Signal Characterization in terms of Sinusoidal and Non-Sinusoidal Components

Geoffroy Peeters, Xavier Rodet
Ircam - Centre Georges-Pompidou Analysis/Synthesis Team, 1, pl. Igor Stravinsky, 75004 Paris, France
Geoffroy.Peeters@ircam.fr, Xavier.Rodet@ircam.fr

Abstract

This paper addresses the problem of signal characterization in terms of sinusoidal and non-sinusoidal components. A first measure of sinusoidality is reviewed. Drawbacks of this sinusoidal estimator are investigated and solutions proposed. Estimation of sinusoidality on non-stationary signal is then made on a pre-processed signal. A phase derived sinusoidality measure and the use of Re-estimated Spectra are introduced which allow deriving very precise and local characteristics. Finally, this characterization is used in a new synthesis scheme using Additive and PSOLA synthesis.

Introduction

Signal characterization in terms of sinusoidal/non-sinusoidal (S/NS) components plays an essential role in many applications today, such as speech coding, high-quality synthesis, sound labeling and so on. Depending on the topic, various terms are used : *voicing coefficient, harmonicity coefficient* - are most used in speech processing and refer to a measure of the activity of the vocal folds and often of signal's harmonicity -, *sinusoidality coefficient, tonality coefficient* - are most used in musical synthesis and refer to the closeness of a spectrum component to that of a pure sinusoid, whether or not the signal is harmonic.

Most of musical sounds (including speech) are produced by the repetition of pulses (this is the interpretation of PSOLA synthesis [3]). These pulses, depending on their similarity along time, can be represented, on long term, by sinusoids at various frequencies (this is the interpretation of Additive synthesis [9]).

In general, S/NS characterization can be a function of time or of time and frequency. According to one or the other choice and to assumptions on signal properties, many different methods have been proposed (see [8] for a review).

In this paper, we focus on IRCAM measure of sinusoidality and on the problems encountered while evaluating it. Notion of sinusoidality is the basis of a new synthesis scheme, which combines the benefits from Additive and PSOLA synthesis.

1 IRCAM measure of sinusoidality

With IRCAM sinusoidal characterization [9], as used for Additive synthesis, no particular assumption is made concerning signal properties. This characterization can be used for inharmonic sounds as well as for speech. The characterization is obtained by computing the complex correlation between the frequency shifted Fourier Transform (FT) of the analysis window and each peak of the Discrete Fourier Transform (DFT) of the signal. The sinusoidality coefficient is given by :

$$\Gamma(\omega) = \left| \sum_{k,|\omega-\omega_k|<W} S^0(\omega_k) H^0(\omega-\omega_k) \right| \quad (1)$$

where $H^0(\omega-\omega_k)$ denotes the normalized FT of the analysis window centered at frequency ω and sampled at discrete frequencies ω_k, $S^0(\omega_k)$ is the normalized DFT of the signal and W is the half-bandwidth of the analysis window's main-lobe. Γ ranges between 0 and 1. The value 1 is obtained for a noiseless steady sinusoidal component, while lower values indicate the presence of noise or of time-variable components.

Noting Ω a frequency for which Γ is close to 1, the amplitude, in a least square sense, of this peak is given by

$$A_\Omega = \frac{\Gamma(\Omega)||S||_\Omega}{||H||_\Omega} \quad (2)$$

where

$$\begin{cases} ||H||_\Omega^2 & = \sum_{k,|\Omega-\omega_k|<W} |H(\Omega-\omega_k)|^2 \\ ||S||_\Omega^2 & = \sum_{k,|\Omega-\omega_k|<W} |S(\omega_k)|^2 \end{cases} \quad (3)$$

while the phase is given by

$$\phi_\Omega = Arg[\Gamma(\Omega)] \quad (4)$$

It is interesting to see that Griffin and Lim's definition of sinusoidality, named "normalized error" [5], turn exactly into : $2(1-\Gamma(\omega)^2)$ [1].

