LIFT: LIKELIHOOD-FREQUENCY-TIME ANALYSIS FOR PARTIAL TRACKING AND AUTOMATIC TRANSCRIPTION OF MUSIC

Verfaille V.
CNRS-LMA
31, chemin Joseph Aiguier
13402 Marseille Cedex 20
FRANCE

Duhamel P.
CNRS-LSS, Supélec
Plateau de Moulon
91192 Gif-sur-Yvette Cedex
FRANCE

Charbit M.
ENST, TSI,
46, rue Barrault
75634 Paris Cedex 13
FRANCE

(verfaille@lma.cnrs-mrs.fr)
(charbit@tsi.enst.fr)
(pierre.duhamel@lss.supelec.fr)

ABSTRACT

We propose a new method for analysing the time-frequency domain, called LiFT. It is especially designed for partial tracking in a polyphonic automatic transcription model. After the signal passes through a Q-constant filter bank, composed of twenty four quarter-tone filters, it is analysed thanks to a generalized maximum likelihood approach. Two hypotheses are tested: the first one is that the output signal of a filter is a cosine plus noise, the second one is that it corresponds to colored noise. This likelihood analysis is developed in two ways: temporally treating the samples and frequentially treating the short time Fourier transform of the signal. For these two approaches, we have tested the robustness to noise and the cosine detection power.

1. INTRODUCTION

The automatic transcription of music is an active field of research in musical signal processing. Several approaches exist, all of them using pitch detection algorithms. The first class of methods looks for a periodicity in the time domain after the input signal is filtered [1]. The second class uses the frequency domain, to detect the harmonic peaks in the short term Fourier transform and then determine the corresponding fundamental frequency or pitch ([2], [3]), or thanks to cepstrum techniques [4]. Other methods exists, using both time and frequency methods, and pattern matching, for example psychoacoustic models [5].

A limit due to the short term Fourier transform is that the frequency resolution is the same for low and high frequency, whereas it is not in the way the ear functions. A way to solve this problem is to use a Q-constant analysis, such as described in [6] (one could also use wavelets). Like this, any harmonic pattern (corresponding to an equidistant repetition in the frequency domain) becomes the translation of an original pattern in the log-frequency domain. Separating harmonic sounds from a polyphony will be easier to do.

To this aim, we developed the generalized maximum likelihood-time-frequency analysis proposed in this paper. It consists of a Q-constant filter bank analysis of the time-frequency domain without aliasing effect, combined with a general maximum likelihood analysis, both described in this article, and a partial tracking for polyphonic music with pattern recognition methods (work in progress).

The likelihood analysis is developed in two ways. The first one is temporal, directly treating the samples. The second one is frequential, treating the short time Fourier transform of the signal. For these two approaches, we have tested the robustness to noise and the efficiency of cosine detection.

2. METHODS USED

First, the signal is analysed thanks to a Q-constant filter bank. The 24 quarter-tone F.I.R. filters are accurately calibrated in order to avoid aliasing during the whole analysis. Then, the time-frequency domain obtained out of this filter bank is analysed thanks to a strong statistical tool: the generalized maximum likelihood ratio.

2.1. Q-constant filter bank

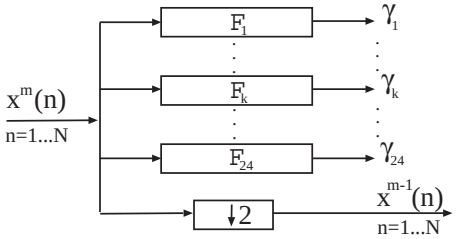

Figure 1: *Structure of the analysis of the input signal $x^m(n)$ at the m^{th} octave: the signal passes through each of the 24 filters of the filter bank, after which γ_k is calculated. Then $x^m(n)$ is decimated, the resulting signal $x^{m-1}(n)$ will be analysed according to the same technique.*

A description of a quarter-tone Q-constant filter bank can be found in [6]. The main idea is to keep the same analysis structure of a signal for every octave (cf. fig.1), while avoiding aliasing. After the analysis of the m^{th} octave, one should make a low-pass filtering to assure having no aliasing on the lower octave. Then,

the $(m-1)^{th}$ octave can be analysed. Time separation for on-sets and off-sets get half accuracy for each lower octave.

