Modèle sinusoïdale : Estimation de la qualité de jeu d'un musicien, détection de certains effets d'interprétation

Antony Schutz, Dirk Slock

Institut EURECOM
2229 route des Crêtes BP 193 - 06904 Sophia Antipolis Cedex, France
antony.schutz@eurecom.fr, dirk.slock@eurecom.fr

Résumé − L'estimation des paramètres d'un signal sinusoïdale est un problème ancien et est largement décrit dans la littérature. Nous proposons ici d'utiliser une méthode bien connue, l'estimation des paramètres par le biais des transformées de Fourier rapide (FFT), en prenant en compte l'influence mutuelle des pics fréquentiels. L'estimation du rapport signal sur bruit (RSB) nous permettra, dans le cas mono-instrumental, de juger la qualité de jeu d'un musicien sur certains instruments acoustiques ainsi que de détecter certains effets d'interprétation.

Abstract − The estimation of the parameters of a sinusoidal signal is an old problem and is largely described in the literature. We propose here to use of a well known method, the estimation of the parameters by using the Fast Fourier Transform (FFT), by taking account the mutual influence of the frequential peaks. The estimate of the signal to noise ratio (SNR) will allow us, in the mono-instrumental case, to judge the quality of play of a musician on certain acoustic instruments or detecting certain effects of interpretation.

1 Introduction

Le modèle sinusoïdale additif est une technique très employée [1, 2, 3]. Nous considérons, ici, les paramètres d'un signal sinusoïdale $s(t)$ définit par :

$$s(t) = x(t) + n(t), \tag{1}$$

$$x(t) = \sum_{n=0}^{N-1} A_n(t) \cos(2\pi \frac{f_n(t)}{f_e} + \phi_n(t)) \tag{2}$$

Où $x(t)$ représente la partie sinusoïdale du signal définie par les paramètres $A_n(t)$, $f_n(t)$ et $\phi_n(t)$ qui sont respectivement l'amplitude, la fréquence et la phase de la n^{ieme} harmonique à l'instant t. f_e est la fréquence d'échantillonnage et $n(t)$ représente un bruit blanc Gaussien additif de variance σ^2. Le Signal audio, qui par nature est non stationnaire, est analysé par morceaux. La méthode de synthèse consiste donc à extraire les paramètres de chaque fenêtre, en générer autant de signaux partiels puis de recoller les morceaux en utilisant la méthode d'addition et recouvrement. Le bruit est extrait par une simple soustraction du signal bruité par le signal synthétisé :

$$\hat{x}(t) = \sum_{n=1}^{N} A_n^{est}(t) \cos(2\pi \frac{f_n^{est}(t)}{f_e} + \phi_n^{est}(t)) \tag{3}$$

$$n^{est}(t) = s(t) - \hat{x}(t), \tag{4}$$

Ne soustraire que la partie sinusoïdale implique que le bruit estimé sera constitué de tout ce qui sort du modèle comme le bruit d'enregistrement et environnent mais surtout le bruit instrumental qui a reçu un intérêt particulier pour certaines applications audio [4].

2 Méthode de synthèse du son Extraction du bruit

2.1 Motivation - Méthode

Le besoin d'algorithme de faible complexité pour les applications temps réel incite souvent l'utilisation des méthodes basées sur l'étude des spectres à court terme, comme ceux obtenue par une transformée de Fourier à court terme (TFCT) [5]. Le signal est analysé par trame en utilisant une fenêtre glissante avec recouvrement. Le choix de cette fenêtre d'analyse est soumis à plusieurs contraintes temporelles et spectrales :

- Premièrement, nous utilisons un modèle stationnaire ce qui implique que la fenêtre doit être la plus courte possible pour pouvoir considérer que les paramètres sont constants. La taille de la fenêtre est limitée par l'incertitude temps-fréquence.
- Si la fenêtre n'est pas rectangulaire, l'énergie doit être concentrée au centre afin de favoriser ces échantillons et de limiter l'effet de la non stationnarité.
- La fenêtre doit être symétrique et autoriser une reconstruction parfaite.
- Pour les contraintes spectrales, le signal est considéré être une somme de sinusoïde analysé par une fenêtre temporellement bornée il s'en suit que le spectre sera le résultat de la convolution d'une somme de delta de Dirac avec la FFT de la fenêtre, qui, par nature possède un support infini. Outre la largeur du lobe principale il faut que la FFT de la fenêtre possède une atténuation importante à partir d'une certaine distance du lobe afin de rendre négligeable l'effet des interférences entre pics fréquentiels [6].

