SINUSOIDAL PARAMETER EXTRACTION AND COMPONENT SELECTION IN A NON STATIONARY MODEL

Mathieu Lagrange [†], *Sylvain Marchand* [‡], *and Jean-Bernard Rault* [†]

[†] France Telecom R&D
4, rue du Clos Courtel, BP 59
F-35512 Cesson Sevigné cedex, France
firstname.name@rd.francetelecom.com

[‡] SCRIME - LaBRI, Université Bordeaux 1
351, cours de la Libération,
F-33405 Talence cedex, France
sm@labri.u-bordeaux.fr

ABSTRACT

In this paper, we introduce a new analysis technique particularly suitable for the sinusoidal modeling of non-stationary signals. This method, based on amplitude and frequency modulation estimation, aims at improving traditional Fourier parameters and enables us to introduce a new peak selection process, so that only peaks having coherent parameters are considered in subsequent stages (*e.g.* partial tracking, synthesis). This allows our spectral model to better handle natural sounds.

1. INTRODUCTION

Spectral sound models provide general representations for many applications such as compression, content extraction and transformation. Most of these models, such as additive synthesis, are based on the Fourier analysis which has proven to be accurate under the condition of local stationarity.

Unfortunately, most natural signals are not stable enough to be considered as locally stationary with typical analyzing frame lengths, *i.e.* $\sim$ 20 ms (close to the perceptual sensibility of the human auditory system). To address this issue, the window length can be shortened but, according to the well known time / frequency resolution tradeoff, this will lead to poor frequency resolution. In the context of voiced speech analysis, McAulay and Quatieri [1] proposed to adapt the analysis window size to the pitch of the voice, although this cannot be applied to polyphonic sources. This paper introduces an alternative that relies on amplitude and frequency modulation based modeling of the audio signal during the analysis time slot. Intra-frame parameters variations are extracted and the bias introduced when estimating the stationary parameters can be compensated. Thanks to these modulation measures and to a more accurate stationary parameter extraction, we are able to develop an efficient peak selection process that better separates "noisy" peaks and modulated ones.

After a brief introduction in Section 2 to the non-stationary model used in this paper, we extend in Section 3 the studies of Masri [2] to the case of the Hann window in order to estimate intra-frame variations. Section 4 is dedicated to the correction of the stationary parameters with two different approaches: time reassignment and the computation of the spectrum of a modulated sinusoid. After an introduction on peak selection processing and sinusoidal characterization, a new peak selection process is presented in Section 5. Possible applications as well as comparative results follow.

2. SINUSOIDAL MODELING

2.1. Stationary Case

Additive synthesis is the original spectrum modeling technique. It is rooted in Fourier's theorem, which states that any periodic function can be modeled as a sum of sinusoids at various amplitudes and harmonic frequencies. For stationary pseudo-periodic sounds, these amplitudes and frequencies continuously evolve slowly with time, controlling a set of pseudo-sinusoidal oscillators commonly called *partials*. The audio signal a can be calculated from the additive parameters using Equations 1 and 2, where n is the number of partials and the functions f_p, a_p, and ϕ_p are the instantaneous frequency, amplitude, and phase of the p-th partial, respectively. The n pairs (f_p, a_p) are the parameters of the additive model and represent points in the frequency-amplitude plane at time t. This representation is used in many analysis / synthesis programs such as SMS [3] or *InSpect* [4].

$$a(t) = \sum_{p=1}^{n} a_p(t)\ \cos(\phi_p(t)) \tag{1}$$

$$\phi_p(t) = \phi_p(0) + 2\pi \int_0^t f_p(u)\ du \tag{2}$$

2.2. Non-Stationary Case

For non-stationary signals, amplitude and frequency parameters a_p and f_p appear as the mean of the amplitude / frequency evolutions in the analysis frame. In our model, we consider that the parameters can evolve in the analysis window. Since the human ear – as every sensory organ – perceives amplitude variation as the logarithm of the excitation, it is convenient to express the amplitude modulation Δ_p^a in dB (decibels). Although a similar logarithmic scale would be appropriate for frequency variations (Δ_p^f) as well, they will be considered as linear for the sake of simplicity. Thus, the audio signal a is given by the following equations:

