Application du critère BIC pour la segmentation en tours de chant

Marwa Thlithi, Thomas Pellegrini, Julien Pinquier, Régine André-Obrecht, Patrice Guyot

Application du critère BIC pour la segmentation en tours de chant

Marwa Thlithi, Thomas Pellegrini, Julien Pinquier, Régine André-Obrecht, Patrice Guyot
IRIT – Université de Toulouse – 118, route de Narbonne – 31062 Toulouse Cedex 9 – France
{thlithi, pellegrini, pinquier, obrecht, guyot}@irit.fr

RESUME

Dans le cadre d'un projet sur l'indexation de documents ethnomusicologiques sonores (ANR CONTINT DIADEMS), le repérage des chanteurs et des chœurs est apparu comme essentiel et nous a amené à s'interroger sur la notion de « tours de chant ». Dans cet article, nous présentons nos premiers pas dans le domaine en proposant une méthode de segmentation fondée sur le Critère d'Information Bayésien (BIC) qui vise à détecter des changements de chanteurs dans des enregistrements musicaux. Le corpus de cette étude est composé d'enregistrements musicaux fournis par des ethnomusicologues et il nous permet d'illustrer l'importance du coefficient de pénalité du critère BIC : sa valeur optimale varie en fonction du contenu des enregistrements. Pour s'affranchir de l'apprentissage d'une unique valeur de ce paramètre, nous proposons de recueillir plusieurs segmentations pour plusieurs valeurs du paramètre et de consolider la détection *a posteriori*. Un gain relatif en termes de F-mesure, de 15% (7% absolu) est obtenu entre cette décision *a posteriori* et une décision prise après apprentissage du coefficient de pénalité.

ABSTRACT

Segmentation in singer turns with the Bayesian Information Criterion

As part of a project on indexing ethno-musicological audio recordings (ANR CONTINT DIADEMS), determining singers and choirs automatically appeared to be essential and led us to reflect about the notion of "singer turns". In this article, we report our first experiments in this direction by exploring a method based on the Bayesian Information Criterion (BIC) to detect singer turns. The BIC penalty coefficient was shown to vary when determining its value to achieve the best performance for each recording. In order to avoid the decision about which single value is best for all the documents, we propose a combination of several segmentations obtained with different values of this parameter. This method uses majority voting. A gain of 15% relative (7% absolute) in terms of F-measure was obtained compared to a single coefficient determined on a development sub-corpus.

MOTS-CLES : segmentation audio, tours de chant, critère BIC.
KEYWORDS: audio segmentation, singer turns, BIC criterion.

1 Introduction

Un document sonore peut être structuré automatiquement de bien des manières en fonction de l'objectif final. S'il s'agit d'indexation d'un fond de documents par exemple, on se posera vraisemblablement les questions : est-on en présence de parole, de musique, à quels moments, qui parle, qui chante, etc. L'une des premières étapes d'inférence de cette structure est de découper

puis d'étiqueter des zones ou segments dits « homogènes acoustiquement ». En traitement automatique de la parole, il s'agira d'identifier les changements de locuteurs ou tours de parole pour savoir qui parle et à quels moments dans le document, pour ensuite faciliter éventuellement d'autres traitements qui pourraient suivre comme la transcription automatique de ce qui est dit (Anguera, 2012). Dans le contexte du projet ANR CONTINT DIADEMS [1] (Description, Indexation, Accès aux Documents EthnoMusicologiques et Sonores) sur l'indexation de documents ethnomusicologiques sonores, nous nous sommes posés la question de savoir si nous pouvions réaliser le même type de structuration en identifiant les changements de chanteurs (solistes et/ou chœurs) au sein des enregistrements musicaux qui sont l'objet d'étude du projet. Nous utiliserons l'expression « segmentation en tours de chant » par analogie à la segmentation en tours de parole. Par chant, nous désignons la voix chantée au sens large, accompagnée ou non par des instruments, en groupe ou en soliste.

