Intérêt du suivi de fréquences pour la détection de sources harmoniques multiples

Maxime LE COZ, Julien PINQUIER, Régine ANDRÉ-OBRECHT

IRIT - 118, route de Narbonne
31062 Toulouse Cedex 9, FRANCE
{lecoz, pinquier, obrecht}@irit.fr

Résumé – Dans cet article, nous présentons une nouvelle approche pour la localisation de superposition de sources harmoniques. Notre méthode est basée sur le suivi des fréquence prédominantes du signal afin de former des *segments sinusoïdaux*. Les relations entre les fréquences de ces derniers sont ensuite étudiées afin de regrouper les *segments sinusoïdaux* appartenant à une même source. Les sources étant localisées sur le plan temps-fréquence, les zones où coexistent différentes sources sont finalement extraites. Notre approche à été testée à la fois sur des contenus de parole et de musique avec des résultats prometteurs.

Abstract – In this article, we present a new approach for the localization of superposed harmonic sources. Our method relies on a tracking of the predominant frequencies of the signal in order to form *Sinusoidal segments*. The ratios between the frequencies of those *Sinusoidal segments* are studied in order to group together those belonging to the same source. Once the source localized on the time-frequency plan, the instant where multiple sources coexist are labeled as containing superposition. The approach has ben tested on both speech and musical contents and shows promissing results.

1 Introduction

1.1 État de l'art

La détection de zones où plusieurs sources harmoniques sont présentes est une étape importante pour l'amélioration des systèmes de transcription, que ce soit en musique ou en parole. En effet, les sources harmoniques interagissent de manière extrêmement complexe et des stratégies spécifiques doivent être envisagées dans un tel contexte.

La communauté parole étudie des contenus de parole lue, préparée, voire spontanée en supposant que le locuteur est seul ou éventuellement en contexte bruité. Des recherches plus récentes visent la transcription de flux de parole moins maîtrisée en présence notamment de parole superposée [8]. D'autres approches basées sur des filtrages par peignes harmoniques ont également été proposées par [5] pour une estimation de plusieurs fréquences fondamentales en contexte de parole ou de musique.

En musique, l'existence de plusieurs sources harmoniques est un problème présent dans la grande majorité des cas pratiques. La transcription multi-pitch est d'ailleurs un sujet très actif de ces dernières années et a donné lieux à plusieurs tâches dans de grandes campagnes d'évaluation comme MIREX [1] ou QUAERO [2]. Diverses approches ont été proposées. On citera par exemple [4] dont la méthode consiste à analyser un critère de saillance des fréquences du spectre afin de détecter dans une trame, la fréquence prédominante et ses harmoniques. Un motif de décroissance de la saillance, défini par la fréquence fondamentale, est ensuite soustrait de la fonction de saillance et une nouvelle fréquence est estimée tant que le résidu reste différent du bruit. Une autre approche basée sur des motifs de décroissance est proposée par [1] sur une analyse fréquentielle multi-résolution afin de limiter l'influence du bruit.

Enfin, les approches basées sur le suivi commencent à être utilisées, mais dans un objectif monopitch pour extraire, parmi un ensemble de candidats par trame, le meilleur contour représentant la mélodie principale [7].

1.2 Ce que nous cherchons

Le système que nous présentons vise à localiser temporellement plusieurs sources harmoniques par le suivi des fréquences prédominantes dans le signal. Il vise aussi à attribuer les fréquences prédominantes détectées à l'une des sources présente. Cette localisation s'effectue sans *a priori* sur le nombre de sources, ni modèle acoustique.

2 Système

Notre système est basée sur l'analyse en contexte des événements harmoniques à l'aide d'un suivi de fréquences. En effet les sources pouvant être difficiles à discerner dans les instants de recouvrement, pouvoir estimer la façon dont se comportent les sources dans le contexte des zones problématiques permet de lever l'ambiguïté.

1. http ://www.music-ir.org/mirex/
2. http ://www.quaero.org/

La figure 1 illustre bien sur un exemple de parole superposée, la difficulté d'analyse de la zone de superposition. La présence de deux sources différentes est rendu évidente par la prolongation des deux sources de part et d'autre du phénomène de superposition (en bleu et en vert).

