2,5$, $\lambda_3 = 17$, $\lambda_4 = 50ms$. λ_5 et λ_6 sont déterminés expérimentalement (voir 4.1).

Remarque : nous imposons $Chant \subset Musique$, et $Parole \cap Chant = \emptyset$.

4 Tests

4.1 Corpus

Pour tester notre algorithme, nous avons constitué un corpus le plus varié possible : de la parole, de la musique instrumentale, du chant a capella ou accompagné (voir tableau. 1) avec des styles, instruments et effectifs variés.

TAB. 1 – Répartition du corpus.

Type	Fichiers		Durée	
	Seuils	Tests	Seuils	Test
Chant a capella	2	9	8'	22'
Parole	3	12	25'	2h
Musique instru.	8	32	25'	2h
Chant accompagné	9	36	45'	3h
Total	**22**	**89**	**1h44'**	**7h22'**

Nous avons utilisé une partie du corpus (1/4 des fichiers, soit 1h30) pour fixer les seuils : $\lambda_5 = 0,7$, $\lambda_6 = 0,08$.

4.2 Identification

Nous testons d'abord la capacité d'identification du système : pour un extrait audio homogène donné, il doit déterminer si c'est de la parole ou de la musique et, dans ce cas, s'il y a du chant. La décision parole/musique est prise par vote majoritaire. Il n'y a jamais d'ambiguïté puisque pour la classification parole/musique, le taux d'erreur est inférieur à 10% (voir [4]). Si c'est un extrait musical, la présence de chant est caractérisée par le fait qu'on en détecte pendant au moins 1/4 du temps. Cela permet de prendre en compte les intermèdes instrumentaux parfois longs.

Tab. 2 – Classification des extraits.

	Chant	Parole	Musique	Total
Chant a capella	9	0	0	**9**
Parole	0	12	0	**12**
Musique instru.	3	0	29	**32**
Musique+Chant	27	0	9	**36**

Les résultats pour cette tâche (tableau 2) sont parfaits pour la parole et la musique. La détection de chant dans des extraits instrumentaux (3 sur 32) est due à des instruments rares (flûte de pan, accordéon, ...) qui ont le même vibrato que la voix. La non détection du chants est plus courante (9 sur 36). Dans ces cas, le chanteur chante peu, ou sa voix est parfois quasi-totalement masquée.

4.3 Détection

Nous avons ensuite testé les performances de détection : déterminer à chaque instant quelles sont les composantes présentes. Nous avons des informations toutes les secondes, mais cette échelle n'est pas adaptée au chant car elle ne permet pas de prendre en compte des interruptions courtes, par exemple les respirations. Nous lissons donc les résultats obtenus pour le chant après l'étape de décision : il y a du chant si on en détecte pendant au moins 2 secondes sur 3 consécutives.

Pour évaluer les résultats, nous comparons notre système à un système « classique » : extraction de 18 MFCC toutes les 10 ms, puis construction de 2 modèles GMM (32 gaussiennes) pour représenter les classes chant et non-chant. Pour les modèles GMM de musique/non musique et de parole/non parole, voir [4].

Afin de pouvoir comparer les deux systèmes, nous avons réalisé l'apprentissage des GMM avec la même partie du corpus qui nous avait permis de régler les seuils λ_5 et λ_6. Les résultats issus du système classique pour le chant sont également lissés sur 3 secondes.

Tab. 3 – Taux de bonne détection.

Type audio	Notre système	GMM
Parole	89,5%	94%
Musique	93%	91%
Chant	70%	70,3%

Les résultats (tableau 3) pour la détection de la parole et de la musique sont bons (89,5% et 93%) et sont compétitifs avec ceux obtenus avec des GMM.

Pour le chant, la détection est moins performante (70%), mais reste malgré tout comparable à un système classique. Les erreurs ont les mêmes causes que dans la tâche d'identification : les non détections sont dues soit au masquage du chant par les instruments, soit à la présence d'instruments (rares) tels une cornemuse. Le chant est alors classé comme de la musique (3/4) ou du silence (1/4). Le taux de confusion (instruments pris pour du chant) est faible : 8,5%, contre 19,5% pour les GMM. Ces erreurs sont dûes à des instruments qui ont le même vibrato que la voix.

5 Conclusion

Dans cet article, nous avons présenté une méthode pour la détection du chant, basée sur deux paramètres simples : le vibrato et le coefficient harmonique, ainsi que sur une segmentation originale du signal. En fusionnant les informations issues de ces paramètres avec celles de la segmentation Parole/Musique [4], nous savons quelles composantes sont présentes : parole, musique, chant. Les performances de notre système sont comparables à celles d'un système classique (GMM et MFCC), avec l'avantage qu'il ne nécessite aucun apprentissage.

Nous allons maintenant essayer d'améliorer nos performances en exploitant deux pistes : améliorer la segmentation temporelle, et combiner notre système et le classique.

Références

[1] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature speech/music discriminator," in *ICASSP*. IEEE, 1997, vol. 2, pp. 1331–1334.

[2] M. Karjalainen and T. Tolonen, "Multi-pitch and periodicity analysis model for sound separation and auditory scene analysis," in *ICASSP*. IEEE, 1999, vol. 2, pp. 929–932.

[3] I. Arroabarren, M. Zivanovic, X. Rodet, and A. Carlosena, "Instantaneous frequency and amplitude of vibrato in singing voice," in *ICASSP*. IEEE, 2003, vol. 5, pp. 537–540.

[4] J. Pinquier, J.L. Rouas, and R. Andre-Obrecht, "A fusion study in speech / music classification," in *ICASSP*. IEEE, 2003, vol. 2, pp. 17–20.

[5] Toru Taniguchi, Akishige Adachi, Shigeki Okawa, Masaaki Honda, and Katsuhiko Shirai, "Discrimination of Speech, Musical Instruments and Singing Voices Using the Temporal Patterns of Sinusoidal Segments in Audio Signals," in *Interspeech - European Conference on Speech Communication and Technology*. ISCA, 2005, pp. 589–592.

[6] Wu Chou and Liang Gu, "Robust Singing Detection in Speech/Music Discriminator Design," in *ICASSP*. IEEE, 2001, vol. 2, pp. 865–868.

[7] I. Arroabarren and A. Carlosena, "Voice production mechanisms of vocal vibrato in male singers," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 1, pp. 320–332, Jan 2007.

[8] R. Timmers and P. Desain, "Vibrato : questions and answers from musicians and science," in *Proc. Int. Conf. on Music Perception and Cognition*, 2000.

[9] Y.D. Cho, M.Y. Kim, and S.R. Kim, "A spectrally mixed excitation (SMX) vocoder with robust parameter determination," in *ICASSP*. IEEE, 1998, vol. 2, pp. 601–604.

[10] R. André-Obrecht, "A new statistical approach for the automatic segmentation of continuous speech," *IEEE Transaction on Acoustics, Speech, and Signal Processing*, vol. 36, pp. 29–40, 1988.