La Figure **1** illustre notre problématique de segmentation en tours de chant. La vérité « terrain » consiste en une annotation manuelle des tours de chant et entrées/sorties d'instruments éventuelles. Dans cet article, nous présentons une méthode de segmentation en tours de chant qui précèderait une étape ultérieure de regroupement de segments figurant un même chanteur ou groupe de chanteurs.

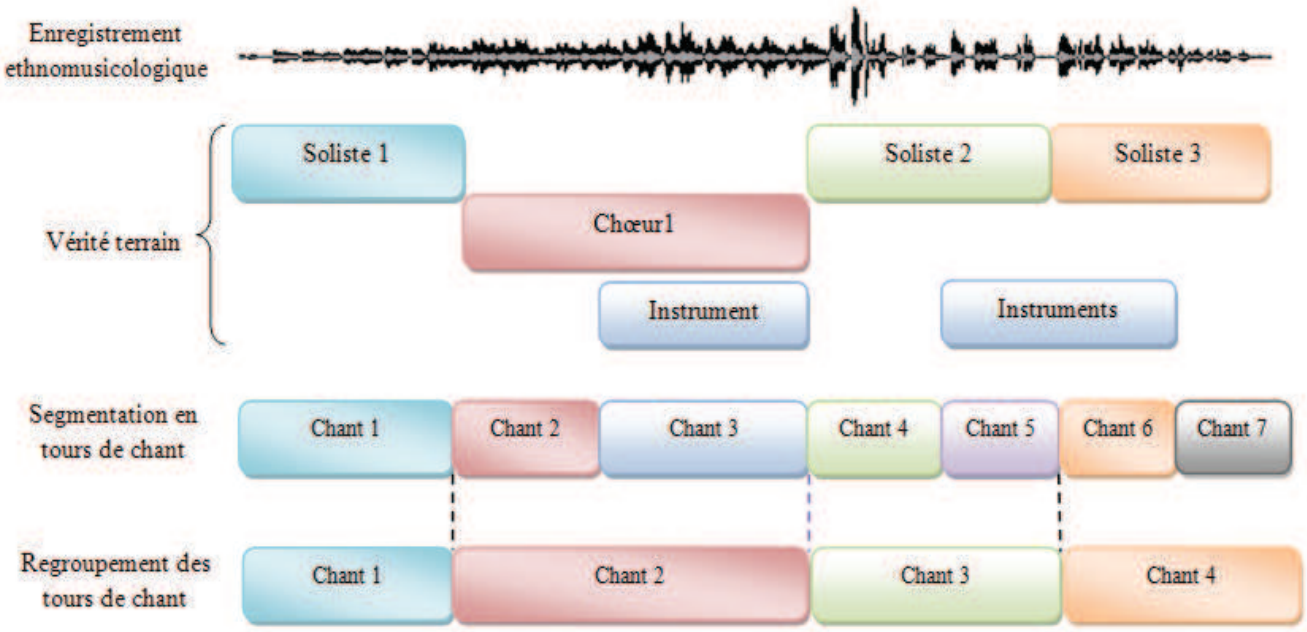

Figure 1 : Illustration de la segmentation en tours de chant et de l'étape de regroupement.

Depuis quelques années notre équipe travaille sur des questions liées à la voix chantée, en particulier sur la détection de voix chantée (Lachambre, 2009) et la segmentation chant solo/chœur (Le Coz, 2012) dans un contexte musical. Ce dernier traitement est réalisé à l'aide de critères similaires à la détection de parole superposée. Dans cette étude, nous continuons à utiliser une analogie avec la parole en testant la méthode de segmentation fondée sur le Critère d'Information Bayésien (BIC), très utilisée en segmentation en tours de parole (Zhu, 2005). Son application sur des enregistrements de chant nécessite une revisite de ce critère et une adaptation appropriée de ses paramètres. En particulier, nous avons observé qu'il était difficile de déterminer une valeur optimale du coefficient de pénalité du critère selon les enregistrements. Ce qui nous a amené à

[1] http://www.irit.fr/recherches/SAMOVA/DIADEMS/

proposer une méthode de type vote majoritaire impliquant plusieurs segmentations réalisées avec plusieurs valeurs de ce paramètre.

