Gabor frames for the analysis of impact sounds using ESPRIT

Adrien Sirdey[1], Olivier Derrien[1], Richard Kronland-Martinet[1], and Mitsuko Aramaki[1]

[1]*Laboratoire de Mécanique et d'Acoustique, CNRS, Marseille, France*

Correspondence should be addressed to Adrien Sirdey (sirdey<at>lma.cnrs-mrs.fr)

ABSTRACT

This articles tackles the estimation of mode parameters in recorded impact sounds obtained by hitting resonant objects. It is shown in this article that the ESPRIT algorithm can be efficiently applied on time-frequency representations of the signal computed using Gabor frames. An experimental study on artificial signals has been conducted in order to highlight the advantages of such an approach, and to compare the performances of ESPRIT and Steiglitz-McBride-based estimation algorithm. A real case analysis situation over a 341 impact sounds database is also discussed.

1. INTRODUCTION

The context of this study is the identification of acoustical modes which characterize a resonant object. This is of great use when building an environmental sound synthesizer (see [1] or [2] for an insight on such synthesizers). Practically, the analysis is made from recorded impact sounds, where the resonant object is hit by another solid object (e.g. a hammer). Assuming that the impact sound is approximately the acoustical impulse response of the resonant object, and under the assumption of small perturbations and linear elasticity, each mode corresponds to an exponentially damped sinusoid (EDS). The modal analysis thus consists of estimating the parameters of each sinusoidal component (amplitude, phase, frequency and damping). These parameters will be stored, and eventually modified, before further re-synthesis. In this paper, only the analysis part will be considered.

In the past decades, significant advances have been made in the field of system identification, especially for estimating EDS parameters in a background noise. The involved analysis methods can be divided in three categories: short-time Fourier-like transforms methods [3], resonant filter methods [4, 5] relying on the Steiglitz-McBride algorithm [6], and high resolution methods like Prony [7], MUSIC [8] or ESPRIT [9]. Among these three high-resolution methods, ESPRIT has proven to have the better efficiency [10, 11, 12], as well as offering a convenient direct estimation of the poles, whereas MUSIC requires a scanning of all possible solutions. This paper will therefore only consider the ESPRIT method in the high-resolution family. In order to reduce the number of components to estimate and the computational cost, a prior sub-band decomposition has already been presented in [13] and [14], and has also been shown to improve the estimation quality. This can also be combined with a prior segmentation of the original signal in the time domain as shown in [7]. The estimation of the model order (i.e. the number of modes) is an important issue. Various methods have been proposed for automatic estimation of the order, e.g. ESTER [15], or [16] which relies on angle measures between subspaces. However, in practical situations this parameter is often deliberately over-estimated.

In this paper, the ESPRIT algorithm is applied on a time-frequency representation of the original sound. The time-frequency representation is here computed with a *Gabor transform* (GT). Under certain conditions, the *inverse Gabor transform* provides a perfect reconstruction of the signal, although only the analysis part will be considered in this paper. The Gabor transform is equivalent to a filter-bank, but offers a convenient formalism through the concept of *Gabor frame*, which allows a straightforward time subsampling and sub-band division of the time-frequency plane. The size of each sub-band and the subsampling parameter depend on the choice of the frame. Contrarily to other filter banks designed for ESPRIT in the past, the objective here is not to minimise the overlap between adjacent channels. As a matter of fact, impact sounds can have short durations and there-

fore require a good resolution in time. For this purpose, the energy spreading properties of the Gabor transform as well as the flexibility offered by the concept of frame is of great use. The transform being linear, an EDS in the original sound is still an EDS inside each frequency channel; ESPRIT can therefore be applied in each of these channels, and a straightforward relation exists between the parameters in the time-frequency domain and the sought parameters in the original time domain. As opposed to the sub-band approach presented in [13], the decimation here is not critical. Although the noise is no longer white in such cases, it is shown through numerical experiments that better estimations can be achieved than with critical sub-sampling. A method relying on psychoacoustical considerations to discard insignificant modes *a posteriori* is also proposed.

The paper is organised as follows: first, a brief state-of-the-art covers the signal model, the ESPRIT algorithm and the Gabor transform. Then, it is shown that original EDS parameters can be recovered by applying the ESPRIT algorithm in each frequency channel of the Gabor transform. The next part describes numerical tests that have been conducted in order to study the behaviour of ESPRIT compared to the Steiglitz-McBride algorithm. Then, the results of a database analysis (consisting of 341 impact sounds) with different methods are presented. Further possible improvements are finally discussed.

2. STATE OF THE ART

2.1. The signal model and the ESPRIT algorithm

It is considered that the discrete signal to be analysed can be correctly modelled by:

$$x[l] = s[l] + w[l] \tag{1}$$

where the deterministic part $s[l]$ is a sum of K damped sinusoids:

$$s[l] = \sum_{k=0}^{K-1} \alpha_k z_k^l. \tag{2}$$

Here the complex amplitudes are defined as $\alpha_k = a_k\, e^{i\phi_k}$ (containing the initial amplitude a_k and the phase ϕ_k), and the poles are defined as $z_k = e^{-d_k + 2i\pi\nu_k}$ (containing the damping d_k and the frequency ν_k). The stochastic part $w[l]$ is a central gaussian white noise of variance σ^2.

The ESPRIT algorithm was originally described by Roy *et. al.* [9]. The principle consists in performing a singular value decomposition (SVD) on an estimate of the signal correlation matrix. The eigenvectors corresponding to the K highest eigenvalues generate the so called *signal subspace*, while the remaining vectors generate the so called *noise subspace*. The shift invariance property of the signal subspace, which basically means that for any component s_k and at all times l one has $s_k(l+1)/s_k(l) = z_k$, allows a simple solution for estimating the optimal poles values z_k. Then, the amplitudes α_k can be recovered by solving a least square problem. The algorithm can be described briefly as follows:

The signal vector is defined as:

$$\boldsymbol{x} = \left[\begin{array}{cccc} x[0] & x[1] & \dots & x[L-1] \end{array}\right]^T, \tag{3}$$

where L is the length of the signal to be analysed. The Hankel signal matrix is defined as:

$$\boldsymbol{X} = \left[\begin{array}{cccc} x[0] & x[1] & \dots & x[Q-1] \\ x[1] & x[2] & \dots & x[Q] \\ \vdots & \vdots & & \vdots \\ x[R-1] & x[R] & \dots & x[L-1], \end{array}\right] \tag{4}$$

where $Q, R > K$ and $Q + R - 1 = L$. The amplitude vector is defined as:

$$\boldsymbol{\alpha} = \left[\begin{array}{cccc} \alpha_0 & \alpha_1 & \dots & \alpha_{K-1} \end{array}\right]^T, \tag{5}$$

and the Vandermonde matrix of the poles:

$$\boldsymbol{Z}^L = \left[\begin{array}{cccc} 1 & 1 & \dots & 1 \\ z_0 & z_1 & \cdots & z_{K-1} \\ \vdots & \vdots & \vdots & \vdots \\ z_0^{L-1} & z_1^{L-1} & \cdots & z_{K-1}^{L-1} \end{array}\right]. \tag{6}$$