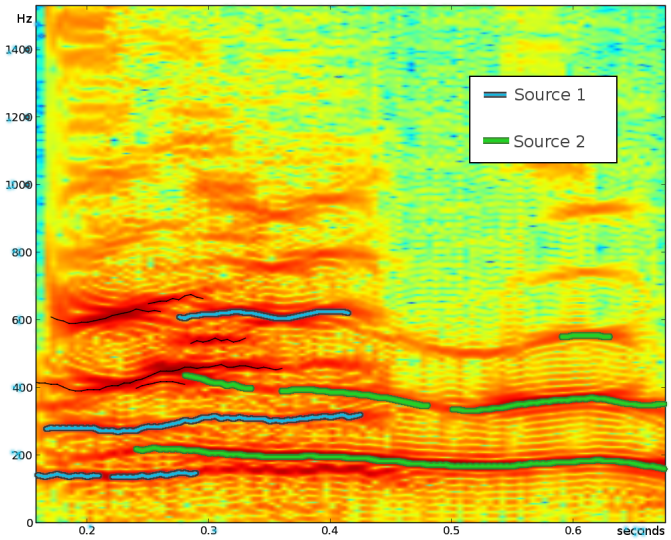

FIGURE 1 – Exemple de phénomène de superposition en parole. Le suivi des fréquences à partir et après la zone de superposition permet de mettre en évidence l'existence des deux sources (en bleu et vert).

Le système que nous présentons s'articule autour de cinq étapes principales :
- sélection de zones d'intérêt,
- sélection des fréquences candidates,
- suivi de fréquences,
- rassemblement harmonique,
- localisation de superpositions.

2.1 Sélection de zones d'intérêt

Un spectrogramme classique est calculé sur des segments isolés de parole ou de musique. La sélection de ces zones est une étape très importante du processus puisque notre approche prend l'hypothèse de la présence d'au moins une source harmonique. La sélection des zones de parole et de musique peut être effectuée *a priori* en utilisant la méthode décrite dans [6].

Sur chaque trame d'analyse, la transformée de Fourier rapide est calculée seulement sur les fréquences inférieures à une valeur de coupure F_{max}. Ce seuil permet de se concentrer sur les zones où les amplitudes des harmoniques liées aux sources sont les plus énergétiques et les plus clairement dissociées du bruit. Cette valeur a été fixée empiriquement à 3000 Hz.

2.2 Sélection de fréquences potentiellement liées aux sources

Sur chaque zone d'intérêt sélectionnée, une recherche des fréquences prédominantes est effectuée.

Le pic maximal de chaque trame analysée est sélectionné. Le couple de valeurs (fréquence, amplitude) de ce pic est noté (f_{max}, amp_{max}).

Les pics d'amplitude supérieure à un seuil défini par une fonction linéaire par morceau paramétrée définie comme suit :

$$th(f) = \begin{cases} a(\frac{(r_{max}-r_{deb})}{f_{p_{max}}} + r_{deb}) \textbf{ pour } f \in [0, f_{p_{max}}] \\ a(\frac{(r_{fin}-r_{max})}{f_{max}-f_{p_{max}}} + r_{deb}) \textbf{ pour } f \in [f_{p_{max}}, f_{max}] \end{cases} \quad (1)$$

Ce seuil sur l'amplitude permet de prendre en compte l'affaiblissement du rapport signal sur bruit lors de l'augmentation des fréquences. Cette décision est également motivée par la nécessité de garder un nombre élevé de candidats pour les étapes suivantes du traitement, et ce sur l'ensemble de la bande de fréquence analysée.

La Figure 2 illustre la façon dont les pics sont sélectionnés afin de tenir compte de l'amplitude locale du signal.

