Caractérisation de la voix chantée en contexte monophonique et polyphonique

Hélène LACHAMBRE, Régine ANDRÉ-OBRECHT, Julien PINQUIER

IRIT - Université de Toulouse, 118 route de Narbonne, 31062 Toulouse Cedex 9, France
{lachambre, obrecht, pinquier}@irit.fr

Résumé – Nous présentons ici une amélioration de notre système de détection de la voix chantée. La détection se fait en deux étapes. Tout d'abord, nous séparons les sons monophoniques des sons polyphoniques. Cette distinction se fonde sur le fait que la valeur estimée de la fréquence fondamentale d'un son monophonique est plus fiable que celle d'un son polyphonique. Nous étudions la moyenne et la variance à court terme d'un indice de confiance, et modélisons leur répartition bivariée avec des lois de Weibull bivariées. Nous présentons une nouvelle façon d'estimer les paramètres de ces lois par la méthode des moments. Nous passons ensuite à la détection du chant proprement dite. Celle-ci se base sur la détection de vibrato, une oscillation de la fréquence fondamentale à un taux entre 4 et 8 Hz. Dans le cas d'un son monophonique, nous recherchons effectivement le vibrato sur la fréquence fondamentale. Dans le cas d'un son polyphonique, nous recherchons le vibrato sur les suivis des maxima du spectre. Les résultats sont encourageants, puisque nous passons d'un taux d'erreur de 29,7 % (ancienne méthode) à un taux d'erreur de 25 %, en prenant en compte le contexte (monophonique ou polyphonique).

Abstract – In this article, we present an improvement of our singing voice detector. The system is in two steps. First, we separate monophonic sounds from polyphonic sounds. This is based on the fact that the estimated pitch of a monophonic sound is more reliable than the one of a polyphonic sound. We study the short term mean and variance of a confidence indicator, and model their bivariate repartition with Weibull bivariate models. We present a new method for the estimation of the parameters of the law with the moment method. Then, the singing voice detection is based on the presence of vibrato, which is a oscillation of the pitch at a rate between 4 and 8 Hz. For a monophonic sound, vibrato is looked for on the pitch. For a polyponic sound, vibrato is looked for on the maxima of the spectrum. Results are promising: we have an error rate of 25 %, compared to 29.7 % without the knowledge on the context (monophonic or polyphonic).

1 Introduction

L'indexation de morceaux de musique est un sujet très étudié actuellement, sous divers points de vue. Des recherches sont notament menées sur la détermination de fréquences fondamentales multiples, du rythme, la reconnaissance d'instruments, du style, de l'ambiance,...

Ici, nous nous intéressons à la détection du chant. Ce sujet, relativement récent, a été étudié sous divers points de vue : quels sont les meilleurs paramètres [8], comment détecter le chant accompagné [9]...

Nous séparons tout d'abord les zones monophoniques de celles polyphoniques (partie 2). Puis le chant est caractérisé par la présence de vibrato. Nous présentons ce paramètre, ainsi que son adaptation pour le cadre polyphonique dans la partie 3. Dans la partie 4, nous présentons la fusion des deux étapes. Enfin, nous présentons les résultats dans la partie 5.

2 Séparation monophonie/Polyphonie

2.1 Paramètres

Dans leur article [1], de Cheveigné et Kahawara présentent un estimateur de fréquence fondamentale nommé YIN. L'estimation nécessite tout d'abord de calculer la « fonction de différence » $d_t(\tau)$, pour chaque trame t de signal :

$$d_t(\tau) = \sum_{k=1}^{N} (x_k - x_{k+\tau})^2 \quad (1)$$

avec x le signal, N la taille de la fenêtre d'analyse et τ le décalage temporel.

La période devrait alors être donnée par la position du premier minimum de cette fonction. Cependant, des bruits à haute fréquence, ainsi que des périodicités imparfaites créent des minima pour des indice de τ trop faible. Les auteurs proposent alors d'utiliser la « Cumulative Mean Normalised Difference » :

$$d'_t(\tau) = \begin{cases} 1 & si\ \tau = 0 \\ d_t(\tau) / \left[{}^1\!/_\tau . \sum_{k=1}^{\tau} d_t(k) \right] & sinon \end{cases} \quad (2)$$

La position T du premier minimum de $d'_t(\tau)$ donne maintenant la période. Ici, nous analysons la valeur de $d'_t(T)$, que nous appelons $cmnd(t)$. Cette valeur peut être interprétée comme un indice de confiance : plus $cmnd(t)$ est faible, plus la valeur de T est certaine.