$$a(t) = \sum_{p=1}^{P} a_p(t).10^{\Delta_p^a(t)\cdot t/20}\ \cos(\phi_p(t)) \tag{3}$$

$$\phi_p(t) = \phi_p(0) + 2\pi \int_0^t \left(f_p(u) + \Delta_p^f(u)\cdot u \right)\ du \tag{4}$$

where a_p, f_p, Δ_p^a, and Δ_p^f are considered as constant during the analysis window. The following section is dedicated to the estimation of the intra-frame modulations.

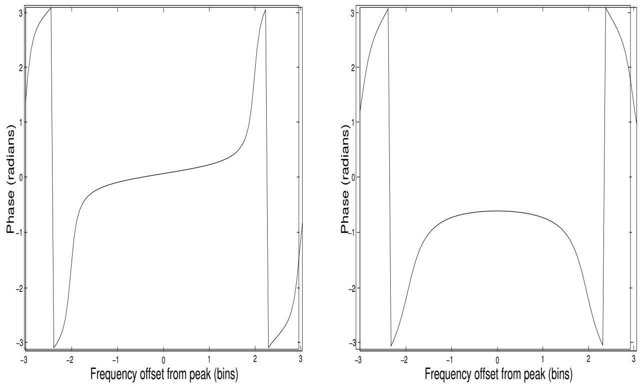

Figure 1: *Influence of amplitude (left) and frequency (right) modulations on zero-padded phase spectrum.*

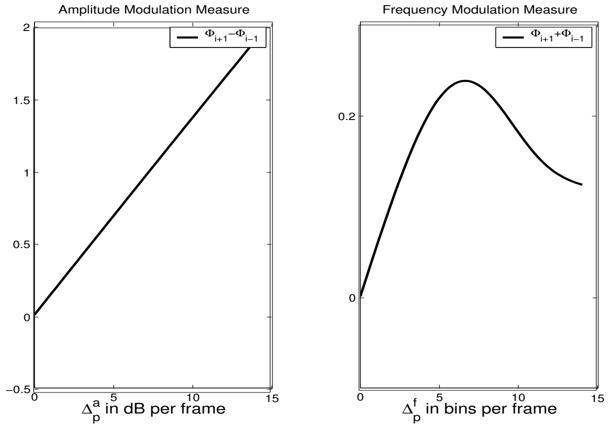

Figure 2: *Evolution of empirical measures ($\Phi_{i+1} \pm \Phi_{i-1}$) as a function of actual modulations.*

3. MODULATED SINUSOIDS

3.1. Modulation Estimation

Several techniques were proposed in order to estimate intra-frame modulations. Marques [5] and Peeters [6] studies were based on the use of truncated Gaussian windows, known for their good theoretical properties. Indeed, the Fourier transform of a Gaussian window is a Gaussian function. However, Gaussian windows have poor frequency resolution [7]. The Hann window, with its prominent and narrow main lobe, has proven to be better for our purposes.

We use the empirical studies of Masri [2] based on phase distortion analysis. For a peak p located in the i-th bin of the zero-padded phase spectrum Φ, picking the values of Φ at indices $i-1$ and $i+1$ allows us to estimate the frequency / amplitude modulations (see Figure 1). Indeed, a constant relationship between Δ_p^a and $\Phi_{i+1}-\Phi_{i-1}$ was found, as well as a more complex one between Δ_p^f and $\Phi_{i+1}+\Phi_{i-1}$ (see Figure 2).