2.2 Estimation de la fréquence

Le maximum de vraisemblance amène à rechercher le maximum du périodogramme [7], la précision sera donnée pour chaque maximum k_m par $\hat{f}_m = k_m \frac{f_s}{N}$, où N est la taille de la fenêtre. La précision peut être améliorée par zéro padding. Les méthodes couramment rencontrées sont les méthodes basées sur la phase [8], l'interpolation [1]. Dans [9], l'équivalence théorique des méthodes basées sur la phase est montrée.

La figure 1 montre l'écart-type de l'erreur d'estimation de la fréquence pour une simple cisoïde, pour le vocodeur de phase et l'interpolation parabolique, comparé à la borne de Cramer-Rao (CRB)[7, 9] donné par :

$$CRB_c = \frac{6}{P^2\ L\ (L^2-1)}\ 10^{-\frac{RSB}{10}} \qquad (5)$$

P est l'amplitude de la cisoïde et est égale à un et L la taille des données.

On y voit que l'erreur d'estimation de l'interpolation sature à partir de 40 dB, cependant il faut souligner que dans notre cas le RSB est généralement faible du fait de la présence du bruit instrumentale. L'utilisation d'une méthode Haute Résolution de type ESPRIT [10, 11] nous a permis de conclure que dans ce cadre d'application le RSB était généralement inférieure à cette valeur.

La figure 2 (deux premières courbes) montre l'écart-type de l'erreur d'estimation de la fréquence pour une sinusoïde (10000 réalisations d'une sinusoïde dont la fréquence est uniformément distribuée de 40 à 500 Hz, la phase de 0 à 2 π et d'amplitude égale à un), pour le vocodeur de phase et l'interpolation parabolique comparés à la CRB notons que pour un signal réel la CRB est donnée par : $CRB_r = 2\ CRB_c$ [12]. La différence entre les deux études réside dans la présence de la fréquence négative du signal réel, en se basant en basse fréquence (tout comme en haute fréquence) l'étude effectuée sur le pic est perturbée par son reflet négatif. Il s'agit cependant d'une perturbation à courte portée qui est conditionnée par la fenêtre d'analyse.

La figure 3 montre l'erreur d'estimation pour une cisoïde dont la fréquence varie par blocs (dans chaque fenêtre la fréquence est constante) et pour deux cisoïdes l'une fixe et l'autre variant, en assumant la position des maximums connus). Nous avons choisi d'utiliser une fenêtre de Hann pour l'étude et on peut dire qu'à partir d'une distance $\Delta_f = 500\ Hz$ la perturbation n'aggrave pas l'erreur d'estimation.

La fenêtre est connue et on peut réduire l'influence des interférences en estimant les paramètres des pics interférants et en soustrayant leurs contributions. La figure 2 présente l'erreur pour les deux méthodes, ainsi que celles obtenues en réduisant les effets des perturbations (la sinusoïde est générée comme précédemment) en combinant les deux méthodes pour la correction et l'estimation. Nous pouvons conclure que la correction est nécessaire mais que dans la tranche de RSB qui nous intéresse les méthodes avec correction sont équivalentes. La taille des fenêtres est fixée à 1024 échantillons, la fréquence d'échantillonage à 44100 Hz (la durée de chaque fenêtre est donc de 23.22 ms). Le recouvrement est de 50 % et un zéro-padding de facteur 4. Le dernier paramètre à choisir demeure le nombre de pics recherchés nous l'avons choisi fixe et dans le cas d'un son instrumental nous avons fixé la valeur à $N_b = 32$. Chercher un nombre fixe de pics dans le

