Modal Analysis of Impact Sounds with ESPRIT in Gabor Transforms

Adrien Sirdey, Olivier Derrien, Richard Kronland-Martinet, Mitsuko Aramaki

MODAL ANALYSIS OF IMPACT SOUNDS WITH ESPRIT IN GABOR TRANSFORMS

A. Sirdey, O. Derrien, R. Kronland-Martinet,
Laboratoire de Mécanique et d'Acoustique
CNRS
Marseille, France
<name>@lma.cnrs-mrs.fr

M. Aramaki,
Institut de Neurosciences cognitives de la Méditerranée
CNRS
Marseille, France
<name>@incm.cnrs-mrs.fr

ABSTRACT

Identifying the acoustical modes of a resonant object can be achieved by expanding a recorded impact sound in a sum of damped sinusoids. High-resolution methods, e.g. the ESPRIT algorithm, can be used, but the time-length of the signal often requires a sub-band decomposition. This ensures, thanks to sub-sampling, that the signal is analysed over a significant duration so that the damping coefficient of each mode is estimated properly, and that no frequency band is neglected. In this article, we show that the ESPRIT algorithm can be efficiently applied in a Gabor transform (similar to a sub-sampled short-time Fourier transform). The combined use of a time-frequency transform and a high-resolution analysis allows selective and sharp analysis over selected areas of the time-frequency plane. Finally, we show that this method produces high-quality re-synthesized impact sounds which are perceptually very close to the original sounds.

1. INTRODUCTION

The context of this study is the identification of acoustical modes which characterize a resonant object, in the perspective of building an environmental sound synthesizer. 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, 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, we consider only the analysis part.

In the past decades, significant advances have been made in the field of system identification, especially for estimating EDS parameters in a background noise. Although the so-called *high-resolution methods* or *subspace methods* (MUSIC, ESPRIT) [1, 2] were proved to be more efficient than spectral peak-picking and iterative analysis-by-synthesis methods [3], few applications have been proposed. One can suppose that the high computational complexity of these methods is a major drawback to their wide use: on a standard modern computer, the ESPRIT algorithm can hardly analyse more than 10^4 samples, which corresponds roughly to 200 ms sampled at 44100 Hz. This is usually too short for analysing properly impact sounds which can last up to 10 s. Sub-band decomposition with critical sub-sampling in each band seems to be a natural solution to overcome the complexity problem, as it has already been shown in [4] and [5]. Another drawback is that ESPRIT gives accurate estimates when the background noise is white, which is usually not the case in practical situations. This problem can be overcome by the use of whitening filters. The estimation of the model order (i.e. the number of modes) is also an important issue. Various methods have been proposed for automatic estimation of the order, e.g. ESTER [6], but this parameter is often deliberately over-estimated in most practical situation.

In this paper, we propose a novel method for estimating the modes with ESPRIT algorithm: we first apply a Gabor Transform (GT), which is basically a sub-sampled version of the short-time Discrete Fourier Transform (DFT), to the original sound in order to perform a sub-band decomposition. The number of channels and the sub-sampling factor depend on the Gabor frame associated to the transform. We show that an EDS in the original sound is still an EDS inside each band, and the original parameters can be recovered from a sub-band analysis using ESPRIT. Furthermore, if the number of frequency sub-bands is high enough, it is reasonable to assume that the noise is white inside each sub-band. We also propose a method to discard insignificant modes a posteriori in each sub-band.

The paper is organised as follows: first, in a brief state-of-the-art, we describe the signal model, the ESPRIT algorithm and the Gabor transform. Then, we show that original EDS parameters can be recovered by applying the ESPRIT algorithm in each frequency band of the Gabor transform. In the next part, we describe an experimentation on a real metal sound, and show the efficiency of our method. Finally, we discuss further improvements.

2. STATE OF THE ART

2.1. The signal model and the ESPRIT algorithm

The discrete signal to be analysed is written:

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

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

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

where 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 gaussian white noise of variance σ^2.