More formally:

- $\Delta_p^a = c \cdot \Phi_a$

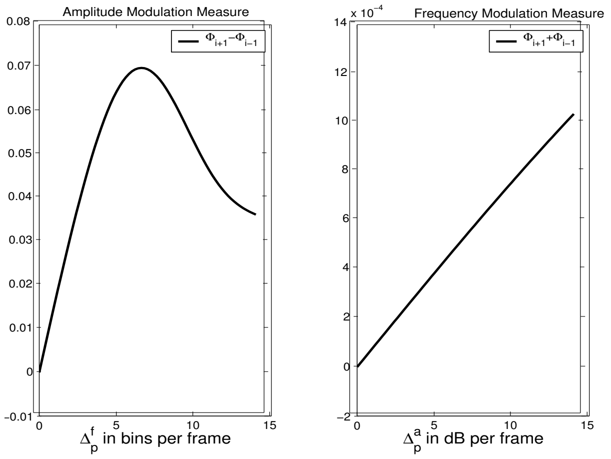

Figure 3: *Influence of frequency modulation on the amplitude modulation measure when there is no amplitude modulation and vice versa.*

- $\Delta_p^f = G(\Phi_f)$

where Φ_a and Φ_f are expressed by:

- $\Phi_a = \Phi_{i+1} - \Phi_{i-1}$
- $\Phi_f = \Phi_{i+1} + \Phi_{i-1}$

The value of c and those of the coefficients of G depend on the window, its size and the zero-padding level. These empirical measures come from the fact that the influence of frequency modulation on phase was found to be symmetrical whereas that of amplitude was found to be anti-symmetrical, as can be seen in Figure 1. It theoretically guarantees the independence of the two estimations. However, this independence is not perfect, as shown in Figure 3, but quite sufficient for our purposes.

Zero-padding, because it interpolates the spectrum by artificially adding zeros to the frame buffer before the Fast Fourier Transform (FFT), is a computation-expensive prerequisite. Indeed, by interpolating the spectrum, it reduces but not totally excludes the bad case where only one of the phase spectrum bins used is shifted by a 2π factor.

3.2. Computation of Modulated Spectrum

The spectrum, at frequency f, of a modulated sinusoid whose set of parameters is $s_p = \{a_p, f_p, \phi_p, \Delta_p^a, \Delta_p^f\}$ is given by the classic short-time Fourier (STFT) formula:

$$A_{f_p}(f) = \sum_{n=-N/2+1}^{N/2-1} w[n] \cdot a[n] \cdot e^{\frac{-2\pi j f n}{N}} \tag{5}$$

where $a[n]$ is the discrete version of a (see Equation 3), $w[n]$ is the analysis window, and N is the size of the STFT.

4. STATIONARY PARAMETER CORRECTION

The complex spectrum is put out of shape by the intra-frame modulations Δ_p^a and Δ_p^f, thus severely corrupting the estimation of the stationary parameters a_p, Φ_p and f_p. This section deals with different techniques used in order to correct the bias in extracting the

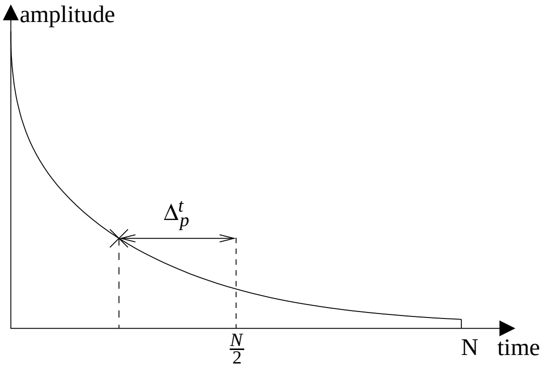

Figure 4: *Time coordinate evolution in a non linear amplitude model.*

amplitude a_p and phase Φ_p, provided that the modulations were correctly measured. In the remainder of this paper, the estimation of f'_p will not be addressed because the analysis method proposed in [8] already performs a frequency correction by considering the signal derivative.

4.1. Time Reassignment

Using a zero-phase window, in a stationary context, it is convenient to consider that the amplitude parameter a_p is equal to $a(t)$ where t is the time position of the center of the analysis window. In a non linear amplitude context, the point of maximum spectral energy (center of gravity) for a component p has a changing time coordinate (see Figure 4). The difference Δ_p^t between this point and the center of the window can be estimated using the time reassignment method. This method was presented by Auger and Flandrin [9] for a large variety of known time-frequency and time-scale distributions. It was introduced in the sinusoidal modeling context in [10, 6]. More precisely, we have:

$$\Delta_p^t = \Re\left\{\frac{X_{th;p}X^*_{h;p}}{|X_{h;p}|^2}\right\} \tag{6}$$

where $X_{th;p}$ denotes the short-time Fourier transform computed using the window multiplied by a time ramp ($w_{th} = w_h \cdot t$) and $X_{h;p}$ is the short-time Fourier transform computed using the original window w. The corrected amplitude is then given by:

$$a'_p = a_p \cdot 10^{\Delta_p^a \Delta_p^t/20} \tag{7}$$

During the time interval Δ_p^t, there is a phase travel due to periodic oscillation. The phase can be corrected in this way:

$$\phi'_p = \phi_p + 2\pi \cdot f'_p \cdot \Delta_p^t \tag{8}$$

Unfortunately, this method only takes the amplitude modulation into account. The following section presents an alternative solution that takes advantage of both the amplitude and frequency modulations.

4.2. Modulated Spectrum

In [8], the amplitude parameter is corrected by considering the continuous power spectrum of the analysis window:

$$a'_p = \frac{a_p}{W(|f'_p - f_p|)} \tag{9}$$

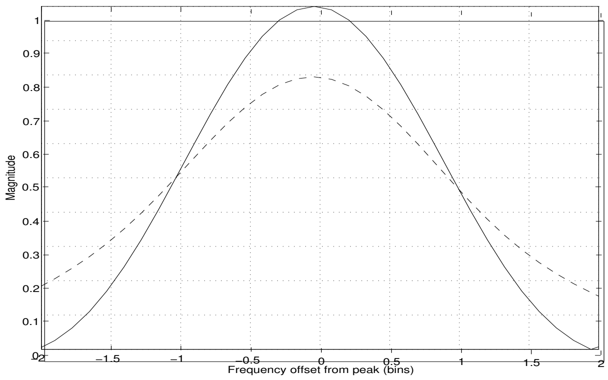

Figure 5: *Window power spectrum (line) versus modulated sinusoid power spectrum (dashed).*

where W is the power spectrum of the analysis window w and f_p is the frequency of the spectrum local maximum. Since the frequency has changed, we have then to estimate the new amplitude. Fortunately, in a stationary model, the power spectrum of a sinusoid has the shape of the window power spectrum. As can be seen on Figure 5, for a frequency correction of half a bin, the amplitude factor is approximately 0.8.

In the case of a strong modulation, the main lobe is "flattened" and so very different from the main lobe of the window power spectrum, inducing a bad estimation of the amplitude correction to be done. It is possible to improve this correction by considering the power spectrum $A_{f'_p}$ of a modulated sinusoid computed with Equation 5 instead of $W(|f'_p - f_p|)$. The corrected amplitude is then expressed by:

$$a'_p = \frac{a_p}{A_{f'_p}(f_p)} \tag{10}$$

For a frequency correction of half a bin and $\Delta_p^a = -2$ and $\Delta_p^f = 4$, the amplitude factor is of 0.7.

4.3. Comparative Results

This section presents comparative results for the different methods exposed. The modulations parameters Δ_p^a and Δ_p^f used to compute the corrections are extracted, *i.e.* errors presented below are not only attributable to the correction process but also to the modulations extraction process. The first part of Figure 6 shows errors as function of amplitude modulation and the second part as function of frequency modulation. Since the reassignment method does not handle frequency modulation, its error is not plotted.

Thanks to these modulation measures and more accurate stationary parameter extraction, we are now able to develop an efficient peak selection process.

5. PEAK SELECTION

Our peak selection process retains a bin if it is a local maximum and its frequency correction as defined in [8] is below one bin.