Performing a SVD on $\boldsymbol{X}$ leads to:

$$\boldsymbol{X} = [\boldsymbol{U}_1 \boldsymbol{U}_2] \left[\begin{array}{cc} \boldsymbol{\Sigma}_1 & 0 \\ 0 & \boldsymbol{\Sigma}_2 \end{array}\right] \left[\begin{array}{c} \boldsymbol{V}_1 \\ \boldsymbol{V}_2 \end{array}\right], \tag{7}$$

where $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are diagonal matrices containing respectively the K largest singular values, and the smallest singular values; $[\boldsymbol{U_1 U_2}]$ and $[\boldsymbol{V_1 V_2}]$ are respectively the corresponding left and right singular vectors. Using the shift-invariance property of the signal subspace, it can be proved that the eigenvalues of the matrices $\boldsymbol{\Phi}_1$ and $\boldsymbol{\Phi}_2$ defined such that:

$$\boldsymbol{U}_1^{\downarrow} \boldsymbol{\Phi}_1 = \boldsymbol{U}_1^{\uparrow} \quad \text{and} \quad \boldsymbol{V}_1^{\downarrow} \boldsymbol{\Phi}_2 = \boldsymbol{V}_1^{\uparrow} \tag{8}$$

provide estimation of the poles z_k. $(.)^{\uparrow}$ and $(.)^{\downarrow}$ respectively stand for the operators discarding the first line and the last line of a matrix. Thus, z_k can be estimated by diagonalizing the matrix $\mathbf{\Phi}_1$ or $\mathbf{\Phi}_2$. The associated Vandermonde matrix $\boldsymbol{Z}^L$ is then computed, and the optimal amplitudes with respect to the least square criterion are finally obtained by:

$$\boldsymbol{\alpha} = (\boldsymbol{Z}^L)^{\dagger}\boldsymbol{x}, \tag{9}$$

where $(.)^{\dagger}$ denotes the pseudoinverse operator.

2.2. The Gabor Transform

The Gabor transform allows the expression of $x[l]$ in a given Gabor frame. A Gabor frame $\{g,a,M\}$ is characterised by a window g, a time-step parameter a, and a number of frequency channels M. The expression $\chi[m,n]$ of $x[l]$ in the Gabor frame $\{g,a,M\}$ is written:

$$\chi[m,n] = \sum_{l=0}^{L-1} \overline{g}[l-an]x[l]\, e^{-2i\pi l\frac{m}{M}}, \tag{10}$$

where $\overline{(.)}$ denotes the complex conjugate. m is a discrete frequency index and n a discrete time-index. One can see that this corresponds to a discretised version of the standard short-time Fourier transform. This transform can generally be inverted (for more details, see for instance [17]). The signal $\chi[m,n]$ for a fixed index m can be seen as a sub-sampled and band-pass filtered version of the signal $x[l]$. As the sub-sampling reduces the length of the data by a factor a, the ESPRIT algorithm can be applied to each frequency channel in order to analyse longer signals.

3. ESPRIT IN A GABOR FRAME

This section covers the application of the ESPRIT algorithm to a single channel in a Gabor frame. The analysed signals are therefore composed of the GT coefficients at a given frequency index m. As the GT is linear, the contribution of the deterministic part $s[l]$ can be separated from the contribution of the noise $w[l]$.

3.1. Deterministic part

$c[m,n]$ denotes the GT of $s[l]$ in channel m and time index n, whereas $c_k[m,n]$ denotes the GT of the signal z_k^l associated to the pole z_k:

$$c_k[m,n] = \sum_{l=0}^{L-1} \overline{g}[l-an]z_k^l\, e^{-2i\pi l\frac{m}{M}}. \tag{11}$$

According to the signal model (2), it can be easily proved that:

$$c[m,n] = \sum_{k=0}^{K-1} \tilde{\alpha}_{k,m}\tilde{z}_{k,m}^n, \tag{12}$$

where the apparent pole $\tilde{z}_{k,m}$ can be written as:

$$\tilde{z}_{k,m} = z_k^a\, e^{-2i\pi a\frac{m}{M}}, \tag{13}$$

and the apparent amplitude:

$$\tilde{\alpha}_{k,m} = \alpha_k\, c_k[m,0]. \tag{14}$$

In other words, the deterministic part of the signal in each channel is still a sum of exponentially damped sinusoids. However, their poles and amplitudes are modified according to the Gabor frame time-step and frequency parameters.

3.2. Stochastic part

It has been proved that the Gabor transform of a gaussian noise is a 2D complex gaussian noise [18]. However, the noise is white only if the decimation factor is critical ($M/a = 2$). When it is not, the noise in each channel is a white noise filtered by a low-pass filter (see appendix 8.1). Resulting noise power spectrum are displayed Fig.6 for different values of M/a. When $M/a \neq 2$, the whiteness hypothesis under which the ESPRIT method is valid is no longer fulfilled. In spite of that, experimental results indicate that pole estimations are still valid, and achieved with an even better resolution than when the decimation is critical. In order to explain this, two facts can be highlighted:

1. The more samples are considered when the signal is significantly above the noise level, the better the estimation. This condition is favoured for low a values (see Fig. 5).
2. The noise level in the neighbourhood of the component to be analysed has more influence on the estimation error than the overall noise spectral shape (see Fig. 7), on which depends the whiteness condition.

The corresponding experiments are deeply commented in section 4.

3.3. Recovering the signal parameters

Assuming that the signal model is still valid, or that its divergence from the theoretical model is negligible, it is

reasonable to apply ESPRIT on $c[m,n]$. $\boldsymbol{c}_m$ denotes the vector of GT coefficients in the channel m and $\boldsymbol{S}_m$ the Hankel matrix built from $c[m,n]$. Applying the ESPRIT algorithm to $\boldsymbol{S}_m$ leads to the estimation of the apparent poles $\tilde{z}_{k,m}$. Inverting equation (13) leads to:

$$z_k = e^{2i\pi\frac{m}{M}}(\tilde{z}_{k,m})^{\frac{1}{a}}. \tag{15}$$

Because of the sub-sampling introduced by the GT, it can be seen from equation (13) that aliasing will occur when the frequency of a pole is outside the interval $\left[\frac{m}{M}-\frac{1}{2a},\frac{m}{M}+\frac{1}{2a}\right]$. To avoid aliasing, a and and the analysis window $g[l]$ must be such that the bandwidth of $g[l]$ is smaller than $\frac{1}{a}$. That way, the possible aliasing components will be attenuated by the band-pass effect of the Gabor transform.