The ESPRIT algorithm was originally described by Roy *et. al.* [2], but many improvements have been proposed. Here, we use the Total Least Square method by Van Huffel *et. al* [7]. The principle consists of performing a SVD on an estimate of the signal correlation matrix. The eigenvectors corresponding to the K highest

eigenvalues correspond to the so called *signal space*, while the remaining vectors correspond to the so called *noise space*. The shift invariance property of the signal space allows a simple solution for 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:

We define the signal vector:

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

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

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

where $Q, R > K$ and $Q + R - 1 = L$. We also define the amplitude vector:

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

and the Vandermonde matrix of the poles:

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

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

$$\boldsymbol{X} = [\boldsymbol{U_1 U_2}] \left[\begin{array}{cc} \boldsymbol{\Sigma_1} & 0 \\ 0 & \boldsymbol{\Sigma_2} \end{array} \right] \left[\begin{array}{c} \boldsymbol{V_1} \\ \boldsymbol{V_2} \end{array} \right], \tag{7}$$

where $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are diagonal matrix 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. The shift-invariance property of the signal space yields to:

$$\boldsymbol{U}_1^{\downarrow} \boldsymbol{\Phi}_1 = \boldsymbol{U}_1^{\uparrow}, \qquad \boldsymbol{V}_1^{\downarrow} \boldsymbol{\Phi}_2 = \boldsymbol{V}_1^{\uparrow}, \tag{8}$$

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

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

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

2.2. The Gabor Transform

The Gabor transform of signal $x[l]$ can be written as:

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

where $g[l]$ is the analysis window, a is the time-step and M the number of frequency channels. $\overline{(.)}$ denotes the complex conjugate. m is a discrete frequency index and n a discrete time-index. $\{g, a, K\}$ is called a *Gabor frame*. For some frames, this transform can be inverted. A necessary condition is $a \leq M$. 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, we apply the ESPRIT algorithm to each frequency channel in order to analyse longer signals.

3. ESPRIT IN A GABOR TRANSFORM

In this section, we investigate the application of the ESPRIT algorithm to a single channel of the GT. As the GT is linear, we separate the contribution of the deterministic part $s[l]$ and the contribution of the noise $w[l]$.

3.1. Deterministic part

We denote $c[m,n]$ the GT of $s[l]$ in channel m and time index n. We also note $c_k[m,n]$ 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), is 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, but the apparent amplitudes and phases are modified.

3.2. Stochastic part

Assuming that the time-step a is close to M ensures that the GT of the noise in each channel is approximately white. Furthermore, it has been proved that the Gabor transform of a gaussian noise is a complex gaussian noise [8]. So we assume that the GT of $w[l]$ in each channel is a complex white gaussian noise.

3.3. Recovering the signal parameters

As the signal model is still valid, it is reasonable to apply ESPRIT on $c[m,n]$. We note $\boldsymbol{c}_m$ the vector of GT coefficients in the channel m and $\boldsymbol{S}_m$ the Hankel matrix build 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, we choose the analysis window $g[l]$ so that its bandwidth is smaller than $\frac{1}{a}$. That way, the possible aliasing components will be attenuated by the band-pass effect of the Gabor transform.

We note $\tilde{\boldsymbol{V}}_m^N$ 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:

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

Without noise, according to equation (12), each 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 considered as noise. Thus practically, the exact number of detectable components in each channel is unknown. So we set the model order in each channel with the ESTER criterion (see section 4.3 for implementation details).

4. EXPERIMENTATION

When applied on synthetical sounds that strictly verify the signal model (1), the full-band ESPRIT algorithm, as well as the ESTER criteria, estimate the model parameters with an excellent precision (see [4], [6]). Estimation errors are observed when dealing with real-life sounds. Therefore this section does not consider the analysis of synthetical sounds, but focuses on the analysis/synthesis of a real metal sound `m5` (which can be listened to at [9]). `m5` has been produced hitting a metal plate with a drum stick. Observing its waveform, Fourier transform and spectrogram (Fig. 4a, 4e and 1) one can see that it presents a rich spectral content and significant lasting energy up to 6 s.