Unfortunately, this simple peak selection process retains also so-called "noisy" peaks (spectral manifestation of a process that can hardly be modeled by sinusoidal additive algorithms). Studies [6] have been made in order to decide if a peak is tonal (if it has

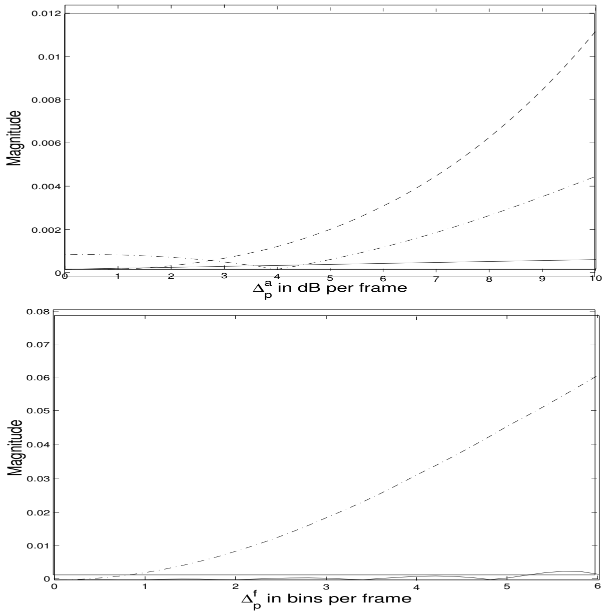

Figure 6: *Errors in amplitude estimation in function of amplitude (top) and frequency (bottom) modulations. Standard method (dot dashed), reassignment method (dashed) and proposed method (line).*

to be considered as a sinusoidal contribution) or not. In a way, these methods could seem redundant because spectral peaks are by nature sinusoids. Inversely, it is possible to synthesize "noisy" sounds with additive algorithms [11]. Thus, instead of having a tonal criterion, one would like to know if the set of parameters s_p is coherent and reliable, *i.e.* is a spectral manifestation of a process that can correctly be modeled by sinusoidal additive algorithms.

The first subsection presents frequency prediction based criteria for sinusoidal characterization. The second one is dedicated to spectrum shape criteria. In the second section, after an introduction on a simple criterion used in MPEG Layer II, we present the cross-correlation and its use in the domain. Its limitation lead us to introduce a new criterion that, thanks to modulation estimation, better separates "noisy" peaks and modulated peaks.

5.1. Frequency Prediction Criterion

Sinusoids, in a stationary model, should have coherent phase and frequency evolutions in time. In the general Advanced Audio Coding (AAC) Standard [12] and the Phase Derived Sinusoidality Measure (PDSM) in [6], frame to frame informations are used to separate spectral peaks. Since frequency is the derivative of the phase, a sinusoid should have coherent phase and frequency evolutions. This closeness of phase measurements evolution and frequency ones is evaluated and used as a peak selection criterion. It assumes a stationary behavior for the sinusoid so that its spectral contribution will stay in the same bin during the measurement process (2 to 3 frames).

5.2. Spectrum Shape Criterion

In the MPEG Layer I and II psychoacoustic model [13], an amplitude criterion is used to establish the tonality of a peak using a fixed amplitude relation between surrounding bins. More formally, a peak is a tonal component if the following relation is satisfied:

$$X(k) - X(k+j) \geq 7\text{dB} \qquad (11)$$

where j and k are bins index of an 1024 length analysis frame. For MPEG Layer I, j is chosen according to

$$j = \begin{cases} -2, 2 & \text{for} \quad 2 < k < 63 \\ -3, -2, 2, 3 & \text{for} \quad 63 \leq k < 127 \\ -6, -3, -2, 2, 3, 6 & \text{for} \quad 127 \leq k \leq 250 \end{cases} \qquad (12)$$

This criterion, despite its quite empirical values, has shown good results in frequency masking curve computing where precision is not so important.

The cross-correlation method has been used successfully in speech coding [14] as a voicing index and in sinusoidal modeling as a sinusoidal likeness measure [6]. These approaches have stationary assumption because the model used is a set of pure steady sinusoids (harmonic for the first case). Indeed, if a peak and its surrounding bins $H(\omega_k)$ have the same values as those of the analysis window translated in frequency and shifted in phase, they come from a pure steady sinusoid.