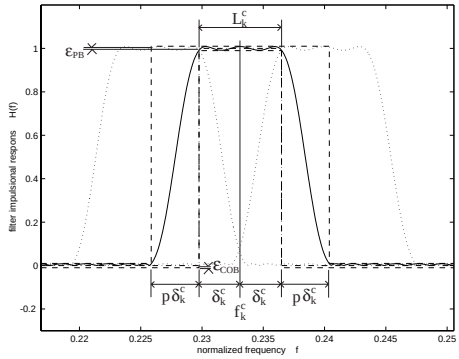

Figure 2: *Full line: first filter of the bank, with highly selective properties ($\varepsilon_{PB} = 0.001$, $\varepsilon_{COB} = 0.01$ and $p = 1$). Each other filter (plotted lines) correspond to this one, after translation to higher frequencies and dilatation. Note that the decreasing band of the filter are smaller than half the pass band width ($p = 1$).*

We define the filter $\mathcal{F}_k$ (cf. fig.2) by its central normalized frequency f_k^c and its bandwidth:

$$L_k^c = 2^{1/48} f_k^c - 2^{-1/48} f_k^c = 2\delta_k^c$$

corresponding to a quarter-tone interval. The maximum deviation for the passing-band ε_{PB}, the maximum deviation for the cut-off band ε_{COB} and the decay bandwidth $d_k^c = p.\delta_k^c$ define the accuracy and the separation power of the filter bank. The value for the quality factor Q is given by $Q = \frac{f_k^c}{L_k^c} \approx 34$ for $p = 1$, which is highly selective.

2.2. Non aliasing conditions

Two conditions are to be respected in order to avoid aliasing. First, the higher frequency analysed should not go over the aliasing limit, which is half the sampling frequency. Each filter been designed according to it central frequency, it means that $f_+ \leq \frac{F_s}{2}$, with the maximum frequency analysed :

$$\begin{aligned} f_+ &= f_{24}^c + (1+p)\delta_{24}^c \\ &= 2^{-1/24}\, f_1^c \left(2 + (2^{1/48} - 2^{-1/48})(1+p)\right) \end{aligned}$$

Note that frequency are normalized, so $F_s = 1$. Reporting this condition to the filter bank calibrating frequency f_1^c, it becomes:

$$f_1^c \leq 2^{-23/24} \left[2 + (2^{1/48} - 2^{-1/48})(1+p)\right]^{-1} \qquad (1)$$

Secondly, when decimating the signal we must apply a low-pass filter (cf. fig. 3). The shorter δ_{24}^c the desired decay bandwidth of the filter $\mathcal{F}_{24}$, the greater n_c the number of coefficients of the low-pass filter. However, we can reduce this constraint by accepting aliasing on the frequency band which has already been analysed, namely $\left[\frac{f_+}{2}; \frac{F_s}{2} - \frac{f_+}{2}\right]$ (that is to say a $(\frac{F_s}{2} - f_+)$ width interval around $\frac{F_s}{4}$). For example with an FIR Hamming filter, the half bandwidth is $\delta_{24}^c \approx \frac{4}{n_c}$; the second condition is:

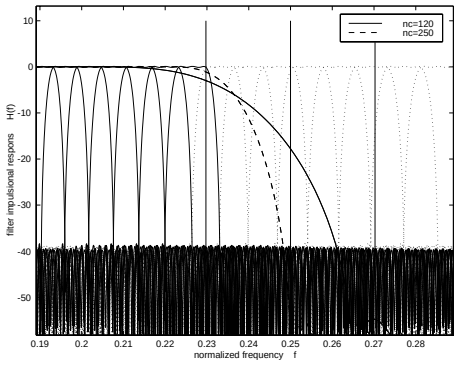

Figure 3: *Low pass filtering with respect of the two non aliasing conditions, for two n_c values: the left filters (full lines) are the upper ones of the $(m-1)^{th}$ octave, the right filters (dashed lines) are the lower of the m^{th} octave, vertical bars represent (from left to right) $f_+/2$ the higher frequency analysed for the $(m-1)^{th}$ octave, $F_s/4$ and $F_s/2 - f_+$ the higher frequency for which aliasing will not affect the analysis of the lower octave. For a $n_c = 120$ coefficients low pass filter (full line), non-aliasing conditions are respected, but the power of the two upper filters are too much modified (a maximum loss of $5\ dB$): this will affect the LiFT detection. To avoid this, it is recommended to use higher values of n_c: for example with 250 coefficients (dashed line: loss of $1\ dB$), only a small part of the last filter is affected.*