1.1 Drawbacks of sinusoidality estimators

- Accuracy of sinusoidality evaluation relies on the assumption that the signal is stationary in the analyzed frame, which is rarely the case in audio signals. Variation of the signal in terms of fundamental frequency, local amplitude and spectral envelope inside the analyzed frame can partially or completely hide the presence of sinusoids. Frequency modulation increases the width of spectral lines (especially in high frequencies), while variation of spectral envelope and amplitude modifies the shape of spectrum lines. In section 2, we propose the evaluation of sinusoidality on a pre-processed signal. The pre-processing consists of giving the signal a constant fundamental frequency, a constant short term energy and a flat spectrum.

- Using correlation criterion (1), experiments have shown $\Gamma(\omega)$ reaching values close to 1 for non-sinusoidal components (side-lobes of sinusoidal components and some noisy peaks). Hence a function decision S/NS is hard to build. In section 3.1, we introduce a method based on phase derivative which avoids wrong peak detection.
- The use of Short Time Fourier Transform (STFT), lends to the usual problem of temporal versus frequency resolution. Short windows give little information about main-lobe shape so that $\Gamma(\omega)$ is not reliable. In section 3.3 we solve this problem by computing a short-time Re-estimated Spectrum.

2 Signal Normalization

Normalization has only to be applied in mixed portions of the signal where both sinusoidal and non-sinusoidal components coexist.

2.1 Normalization of fundamental frequency

The signal is processed so as to get rid of frequency modulation. It is assumed that frequency modulation of all frequencies is correlated to that of fundamental frequency : $\Delta(kf_0(t)) = k\Delta(f_0(t))$. In a first step, fundamental frequency $f_0(t)$ is estimated [4].

The signal is given a constant fundamental frequency equal to its mean value along time , $\overline{f_0(t)}$, using time-variable resampling [10]:

$$\tilde{s}_{re}(n') = \sum_{n=-\infty}^{+\infty} s(nT_e) \cdot \frac{F_{re}}{F_e} \cdot \mathrm{sinc}\left(\pi F_{re}(n'T_{re} - nT_e)\right) \tag{5}$$

where F_e and T_e are the original sampling rate and sampling period, T_{re} is the new sampling period which depends on the local fundamental frequency $T_{re} = T_e \cdot \overline{f_0(t)}/f_0(t)$, and where F_{re} is equal to

- F_e if $T_{re} < T_e$ (up-sampling),
- $1/T_{re}$ if $T_{re} > T_e$ (down-sampling).

2.2 Flattening of spectral envelope

The signal is processed so as to get a flat spectrum and a constant short term energy. This is done by applying the Inverse Filter and the gain factor obtained by Linear Prediction (LP). For better prediction, LP coefficients are computed on pitch synchronous windows.

2.3 Results

In Figure 1 we compare the measure of $\Gamma(\omega, t)$ on the original signal, and on the pre-processed signal. Changes in pitch (at time 0.15 s) and of formant position (at time 0.9 and 1.5 s) hide the presence of sinusoidal components on the original signal, while those sinusoidal components are, for the most part, detected on the pre-processed signal.

3 Improved sinusoidality measure

3.1 Phase derived sinusoidality measure

It is easy to see that for a secondary lobe, centered around ω, of a sinusoidal component and for a non-sinusoidal component ("noise"), centered around ω, the phase difference of the STFT at two instants is not proportional to ω. This property is the basis of an improved sinusoidality measure. To also take into account non-constant-frequency sinusoids, a linear model of frequency variation, i.e. a quadratic model of phase variation, is used to compute this measure of sinusoidality.