Dans la section 2, nous commençons par rappeler brièvement le critère BIC théorique et l'algorithme de référence que nous avons utilisé. Ensuite, le contexte applicatif est présenté en détail puis nous détaillons la mise en œuvre spécifique à notre tâche et les premiers résultats. Dans la section 4, le problème de la variabilité du facteur de pénalité est discuté et la méthode de combinaison est décrite. Enfin, nous comparons les performances globales obtenues.

2 Critère BIC pour la segmentation audio

2.1 Présentation générale du critère

Le Critère d'Information Bayésien, comme son nom l'indique, se place dans un contexte bayésien de sélection de modèles. Variante du critère d'Akaïke, il est utilisé dans de nombreux contextes applicatifs et ce depuis très longtemps (Akaïke, 1974), (Schwarz, 1978). Ces dernières années, il est au cœur de nombreux travaux de segmentation audio (Cettolo, 2005), (Siu, 1991), (Chen, 1998), (Delacourt, 2000) et bien des systèmes de segmentation et de regroupement en locuteurs, actuels et performants, implique le BIC.

La segmentation audio consiste à diviser le flux audio en des segments homogènes en effectuant un test d'hypothèses. Pour chaque point de changement potentiel, il y a deux hypothèses possibles : la première suppose que, de part et d'autre de ce point, le signal obéit au même modèle probabiliste, noté M_0 (H_0), la deuxième suppose qu'il y a un changement de modèle et que deux modèles différents M_1 et M_2 sont nécessaires (H_1). Dans la pratique, les modèles sont estimés sur trois fenêtres d'analyse et des critères tels que le critère BIC sont utilisés pour déterminer si le signal est « mieux » représenté par deux modèles distincts ou par un modèle unique ; un seuil est déterminé de manière empirique ou adapté dynamiquement.

Il s'en suit que si le signal analysé correspond à une séquence de N vecteurs d'observation (vecteurs acoustiques ou trames) de dimension d, noté $X_0(x_1, x_2, \dots, x_N)$; un point de changement potentiel positionné après la trame t induit deux sous suites consécutives $X_1(x_1, x_2, \dots, x_t)$ et $X_2(x_{t+1}, x_2, \dots, x_N)$. En supposant que X_0, X_1 et X_2 suivent des lois gaussiennes données respectivement par $M_0(\mu_0, \Sigma_0)$, $M_1(\mu_1, \Sigma_1)$ et $M_2(\mu_2, \Sigma_2)$, le critère ΔBIC à l'instant t est donné par :

$$\Delta BIC(t) = R(t) - \lambda P$$

$R(t)$ est le rapport de log vraisemblance entre les deux hypothèses $(LL(H_1)/LL(H_0)$ et s'écrit :

$$R(t) = \frac{1}{2}\left(N \log\left(\left|\widehat{\Sigma_0}\right|\right) - t \log\left(\left|\widehat{\Sigma_1}\right|\right) - (N-t)\log\left(\left|\widehat{\Sigma_2}\right|\right)\right.$$

avec $\left|\widehat{\Sigma_i}\right|$, déterminant de la matrice de covariance Σ_i, estimée sur la suite X_i.

P est proportionnel à la différence des nombres de paramètres estimés pour chaque hypothèse et vaut, dans le cas de matrices de covariance pleines :

$$P = \frac{1}{2}\left(d + \frac{1}{2}d(d+1)\right)\log N$$

Le facteur de pénalité λ est appris de telle sorte que le critère ΔBIC soit positif dès lors que l'hypothèse H_1 est vérifiée, indiquant l'existence de deux modèles différents. Sinon, l'hypothèse

H_0 est validée ainsi que l'existence d'un seul modèle pour la fenêtre X_0.

2.2 Algorithme de référence

La mise en œuvre du BIC implique de déterminer deux paramètres importants : la taille N de la fenêtre de signal dans laquelle une frontière de segment est recherchée, et le facteur de pénalité λ du critère. Dans un premier temps, nous avons cherché à fixer la valeur de ces deux paramètres sur un sous-ensemble de notre corpus décrit dans la prochaine section. Une première version de l'algorithme faisait suite aux travaux d'El-Khoury dans laquelle la taille de la fenêtre était constante (El-Khoury, 2009). Cependant, la recherche d'une taille optimale de cette fenêtre, indépendante de l'enregistrement considéré, s'est révélée difficile. Nous avons alors implémenté une autre version de l'algorithme dans laquelle la taille de fenêtre d'analyse augmente tant qu'aucune frontière potentielle n'est trouvée. Ce procédé s'inspire d'études en segmentation de la parole (Andre-Obrecht, 1988), et il est illustré dans la Figure **2**.