Dans le cas d'un son monophonique, la fréquence fondamentale est facile à estimer, $cmnd(t)$ est faible et varie peu.

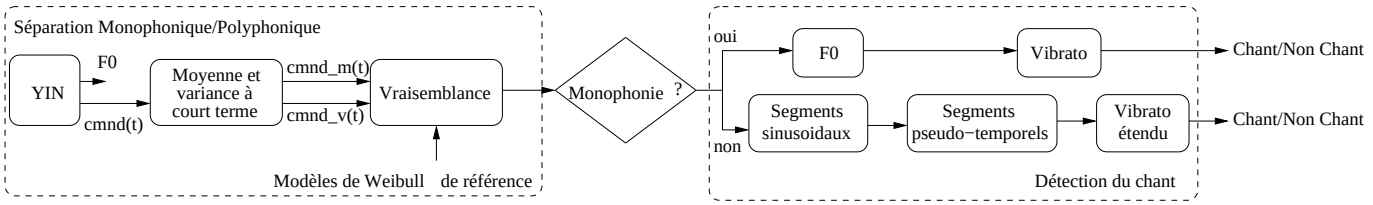

FIG. 1 – Schéma du système proposé.

À l'inverse, dans le cas d'un son polphonique, toutes les fréquences sont mélangées, la notion de « fréquence fondamentale » n'a plus de sens, $cmnd(t)$ est plus élevé, et varie plus.

Ainsi, les paramètres que nous utilisons pour la séparation monphonique / polyphonique sont la moyenne et la variance court terme de $cmnd(t)$, notées $cmnd_{moy}(t)$ et $cmnd_{var}(t)$.

2.2 Modélisation

Nous modélisons la répartition bivariée de ces deux paramètres $(cmnd_{moy}, cmnd_{var})$ en utilisant une extention de la loi de Weibull bivariée proposée par Hougaard [3], extension proposée par Lu et Bhattacharyya [6]. La fonction de répartition de cette distribution est donnée par la fonction suivante :

$$F(x,y) = 1 - \exp\left(-\left[\left(\frac{x}{\theta_1}\right)^{\frac{\beta_1}{\delta}} + \left(\frac{y}{\theta_2}\right)^{\frac{\beta_2}{\delta}}\right]^{\delta}\right) \quad (3)$$

pour $(x,y) \in \mathbb{R}^+ \times \mathbb{R}^+$, avec $(\theta_1, \theta_2) \in \mathbb{R}^+ \times \mathbb{R}^+$ les paramètres d'échelle, $(\beta_1, \beta_2) \in \mathbb{R}^+ \times \mathbb{R}^+$ les paramètres de forme et $\delta \in]0,1]$ le paramètre de corrélation.

Nous estimons les paramètres via la méthode des moments. L'estimation de θ_1, θ_2, β_1 et β_2 est un problème qui a déjà été longuement étudié, puisqu'il s'agit de l'estimation des paramètres des deux lois marginales. On pourra par exemple se reporter à l'article Morice de [7], qui présente entre autre l'estimation par la méthode des moments.

Pour l'estimation de δ, nous devons nous baser sur l'expression du moment croisé d'ordre 1, qui est donné dans [6] :

$$\begin{aligned} Cov(X,Y) = \theta_1\theta_2 . \\ [\Gamma\left({}^{\delta}/_{\beta_1}+1\right)\Gamma\left({}^{\delta}/_{\beta_2}+1\right)\Gamma\left({}^{1}/_{\beta_1}+{}^{1}/_{\beta_2}+1\right) \\ -\Gamma\left({}^{1}/_{\beta_1}+1\right)\Gamma\left({}^{1}/_{\beta_2}+1\right)\Gamma\left({}^{\delta}/_{\beta_1}+{}^{\delta}/_{\beta_2}+1\right)] \\ \div\Gamma\left({}^{\delta}/_{\beta_1}+{}^{\delta}/_{\beta_2}+1\right) \end{aligned} \quad (4)$$

avec $\Gamma(x)$ la fonction gamma.