Denoting $\tilde{\boldsymbol{Z}}_m^N$ the Vandermonde matrix of the apparent poles $\tilde{z}_{k,m}$ (N is the time-length of signal $c[m,n]$), the least square method for estimating the amplitudes leads to:

$$\boldsymbol{\alpha} = \frac{(\tilde{\boldsymbol{Z}}_m^N)^\dagger \boldsymbol{c}_m}{c_k[m,0]}. \tag{16}$$

Without noise, according to equation (12), every EDS should be detected in each channel, which generates multiple estimations of the same modes. Theoretically, the model order should be set to K in each channel. However, this is usually a large over-estimation. Because each channel of the GT behaves like a band-pass filter, an EDS with a frequency far from $\frac{m}{M}$ will be attenuated and buried in the noise. Thus practically, the optimal model order has to be chosen for each channel on which the analysis is performed. It can be determined using the ESTER criterion [15], or deliberately over-estimated.

3.4. Choice of the analysis channels

All channels of the Gabor transform are not likely to contain energy relative to a vibration mode. In order to diminish the computational time as well as the number of components obtained in the end of the analysis, it is useful only to perform the analysis on channels in which deterministic energy is likely to be found. This choice can be made "by hand", observing the Fourier transform or the spectrogram. Or, assuming that the deterministic energy is clearly above the noise level, the channels containing one ore several partials will exhibit an energy peak compared to their neighbours. A peak detection algorithm can therefore be applied on the energy of the Gabor transform channels (see Fig.2-b).

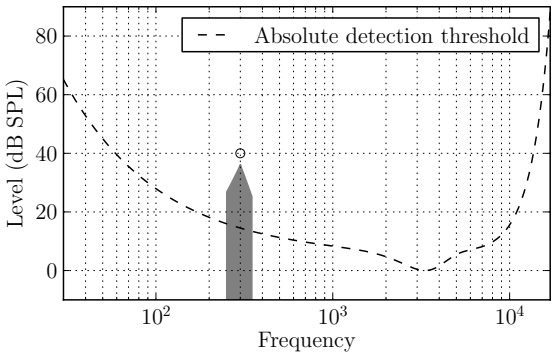

Fig. 1: Illustration of the masking phenomenon considered in the analysis protocol. The dashed line models the absolute detection level of the human hearing, and the grey area correponds to the masking domain of the component represented by a white dot.

3.5. Discarding multiple components

If the distance between a set of channels on which an analysis has been performed is smaller than the bandwidth of the analysis window $g[l]$, the same components are likely to appear in all of these channels. These multiple estimations of the same component (hereafter named replicas) have to be identified. The only one that will be kept for the final re-synthesis is the one which frequency is the closest to the central frequency of the channel where it has been detected, for this is where the signal-to-noise ratio is optimal and therefore where the estimation is likely to have the lowest error. A component c_r (with frequency f_r) is considered a replica of a component c_o (with frequency f_o) if the following conditions are fulfilled:

$$|f_r - f_o| < \varepsilon_f \tag{17}$$
$$|f_r - f_o| < |f_o - f_i| \tag{18}$$

Here ε_f is a frequency confidence interval and f_i is the closest frequency to f_r among the components detected in the same channel as c_r.

3.6. Discarding irrelevant components

Practical tests have shown that some of the modes detected using the previously described approach are not relevant for they have an insignificant energy. What's more, the only components of interest for the given pur-

pose are the one that can be heard by a potential listener. It is therefore possible to rely on psychoacoustical considerations to discard components. To do so, two phenomenons can be taken into account : the absolute detection threshold of the human hearing system, and the masking of a component by another. For stationary sounds, these masking processes have been deeply described, and convincingly approximated by analytical laws exposed in the following sections 3.6.1 and 3.6.2. For non-stationary sounds such as damped sinusoids, no straightforward masking description has been proposed. However they can be considered as a succession of stationary processes, on which the pre-mentioned masking phenomenon can be applied. This last point is considered in section 3.6.3.

3.6.1. The absolute masking threshold

The absolute threshold level in dB can be estimated (see [19]) by :

$$3.64 f_{\mathrm{kHz}}^{-0.8} - 6.5 e^{-0.6(f_{\mathrm{kHz}}-3.3)^2} + 10^{-3} f_{\mathrm{kHz}}^4 \qquad (19)$$

Where f_{kHz} is the frequency in kHz. When dealing with recorded sounds, the output level is *a priori* unknown. Therefore the function is usually modified so that its minimum matches the minimum encodable value in the considered audio format (for instance $20\log_{10}(1) = 0$, in wav format encoded as 16 bits integers). The corresponding function is plotted Fig.1.

3.6.2. Masking of a component by another

This phenomenon occurs when a sinusoid (the *masker*) reduces the perceived loudness of another component of smaller amplitude (the *masked*) to the point that it becomes undetectable by the human auditory system. In [20], the masking threshold generated by a masker sinusoid is expressed as a function of its amplitude and frequency. It is modelled by two slopes which are linear when the energy is in dB and the frequency in barks: a 18 dB/bark slope for lower frequencies, and a -22 dB/bark for higher frequencies. A masking offset of -4 dB from the masker amplitude is also assumed. Finally, the masking domain in frequency corresponds to the critical band associated to the masker (see Fig.1).

3.6.3. Masking between damped sinusoids

A damped sinusoid can be considered as a sinusoid which amplitude varies continuously with time. Under the assumption that the variation is small enough, it can therefore be approximated by a succession of sinusoids which amplitude is constant over a given duration Δ_t. Doing so, one approximates the EDS model into another one which is compatible with the aforementioned masking phenomenons. For all constant-amplitude portions of the approximating signal, is is firstly determined for each component whether it is above or below the absolute detection threshold ; secondly the masking domain of each component is computed. After each computation, other components which may fall into the masking domain are labelled. Once this has been done for each signal portion, the components which are always below the detection threshold are discarded, as well as the components which are labelled as masked for every portion in which they are above the detection threshold.

Is is legitimate to wonder about the pertinence of such an approximation of masking processes. Unfortunately, no psychoacoustical model describing masking between damped sinusoids has been proposed yet. Two sensible points can be highlighted : first, the one-to-one (masker-masked) relation between components considered here is a very simplified scenario. In reality, a single mask can be formed by a masking effects addition of several different components (see [21] for a description of this phenomenon); here the additivity of masking effects has not been considered. Secondly, the constant amplitude approximation in the time domain is equivalent to a peak approximation in the frequency domain. In reality the components have a wider bandwidth due to their damping (see Fig.3.4 in [22] for a clear illustration of masking pattern widening). A masker component is therefore likely to have a wider masking range, and a masked component is likely to stick out of a hypothetical masking domain wherein its peak approximation is confined.