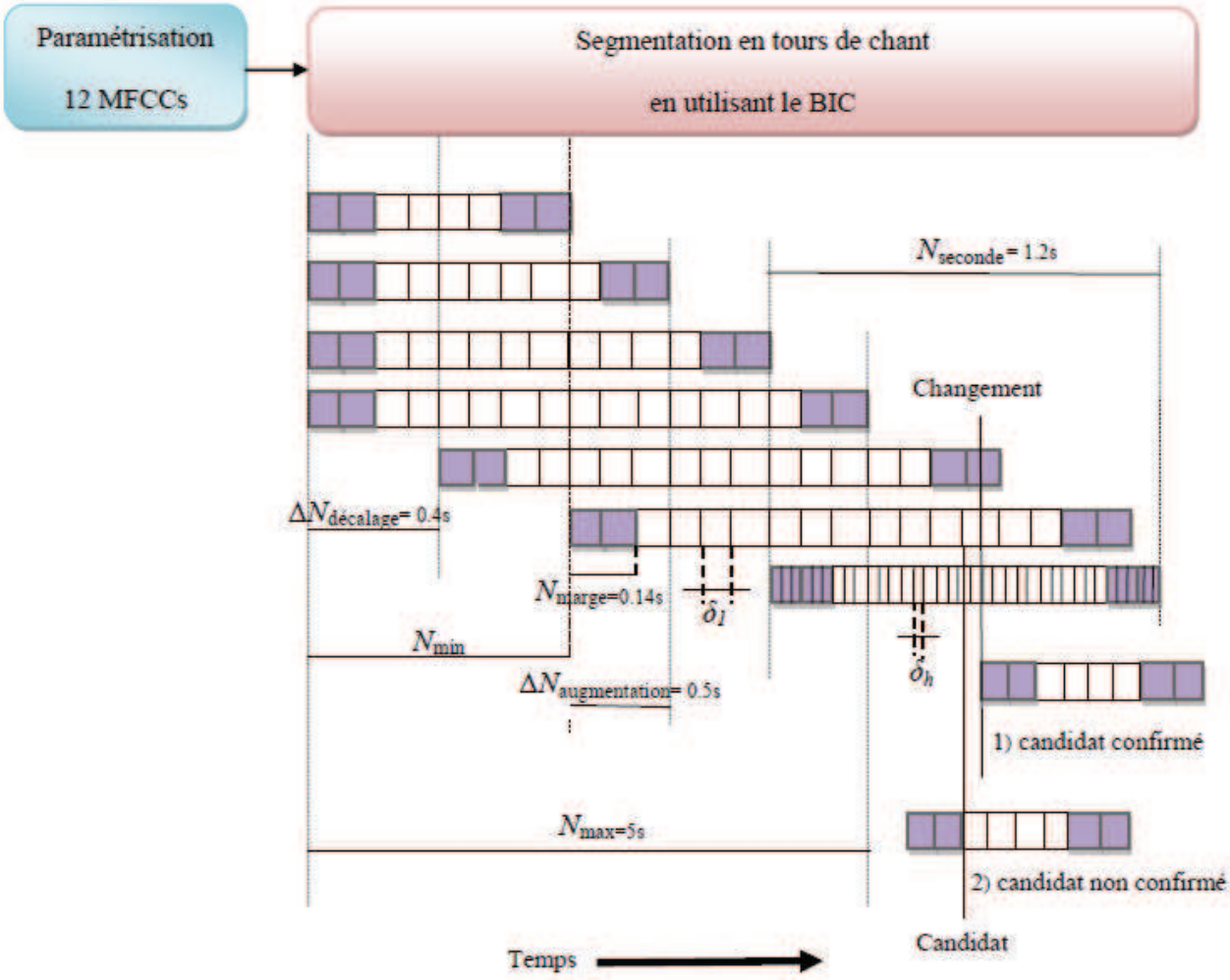

Figure 2 : Illustration de l'algorithme de notre système de segmentation en tours de chant. Figure adaptée de (Cettolo, 2005).