$$2\ \delta_{24}^c \leq 2\ \left(\frac{F_s}{4} - f_+\right)$$

Reported to the bank filter calibration frequency f_1^c, to the number of coefficients n_c and to the dilatation coefficient p, we finally obtain the second condition:

$$f_1^c \leq \frac{2^{1/24}\ (F_s/2 - 4/n_c)}{[2 + (2^{1/48} - 2^{-1/48})(1+p)]} \qquad (2)$$

This second condition is stronger than the first one: when it is verified, the first one is also verified.

2.3. Generalized Maximum Likelihood Ratio analysis (GLR)

Let us consider the signal $x(n)$ as the sum of cosines $x_0(n)$ and a white noise $b(n)$ with variance σ^2 (which is supposed to be known).

$$\begin{aligned} x_0(n) &= \sum_j \alpha_j\ \cos\left(2\pi f_{0,j} n + \Phi_j\right) \\ &= \sum_j c_{0,j}\ \cos\left(2\pi f_{0,j} n\right) + s_{0,j}\ \sin\left(2\pi f_{j,0} n\right) \end{aligned}$$

Out of the filter $\mathcal{F}_k$, the signal $y_k(n)$ is the sum of a filtered cosine $y_{0,k}(n)$ defined by(f_0, c_0, s_0) and a filtered noise $b_k(n)$ (we consider that no more than one partial exists in one filter bandwidth), or just a filtered noise $b_k(n)$. This output is analysed statistically over a sliding window (a 512 points window is enough), evaluating the GLR upon two hypothesis for each window.

The first hypothesis H_0 to be tested is that there is only noise in the signal out of the filter. The second hypothesis H_1 is that there is a sine wave plus noise. Under each of both hypotheses, we perform the maximum of the probability density function, and give the GLR of H_1 against H_0, defined as:

$$\Gamma = \frac{\max_{\theta \in H_1} P_{H_1}}{\max_{\theta \in H_1} P_{H_0}} \tag{3}$$

Since Γ varies exponentially, we also give $\gamma = \log \Gamma$ the log-likelihood function [7]. We used two approaches: the first one deduced from the time representation and the second one from the frequency representation of the output signal.

2.4. First approach: temporal representation

Let us consider N observations of the output signal $y_k(n)$, $n = 0, ..., N-1$. We note $y = (y_k(0), \ldots, y_k(N-1))^T \in \mathbb{R}^N$ and $\theta = (c_0, s_0)^T \in \mathbb{R}^2$ the cosine amplitude vector. Under the hypothesis H_0, y is gaussian with $\theta = \theta_0 = (0,0)^T$. Under the hypothesis H_1, y is gaussian with $\theta \neq (0,0)^T$. H_0 is simple because the hypothesis is completely described by the value of θ, whereas H_1 is not.

The density probability is:

$$P_{H_i}(y; \theta, \sigma^2) = \frac{\exp\left(-\frac{1}{2} \varepsilon^H \, {R_b}^{-1} \, \varepsilon\right)}{(2\pi)^{n/2} \det(R_b)^{1/2}}$$

with R_b the correlation matrix and $\varepsilon(f_0, \theta) = y - D(f_0)\theta$. The covariance matrix R_b cannot be inversed since it is badly conditioned. A solution is to consider that the noise out of the filter is still white (in the bandwidth concerned). In that case, $R_b \approx \frac{\sigma^2}{2} I_N$ in the bandwidth concerned.