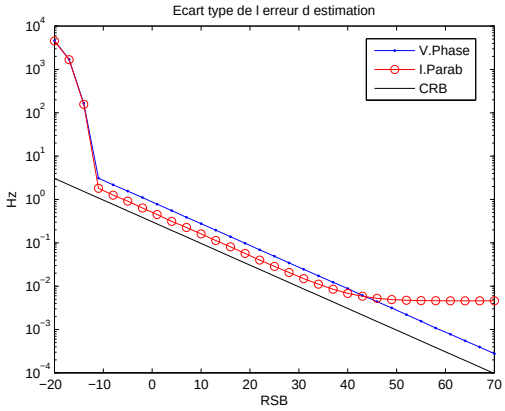

FIG. 1 – *Ecart type de l'erreur d'estimation en fréquence pour une cisoïde*

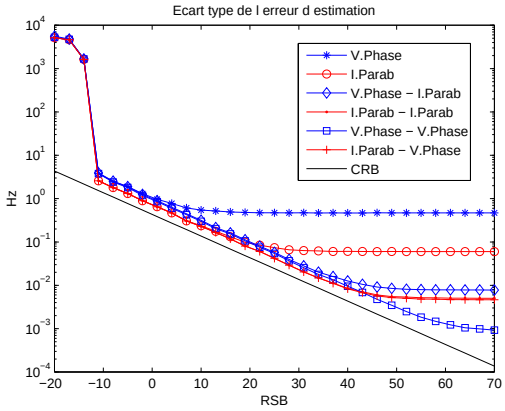

FIG. 2 – *Ecart type de l'erreur d'estimation en fréquence pour une sinusoïde*

spectre contraint quelque peu les résultats. Dans le cas où le signal n'est constitué que de bruit, le signal sera synthétisé par les paramètres liés aux pics dominants du bruit et le RSB estimé aura une borne minimum. De manière équivalente si le spectre est riche, les harmoniques les plus faibles seront interprétées comme appartenant au bruit et on sous-estimera le RSB.

2.3 Interpolation Correction des amplitudes

Nous utilisons l'interpolation parabolique, pour chaque maximum f_m détecté, on définit :

$$Y_{m'} = S_{dB}(f_m + m'), \qquad m' = -1, 0, 1 \qquad (6)$$

Où $S_{dB}(f) = 20log_{10}(|X(f)|)$, et $X(f)$ est la transformée de Fourier du signal $x(t)$. La fréquence estimée est donnée par :

$$f_m^{est} = f_m + \frac{1}{2}\ \frac{Y_{+1} - Y_{-1}}{Y_{-1} + Y_{+1} - 2Y_0} \qquad (7)$$

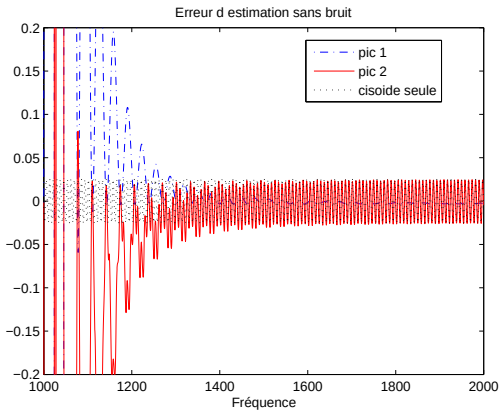

FIG. 3 – *Erreur d'estimation en fréquence, avec et sans perturbation*

Et l'amplitude correspondante par :

$$S_{dB}^{est} = Y_0 - \frac{f_m^{est}}{4}\,(Y_{-1} - Y_{+1}),\ A_m^{est} = 10^{\frac{1}{20}\ S_{dB}^{est}} \qquad (8)$$

Pour calculer l'influence dûe aux autres pics il faut obtenir une expression des perturbations engendrées par leur présence dans le voisinage du pic étudié. Dans notre cas la fenêtre temporelle utilisée est une fenêtre de Hann de taille N définie par :

$$w(n) \ = \ 0.5 \ - \ 0.5 \ \cos(\ 2\pi\frac{n}{N}\ ), \qquad 0 \leq n < N \qquad (9)$$

La fenêtre de Hann est temporellement bornée, elle est multipliée par une fonction rectangle, on peut réécrire la fenêtre de Hann :

$$\begin{aligned} w(n) \ &= \ [\ 0.5 \ - \ 0.5 \ \cos(\ 2\pi\frac{n}{N}\ )\ ]\ \ r(n) \qquad (10)\\ &= \ 0.5\ r(n) - 0.25\ e^{2i\pi\frac{n}{N}}\ r(n) - 0.25\ e^{-2i\pi\frac{n}{N}}\ r(n) \end{aligned}$$

La TFD de la fonction rectangle s'écrivant :

$$R(f) = \sum_{t=0}^{T-1}(e^{-2i\pi ft}) = e^{-i\pi f(T-1)}\ \frac{sin(\pi fT)}{sin(\pi f)} \qquad (11)$$

Il s'ensuit pour la fenêtre de Hamming :

$$W(f) = 0.5\ R(f) - 0.25\ R(f - \frac{1}{T}) - 0.25\ R(f + \frac{1}{T}) \qquad (12)$$

Pour chaque pic nous soustrayons la contribution [6] des interférants définit par :

$$W_m^{est}(f) = \sum_{\substack{n=1\\ n\neq m}}^{Nb_{\in\Delta f}} A_n^{est} W(f - f_n^{est}) e^{i\phi_n^{est}} + ... \qquad (13)$$

$$+ \sum_{n=1}^{Nb_{\in\Delta f}} A_n^{est} W(f + f_n^{est}) e^{i\phi_n^{est}},\ f = [f_m - 1, f_m, f_m + 1]$$

Où A_n^{est},f_n^{est} et ϕ_n^{est} sont les paramètres estimées et f_m la fréquence entière, un maximum. Le deuxième terme de l'équation ne sert qu'en basse et haute fréquence. Une fois les contributions soustraites nous ré interpolons le pic afin d'en obtenir la fréquence et l'amplitude puis nous calculons la phase (interpolation linéaire de la phase déroulée).