3.7. Illustration

As an application, one can observe Fig.2-a the spectrogram of sound that has been analysed with the aforementioned method. The ESPRIT algorithm has been applied on each of its Gabor transform channels determined by a peak detection algorithm (see section 3.4) and displayed as white dots Fig.2-b. The analysis order was over-estimated to 6 in each of these channels. This led to 780 components, of which 176 presented a negative damping, and had to be discarded before resynthesis in order to avoid diverging signals. The resynthesis spectrogram is displayed Fig.2-c. The spectrogram Fig.2-d corresponds to the resynthesis after discarding the irrelevant components as described in section 3.6.

The corresponding sounds can be listened to at **??**. One can notice the perceptual quality of the resynthesis, as

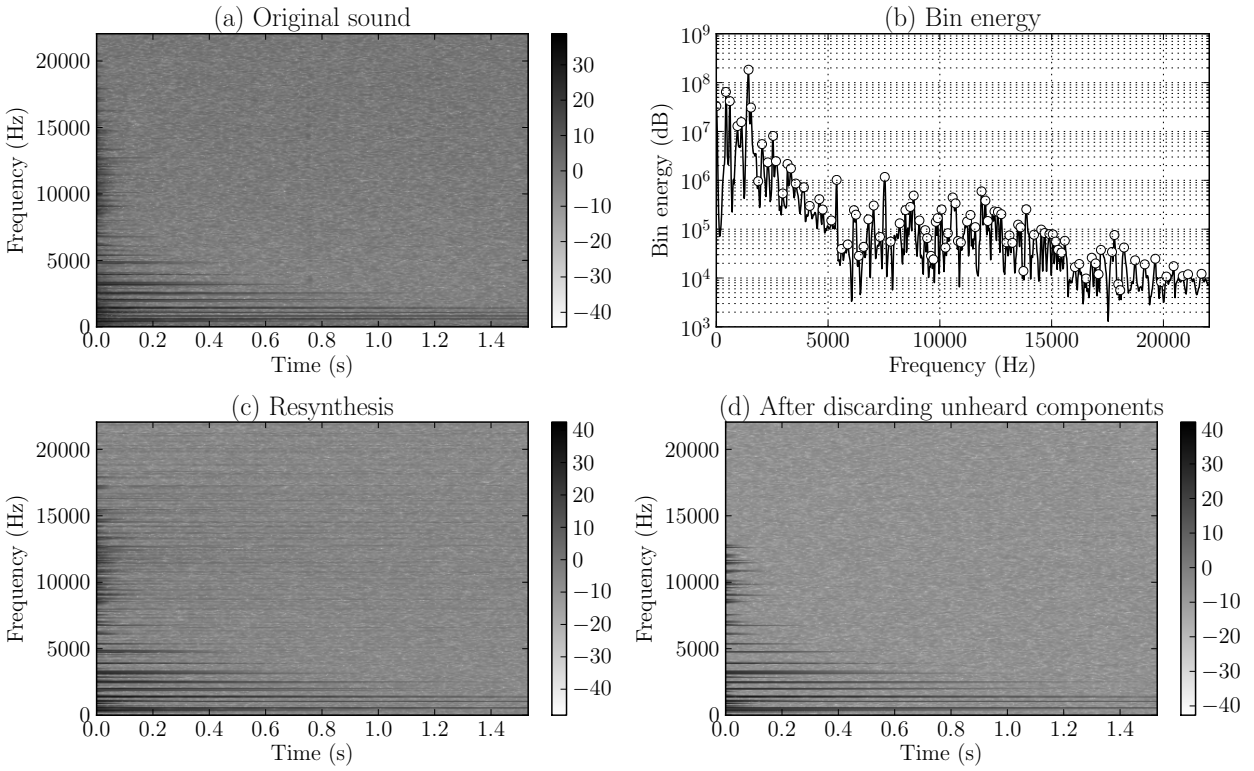

Fig. 2: Analysis of a sound which spectrogram is displayed on the plot *a*. The ESPRIT algorithm is applied on each of the bin displayed as white dots plot *b*, with an order forced to 6. The corresponding resynthesis containg 604 components is displayed plot *c*. After discarding unheard components as described in section 3.6, only 57 components remain. The corresponding resynthesis spectrogram is displayed on plot *d*.

well as the low impact that the psychoacoustical discarding procedure has on the final rendering; as a matter of fact, no difference at all can be noticed. The noisy background of both resynthesis spectrograms is only a quantification artefact due to the 16 bits integers data format that was used.

4. NUMERICAL TESTS

Here are described some numerical tests that have been conducted to observe the behaviour of the ESPRIT method and the Steiglitz-McBride algorithm in controlled conditions. The two methods were studied in the full-band case only, in order to highlight their inner characteristics and not to take into account any preprocessing effect. The tests mainly consist in the observation of the poles estimation errors as experimental conditions evolve. The amplitudes have not been considered here, since their estimation relies on the pole estimation, and only consist in a least-square method.

4.1. Robustness to noise

Strictly speaking, a signal-to-noise ratio is the ratio between the signal and noise powers. In the case of impact sounds following the EDS model, this value evolves in time. The estimation quality is not entirely determined by the SNR. As exposed in section 4.3, the number of samples for which the signal amplitude is above the noise standard deviation is of great influence as well. It has therefore been chosen here to display the results of the experiments directly in function of the noise variance.

The deterministic part of the analysed signal was a damped sinusoid of frequency 100 Hz, damping 40 s^{-1}, amplitude 1 and length 2500 samples. The mean error on damping and frequency has been computed for 100

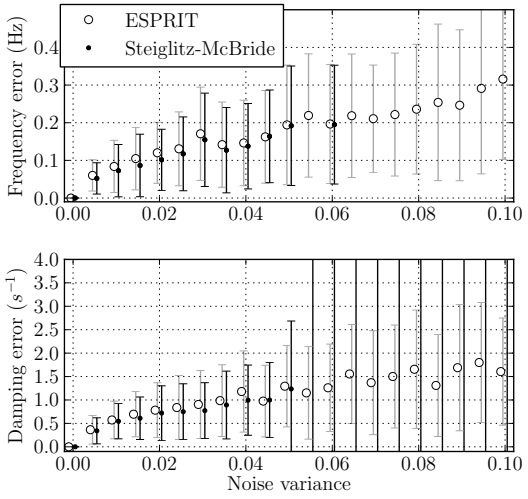

Fig. 3: Robustness to increasing noise variance, for a damped sinusoid with an initial amplitude equal to 1, a frequency of 1000 Hz and a damping of 40 s^{-1}. For each variance, 100 estimations have been conducted.

estimations at different values of the noise variance, and the results are displayed Fig.3.