1. Consider STFTs evaluated at successive times. Peaks detected by use of (1) in three successive STFTs at t_1, t_2 and t_3 are grouped into tracks, according to a frequency distance criterion. Let $f(t_1), f(t_2), f(t_3)$ and $\phi(t_1), \phi(t_2), \phi(t_3)$ be the estimated frequency and phase of three peaks grouped in a given track.
2. The 2nd order phase polynomial of t which passes through the estimates $\phi(t_1), \phi(t_2), \phi(t_3)$ is determined.
3. The first derivative of this polynomial is used to compute an estimate of phase-derived instantaneous frequency f_p at time $t1, t2, t3$:
$$\begin{cases} f_p(t_1) &= \frac{1}{4\pi d}(4\phi_j(t_2) - 3\phi_i(t_1) - \phi_k(t_3)) \\ f_p(t_2) &= \frac{1}{4\pi d}(\phi_k(t_3) - \phi_i(t_1)) \\ f_p(t_3) &= \frac{1}{4\pi d}(\phi_i(t_1) - 4\phi_j(t_2) + 3\phi_k(t_3)) \end{cases} \tag{6}$$
where d is the distance in time between successive STFTs.
4. The Euclidean distance (ED) is then computed between estimated frequencies and phase-derived instantaneous frequencies :
$$e = \sqrt{\sum_{i=1}^{3}(f(t_i) - f_p(t_i))^2} \tag{7}$$
The ED e gives a measure of sinusoidality according to a linear frequency variation model.

As opposed to correlation criterion (1) which has a constant resolution on the frequency axis, the resolution obtained with ED criterion (7) decreases with frequency (as does phase precision). Therefore the final measure

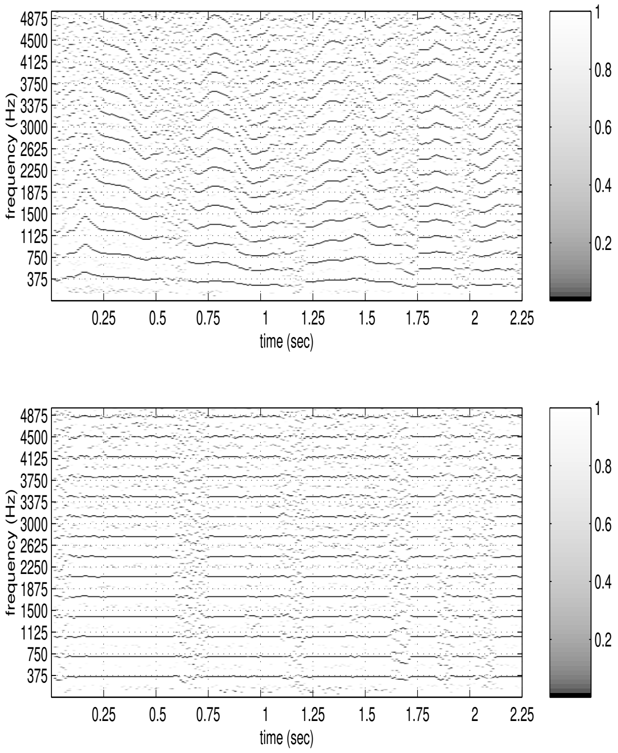

Figure 1: $1 - \Gamma(\omega, t)$ computed on a) original signal, b) pre-processed signal. Darker values indicate $\Gamma(\omega, t)$ close to 1. Signal : Tibetan female singing voice ($\overline{f_0}(t)$=345 Hz, LPC order 40, Blackman-Harris-3 window, 20ms)

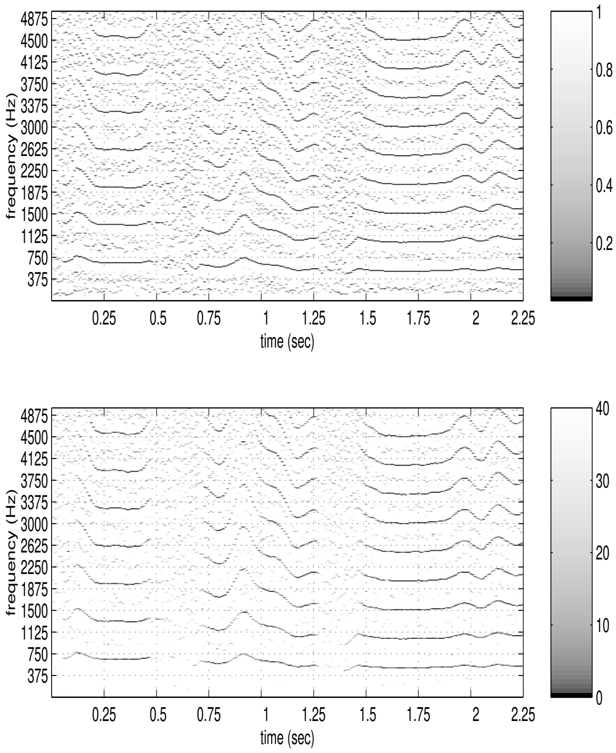

Figure 2: Sinusoidality estimation by a) $\Gamma(\omega, t)$, b) $e(\omega, t)/f$ with frequency reassignment. Signal : Chinese female singing voice ($\overline{f_0}(t)$=565 Hz, Blackman-Harris-3 window, 20ms, d=3.3ms)