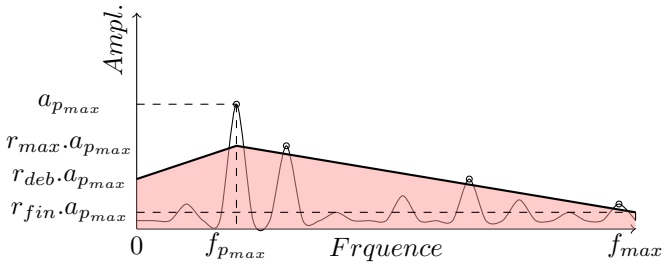

FIGURE 2 – Sélection des pics candidats par seuil dynamique. La fonction linéaire par morceaux utilisée comme seuil est définie par les coordonnées du pic principal. Seuls les pics d'amplitude supérieure au seuil sont sélectionnés.

2.3 Suivi de fréquences

Une fois les pics extraits sur les trames, un suivi de fréquence est effectué afin de suivre l'évolution temporelle des fréquences prédominantes. La méthode de suivi a été créée par Tanigushi [9]. Celle-ci permet de relier les pics des différentes trames en *segments sinusoïdaux*.

Afin d'être reliés, deux pics de deux trames adjacentes respectivement $p_t^i = (f_t^i, amp_t^i)$ et $p2 = (f_{t+1}^j, amp_{t+1}^j)$ doivent remplir la condition $d_{Tani}(p1, p2) < Th_{Tani}$.

La distance $d_{Tani}(p1, p2)$ est calculée selon la formule suivante :

$$d_{Tani}(p1, p2) = \sqrt{\left(\frac{f_t^i - f_{t+1}^j}{C_f}\right)^2 \times \left(\frac{amp_t^i - amp_{t+1}^j}{C_p}\right)^2} \quad (2)$$

Pour expliciter ces valeurs, on notera que pour une valeur de seuil Th_{Tani} fixé à 1, ceci revient à trouver un pic dans un voisinage elliptique du plan amplitude-fréquence avec un rayon sur les fréquences de valeur C_f et un rayon sur l'amplitude C_p.

Si la condition est remplie, alors les pics des deux trames sont liés au sein d'un même *segments sinusoïdal* et le processus continue jusqu'à l'analyse complète des trames.

Les *Segments Sinusoïdaux* (SS) extraits suivent l'évolution en temps-fréquences des principales harmoniques présentes sur le segment temporel. Afin de ne garder que les *SS* significatifs, tout ceux de longueur inférieure à un seuil $nLenMin$ ne sont pas conservés. Nous estimons, en effet, que plus une suite de pics liés est longue, plus elle est significative et moins elle ne peut être le fruit du hasard.

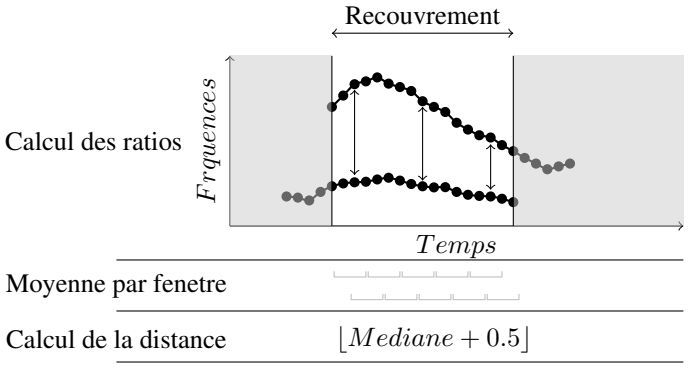

FIGURE 3 – Chaine de calcul du critère harmonique entre deux *Segments Sinusoïdaux*. Ces différentes étapes permettent d'assurer un ratio entier constant entre les valeurs de fréquence pour une valeur de distance proche de 0.

2.4 Regroupement harmonique

À partir d'un ensemble de *SS*, nous cherchons à estimer ceux qui sont liés à la même source harmonique. Pour cela, nous utilisons le fait qu'il existe des rapports entiers entres les fréquences issues d'une même source. Nous calculons la distance d_{clus} entre tous les couples de *SS* extraits ayant un minimum de $th_{minOverlap}$ trames en commun.

Pour chaque couple (ss_1, ss_2), nous analysons la liste des fréquences $lf(ss_1)$ et $lf(ss_2)$ sur leur section de recouvrement temporel.