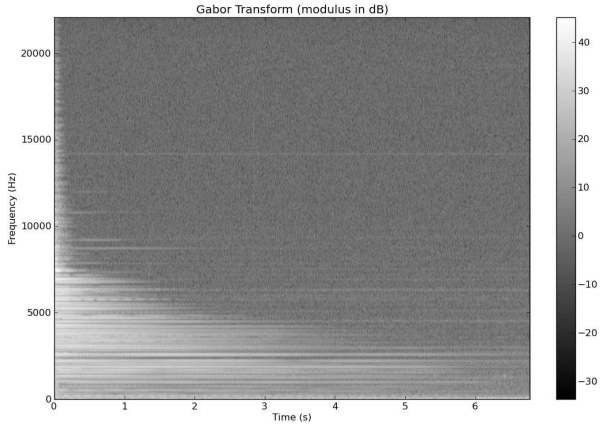


Figure 1: *Spectrogram of* `m5`.

4.1. Analysis with full-band ESPRIT method

Considering the size of the Hankel matrix corresponding the whole sound (around 150000×150000), only a part of the original signal can be analysed with the full-band ESPRIT algorithm. Fig. 2 shows the ESTER criteria cost function computed for the 10000 first samples of `m5`. The optimal model order theoretically corresponds to the maximum of this function, which is reached here for $K = 4$ modes. This value is obviously not consistent, as one can see on the spectrogram of `m5`: the spectral content is obviously much more complex. A reasonable compromise would be to choose the maximum order for which the cost function is above a given threshold. For instance, this threshold can be set to 100. The corresponding model order is $K = 206$. After applying the ESPRIT algorithm, 29 EDS appear to have a negative damping, which will form diverging components at the re-synthesis. Since they do not describe physical modes, they must be discarded. The resulting synthesised sound `m5_std_esprit` ([9]) is unsatisfying from a perceptual point of view, and reveals that the damping behaviour of some modes has been wrongly estimated as well. Furthermore there is a significant difference in the spectral content of the original and the re-synthesized sound above 12000 Hz, as shown by Fig. 3 and 4e.

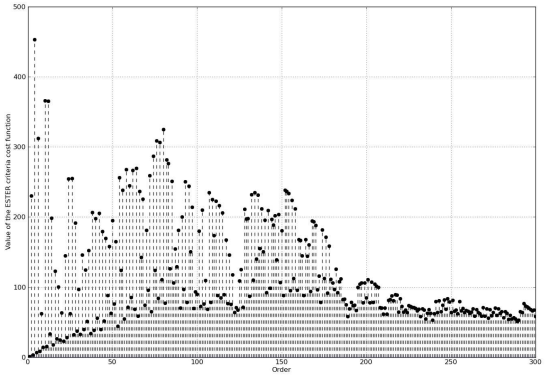


Figure 2: *ESTER criteria cost function computed for the 10000 first samples of the full-band signal* `m5`.