L'estimation de δ semble très ardue, au vu de l'équation 4. Nous avons montré [4] que cette équation est équivalente à :

$$f(\delta) = \delta B\left({}^{\delta}/_{\beta_1}, {}^{\delta}/_{\beta_2}\right) = C \quad (5)$$

avec $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ la fonction Beta et C une constante dépendante de θ_1, θ_2, β_1, β_2 et $Cov(X,Y)$, qui sont connus.

Ainsi, trouver δ est équivalent à déterminer les zéros de la fonction $f(\delta) - C$. Nous avons également montré que $f(\delta)$ est strictement décroissante, donc il nous suffit de trouver l'unique zéro, ici avec une méthode par dichotomie.

Une loi de Weibull bivariée est apprise pour chaque classe, avec les données d'apprentissage présentées dans le tableau 1 (une seconde correspond à 100 vecteurs pour l'apprentissage).

2.3 Classification et résultats

$cmnd(t)$, $cmnd_{var}(t)$ et $cmnd_{moy}(t)$ sont calculés toutes les 10 ms,soit 100 couples $(cmnd_{var}(t), cmnd_{moy}(t))$ par seconde. La classification est faite chaque seconde, en calculant la vraisemblance des 100 couples $(cmnd_{var}(t), cmnd_{moy}(t))$ par rapport aux modèles de Weibull bivariés de référence estimés pour chaque classe.

Les résultats sont très encourangeants : le taux d'erreur est de **6,3 %**. Au vu des données d'apprentissage et de test (voir partie 5.1), il est pertinent de considérer cette méthode comme robuste et indépendante du corpus.

3 Détection du chant

3.1 Vibrato

Le vibrato est une oscillation périodique de la fréquence fondamentale. Pour le chant, il a la particularité d'être toujours présent [10, 12], à une fréquence comprise entre 4 et 8 Hz.

Pour détecter la présence de chant sur un vecteur de fréquence F, nous utilisons la caractérisation proposée par Gehrard [2] : il y a du vibrato si la Transformée de Fourier de F présente un maximum entre 4 et 8 Hz.

La notion de vibrato est intrinséquement liée à la notion de fréquence fondamentale et est donc définie pour des sons monphoniques. Nous proposons de l'étendre par le biais de deux segmentations que nous proposons ci-dessous.

3.2 Segmentations

3.2.1 Segmentation sinusoïdale

Cette segmentation, qui a pour but de réaliser un suivi de fréquences, a été proposée par Taniguchi *et al.* [11].

Un segment sinusoïdal est défini par quatre paramètres : l'indice de début, l'indice de fin, le vecteur des fréquences et le vecteur des amplitudes. La taille de ces deux vecteurs est déterminée par la durée du segment.

Les auteurs proposent l'algorithme suivant :

1. Calculer le spectre (ici toutes les 10 ms, avec une fenêtre de Hamming de 20 ms).

2. Lisser le spectre (ici un filtre moyenneur sur trois points) et convertir les fréquences en $cent$ ($100\,cent = 1/2\,ton$) :

$$f_{cent} = 1200 log_2(\frac{f_{Hz}}{440.2^{\frac{3}{11}-5}}). \quad (6)$$

3. Détecter les maxima du spectre. Pour la trame t, nous avons deux ensembles : $(f_t^i)_{i=1..N}$ et $(p_t^i)_{i=1..N}$, les fréquences et amplitudes, avec N le nombre de maxima.
4. Calculer les distances entre les différents maxima du spectre de deux instants successifs :

$$d_{i_1,i_2}(t) = \sqrt{\left(\frac{f_t^{i_1} - f_{t-1}^{i_2}}{C_f}\right)^2 + \left(\frac{p_t^{i_1} - p_{t-1}^{i_2}}{C_p}\right)^2} \quad (7)$$

5. Deux points $(t, f_t^{i_1})$ et $(t+1, f_{t+1}^{i_2})$ appartiennent au même segment 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 [11]).

3.2.2 Segmentation pseudo-temporelle

Cette segmentation « pseudo-temporelle », que nous avons proposée dans un article précédent [5], nous permet d'analyser les relations temporelles entre les débuts et fins de segments sinusoïdaux. Elle est réalisée à partir des segments sinusoïdaux :

1. Trouver toutes les extrémités temporelles des segments sinusoïdaux, en distinguant les débuts des fins.
2. Placer une limite à l'instant t s'il y a au moins 2 extrémités à t ET 3 débuts ou 3 fins entre t et $t+1$.