In real signals, this is rarely true so it is natural to use the cross-correlation $\Gamma_s(p)$ (s stands for stationary) to measure correlation between normalized $H(\omega_k)$ (real lobe computed via STFT, plotted with crosses on Figure 7) and $W(\omega_k)$ (plotted with a solid line), the normalized STFT of the analysis window w using a narrow bandwidth $[-B, B]$:

$$\Gamma_s(p) = \left| \sum_{k, f_k \in [f_p - B, f_p + B]} H(f_k) \cdot W(f_k) \right| \qquad (13)$$

where f_k and f_p are expressed in bins.

For harmonic sounds, if the fundamental frequency is modulated by a frequency factor Δ^f_p, the i-th harmonic is modulated by a factor $i\Delta^f_p$ corrupting so much the spectrum that it leads Marques in [5] to wonder if there were really harmonics in the high frequency of a speech spectrum. Indeed, the frequency modulation spreads the energy over a large number of bins, this phenomenon can be observed on the spectrogram of the highest harmonics in Figure 8. A solution to this problem is to "flatten" the spectrum by filtering the input signal to give it a constant fundamental frequency equal to its mean value along time to suppress the frequency modulation. This method relies on a good fundamental frequency estimation and a monophonic source context.

With an estimation of the modulations, we are able to get rid of these constraints. To estimate the coherence of the parameter set, instead of considering $W(\omega_k)$, we compare $H(\omega_k)$ to $A_{f'_p}(f_k)$ (modulated lobe plotted with circles on Figure 7), the spectrum of a modulated sinusoid computed with Equation 5 with extracted set of parameters s_p. $\Gamma_n(\omega)$ (n stands for non stationary) is then defined as:

$$\Gamma_n(f_p) = \left| \sum_{k, \omega_k \in [-B, B]} H(f_p + \omega_k) \cdot A'_{f_p}(f_p + \omega_k) \right| \qquad (14)$$

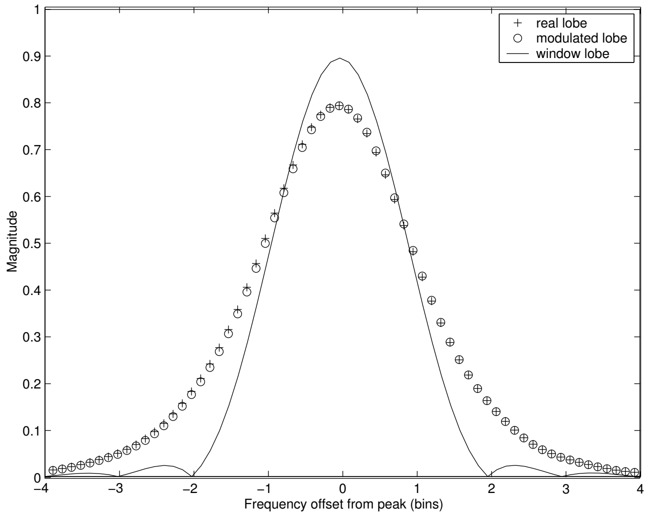

Figure 7: *Real modulated spectrum main lobe (crosses), compared with window lobe (solid line) and modulated lobe (circles) computed from Eq. 5 with measured parameters* $\{a_p, f_p, \Phi_p, \Delta_p^a, \Delta_p^f\}$.

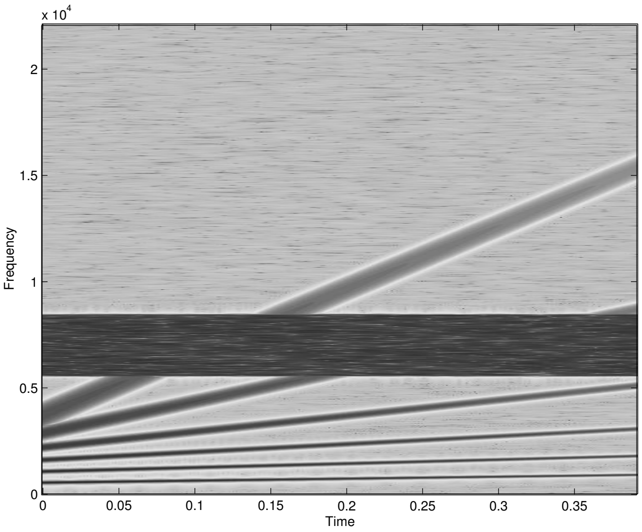

Figure 8: *Spectrogram of the analyzed signal*

5.3. Applications

Since the Γ_n criterion denotes the coherence and the reliability of the peak (see discussion at introduction of Section 5), we are able to robustly classify peaks. The applications can then be:

- **noise reduction** by setting a threshold below which the peak is considered as noisy,
- **component selection** in a sinusoidal coding framework,
- **ordering peaks** under a "trust" criterion to the next analysis stage.