is $e/\overline{f(t_i)}$. In Figure 2, we compare sinusoidality estimation using $\Gamma(\omega, t)$ and $e(\omega, t)/f$. Experiments have shown that $\Gamma(\omega)$ and e have different distributions (see Figure 3) which facilitates S/NS separation.

3.2 Frequency estimation improvement

Peaks detected as sinusoidal are then reassigned to the instantaneous frequency computed with (6) [2].

On Figure 4-a), we compare estimation of frequencies using correlation method and phase difference method for a signal composed of 3 harmonic sinusoids of linearly decreasing frequencies. The analysis is made using a 10 ms constant size Blackman-Harris-3 window. With time, fundamental period becomes greater in comparison to window size. Before time 0.23s (at time 0.23s main-lobes intersect at -13dB), estimation of frequencies using phase difference is more precise than estimation using correlation. After this time, due to the superposition of the main-lobes, both estimates suffer from the same imprecision ("-6dB intersection of main-lobes" occurs at time 0.47s). According to this, suggested size of analysis window is $3T_0$.

3.3 Neighboring peak subtraction

Finally, peaks are re-estimated by computing a complex Re-estimated Spectrum (RS). This RS is obtained by subtracting from each peak the influence of the neighboring peaks [6]. Subtraction should only be applied to sinusoidal peaks and thus a first knowledge of the peaks' sinusoidality is required. RS is computed by :

$$\mathrm{RS}_j(\omega_k) = S(w_k) - \sum_{i \in E_V - \{j\}} H(\omega_i - \omega_k) A_i e^{j\phi_i} \tag{8}$$

where $\mathrm{RS}_j(\omega_k)$ is the local RS of the j^{th} peak, $H(\omega)$ is the FT of the analysis window, ω_i, A_i and ϕ_i are the frequency, amplitude and phase of the i^{th} peak. E_V is the set of peaks detected as sinusoidal. The use of RS allows for detection of peaks which are hidden when analysis window size is insufficient to separate adjacent spectral lines.

Figure 4-b), shows the same comparison as in section 3.2 but using Re-estimated Spectra. Both estimation of frequency, using the correlation and the phase difference methods, are better. Time-limit of accurate detection is now at 0.33s where main-lobes intersects at -9dB. Suggested size of analysis window is $2.5T_0$.

4 Synthesis scheme using S/NS characterization

According to our definition of sinusoidality, we propose a new synthesis scheme. The signal is first separated

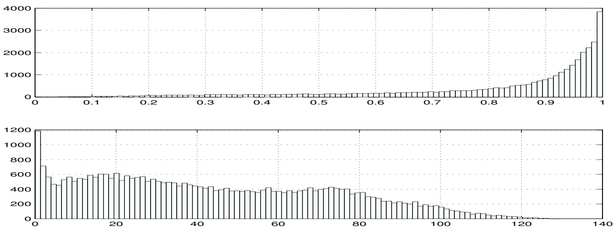

Figure 3: Distribution of a) $\Gamma(\omega, t)$, b) $e(\omega, t)/f$ for signal of Figure 2

into its sinusoidal and non-sinusoidal parts using time varying filtering.

- Additive synthesis is used for modification of the sinusoidal part (steady periodic),
- PSOLA method is used for modification of the non-sinusoidal part (noise and non periodic glottal pulses).

Importance is given to the synchrony between both signals. Modification of the noisy part superimposed on the harmonic part is performed with a method derived from the PSOLA scheme by slightly randomizing waveform positions during synthesis [7].

Examples of this synthesis will be given during the presentation of this paper.