From a null variance up until a variance of 0.05, the two methods provide similar errors, or slightly better for Steiglitz-McBride. Beyond, aberrant estimations dramatically modify the mean and standard deviation of the Steiglitz-McBride estimations, whereas the ESPRIT method still provides accurate results.

4.2. Resolution limits

Some impacts sounds present a high density of modes; this is typically the case for metallic sounds. In such a situation, each canal of the Gabor transform may contain many partials. As it will be shown in the following section, this diminishes the estimation precision of the method. On Fig.4, one can compare the error committed on the pole estimation as the modal density increases for two methods: ESPRIT and a standard Steiglitz-McBride algorithm which estimates the coefficients of an ARMA filter such as proposed in [4]. The analysed signals consisted in an increasing number of damped sinusoids, which frequencies were randomly distributed between 2000 and 2200 Hz. Their amplitudes were arbitrarily set to 100, and their dampings to 40 s^{-1}.

For both methods, two observations can be made: first, the frequency error estimation increases as the number of components gets higher. Secondly, there is a maximum number of components (different for each method) that can be "correctly" estimated. "Correctly" meaning here an estimation for which the frequency and damping errors are below 1 Hz[1]. Globally the ESPRIT method exhibits smaller frequency error estimations as well as a higher number of components which can be correctly estimated. One can also notice that the Steiglitz-McBride results regroup in two scenarios: either all components are correctly estimated, or no component is correctly estimated at all. With ESPRIT, however, even when all components are not correctly estimated, some still are. Comparing the left figure for which the analysed signal were made of 1500 samples with the right figure for which 4500 samples were analysed, one can notice how the estimation quality of ESPRIT increases for a large number of samples, whereas this augmentation does not affect the estimation quality of the Steiglitz-McBride approach: the number of correctly estimated components goes from 6 to 16 with ESPRIT, while it sticks to 3 with Steiglitz-McBride[2]. As the computational possibilities will grow in the future, the ESPRIT method resolution will get better and better, since it will become possible to analyse longer signals.

4.3. Influence of the analysis horizon

The experiment described here is meant to observe the behaviour of the ESPRIT method as the number of analysed samples changes. This is of matter of importance in practical situations, since the length of the signal directly conditions the computational complexity of the problem, and therefore the computational time. The first experiment, which results are displayed Fig. 5, shows the evolution of the mean error of 100 estimations as the analysis horizon is progressively extended from the beginning of the signal. The analysed signal consisted in a damped sinusoid of frequency 400 Hz, with a damping of 60 s^{-1} and an amplitude of 1, to which was added a gaussian white noise of variance 10^{-4}. One can observe that the estimation resolution increases with the analysed length,

[1] This value has been arbitrarily chosen for the purposes of the analysis comparison.

[2] This explains the different conclusion than the one obtained in [23], where the Steiglitz-McBride estimations were shown to be more accurate: possibly because of the computational limitations at that time, ESPRIT was applied on only 295 samples. With the nowadays computers, longer signals can be analysed, and the ESPRIT performances have greatly improved.

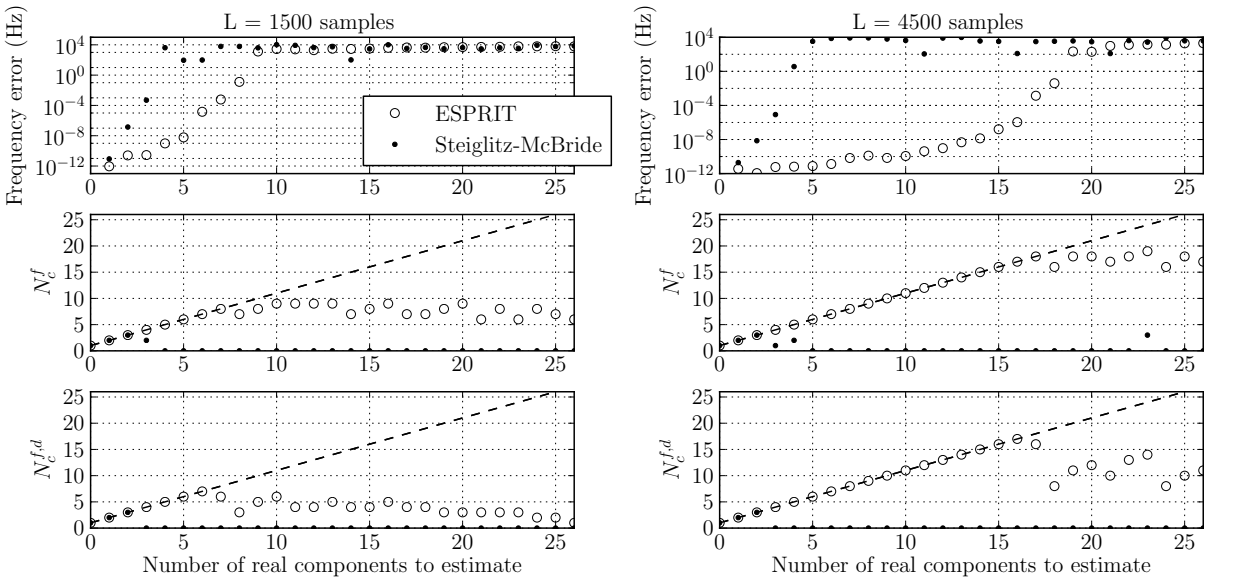

Fig. 4: Comparison of the estimations between ESPRIT and Steiglitz-McBride, for an increasing number of components randomly distributed between 2000 and 2200 Hz, with a random phase, an amplitude of 100, and a damping of 40 s^{-1}. N_c^f is the number of frequency estimations with an error inferior to 1 Hz, and $N_c^{f,d}$ the number of estimations with an error below 1 for both damping and frequency.

up until a threshold which is roughly the effective signal length, i.e. the number of samples for which the deterministic part of the signal is significantly above the noise level. Hence the more samples significantly above the noise level are available for the analysis, the better the estimation will be.