Pour chaque suite de $th_{minOverlap}$ couples de fréquences consécutive sur la durée du recouvrement, la moyenne des valeurs de $divrg_(f_1, f_2)$ est calculée. Ceci afin de décrire la relation locale entre le comportement des deux *Segments Sinusoidaux*.

$divrg$ calcule la distance entre le ratio de la plus grande fréquence sur la plus faible et l'entier le plus proche.

Une fois cette valeur extraite pour chaque fenêtre, la valeur $d_{clus}(ss_1, ss_2)$ est extraite comme la valeur médiane parmi les fenêtres.

Ainsi, nous réalisons un graphe où les nœuds sont les SS et nous relions uniquement ceux dont $d_{clus} < th_{clus}$. De ce graphe, chaque composante connexe représente les *SS* liés à une source. Ceci permet d'assurer une transitivité pour le groupement.

Un exemple de graphe représentant les sources pour la musique et pour la parole est présenté sur la figure 2.4.

2.5 Localisation de zones de superposition

Une fois les différents regroupements établis, la recherche de zones de sources simultanées se fait sur l'ensemble d'un segment temporel étudié. Nous recherchons alors la présence simultanée d'au moins deux sources.

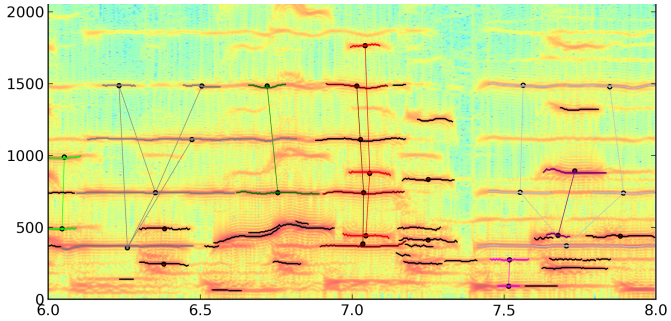

FIGURE 4 – Exemple de regroupements harmoniques sur un extrait de musique polyphonique. Sur la deuxième partie de l'image, nous pouvons observer la transitivité permettant de conserver l'unité de la source grise.

3 Expérience

3.1 Description

Une expérience qualitative a permis de donner de bons résultats sur deux types de contextes musicaux. Le premier correspond à de la pop européenne : une chanson du concours de l'*Eurovision*, enregistrée en studio comportant des solos et des parties accompagnées. Le second est un chœur ethnique enregistré en extérieur avec un fichier de chant africain extrait de la base sonore du musée de l'homme [2]. Dans ce dernier fichier, nous cherchons à différentier les passages de chœur et de solo. Malgré la différence de contenu et de qualité de ces deux enregistrement, nous avons choisi de tester la robustesse de notre approche en paramétrant notre système de la même façon pour ces deux analyses. Chaque segment de 2 secondes d'enregistrement est classé suivant notre méthode.

Notre système à également été évalué sur la parole : des fichiers extraits de la campagne d'évaluation *ANR ETAPE* [3]. Des zones, contenants respectivement des segments de parole superposée et non superposée dans des proportions équivalentes, ont été extraites de ces fichiers afin d'évaluer notre système

dans des conditions comparables. Ce corpus comprends 60 extraits d'une durée d'une seconde en moyenne.

3.2 Résultats

L'expérience sur les contenus musicaux obtient un taux d'accuracy (bonne classification) de 71% sur le fichier d'*Eurovision* et de 69% pour l'enregistrement de musique ethnique. Ces résultats sont obtenus en gardant les mêmes paramètres pour les deux types de musique. Bien sûr, une paramétrisation spécifique à ces deux types de musique très différents est possible, mais l'objectif de notre méthode est de garder au maximum une approche générique pour un type de contenu donné. Les erreurs sont principalement dues à une différence d'amplitude trop élevée entre deux sources qui tend parfois à masquer l'une des sources, ceci conduisant à une décision erronée de la présence d'une unique source. Inversement, certains cas de recherche de sources dans des zones extrêmement bruitées peut conduire à un suivi chaotique des pics appartenant au bruit, et menant à une fausse détection de plusieurs sources.