La recherche se déroule en deux temps, impliquant deux résolutions temporelles différentes :

1. La longueur initiale de la fenêtre d'analyse est fixée à N_{min} . Elle augmente de $\Delta N_{augmentation}$ tant que le test d'hypothèse ΔBIC ne valide aucune frontière interne, jusqu'à une valeur maximale N_{max} . Les valeurs de ΔBIC sont calculées à intervalles

réguliers pour des valeurs échantillonnées de t, à savoir une fois toutes les $\delta 1$ observations. Si aucune frontière n'est détectée lorsque N_{max} est atteint, la fenêtre d'analyse est décalée de $\Delta N_{décalage}$ et l'analyse est réinitialisée.

2. Si une frontière potentielle est détectée, une fenêtre de longueur $N_{seconde}$ est centrée sur cette frontière et les valeurs de ΔBIC sont calculées au sein de cette fenêtre à une résolution haute, toutes les δh observations, afin d'affiner la position de cette frontière.

Nous imposons que toute frontière ne produise aucun segment d'une durée inférieure à N_{marge}, ce qui implique qu'aucune frontière n'est recherchée entre les zones $[1, N_{marge}]$et $[N - N_{marge}, N]$.

3 Performances du système de référence

3.1 Contexte applicatif et corpus d'étude

Nous avons effectué nos expériences sur un corpus d'enregistrements sélectionnés spécifiquement pour la détection des tours de chant par les ethnomusicologues partenaires du projet DIADEMS. Des exemples sont accessibles en ligne[2]. Ces enregistrements ont été réalisés dans plusieurs pays sub-sahariens (Congo, Gabon, Cameroun), avec une qualité acoustique variable (enregistrements en extérieur en général, présence de bruits de fonds et d'évènements sonores autres que la musique). Ils contiennent principalement des tours de chant soliste/chœur, des zones de voix chantée en alternance ou superposée avec des instruments ou de la parole. Les fichiers ont été annotés manuellement en tours de chant. Une frontière est posée dans les situations suivantes :

- Passage d'un seul chanteur (soliste) à plusieurs chanteurs (chœur) et vice versa,
- Passage d'un chanteur A à un chanteur B,
- Passage du chant à la parole et vice versa.

Ce corpus est constitué de 9 enregistrements dont la durée totale est de 20 minutes que nous avons divisées en un corpus de développement (DEV) et un corpus d'évaluation (EVAL) dans les proportions 20% et 80% respectivement.

3.2 Mise en œuvre de l'algorithme

3.2.1 Choix des paramètres acoustiques

Toujours par analogie entre tours de parole et tours de chant, nous avons travaillé avec des paramètres acoustiques utilisés en segmentation de la parole. Nous avons étudié comme vecteur d'observation, les 12 MFCC, accompagnés éventuellement de l'énergie de la trame, des dérivées premières et secondes de ces coefficients, les MELSPEC, les PLP et les RASTA-PLP. L'estimation est réalisée sur 20 ms, toutes les 10 ms.

3.2.2 Ajustement des paramètres

Les paramètres de l'algorithme ont été déterminés sur le corpus de développement :

- N_{min} : la taille minimale de la fenêtre de recherche d'une frontière est fixée à 0,8s tandis que sa longueur maximale N_{max} est de 5s,
- $\Delta N_{augmentation}$: le nombre d'observations ajoutées à la fenêtre de détection tant qu'il

[2] http://diadems.telemeta.org/archives/fonds/CNRSMH_DIADEMS/

n'y a pas de frontière de segments détectée et tant que la taille maximale n'est pas atteinte, correspond à 0,5s,

- $\Delta N_{décalage}$: le décalage de la fenêtre sur le signal lorsque la taille maximale est atteinte et qu'aucune frontière potentielle n'est découverte correspond à 0,4s,
- N_{marge} : la taille minimale d'un segment de chant est de 0,14s,
- $N_{seconde}$: la taille de la fenêtre d'analyse fine correspond à 1,2s,
- la faible résolution temporelle du calcul du ΔBIC est de $\delta 1 = 5$ (1 trame sur 5, soit 50ms) tandis que la haute résolution considère toutes les trames de paramètres : $\delta h = 1$ (soit 10ms de précision).