4.4. Influence of noise spectral density shape

The section describes numerical experiments meant to determine the influence of a non-white noise on the estimation. They were motivated by the fact that usually, the Gabor transform of a white noise is not a white noise in a given channel. In fact, the spectral density of the noise in a given frequency channel is equal to the square of the window g Fourier transform subsampled , up to a multiplicative constant that depends on the noise variance and the time subsampling parameter a (see 8.1). As an illustration, three power spectra of a white noise Gabor transform are presented Fig. 6, for different ratios M/a. One can observe that the resulting noise is white for a ratio $M/a = 2$. The experiments described in this section consisted in the estimation of a damped sinusoid of frequency 21 Hz, damping 10 s^{-1}, initial amplitude 1 and composed of 2500 samples, added to a background noise corresponding to different M/a ratios varying from 2 to 256. The subsampling operation having an influence on the global noise energy, it has to be normalised in order to isolate the effect of the noise shape only. Three different noise energy normalisations have been considered. One on the whole spectrum, ensuring that his global energy is 0.1. The second is a normalisation to 0.1 over the "window frequency support", defined here as the frequency range over which the first prominent lobe of the window Fourier transform is higher than the second prominent lobe. The third normalisation is made over the "component frequency support", defined as the frequency range centred around the component frequency which contains 95% of the total component energy. One can see Fig.7, that the estimation error increases as the noise is less and less white (i.e for a growing M/a factor), in the overall normalisation case only. For the two other "local" normalisations, the noise power spectrum shape has no influence on the estimation quality. From this it can be deduced that the relevant measure for the estimation precision is the level of the noise in the neighbourhood of the component frequency, and that the global shape of

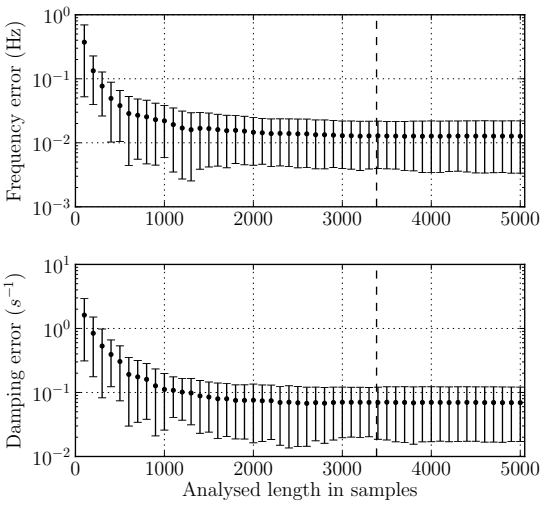

Fig. 5: Evolution of the error mean on 100 realisations as the number of analysed samples evolves. The dashed line corresponds to the number of samples beyond which the component amplitude is below the standard deviation of the background white noise.

the noise has little or no influence at all on the estimation error. This fact comforts the idea that applying ESPRIT over a Gabor transform frequency channel can lead to satisfactory estimations, although the whiteness condition on the background noise is not satisfied.

5. COMPARISON ON RECORDED SOUNDS

The objective here was to compare the global effectiveness of different methods that can be used to decompose a sound as a sum of damped sinusoids. To do so, the resynthesis corresponding to each method is compared to the original sound. The available sound database consisted in 341 sounds recorded in an anechoic room by hitting different objects of the everyday life. 5 analysis methods have been studied. 3 full-band methods, and 2 sub-band methods. The full-band methods consisted in:

- Steiglitz-McBride (STMCB)
- ESPRIT in full-band with the maximum order (ESP MAX)
- ESPRIT with an ESTER-derived order criteria (ESP ESTER)

The maximum order for the full-band ESPRIT analysis is defined as the number of audio samples available divided by 4 (in terms of real components). The ESTER-derived criteria sets the order to the maximum order for which the ESTER cost function is above a fifth of its maximum value. The Steiglitz-McBride algorithm has been applied for 100, 200, 300 and 400 poles. This choice is motivated by the fact that depending on the analysed sound, there is a maximum order above which the Steiglitz-McBride algorithm does not converge. This order is typically between 200 and 300. The best resynthesis was then automatically chosen for comparison with other methods. In order to limit the computational time, the maximal analysable length was limited to 2^{14}.

The considered sub-band methods are:

- FZARMA
- ESPRIT in a Gabor transform (ESP GABOR)

FZARMA (Frequency-Zooming ARMA, [4]) is an analysis methods that combines a "frequency zooming" procedure with the Steiglitz-McBride algorithm. The frequency zooming consists in a band-pass filtering of the original signal followed by a decimation, around each of the frequency range of interest (typically each of the partials). For both analysis methods, FZARMA and ESPRIT, the order has been forced to 6 in each sub-band. The Gabor frame consisted in a blackman-harris window of length 2048, a number of channels $M = 2048$ and a time-step parameter $a = 32$.

In all cases, the components with a negative damping were discarded prior to resynthesis. All other components were kept. As a measure of dissimilarity between sounds, the Itakura-Saito divergence (ISD) has been chosen, as defined in [24]. See Appendix 8.2 for a formal definition. The psychoacoustical discarding procedure described in section 3.6 has not been applied before resynthesis, since it prevents an objective comparison of the original sounds with their corresponding resynthesis. As a matter of fact, discarding a masked component in a resynthesis might have no effect at all from the perceptual point of view, but it will modify the sound spectrum and therefore the Itakura-Saito divergence, which will skew the comparison interpretation.

The results are displayed Fig.8, and the mean divergences for the different methods Table 1. In the following, an analysis is considered "correct", if its corresponding Itakura-Saito distance is below 10. It shows

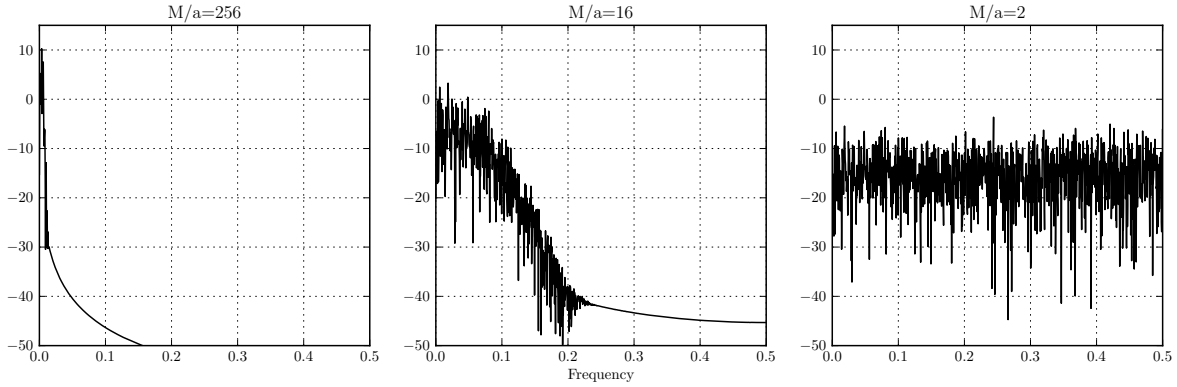

Fig. 6: Power spectrum of a white noise through a single Gabor transform channel, for different M/a ratios. As M/a tends toward 2, the underlying noise tends toward a white noise.