As far as sinusoidal coding is concerned, component selection is mainly done in the literature by amplitude or loudness criteria [15]. This means that the peaks having the highest amplitude, Signal to Mask Ratio (SMR) or loudness are selected. Unfortunately, sinusoidal representations are usually associated in a hybrid framework with other representations (transient and noise). The sinusoidal section, which is often processed at first, will try to model the input signal even if the representation is not relevant. Γ_n selection allows us to select relevant components, *i.e.* which adequately model the deterministic part of the input signal.

5.4. Results

In this section, we present results for three component selection criteria. The analyzed sound is an increasing frequency and decreasing amplitude set of sinusoids mixed with a bandpass filtered white noise signal; its spectrogram is plotted in Figure 8.

The first peak selection process uses a simple amplitude criterion that orders peaks by decreasing amplitude (see Figure 9(a)). The second one uses the stationary correlation criterion (see Figure 9(b)); it sorts peaks in decreasing values of Γ_s. The last one uses the non-stationary criterion (see Figure 9(c)); it sorts peaks in decreasing values of Γ_n. Since the set of peaks is now ordered, we can choose the "best" peaks if we choose to retain only a few peaks at each frame ("best" means here that a peak has a criterion value highest than the other peaks in the set). In the three figures, at each frame, only the highest ten percents of the peaks detected by the analysis stage were plotted.

Because the noise level is very high in the limited band, the amplitude criterion selects peaks generated by the white noise. As far as quasi-stationary sinusoids (fundamental and lowest harmonics) detection is concerned, the standard correlation criterion Γ_s shows good results but fails quickly as the frequency modulation increases. The non-stationary criterion Γ_n handles this problem better, even though for very high harmonics, it does not seems possible to decide if a peak comes from a highly modulated sinusoid or from white noise.

6. CONCLUSION

In this paper, we have presented a way to handle non stationarity in a sinusoidal model. By using the algorithm of modulation extraction proposed by Masri in [2], we first correct traditional Fourier parameters. Thanks to these modulation measures and more accurate Fourier parameters, we proposed a new peak selection process which better differentiate modulated components from stochastic components. These considerations greatly improve the accuracy and robustness of spectral modeling.

7. REFERENCES

[1] Robert J. McAulay and Thomas F. Quatieri, "Speech Analysis/Synthesis Based on a Sinusoidal Representation," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 34, no. 4, pp. 744–754, 1986.

[2] Paul Masri, *Computer Modeling of Sound for Transformation and Synthesis of Musical Signals,* Ph.D. thesis, University of Bristol, 1996.

[3] Xavier Serra, *Musical Signal Processing,* chapter Musical Sound Modeling with Sinusoids plus Noise, pp. 91–122, Studies on New Music Research. Swets & Zeitlinger, Lisse, the Netherlands, 1997.