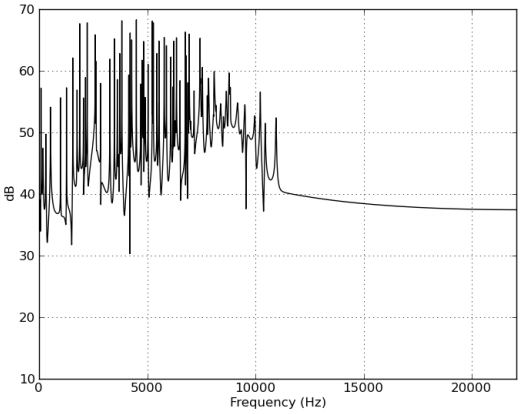
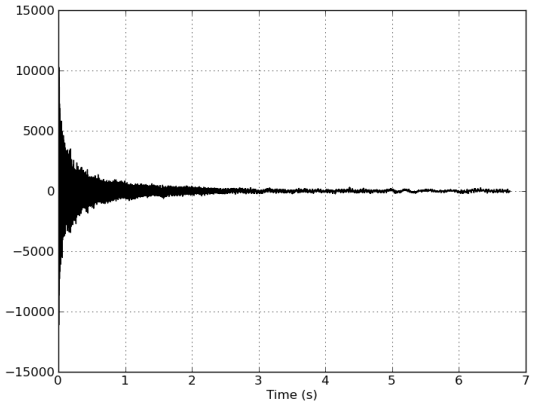
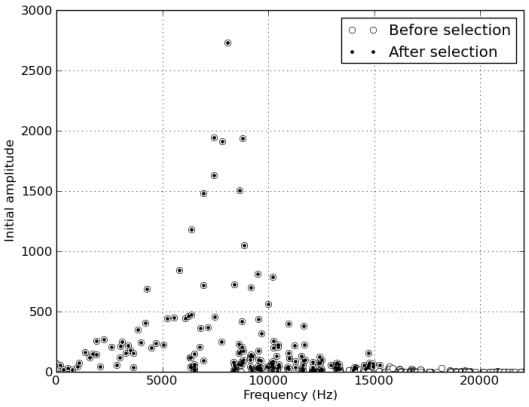


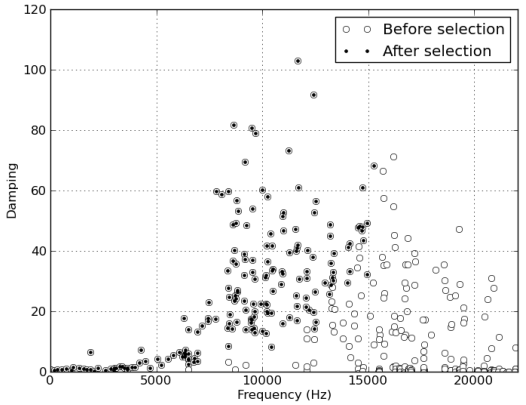
Figure 3: *DFT spectrum of the re-synthesized sound* `m5_std_esprit` *obtained by applying a full band ESPRIT algorithm. The model order is* $K = 206$.