Method	$\overline{\text{ISD}}$	Correct analysis	$\overline{N}_c$
ESP MAX	0.94	83.97 %	3302
ESP GABOR	2.24	98.24 %	193
FZARMA	2.33	95.01 %	193
ESP ESTER	4.12	55.72 %	65
STMCB	4.12	33.43 %	229

Table 1: Statistics on the 341 impact sounds database. $\overline{\text{ISD}}$ stands for the mean Itakura-Saito divergence. The "Correct analysis" are defined as the ones for which the Itakura-Saito divergence is below 10. $\overline{N}_c$ denotes the mean number of components (rounded to the closest integer) used for resynthesis for each sound.

that ESPRIT in Gabor frames offers the best reliability (98 % of correctly analysed signals), closely followed by the FZARMA method (95 % of correctly analysed signals). Among all methods, ESPRIT in full-band with a maximum order analysis provides the resynthesis closest to the originals, although only 84 % of the signals were correctly synthesised. Furthermore, the number of components used for the ESPRIT full-band resynthesis is much higer than with the other methods. Full-band ESPRIT with an ESTER determined order led to a correct analysis in 56 % of the cases, whereas only 33 % of the sounds were correctly synthesised using the Steiglitz-McBride algorithm.

6. CONCLUSION

It has been shown that ESPRIT can be applied on time-frequency representations and that it constitutes a reliable way to describe signals as a sum of damped sinusoids. This has been clearly highlighted by numerical experiments, as well as a real case analysis-synthesis

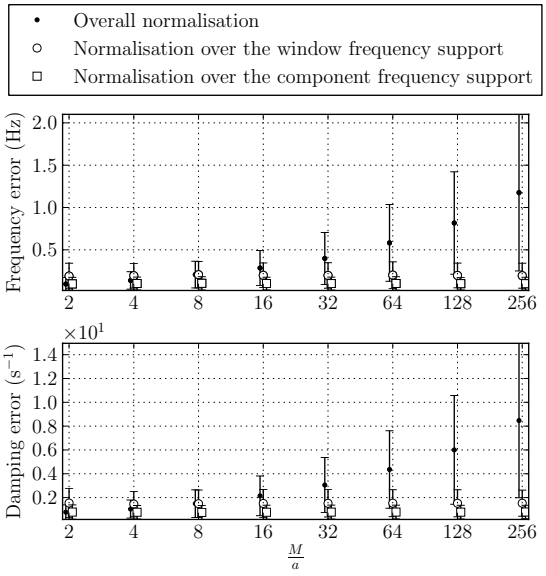

Fig. 7: Mean error over 1000 pole estimations of an exponentially damped sinusoid of frequency 21 Hz, damping 10 s^{-1} and amplitude 1, to which was added a noise corresponding to various M/a ratios.

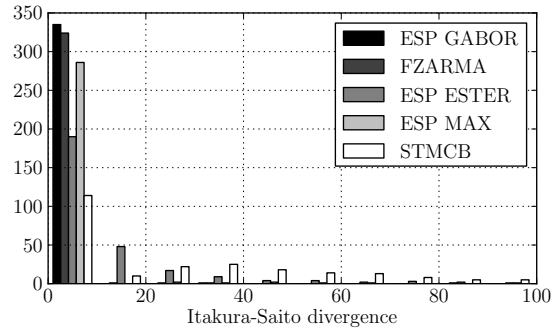

Fig. 8: Histogram over the Itakura-Saito divergence with 10 intervals for the whole sound database and the 5 tested methods.

comparison with other existing methods. The ESPRIT method applied on the full-band signal provided the best estimations. It has to be noted, however, that this comes at the expense of:

- reliability, since the method seems more likely to give diverging results;
- sparsity, since the order is greatly over-estimated;
- computational cost, since the signal has to be analysed on a relatively large number of samples in order to be correctly synthesized.

The computational costs of the different methods have not been considered in the paper, but is has to be pointed out that the FZARMA [4] method, which provided slightly less precise estimations than ESPRIT in Gabor frames, is generally much faster. According to the application, it can appear sometimes more suitable. Both methods have the benefits of the sub-band analysis: an extension of the analysis horizon, and a diminution of the complexity by only considering successive regions in the frequency domain; this appears to be a key-reason for the quality of the estimations that they provide and for their relative sparsity compared to the full-band methods. On top of that, the information given by the time-frequency representation is of great use for targeting the analysis on the time-frequency intervals that contain the desired information. Combined with a selection of the most important components based on psychoacoustical considerations, the proposed method constitute an adapted tool for the analysis-synthesis of impact sounds.

7. AKNOWLEDGMENTS

The authors would like to thank Julien Perron for his role in the calibration of the method, and Vincent Germain for the creation of the impact sounds database. The Gabor computation programs were written in python, highly inspired by the LTFAT toolbox ([25]). This project has been partly supported by the French National Research Agency (ANR-10-CORD-010 “Métaphores sonores”, http://metason.cnrs-mrs.fr/).

8. APPENDICES

8.1. A white noise through a Gabor transform channel

Given $w[l]$ a white noise of variance σ^2. The (m+1)-th channel of its Gabor transform $c_{w,m}[n]$ is:

$$c_{w,m}[n] = \sum_{l=0}^{L-1} w[l]\,\overline{g}[l-an]\,e^{-2i\pi\frac{m}{M}l}, \qquad (20)$$

which is a a-sub-sampled version of the convolution product between g inverted in time, and the modulated noise hereafter denoted $\tilde{w}[l]$:

$$\begin{aligned} c_{w,m}[n] &= \left\{\left(w[l]\,e^{-2i\pi\frac{m}{M}l}\right) * \overline{g}[-l]\right\}[an] \quad (21)\\ &= \left\{\tilde{w}[l] * \overline{g}[-l]\right\}[an] \quad (22)\end{aligned}$$

where $*$ stands for the convolution operator.

Let $\mathscr{F}(.)$ denote the Fourier transform operator. A sub-sampling in time induces an over-sampling in frequency according to the relation:

$$\mathscr{F}\left(x[an]\right)(\nu) = \frac{1}{a}\mathscr{F}\left(x[n]\right)\left(\frac{\nu}{a}\right) \qquad (23)$$

for any x for which $\mathscr{F}(x)$ exists. Since $w[l]$ is a white noise, $\mathscr{F}\left(\tilde{w}[l]\right) = \mathscr{F}\left(w[l]\right)$. Futhermore, $\mathscr{F}\left(\overline{g}[-l]\right) = \overline{\mathscr{F}\left(g[l]\right)}$. The Fourier transform of a convolution product being the product of the Fourier transforms, one finally has:

$$\mathscr{F}\left(c_{w,m}\right) = \frac{1}{a}\left\{\mathscr{F}\left(w\right)\overline{\mathscr{F}\left(g\right)}\right\}[an] \qquad (24)$$

Therefore, the power spectrum of $c_{w,m}[n]$ is:

$$\left|\mathscr{F}\left(c_{w,m}\right)\right|^2 = \frac{\sigma}{a}\left|\mathscr{F}(g)\right|^2 \qquad (25)$$

Usually, $\mathscr{F}(g)$ corresponds to the frequency response of a low-pass filter, as it can be seen Fig.6 in the case of a blackman-harris window.