The density probability becomes:

$$P_{H_i}(y; \theta, \sigma^2) = \frac{\exp\left(-\frac{1}{\sigma^2} \varepsilon^H \, \varepsilon\right)}{\left((2\pi)^n \, N \frac{\sigma^2}{2}\right)^{1/2}}$$

The generalized likelihood ratio is defined by:

$$\Gamma_T = \frac{\max_{\theta \in H_1} P_{H_1}(y; \theta, \sigma^2)}{P_{H_0}(y; \theta_0, \sigma^2)} \tag{4}$$

The value $\overline{\theta}$ that maximize P_{H_1} is obtained by derivating this probability density function:

$$\begin{aligned} \overline{\theta} &= \arg\max_{\theta \in H_1} P_{H_1}(y; \theta, \sigma^2) \\ &= [D^H(\overline{f}_0) \, D(\overline{f_0})]^{-1} \, D^H(\overline{f}_0) \, y \\ &\approx \frac{2}{N} D^H(\overline{f}_0) \, y \end{aligned}$$

where $D(f_0)$ is the cosine-sine matrix:

$$D(f_0) = \begin{pmatrix} \ldots & \cos(2\pi \, k \, f_0) & \ldots \\ \ldots & \sin(2\pi \, k \, f_0) & \ldots \end{pmatrix}^T$$

with $k = 0, \ldots, N-1$ and $\overline{f_0}$ the frequency (in the filter bandwidth) that minimize the mean-square error between the signal and the model $\|\varepsilon(f_0, \theta)\|_2^2$:

$$\overline{f_0} = \arg\max_{f_0} y^H \, D(f_0) \, D^H(f_0) \, y$$

We now know the cosine amplitude vector $\overline{\theta} = (\overline{c_0}, \overline{s_0})^T$ and the frequency $\overline{f_0}$ of the cosine. The log-likelihood γ_T is approximated by:

$$\gamma_T \approx \frac{1}{\sigma^2} \, y^H \, D(\overline{f_0}) \, D^H(\overline{f}_0) \, y \tag{5}$$

According to its value, we can decided either the hypothesis of the presence of a cosine is true or not.

2.5. Frequency representation

In the frequency representation, we know the joint probability density function of the real part α and the imaginary part β of the STFT (short time Fourier Transform). It can be expressed as a function of the modulus ρ and the phase ω. By integrated according to the phase (a 0-order modified Bessel function), we obtain the probability density function of the modulus. We note the STFT:

$$\begin{aligned} \hat{y}(f_i) &= \frac{1}{N} \sum_{n=0}^{N-1} y(n) \, \exp\left(-2j\pi n \frac{i}{N}\right) \\ &= \alpha + j\,\beta \; = \; \rho \, e^{j\omega} \end{aligned}$$

where α and β are gaussian, with their respective means $m \cos\Phi$, $m \sin\Phi$ (m being the modulus of the sinusoidal component and Φ its phase) and their respective variance ${\sigma_A}^2$ and ${\sigma_B}^2$. The covariance matrix is noted $C_2(f_0)$ and the error is $\varepsilon = \begin{pmatrix} \alpha - m \, \cos\Phi \\ \beta - m \, \sin\Phi \end{pmatrix}$. The joint probability density function is:

$$P_{H_i}(\alpha, \beta; m, \Phi, \sigma^2) = \frac{\exp\left(-\frac{1}{2} \, \varepsilon^T \, {C_2(f_0)}^{-1} \, \varepsilon\right)}{2\,\pi\sqrt{\det C_2(f_0)}}$$

since $\sigma_A \, \sigma_B \, I_2 \approx \sigma^2 \, I_2$ for frequencies in the filter bandwidth. Under the hypothesis H_0, α and β are gaussian with $m = m_0 = 0$. Under the hypothesis H_1, α and β are gaussian with $m \neq m_0$.