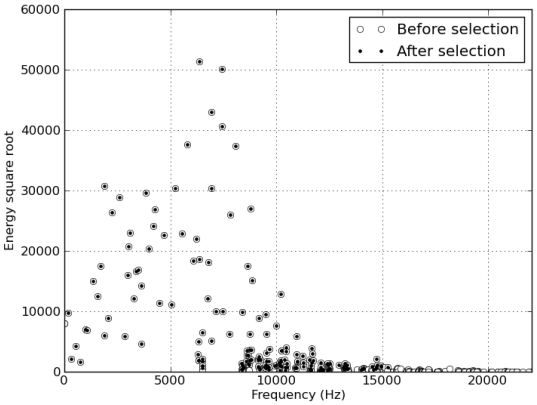
(a) *Waveform of the metal sound* `m5`



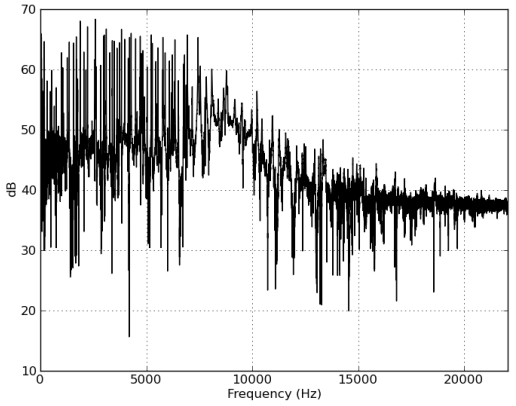
(b) *Initial amplitude*



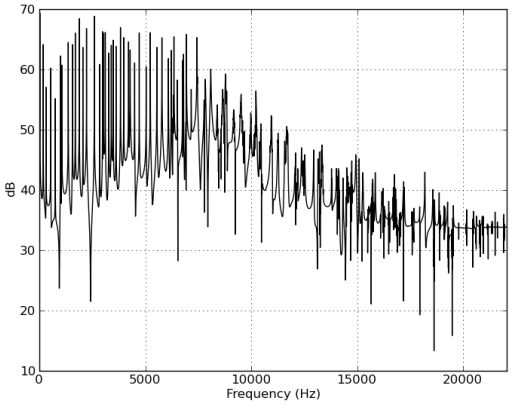
(c) *Damping*



(d) *Energy square root*



(e) *DFT spectrum of* `m5`



(f) *DFT spectrum of the re-synthesized sound* `m5_resyn` *with all components*

Figure 4: *Overview of the analysis of* `m5` *(a) using the ESPRIT algorithm over its Gabor transform. (b), (c) and (d) show the 401 mode parameters which have been initially extracted. (e) and (f) respectively show the DFT spectrum of the original sound* `m5` *and the DFT spectrum of the re-synthesised sound* `m5_resyn`*; both sounds are available at [9]. The 181 modes marked with a black dot are the ones that remain after discarding the modes which initial amplitude is below the absolute detection threshold; the resulting synthesis sound* `m5_resyn_amp_ts` *can be listened to at [9].*

4.2. Analysis with ESPRIT in a Gabor transform

The chosen Gabor frame consists in a Blackman-Harris window of length 1024, a time-step parameter $a = 32$, and a number of channels $M = 1024$. It is unnecessary to apply the ESPRIT algorithm over regions of the time-frequency plane that only contain noise. Since the most important deterministic information is contained in the channels of high energy, those channels can be identified using a peak detection algorithm over the energy profile of the Gabor transform as shown in Fig. 5. In a software environment, the

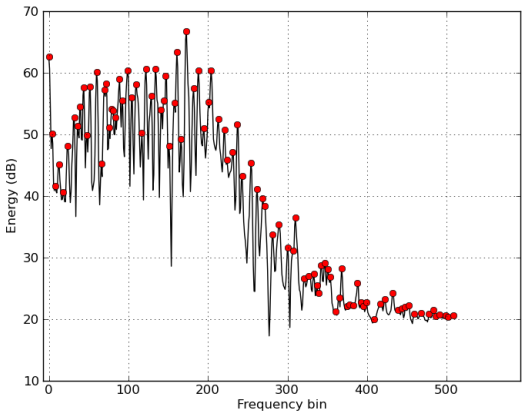


Figure 5: *Energy of the Gabor transform of* `m5` *computed for each of its channels. The dots correspond to the channels identified as peaks.*

choice of which channels will be analysed could be left to the user. It is reasonable to think that the noise whitening induced by the sub-band division of the spectrum makes the ESTER criteria more reliable than in the full-band case, therefore the analysis order is computed for each of the selected channels, and set to the maximum of the ESTER criteria cost function. Doing so, a total number of 430 modes is obtained.

4.3. 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 component is likely to appear in all of these channels. These multiple estimations of the same component have to be identified, and only one will be kept for the final re-synthesis: the one which is the closest to the central frequency of its detection channel. In the example presented here, 29 components have been identified as replicas using a frequency confidence interval of 1 Hz. Fig. 4b, 4c and 4d show the mode parameters (amplitude, damping, energy as function of frequency) that remain after discarding the replicas. The resulting re-synthesized sound `m5_resyn` can be listen to at [9]. Fig. 4f shows the DFT spectrum of `m5_resyn` which can be compared to the DFT spectrum of the original analysed sound Fig. 4e.