Discussion and Conclusion

Sinusoidality estimation on a pre-processed signal permits a great part of hidden sinusoidal components to be recovered. Parameter values for the original signal can be derived from estimated values on pre-processed signal, at least for time and frequency. In the case where amplitude and phase are important (for re-synthesis), evaluation on pre-processed signal can be used as a guideline during evaluation of parameters on the original signal. Re-estimated Spectra have shown to improve parameter estimation, but require a previous knowledge about the sinusoidality of the spectrum's components, which is not always possible. Measure of sinusoidality based on phase-derivative is of great interest because it has properties which differ from those of correlation criterion. While it is less sensitive to noise and secondary lobe detection, it is more sensitive to local variations of the signal. Further work will consider these variations for the phase evaluation used by the model and will use the different methods simultaneoulsy.

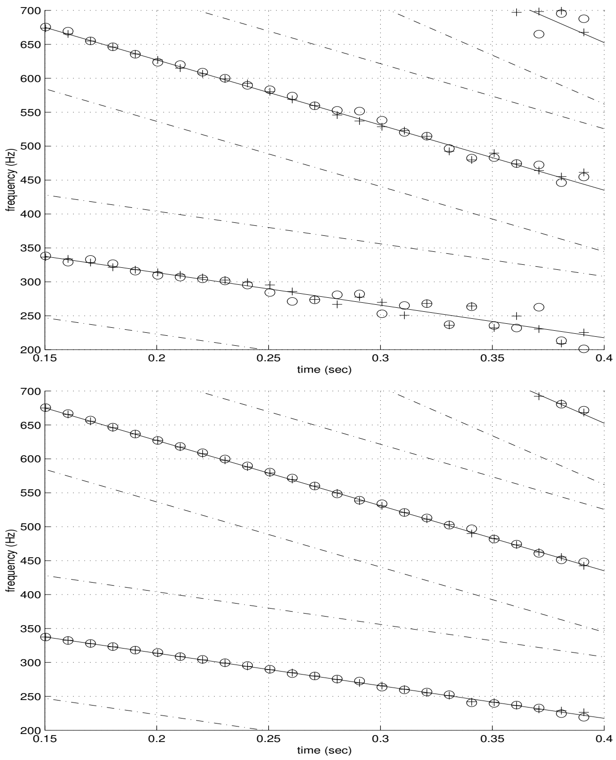

Figure 4: a) Frequency estimation using correlation method (o), phase difference method (+), real frequencies (continuous lines), "-6dB edge of main-lobes" (dash-dot lines), b) Same as a) but using Re-estimated Spectra (see section 3.2 and 3.3)

References

[1] M. Campedel. *Application du modèle sinusoides et bruit au codage, débruitage et à la modification des sons de parole.* PhD thesis, ENST, 1998.

[2] F. Charpentier. Pitch Detection using Short-Term Phase Spectrum. In *ICASSP*, 1986.

[3] F. Charpentier and M. Stella. Diphone Synthesis Using an Overlap-Add Technique for Speech Waveforms Concatenation. In *ICASSP*, 1986.

[4] B. Doval and X. Rodet. Estimation of Fundamental Frequency of Musical Sound Signals. In *ICASSP*, 1991.

[5] D. Griffin and J. Lim. Multiband Excitation Vocoder. *IEEE Trans. Acoust., Speech, Signal Processing*, 1988.

[6] J. Marques and L. Almeida. A Background for Sinusoid Based Representation of Voiced Speech. In *ICASSP*, 1986.

[7] G. Peeters. Analyse-Synthèse des sons musicaux par la méthode PSOLA. In *JIM (Journées Informatique Musicale*, 1998.

[8] G. Richard and C. d'Alessandro. Analysis/Synthesis and Modification of the Speech Aperiodic Component. *Speech Communication*, 1996.

[9] X. Rodet. Musical Sound Signal Analysis/Synthesis : Sinusoidal+Residual and Elementary Waveform Models. In *Proc. IEEE Symp. Time-Freq. and Time-Scale Anal.*, 1997.

[10] J. Smith. Bandlimited Interpolation - Introduction and Algorithm. Technical report, CCRMA, http://cm.stanford.edu/ jos/src/src.htm, 1998.