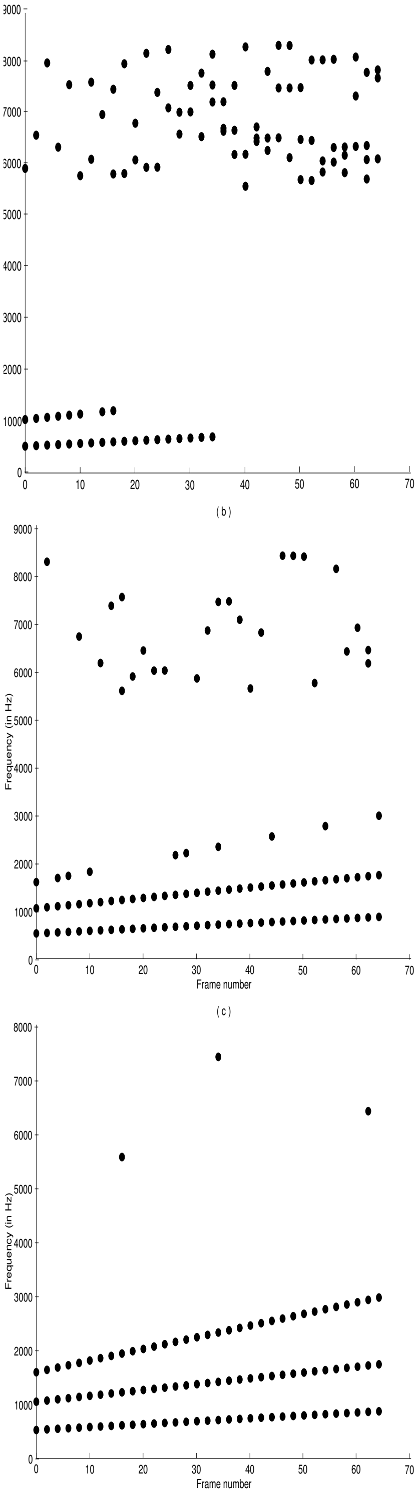

Figure 9: *Spectral peaks retained by three peaks selection process: amplitude criterion (a), stationary correlation criterion (b) and non-stationary one (c).*

[4] Sylvain Marchand and Robert Strandh, "InSpect and ReSpect: Spectral Modeling, Analysis and Real-Time Synthesis Software Tools for Researchers and Composers," in *Proceedings of the International Computer Music Conference (ICMC)*, Beijing, China, October 1999, International Computer Music Association (ICMA), pp. 341–344.

[5] J. Marques and L. Almeida, "A Background for Sinusoid Based Representation of the Voiced speech," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Tokyo, 1986, pp. 1233–1236.

[6] Geoffroy Peeters and Xavier Rodet, "SINOLA: A New Analysis/Synthesis Method using Spectrum Peak Shape Distortion, Phase and Reassigned Spectrum," in *Proceedings of the International Computer Music Conference (ICMC)*, Beijing, China, October 1999, International Computer Music Association (ICMA).

[7] Sylvain Marchand, *Sound Models for Computer Music (analysis, transformation, synthesis)*, Ph.D. thesis, University of Bordeaux 1, LaBRI, December 2000.

[8] Myriam Desainte-Catherine and Sylvain Marchand, "High Precision Fourier Analysis of Sounds Using Signal Derivatives," *Journal of the Audio Engineering Society*, vol. 48, no. 7/8, pp. 654–667, July/August 2000.

[9] François Auger and Patrick Flandrin, "Improving the readability of time-frequency and time-scale representations by the reassgnment method," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 43, pp. 1068–1089, May 1995.

[10] Kelly Raymond Fitz, *The reassigned Bandwith-Enhanced Method of Additive Synthesis*, Ph.D. thesis, University of Illinois, 1999.

[11] Pierre Hanna and Myriam Desainte-Catherine, "Influence Of Frequency Distribution On Intensity Fluctuation of Noise," in *Proceedings of the Digital Audio Effects (DAFx) Conference*. University of Limerick and COST (European Cooperation in the Field of Scientific and Technical Research), December 2001, pp. 120–124.

[12] ISO MPEG4, "ISO/IEC JTC1/SC29/WG11 FDIS 14496 Information technology - Generic Coding of Audio Visual Objects, Part 3 (MPEG-4)," 2001.

[13] ISO MPEG2, "ISO/IEC JTC1/SC29/WG11 Coding of Moving Pictures and Associated Audio for Digital Storage Media at up to About 1.5Mbit/s, standard n°11172, alias 'MPEG-1' ISO-MPEG," November 1992.

[14] Daniel W. Griffin and Jae S. Lim, "A New Model-Based Speech Analysis/Synthesis System," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Tampa, 1985.

[15] Heiko Purnagen, Nikolaus Meine, and Bernd Edler, "Sinusoidal Coding Using Loudness-Based Component Selection," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2002.