The covariance matrix $C_2(f_0)$ can be approached by $\frac{\sigma^2}{2} I_2$ for big enough data vectors (typically, 256 or 512 elements); it gives:

$$P_{H_i}(\alpha, \beta; m, \Phi, \sigma^2) \approx \frac{\sqrt{2}}{2\,\pi\,\sigma} \, \exp\left[-\frac{1}{\sigma^2} \left(\rho^2 + m^2 - 2\,\rho\, m \cos(\omega - \Phi)\right)\right]$$

The relation between the joint probability density function for (ρ, ω) and the joint probability density function for (α, β) is given by $P_{H_i}(\rho, \omega; m, \Phi, \sigma^2) = \rho \, P_{H_i}(\alpha, \beta; m, \Phi, \sigma^2)$. We obtain the modulus probability density function by integrating the joint probability function of (ρ, ω) following the phase ω:

$$\begin{aligned} P_{H_i}(\rho; m, \Phi, \sigma^2) &= \int_0^{2\pi} P_{H_i}(\rho, \omega; m, \Phi, \sigma^2) \, d\omega \\ &= \frac{\sqrt{2}\,\rho}{2\,\pi\,\sigma^2} \exp\left(-\frac{\rho^2 + m^2}{\sigma^2}\right) I_0\left(\frac{2\,\rho\, m}{\sigma^2}\right) \end{aligned}$$

Reminding the generalized maximum likelihood ratio:

$$\Gamma_F = \frac{\max_{\theta \in H_1} P_{H_1}(\rho; m, \Phi, \sigma^2)}{\max_{\theta \in H_0} P_{H_0}(\rho; m, \Phi, \sigma^2)}$$

and given done $\overline{\rho} = \max_{f_0 \in [f_k^c - \delta_k^c ; f_k^c + \delta_k^c]} \rho$ the maximum estimated of the sinusoidal part modulus in the filter bandwidth at the frequency $\overline{f_0} = \arg\max_{f_0 \in [f_k^c - \delta_k^c ; f_k^c + \delta_k^c]} \rho$, we finally obtain:

$$\Gamma_F \quad \approx \quad \exp\left(-\frac{m^2}{\sigma^2}\right) \; I_0\left(2\,\frac{\overline{\rho}\,m}{\sigma^2}\right) \tag{6}$$

with the 0-order modified Bessel function:

$$I_0(x) = \int_0^{\pi} \frac{\exp(x\ \cos\ \omega)}{\pi}\, d\,\omega$$

In order to know the value of $\overline{m}$, we have an implicit equation to solve:

$$\overline{\rho}\, I_1\left(2\,\frac{\overline{\rho}\,m}{\sigma^2}\right) - m\ I_0\left(2\frac{\overline{\rho}\,m}{\sigma^2}\right) = 0 \tag{7}$$

It is easily done by tabuling (cf. fig.4).

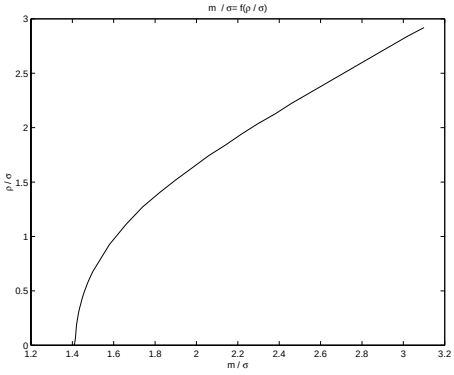

Figure 4: *Value of $\frac{\overline{m}}{\sigma}$, solution of the implicit equation for small values of $\frac{\overline{\rho}}{\sigma}$.*

We now know the cosine amplitude modulus $\overline{\rho}$ and the frequency $\overline{f_0}$ of the cosine (we also know the phase $\overline{\omega}$ from the maximization of ρ). For great values of $\overline{\rho}$, the generalized maximum likelihood ratio is no more calculable, since it increases exponentially. We calculate the log-likelihood function $\gamma_F = \log \Gamma_F$, and approximate it thanks to its series development (cf. [8]):

$$\gamma_F \approx \frac{\overline{m}^2}{\sigma^2} - \frac{1}{2}\log\left(4\,\pi\,\frac{\overline{m}^2}{\sigma^2}\right) \tag{8}$$

For small values of $\overline{\rho}$, we use the general form:

$$\gamma_F \approx -\frac{\overline{m}^2}{\sigma^2} + \log\left[I_0\left(2\,\frac{\overline{\rho}\,\overline{m}}{\sigma^2}\right)\right] \tag{9}$$