3.2.3 Premiers résultats

Les performances ont été évaluées en termes de Rappel, Précision et F-mesure pondérés sur les durées des fichiers. Nous acceptons une tolérance de plus ou moins 30 ms sur les positions des frontières. Les meilleurs résultats ont été obtenus en limitant le vecteur d'observation à 12 coefficients MFCC. Seuls les paramètres RASTA-PLP ont donné des résultats équivalents à ceux obtenus avec les MFCC mais sans gain significatif.

Afin de définir une valeur « optimale » du coefficient de pénalité λ, les performances du système ont été calculées en le faisant varier sur les enregistrements de l'ensemble de développement. La valeur choisie de 1,02 pour présenter les résultats correspond au point de fonctionnement où la précision et le rappel sont égaux sur les enregistrements de DEV. Avec une telle configuration, des F-mesures de 66,6% et de 43,7% ont été obtenues sur DEV et EVAL respectivement. Ces valeurs sont rapportées dans la première partie du Tableau **1** de la section 4. La performance sur EVAL est beaucoup plus faible que sur DEV : une analyse des résultats montre que deux enregistrements qui présentent des alternances de chant solo/chœur très rapides, ont des performances très faibles, avec des F-mesures entre 30% et 40%.

4 Pertinence du coefficient de pénalité λ

L'ajustement du coefficient de pénalité s'est avéré délicat ; nous illustrons et analysons dans la première partie de cette section sa pertinence ainsi que sa grande sensibilité aux variations des conditions et de contenus des enregistrements. Afin de remédier à ce problème de variabilité, il nous est apparu nécessaire de relâcher cette contrainte, en raisonnant systématiquement sur plusieurs valeurs de λ et en confrontant plusieurs résultats de segmentation ; ce nouvel algorithme est présenté dans cette même section.

4.1 Variabilité des performances en fonction de λ

Le rôle du coefficient λ dans le critère BIC est de pénaliser une modélisation trop complexe : dans le cadre gaussien et multi modèles qui est le nôtre, plus la valeur de λ augmente, plus l'hypothèse H_1 est pénalisée et plus l'insertion d'une frontière est difficile : l'algorithme a tendance à moins segmenter. Globalement, choisir la bonne valeur de λ revient à trouver le bon compromis entre Rappel et Précision. La Figure **3** présente la variation de la performance globale de notre système sur le corpus DEV en fonction du coefficient de pénalité dans l'intervalle $[0{,}8 \quad 1{,}2]$. Nous remarquons que la performance varie de manière importante en fonction de ce facteur. Une des raisons majeures est sans nul doute liée à la petite taille de l'ensemble DEV.

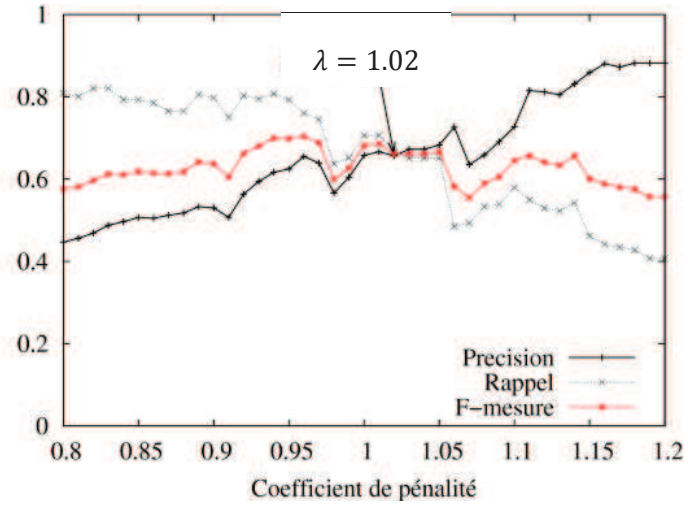

Figure 3 : Variation de la performance du système de segmentation en fonction du coefficient de pénalité sur le corpus DEV.