3 Estimation de la qualité de jeu, détection d'effets d'interprétation

3.1 Introduction

Les signaux audio sont modélisés comme des sommes de sinusoïdes. Cependant ce modèle omet certaines caractéristiques du son de ces instruments [13], un son de flute est composé de souffle, un son de violon est provoqué par la décrochement de la corde par l'archet ou plus généralement pour les instruments à touche et à cordes le son est induit par une attaque. On peut en tirer un avantage, une fois que la principale partie sinusoïdale est extraite il ne demeure que le bruit environnant et instrumental. Dans ces conditions l'étude du RSB devient un précieux allié pour juger la qualité de jeu d'un étudiant, ou encore détecter un effet d'interprétation. Par exemple un bassiste à le choix de jouer les cordes en les attaquants avec les doigts ou en les frappants avec le pouce.

3.2 Application

3.2.1 Detection du « slap » à la basse

Le « slap » est une technique couramment utilisée par les bassistes. L'attaque consiste à frapper les cordes avec la tranche extérieure du pouce sur ou près du manche, comme un marteau. Le son résultant est presque totalement percussif au moment de l'attaque puis le régime sinusoïdale s'installe. On constate que la note « slapée »

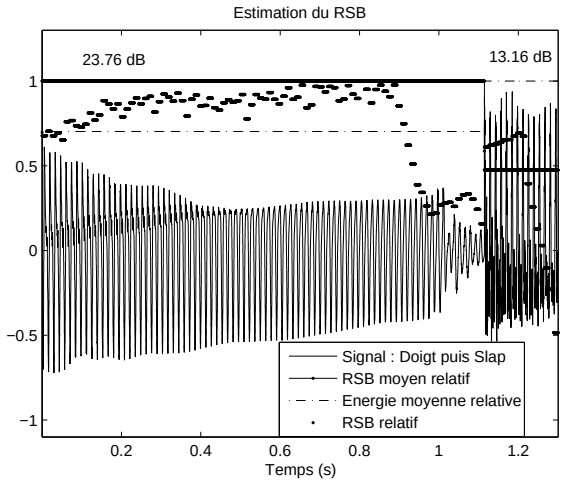

FIG. 4 – Evolution du RSB : La première est jouée au doigt et la deuxième en « slap » sur une basse

possède un RSB moyen deux fois inférieur à celle jouée au doigt. Le RSB n'est pas constant pendant la durée de la note. Au début il est faible ceci est du à l'attaque, puis il atteint son maximum avant de décroître avec l'atténuation des harmoniques (notons que les harmoniques d'une note slapée subissent l'effet du point d'attaque [14], le slap se pratique près du manche et davantage d'harmoniques sont atténuées). On se trouve ici dans un cas monophonique et dans ces conditions la détection est aisée.

3.2.2 Positionnement de l'archet sur un violon

Lorsque l'on joue du violon, le son est provoqué par le déplacement de l'archet. Pour qu'une note soit bien jouée il faut, outre le fait que la vitesse de déplacement et la pression exercée soient constantes, que l'archet soit parfaitement perpendiculaire aux cordes. Dès qu'il s'écarte de sa position le son devient de moins en moins « pur » et le frottement de plus en plus dominant. La figure 5 montre le résultat de l'analyse sur une pièce de violon jouée par un étudiant, on voit clairement les différences du RSB correspondant à une mauvaise position de l'archet et cela indépendamment du volume de la note, ainsi on peut définir des seuils qualifiant la qualité de la note. Dans cet exemple les notes parfaitement jouées ont un RSB supérieur à 18 dB contre un RSB inférieur à 7 dB pour les notes désastreuses. En pratique il s'agit d'ajuster ces seuils en tenant compte du bruit ambiant.