3. RESULTS

3.1. Re-synthesis

Since we know precisely (ρ, Φ, f_0) for the cosine $y_{0,k}$ existing out of the filter, we can calculate a re-synthesized signal y_r. When a cosine exists in a filter bandwidth (cf. fig.5), we obtain the same signal with a very small relative error $\varepsilon = \frac{\|y_{0,k} - y_r\|_2}{\|y_{0,k}\|_2}$ (around $1\ \%$). When only noise exist in the filtered signal, we synthesize a cosine with a very small modulus, but when taking into account the

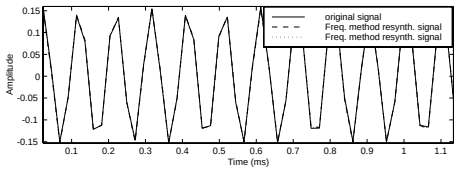

Figure 5: *Re-synthesis of a cosine detected out of a filter. For a cosine detected, both methods give the same curve.*

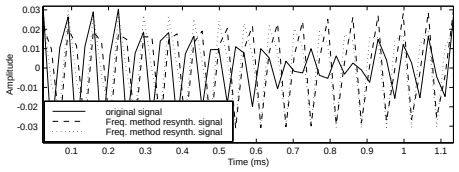

Figure 6: *Re-synthesis of a signal considered as noise after maximum likelihood estimation.*

threshold on γ_T and γ_F,, the decision is taken that it corresponds to noise.

A precise re-synthesis can be done for partials, but the LiFT method is not developed to re-synthesize the residual part of a signal.

3.2. Likelihood-time-frequency smoothing

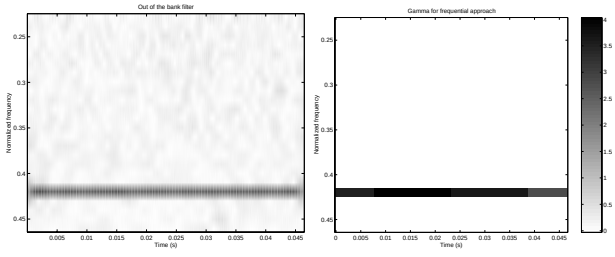

Figure 7: *Time frequency domain of a signal with one cosine plus noise ($SNR = 5$) after the bank filtering (left figure), and with frequency approach (right figure, γ_F). Time is plotted among X and normalized frequency among Y. The noise does not appear anymore and the output of the filter bank is smoothed.*

We represented on figure 7 the output of the bank filter for just one cosine plus noise with a $5\,dB$ signal-to-noise ratio (left figure) and the LiFT analysis (right figure). The filter bank used was a low quality filter bank, with $\varepsilon_{EP} = 0.01$, $\varepsilon_{COB} = 0.01$ and $p = 1.5$. The smoothing effect of our analysis immediately appears: the right figure clearly indicates where a cosine exists, even when there is noise around.

3.3. Robustness to noise

We calculated the log-likelihood functions γ_T and γ_F with given signal-to-noise ratios SNR, and for different estimated SNR in the γ computation (cf. fig.8).

This figure depicts that for sinus emerging from noise, both methods will easily detect its presence if $SNR > 0.$ in the sound,

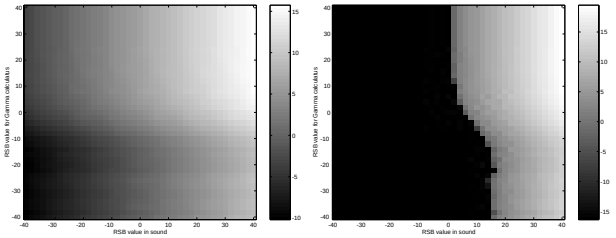

Figure 8: γ *for temporal method (left figure) and frequency method (right figure), for SNR values in* $[-40; 40]$*. Gamma* (z*, grey level) is calculated for both methods with SNR varying* (y)*, with a sound synthesized with different values of* $\sigma^2 = 10^{-SNR/10}$ (x)*.*

even with a bad estimated value of the SNR. However, for the frequency method, if $SNR < 0$ in the sound, nothing will be detected. The temporal method seems more performant to detect low level partials, but needs much more computational time.