La figure 2 présente un exemple de segmentation sinusoïdale et pseudo-temporelle pour un extrait de 23 secondes de chant monophonique.

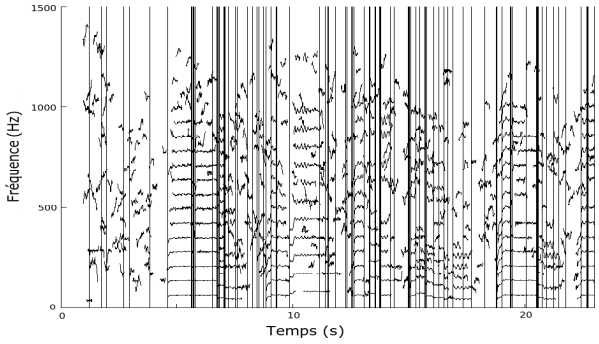

FIG. 2 – Segmentations sinusoïdale (lignes horizontales) et pseudo-temporelle (lignes verticales) d'un extrait de 23 secondes de chant monophonique.

3.3 Vibrato étendu

Avec la notion de segment sinusoïdal et de segment pseudo-temporel, nous sommes maintenant en mesure d'élargir la notion de vibrato. Dans [5], nous avons proposé la notion de vibrato étendu, $vibr$, défini comme suit :

$$vibr = \frac{\sum_{s\in\Gamma} l(s)}{\sum_{s\in\Omega} l(s)} \quad (8)$$

avec :

Ω l'ensemble des segments sinusoïdaux présents dans le segment temporel courant,

Γ l'ensemble des segments sinusoïdaux avec du vibrato,

$l(s)$ la durée du segment sinusoïdal s.

$vibr$ représente donc, pour un segment pseudo-temporel, la proportion de segments sinusoïdaux qui ont du vibrato.

4 Fusion

Une évolution importante de notre méthode est la phase de décision. D'une part, nous avons apporté un prétraitement en distinguant le contexte monophonique du contexte polyphonique, ce qui nous permet de différentier la classification.

D'autre part, nous tenons maintenant compte d'une caractéristique du chant : lorsqu'un chanteur chante, il fait des pauses très courtes (d'une durée inférieure à 0,5 seconde) pour respirer. De plus, s'il y a un accompagnement musical, il y a également des pauses moyennes (1-3 secondes) avec des transitions instrumentales, et/ou des pauses longues (jusqu'a 1 minute) avec des intermèdes instrumentaux. Ces considérations nous ont menés à modifier notre processus de décision : désormais, la présence de chant est avérée si, pendant une certaine durée (a préciser), nous avons détecté du vibrato (ou si le vibrato étendu est suffisamment élevé).

En contexte monophonique : La notion de fréquence fondamentale a un sens, la valeur donnée par l'estimateur YIN est fiable. La recherche de vibrato est donc faite sur la fréquence fondamentale.

En contexte polyphonique : La notion de fréquence fondamentale unique n'a plus de sens. Nous utilisons le vibrato étendu.

5 Expériences

5.1 Corpus

Nous avons réalisé les tests avec un corpus « fait maison », contenant des sons monophoniques et polyphoniques, avec et sans voix chantée. La répartition du corpus en terme de durée, pour l'apprentissage et le test est décrite dans le tableau 1.

Notons que des instruments, chanteurs, et styles du corpus de test n'existent pas dans le corpus d'apprentissage.

TAB. 1 – Répartition du corpus (apprentissage et test).

Classe	App. (durée)	Test (durée)	Test (Nb. de séq.)
Instrument solo	25 s	2 min 57 s	177
Chanteur solo	25 s	4 min 38 s	278
Monophonie	**50 s**	**7 min 35 s**	**455**
Chanteurs et instr.	25 s	3 min 10 s	190
Plusieurs instr.	25 s	3 min 23 s	203
Polyphonie	**50 s**	**6 min 33 s**	**393**
Total	**2 min 5 s**	**18 min 41 s**	**1121**

5.2 Résultats

En réalisant la classification chant / non chant sans l'étape préalable de séparation monophonique / polyphonique, le taux d'erreur était de **29,7 %** [5].

Classification monophonique / polyphonique manuelle
Nous menons deux séries d'expériences. Tout d'abord, nous étudions la pertinence d'utiliser la connaissance monophonie / polyphonie pour la détection du chant. La première étape, la séparation monophonie /polyphonie est faite manuellement, puis la détection du chant est réalisée comme présenté ci-dessus. Les résultats sont présentés dans les tableaux 2 et 3, le taux d'erreur est de **21,7 %**.