Les résultats sur les segments de parole donnent un taux de bonne classification de plus de 78%. Dans cette expérience, les erreurs sont plutôt liées à la sur-détection de sources avec seulement un échantillon de parole superposée classé comme n'ayant qu'une seule source. Comme pour l'expérience sur le contenu musical, les erreurs de détection de « fausses sources » sont principalement dues au suivi dans des zones n'ayant pas de sources harmoniques. En effet, la durée de recouvrement de parole pouvant être extrêmement courte, elle se trouve de l'ordre de longues fricatives ou des zones de silences de réflexion.

4 Conclusion & Perspectives

Notre étude montre la possibilité d'utiliser une méthode de suivi de fréquences pour la détection de sources harmoniques simultanées. Le principal intérêt de cette approche, par rapport aux approches basées sur des apprentissages, réside dans sa souplesse permettant de s'adapter à des contextes et des contenus très différents.

Bien que nous ayons choisi de garder la même paramétrisation pour les deux types de musique, notre approche pourrait également utiliser des informations *a priori* sur le contenu à étudier (parole ou musique, niveau de bruit...). Une paramétrisation différente permettant de rendre le suivi moins sensible au bruits de fond, ou au contraire de chercher des sources ayant une plus grande différence d'amplitudes entre elles.

Afin d'améliorer notre système et d'éviter notamment la surévaluation des sources, nous pourrions utiliser un indicateur de présence de source harmonique à l'échelle de la trame afin de ne pas chercher à suivre l'évolution des pics correspondant à du bruit. Nous planifions également de réaliser une fusion entre notre approche et une méthode plus classique d'extraction multi-pitch en musique afin d'utiliser les forces et faiblesses des deux approches.

L'estimation précise du nombre de sources est également un objectif à plus long terme de notre approche, nécessitant probablement la combinaison avec d'autres méthodes afin de minimiser les erreurs.

Enfin, notre méthode permet une localisation fréquentielle des différentes sources qui pourrait ensuite servir de support à des analyses visant à les caractériser ou les isoler. Citons par exemple l'identification d'instruments à travers le timbre via le profil harmonique ou la séparation de sources.

Références

[1] Karin Dressler. An auditory streaming approach for melody extraction from polyphonic music. In *ISMIR 2011 proceedings*, October 2011.

[2] Telemeta Plateform For Ethnomusicology Exchange. http ://archives.crem-cnrs.fr/items/cnrsmh_i_1957_003_001_03/.

[3] Guillaume Gravier, Gilles Adda, Niklas Paulson, Matthieu Carré, Aude Giraudel, and Olivier Galibert. The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In *International Conference on Language Resources, Evaluation and Corpora*, page na, Turquie, 2012.

[4] Anssi Klapuri. Multipitch analysis of polyphonic music and speech signals using an auditory model. *IEEE Transactions on Audio, Speech & Language Processing*, 16(2) :255–266, 2008.

[5] Jean-Sylvain Liénard, Francois Signol, and Claude Barras. Speech fundamental frequency estimation using the alternate comb. In *INTERSPEECH*, pages 2773–2776, 2007.

[6] J. Pinquier, J. L. Rouas, and R. Andre-Obrecht. A fusion study in speech/music classification. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing (ICASSP '03)*, volume 2, 2003.

[7] Joe Cheri Ross, Vinutha T. P., and Preeti Rao. Detecting melodic motifs from audio for hindustani classical music. In Fabien Gouyon, Perfecto Herrera, Luis Gustavo Martins, and Meinard Müller, editors, *ISMIR*, pages 193–198. FEUP Edicoes, 2012.

[8] Francois Signol. *Automatic multipitch estimation for monaural speech mixture signals.* PhD thesis, LIMSI-CNRS, Université Paris Sud B.P. 133 F-91403 ORSAY CEDEX, 2009.

[9] Toru Taniguchi, Mikio Tohyama, and Katsuhiko Shirai. Detection of speech and music based on spectral tracking. *Speech Commun.*, 50(7) :547–563, July 2008.