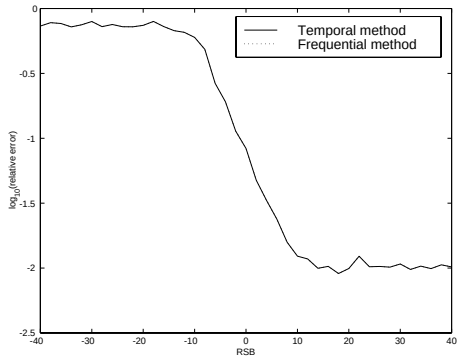

Figure 9: $\log_{10}$ *of the relative error between the* $\mathcal{F}_k$ *output signal and the re-synthesized signal by both methods. For* $SNR > 0$*, the relative error is around* $1\ \%$.

We calculated the relative error between the analysed sound and the synthesized sound, for different SNR values (cf. fig.9). Both methods give the same results: for noisy sounds, the relative error is great (around $10\ \%$ up to $SNR = -10$), and goes down around $1\ \%$ for $SNR > 10$.

3.4. Cosine detection

We compared the cosine detection for the two approaches. For this, we constructed a synthetic signal with one cosine and white noise, and a signal-to-noise ratio fixed to 1. γ is calculated for both methods, with or without frequency refining (cf. fig.10). The detection are similar with both approaches, and have their maximum value for the true frequency f_0 of the given cosine. This means that with frequency refining, we can determine very precisely the frequency of the cosine, if needed for the forthcoming polyphonic pitch analysis.

4. CONCLUSIONS

The likelihood-time-frequency analysis proposed is a good tool for multi-partial detection. The time-frequency domain is represented with the maximum likelihood statistical approach, which is very consistent to noise. The time methods is more accurate than the frequency method for low level partials, but needs much more computational time. With strong selectivity for the Q-constant filters, the LiFT analysis allows a good quality re-synthesis of the sinusoidal part of the signal, and even with lower selectivity, it gives a good representation of the equivalent smoothed sonagram. The good detection power and the robustness to noise promise the feasibility of a strong polyphonic automatic transcription tool, based on this analysis.

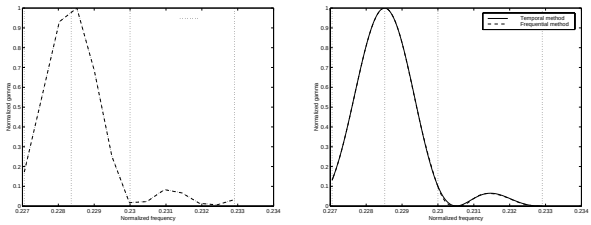

Figure 10: *Detection power for both methods: we represent normalized* γ *versus normalized frequency, i) with 200 frequency points refining (right figure), and ii) without refining and with the cosine frequency between two frequency bins (left figure). Dashed lines represents, from left to right, the minimum analysed frequency, the cosine frequency, the filter central frequency and the maximum analysed frequency.*

5. REFERENCES

[1] M. R. Schroeder, "Period histogram and product spectrum: new methods for fundamental-frequency measurement" Journal of the Acoustical Society of America, vol. 43, pp. 829-834 1968.

[2] E. Terhardt, "Calculating virtual pitch", Hearing Research 1, 155-182, 1979.

[3] B. Doval, X. Rodet, "Fundamental Frequency Estimation and Tracking using Maximum Likelihood Harmonic Matching and HMM's", Proc. IEEE- ICASSP, pp. 221-224, 1993.

[4] A. M. Noll, "Cepstrum Pitch Determination" Journal of the Acoustical Society of America, vol. 41, no. 2, pp. 293-309 1967.

[5] S. Dixon, "Multiphonic Note Identification", Proceedings of the 19th Australasian Computer Science, Melbourne, Australia, January 31 - February 2 1996.

[6] J. C. Brown, "Calculation of a Constant Q Spectral Transform", Journal of the Acoustic Society of America, vol. 89, no. 1, 425-434, 1991.

[7] M. B. Priestley, "Spectral Analysis and Time Series", Academic Press, 305-307, 1981.

[8] I. S. Gradshteyn, I. M. Pyzhik, "Table of Integrals, Series and Products", Academic Press, 1980.