Néanmoins nous avons remarqué que la valeur de λ peut varier énormément d'un enregistrement à l'autre dès lors que l'on cherche à l'optimiser sur l'enregistrement seul. Pour certains enregistrements, nous trouvons de bonnes performances avec des valeurs proches de 1,2 ; or les segments attendus sont longs ; les valeurs moins élevées de l'ordre de 0,8 se révèlent meilleures là où les segments attendus sont plus courts. Cela engendre une variabilité importante de la performance globale de notre système en fonction de λ.

À des fins de comparaison, nous avons, pour chaque enregistrement du corpus DEV et EVAL, déterminé la meilleure valeur de F-mesure obtenue en faisant varier le coefficient de pénalité ; nous appellerons ce système artificiel, le système « *oracle* ». Il en résulte les performances globales données dans le Tableau **1**. Sur DEV et EVAL respectivement, la F-mesure atteint 79,5% et 53,8%. En comparant ces résultats avec ceux obtenus au point de fonctionnement λ constant à 1,02, nous trouvons un écart de performance d'environ 17% relatifs (13,4% absolus) pour DEV et 19% relatifs (10,1% absolus) pour EVAL. Cette différence absolue de 10% en termes de F-mesure confirme la nécessité de ne pas fixer *a priori* le coefficient de pénalité, quitte à inclure un post-traitement pour obtenir une segmentation *a posteriori*.

4.2 Décision *a posteriori* par consolidation

Afin d'éviter le problème de variabilité et le choix *a priori* du facteur de pénalité, nous avons effectué plusieurs segmentations en faisant varier le coefficient λ sur l'intervalle [0,8 1,2] avec un pas de 0,01. Sont ainsi obtenues 41 segmentations d'un même enregistrement. Pour décider quelles sont les frontières retenues, un vote est effectué sur les candidats obtenus : une frontière sera validée si elle est trouvée par au moins S_0 segmentations parmi les 41, en tolérant un écart de 0,3s. Nous parlerons de Décision Consolidée *A Posteriori* (DCAP). La valeur de S_0 comprise entre 1 et 41 a été déterminée en utilisant les enregistrements du DEV et fixée à 17.

4.3 Performances globales

Les résultats expérimentaux sont rassemblés dans le Tableau **1**. La performance globale obtenue par DCAP est de 50,4% en F-Mesure sur le corpus EVAL. L'augmentation est d'environ 6% relatifs (4,0% absolus) pour le corpus DEV et 15% relatifs (6,7% absolus) pour le corpus EVAL,

par rapport à celles trouvées avec une valeur de λ fixée à 1,02.

Naturellement cette performance reste plus faible que la performance du système « *oracle* » : il existe encore une marge de gain possible de 7% relatifs (3,4% absolus), si l'on considère la F-mesure obtenue avec le système « *oracle* » comme la limite supérieure à atteindre.

Système	DEV			EVAL		
	Précision	Rappel	F-mesure	Précision	Rappel	F-mesure
$\lambda = 1{,}02$	65,6	66,6	66,1	39,7	48,8	43,7
oracle	77,3	81,8	79,5	51,1	56,8	53,8
DCAP	65,7	75,2	70,1	59,3	43,8	50,4

Tableau 1 : Performance globale de notre système de segmentation en tours de chant.

Certains enregistrements du corpus EVAL montrent une F-mesure de presque 80% et d'autres restent à 30%. Les erreurs sur ces enregistrements sont principalement des fausses alarmes : leur écoute révèle beaucoup de superpositions de chanteurs, d'alternances très rapides entre les solistes et le chœur, la présence d'instruments percussifs comme des cloches, des frappements des mains, et du bruit de fond (parole, cris). En outre, ces enregistrements se sont avérés être les plus difficiles à annoter manuellement : dans certains cas d'alternances rapides de chanteurs, il n'est pas évident de décider s'il faut réellement insérer une frontière de segment : cette observation nécessite une analyse en termes d'usage pour comprendre les limites de l'attendu en termes de segmentation et nécessitera une discussion avec les ethnomusicologues.