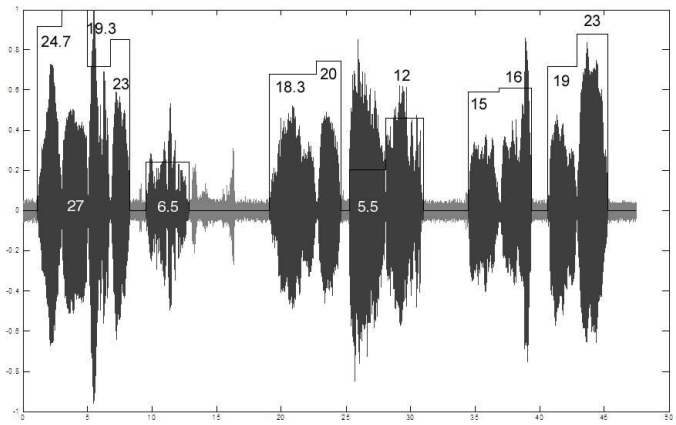

FIG. 5 – Evolution du RSB pour une pièce de violon. Clair : signal, foncé : note, les valeurs indiquées correspondent au RSB estimé

3.3 Discussion pour d'autres instruments

Etudier le bruit est un concept prometteur pour les applications musicales. Dans [4] le bruit est utilisé pour l'estimation du tempo, la détection des attaques y est plus aisée. On peut penser à d'autres applications, par exemple pour les instruments à vent, la qualité de la note dépend de la constance du souffle et de la hauteur de la note. Le souffle étant dans le bruit on peut estimer ses variations en étudiant l'évolution de l'énergie. Pour un hautbois ou une clarinette la hauteur varie en fonction de l'intensité du souffle cependant on peut très bien souffler sans avoir la moindre note. Pour les instruments à cordes, le bruit contenant essentiellement les attaques peut permettre de différencier une note jouée d'un *legato*, une seule attaque pour plusieurs notes.

3.4 Conclusion

Dans cet article, nous avons présenté une méthode rapide de synthèse audio pouvant être implémentée dans un simulateur temps réel. La méthode souffre de quelques défauts comme l'utilisation d'un modèle stationnaire, d'une méthode d'estimation biaisée, d'un ordre de modélisation fixe. Elle donne cependant des résultats satisfaisant pouvant très bien être utilisé à titre pédagogique pour des étudiants voulant parfaire leur maîtrise de jeu. Les concepts abordés dans la 3^{eme} partie peuvent être utilisés pour la transcription automatique où généralement les effets d'interprétation ne sont pas pris en compte.

Références

[1] F. Keiler, S. Marchand. *Survey on extraction of sinusoïdes in stationary sounds*. Proc. Of the 5^{th} Int. Conference on Digital Audio Effects (DAFx - 02).

[2] R .Althoff, F .Keiler et U .Zölzer. *Extracting sinusoïds from harmonic signals*. Proceedings of the 2nd COST G-6 Workshop on Digital Audio Effects (DAFx99), NTNU, Trondheim 1999

[3] E. Aboutanios. *A modified Dichotomous Search Frequency Estimator*. IEEE Signal Processing Letters, vol 11, No 2. February 2004.

[4] M. Alonso, R. Badeau, B. David and G. Richard. *Musical tempo estimation using noise subspace projection*. IEEE Workshop on Application of signal processing to audio and acoustics 2003.

[5] X. Serra. *Musical sound modeling with sinusoïds plus noise*. Musical signal processing. Editor C.Roads et al, Swets and Zeitlinger Publisher, 1997.

[6] J.S. Marques, L.B. Almeida. *A background for sinusoid based representation of voiced speech*. In Proc. Int. Conf. on acoustics, Speech and Signal Processing 1986.

[7] D.C. Rife, R.R. Boorstyn. *Single tone parameter estimation from discrete time observation*. IEEE Transaction on information theory 1974.

[8] F. Auger, P. Flandrin. *Improving the readibility of time frequency and time scale representations by the reassignement method*. IEEE Transaction on Signal Processing 1995.

[9] S. Marchand, M. Lagrange *On the Equivalence of Phase-Based Methods for the Estimation of Instantaneous Frequency*. In Proceedings of the 14th European Conference on Signal Processing 2006.

[10] R. Badeau. *Méthodes à haute résolution pour l'estimation et le suivi de sinusoïdes modulées*. Thèse de Doctorat ENST 2005.

[11] R. Badeau, G. Richard, B. David. *Fast adaptive ESPRIT algorithm*. IEEE Workshop on Statistical Signal Processing SSP'05.

[12] S.M. Kay. *Fundamentals of statistical Signal Processing - Estimation Theory*. Signal Processing Series. Prentice Hall, 1993

[13] N.H. Fletcher et T.D. Rossing *The physics of musical instruments*. Seconde édition, Springer-Verlag, 1998

[14] C. Traube, J.O. Smith *Estimating the plucking point position on a guitar string* Proc. Of the Int. Conference on Digital Audio Effects (DAFx-00), 2000