4.4. Discarding irrelevant components

The estimated set of modes is the one that best fits the signal model (2) with respect to the Total Least Square criterion. However, as shown in Fig 4b, some of those modes are not relevant for they have an insignificant energy. In order to produce perceptually convincing sounds, one can rely on psychoacoustic results in order to discard inaudible modes. For instance, the absolute detection threshold can be used to discard modes by observing their initial amplitude. The black doted modes on Fig. 4b, 4c and 4d represent the modes that remain after applying an absolute detection threshold ([10]) and setting the minimum of the threshold to the minimum amplitude that the sound format can handle (e.g. ± 1 for wav format coded as 16 bits integers). The resulting sound `m5_resyn_amp_ts`, containing 181 modes, can be listened to at [9].

It is also possible to use energy arguments and favour high energy modes over low energy modes. In the directory named 'Cumulative synthesis' available at [9] are stored successive re-synthesis of `m5` computed by successively adding the modes sorted in decrescent order of energy. One can note that there is no significative perceptual difference between the sounds beyond 105 modes.

5. FURTHER IMPROVEMENTS

One of the advantages provided by the use of time-frequency representations is the existence of efficient statistical estimators for the background noise. As it can be seen on Fig. 1, a significant number of Gabor coefficients describing an impact sound correspond to noise, and can therefore be used to estimate the variance of the stochastic part of the signal (see [8]). If the additive noise is coloured, it is even possible to estimate the variance in several selected frequency bands. Knowing the variance of the noise for each frequency channel offers the possibility to use noise masking properties of the human hearing to discard inaudible components, and possibly lead to a more selective criteria than the absolute detection threshold described in section 4.4.

The concept of nonstationary Gabor frames ([11]) makes it also possible to adapt the resolution of the Gabor transform so as to get an optimal compromise between precision and computational cost. It would allow, for instance, to take into account the logarithmical frequency resolution of the human hearing when applying the Gabor transform. Furthermore, it can be observed that the damping usually decreases with frequency; nonstationary Gabor frames would allow to adapt the time-step parameter of the Gabor frame along the frequency scale, so that computational cost is saved while a sufficient number of coefficients are taken for the analysis.

6. CONCLUSION

It has been shown that using the ESPRIT algorithm over time-frequency representations leads to perceptually convincing re-synthesis. The method has the same benefits than the sub-band analysis: it allows an extension of the analysis horizon, and it diminishes the complexity of the problem by only considering successive regions in the frequency domain; but 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, thereby avoiding unnecessary analysis and reducing the global computational cost.

7. REFERENCES

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

[2] 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.

[3] M. Goodwin, "Matching Pursuits with Damped Sinusoids," in *Acoustics, Speech, and Signal Processing, 1997. Proceedings.(ICASSP'97). IEEE International Conference on*. IEEE, 2004, pp. 2037–2040.

[4] R. Badeau, *Méthodes à haute résolution pour l'estimation et le suivi de sinusoïdes modulées*, Ph.D. thesis, Ecole Nationale Supérieure des Télécommunications, 2005.

[5] 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.

[6] R. Badeau, B. David, and G. Richard, "Selecting the modeling order for the ESPRIT high resolution method: an alternative approach," in *Acoustics, Speech, and Signal Processing, 2004. Proceedings.(ICASSP'04). IEEE International Conference on*. IEEE, 2004, vol. 2.

[7] S. Van Huffel, H. Park, and J.B. Rosen, "Formulation and solution of structured total least norm problems for parameter estimation," *Signal Processing, IEEE Transactions on*, vol. 44, no. 10, pp. 2464–2474, 1996.

[8] 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," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 5638–5641.

[9] "Sample sounds," available at: http://www.lma.cnrs-mrs.fr/ kronland/Dafx11/.

[10] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[11] Florent Jaillet, Peter Balazs, and Monika Dörfler, "Nonstationary Gabor frames," in *Proceedings of the* 8^{th} *international conference on Sampling Theory and Applications (SAMPTA'09)*, Marseille, France, May 2009.