TAB. 2 – Matrice de confusion - Monophonies.

	Chant	Non chant
Chant solo	**0,83 %**	0,17 %
Instrument solo	0,20 %	**0,80 %**

TAB. 3 – Matrice de confusion - Polyphonies.

	Chant	Non chant
Chant et instruments	**0,66 %**	0,34 %
Instruments	0,16 %	**0,84 %**

Nous notons que l'apport de la classification monophonique / polyphonique est non négligeable : nous avons une amélioration de 8 %.

Le fait qu'il est plus difficile de détecter la présence de chant dans des extraits polyphoniques s'explique par le fait que par moment la voix est trop faible par rapport aux instruments de musique, qui la « masquent ».

Classification monophonique / polyphonique automatique
Nous testons le système dans son intégralité, en réalisant la segmentation monophonique / polyphonique automatique avec la méthode présentée dans la partie 2. Les résultats sont présentés dans le tableau 4, le taux d'erreur est de **25 %**.

Ainsi, même avec une classification monophonique / polyphonique automatique, les résultats sont améliorés de près de 5 %. Les erreurs sont dûes aux mêmes causes que dans l'expérience précédente, auxquelles il faut ajouter les imprécisions de la première classification.

TAB. 4 – Matrice de confusion - Système entier.

	Chant	Non chant
Chant solo	**0,79 %**	0,21 %
Instrument solo	0,26 %	**0,74 %**
Chant et instruments	**0,65 %**	0,35 %
Instruments	0,18 %	**0,82 %**

6 Conclusion et perspectives

Nous avons présenté un détecteur de chant, qui différentie les zones monophoniques des zones polyphoniques. Ce prétraitement apporte un gain de près de 5 %, et nous pouvons espérer un gain maximum de 8 %.

Nous allons maintenant nous atteler à l'amélioration de notre détecteur, en nous focalisant plus précisément sur les zones polyphoniques, qui sont encore les plus difficiles.

Références

[1] A. de CHEVEIGNÉ et H. KAWAHARA : YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, 111(4):1917–1930, 2002.

[2] D. B. GERHARD : Perceptual Features for a Fuzzy Speech-Song Classification. *In (ICASSP)*, vol. 4, p. 4160–4163. IEEE, 2002.

[3] P. HOUGAARD : A class of multivariate failure time distributions. *Biometrika*, 73(3):671–678, 1986.

[4] H. LACHAMBRE : Estimation of Weibull bivariate distribution parameters via the moment method. Rap. tech., IRIT - SAMoVA, Dec 2008.

[5] H. LACHAMBRE, R. ANDRÉ-OBRECHT et J. PINQUIER : Singing voice characterization for audio indexing. *In 15th European Signal Processing Conference (EUSIPCO)*, p. 1563–1540, 2007.

[6] J. LU et G. BHATTACHARYYA : Some new constructions of bivariate Weibull models. *Annals of Institute of Statistical Mathematics*, 42(3):543–559, 1990.

[7] E. MORICE : Quelques problèmes d'estimation relatifs à la loi de Weibull. *Revue de statistique appliquée*, 16(3):43–63, 1968.

[8] M. ROCAMORA et P. HERRERA : Comparing audio descriptors for singing voice detection in music audio files. *In Brazilian Symposium on Computer Music, 11th. San Pablo, Brazil*, 2007.

[9] S. SANTOSH, S. RAMAKRISHNAN, V. RAO et P. RAO : Improving singing voice detection in presence of pitched accompaniment. *In Proc. of the National Conference on Communications (NCC)*, 2009.

[10] C. E. SEASHORE : *Psychology of Music*. McGraw-Hill Book Company, inc., 1938.

[11] T. TANIGUCHI, A. ADACHI, S. OKAWA, M. HONDA et K. SHIRAI : Discrimination of Speech, Musical Istruments and Singing Voices Using the Temporal Patterns of Sinusoidal Segments in Audio Signals. *In Interspeech - European Conference on Speech Communication and Technology*. ISCA, sept. 2005.

[12] R. TIMMERS et P. DESAIN : Vibrato : questions and answers from musicians and science. *In Proc. Int. Conf. on Music Perception and Cognition*, 2000.