8.2. The Itakura-Saito divergence

Given two signals x_i and x_j with their respective Gabor transforms being $c_i[m,n]$ and $c_j[m,n]$ of size $M \times N$, let $\boldsymbol{m}_{i,j}$ be the point-wise ratio between the two transforms defined as:

$$\boldsymbol{m}_{i,j} = \frac{c_i[m,n] + \lambda}{c_j[m,n] + \lambda} \tag{26}$$

with λ a regularisation term preventing divergence during the computations. The non-symmetrical Itakura-Saito divergence between x_i and x_j is defined as:

$$d_{i,j} = \frac{1}{MN} \sum_{m,n} |\boldsymbol{m}_{i,j}[m,n]| - \log |\boldsymbol{m}_{i,j}[m,n]| - 1 \tag{27}$$

Notice that $d_{i,j}$ is actually equal to 0 when $c_1 = c_2$, but that usually $d_{i,j} \neq d_{j,i}$. A symmetrical version of the Itakura-Saito divergence is therefore defined as:

$$d_{i,j}^{\text{sym}} = d_{j,i}^{\text{sym}} = \frac{1}{2}\left(d_{i,j} + d_{j,i}\right), \tag{28}$$

and is the one used in the paper.

9. REFERENCES

[1] C. Verron, M. Aramaki, R. Kronland-Martinet and G. Pallone, "A 3-D immersive synthesizer for environmental sounds", Audio, Speech, and Language Processing, IEEE Transactions on, vol. 18, no. 6, pp. 1550-1561, 2010.

[2] "SoundDesignToolkit", http://www.soundobject.org/SDT/.

[3] J.B. Allen and L.R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis", Proceedings of the IEEE, vol. 65, no. 11, pp. 1558-1564, 1977, IEEE.

[4] M. Karjalainen, P.A.A. Esquef , P. Antsalo, A. Makivirta and V. Valimaki, "Frequency-zooming ARMA modeling of resonant and reverberant systems", Journal of the Audio Engineering Society, vol. 50, no. 12, pp. 1012-1029, 2002, Audio Engineering Society INC.

[5] M. Aramaki and R. Kronland-Martinet, "Analysis-synthesis of impact sounds by real-time dynamic filtering", Audio, Speech, and Language Processing, IEEE Transactions on, vol. 14, no. 2, 695-705, 2006, IEEE.

[6] K. Steiglitz and L. McBride, "A technique for the identification of linear systems", Automatic Control, IEEE Transactions on, vol. 10, no 4, pp 461–464, 1965, IEEE.

[7] J. Laroche, "A new analysis/synthesis system of musical signals using Prony's method-application to heavily damped percussive sounds", Acoustics, Speech, and Signal Processing, 1989. ICASSP-89., 1989 International Conference on, pp. 2053-2056, 1989.

[8] R. Schmidt, "Multiple emitter location and signal parameter estimation", Antennas and Propagation, IEEE Transactions on, vol. 34, no. 3, pp. 276-280, 1986.

[9] R. Roy and T. Kailath, "ESPRIT - Estimation of Signal Parameters via Rotational Invariance Techniques", Acoustics, Speech and Signal Processing, IEEE Transactions on, vol. 37, no. 7, pp. 984-995, 1989.

[10] A. Kot, S. Parthasarathy, D. Tufts and R. and Vaccaro, "The statistical performance of state-variable balancing and Prony's method in parameter estimation", Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'87, vol. 12, pp. 1549–1552, 1987.

[11] P. Stoica and A. Nehorai, "Study of the statistical performance of the Pisarenko harmonic decomposition method", Communications, Radar and Signal Processing, IEE Proceedings F, vol. 135, no 2, pp. 161–168, 1988.

[12] "On SVD for estimating generalized eigenvalues of singular matrix pencil in noise", Y. Hua and T.K. Sarkar, Signal Processing, IEEE Transactions on, vol. 39, no 4, pp 892–900, 1991.

[13] R. Badeau, "Méthodes haute-résolution pour l'estimation et le suivi de sinusoïdes modulées", Ph.D. Thesis, École Nationale Supérieure des Télécommunications, 2005

[14] K. Ege, X. Boutillon and B. David, "High-resolution modal analysis", Journal of Sound and Vibration, vol. 325, no. 4-5, pp. 852–869, 2009.

[15] R. Badeau, B. David and G. Richard, "A new perturbation analysis for signal enumeration in rotational invariance techniques", Signal Processing, IEEE Transactions on, vol. 54, no. 2, pp. 450–458, 2006.

[16] M.G. Christensen, A. Jakobsson and S.H Jensen. "Sinusoidal order estimation using angles between subspaces", EURASIP Journal on Advances in Signal Processing, vol. 2009, pp. 62, 2009, Hindawi Publishing Corp.

[17] K. Gröchenig, "Foundations of time-frequency analysis", Birkhauser, 2001.

[18] F. Millioz and N. Martin, "Estimation of a white Gaussian noise in the Short Time Fourier Transform based on the spectral kurtosis of the minimal statistics: Application to underwater noise", Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on, pp. 5638-5641, issn 1520-6149, 2010.

[19] Terhardt, E., "Calculating virtual pitch", Hearing research, vol.1, no. 2, pp. 155-182, 1979, Elsevier.

[20] Zwicker, E. and Fastl, H., "Psychoacoustics: Facts and models", vol. 2, 1999, Springer Berlin.

[21] Humes, L.E. and Jesteadt, W. "Models of the additivity of masking", The Journal of the Acoustical Society of America, vol. 85, no. 3, pp. 1285-1294, 1989.

[22] Necciari, T. "Masquage auditif temps-fréquence: mesures psychoacoustiques et application à l'analyse-synthèse des sons", 2010.

[23] Esquef, P.A.A. and Karjalainen, M. and Välimäki, V. "Frequency-zooming ARMA modeling for analysis of noisy string instrument tones", EURASIP Journal on Applied Signal Processing, vol. 2003, pp. 953–967, 2003, Hindawi Publishing Corp.

[24] A. Olivero, "Identification of Time-Frequency Maps for Sounds Timbre Discrimination, In Proc. 14th Int. Conf. on Digital Audio Effects (DAFx-11), pp. 123-125, IRCAM-Paris, France, 2011.

[25] Peter L. Søndergaard and Bruno Torrésani and Peter Balazs, "The Linear Time Frequency Analysis Toolbox", International Journal of Wavelets, Multiresolution Analysis and Information Processing, accepted for publication, 2011