5 Conclusions et perspectives

Dans cet article, nous avons présenté le problème de la segmentation en tours de chant par analogie aux tours de parole, en vue de traitements ultérieurs, avec comme objectif final l'indexation par le contenu d'enregistrements musicaux ethnomusicologiques. Nous avons appliqué une méthode de segmentation fondée sur le critère BIC. Le choix d'une unique valeur du paramètre de pénalité de ce critère, obtenu par ajustement global sur un corpus de développement, s'est avéré non satisfaisant, puisque les performances variaient sensiblement d'un enregistrement à un autre. Afin d'éviter de choisir une unique valeur, nous avons combiné des segmentations obtenues avec différentes valeurs, et obtenu la segmentation finale en ne gardant que les frontières présentes dans plusieurs d'entre elles. Avec cette dernière méthode, un gain en F-mesure de 15% relatifs a pu être obtenu par rapport au système de référence.

De nouvelles versions de cette segmentation sont testées actuellement, comme la combinaison avec des segmentations réalisées avec d'autres paramètres (RASTA-PLP) ainsi que la calibration/fusion discriminante des scores de ΔBIC des différentes segmentations. Nous souhaitons confirmer nos résultats sur une plus grande quantité de données et également sur des données différentes de celles du projet, comme des enregistrements en studio de musique dans des conditions acoustiques plus contrôlées, afin de préciser les limites de ce type d'approche. Dans un avenir plus lointain, il s'agira d'étudier le problème de regroupement de chanteurs.

Remerciements

Ce travail a bénéficié d'une aide de l'Agence Nationale de la Recherche portant la référence (ANR-12-CORD-0022).

Références

André-Obrecht, R. (1998). A new statistical approach for the Automatic Segmentation of Continuous Speech Signals, *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 36-1, pp. 29-40.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic and Control*, AC-19, pp. 716-723.

Anguera, X et Bozonnet, S. et Evants, N. et Fredouille, C. et Friedland, G. et Vinyals, O. (2012). Speaker Diarization: A Review of Recent Research. *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20:2.

Cettolo, M. et Vescovi, M. et Rizzi, R. (2005). Evaluation of BIC-based algorithms for audio segmentation. *In Computer Speech And Language*, pp. 147-170.

Chen, S. Scott et Gopalakrishnan, P. S. (1998). Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion. *In The DARPA Broadcast News Transcription and Understanding Workshop*, Landsdowne, VA, USA, pp. 127-132.

Delacourt, P. et Wellekens, C. (2000). DISTBIC: a speaker-based segmentation for audio data indexing. *In Speech Communication*, vol. 32, pp. 111-126.

El-Khoury, E. et Senac, C. et Pinquier, J. (2009). Improved speaker diarization system for meetings. *In Proc. ICASSP*, Taipei, pp. 4097-4100.

Lachambre, H. et André-Obrecht, R. et Pinquier, J. (2009). Singing voice detection in monophonic and polyphonic contexts. *In Proc. European Signal Processing Conference*, Glasgow, pp. 1344-1348.

Le Coz, M. et André-Obrecht, R. et Pinquier, J. (2012). Feasibility of the Detection of Choirs for Ethnomusicologic Music Indexing. *In Proc. International Workshop on Content-Based Multimedia Indexing*, Annecy, pp. 145-148.

Schwarz, Gideon. (1978). Estimating the dimension of a model. *In The annals of Statistic,* vol. 6, pp. 461-464.

Siu, M-H. Yu, G. et Gish, H. (1991). Segregation of speakers for speech recognition and speaker identification. *In Proc. ICASSP*, Toronto, Canado, pp. 873-876,

Zhu, X. et Barras, C. et Meignier, S. et Gauvain, J.-L. (2005). Combining speaker identification and BIC for speaker diarization. *In Proc. INTERSPEECH*, Lisbon, pp. 2441-2444.