WIDE-BAND HARMONIC SINUSOIDAL MODELING

Jordi Bonada
Music Technology Group,
Universitat Pompeu Fabra
Barcelona, Spain
jordi.bonada@iua.upf.edu

ABSTRACT

In this paper we propose a method to estimate and transform harmonic components in wide-band conditions, out of a single period of the analyzed signal. This method allows estimating harmonic parameters with higher temporal resolution than typical Short Time Fourier Transform (STFT) based methods. We also discuss transformations and synthesis strategies in such context, focusing on the human voice.

1. INTRODUCTION

The concept of narrow or wide-band analysis of a periodic signal relates to the ratio between frequency resolution and fundamental frequency, and therefore to the number of periods covered by the analysis window. Narrow-band analysis takes several periods so that in quasi-stationary conditions harmonics appear as clear and separated peaks in the spectrum. By contrast, wide-band analysis uses one or two periods so that the frequency distance between harmonics is similar to the spectral resolution, and therefore the spectra produced by each harmonic affects significantly its neighbor harmonics, which complicates the estimation of individual frequency components. Moreover, narrow-band analyses perform with lower temporal resolution than wide-band analyses. In general, algorithms based on modeling and tracking spectral peaks use a narrow-band approach to facilitate the detection of individual frequency components. This is the case of phase-locked vocoder [1] and sinusoidal models [2]. On the other hand, typical time-domain algorithms such as Time Domain Pitch-Synchronous Overlap-Add (TD-PSOLA) [3] or Linear Prediction Pitch-Synchronous Overlap-Add (LP-PSOLA) [4] use two period long frames, so they work in wide-band conditions.

As pointed out in [3] regarding the spectral interpretation of the TD-PSOLA algorithm, if a short-time signal $x(n)$ is repeated at a rate of f_0 then it can be shown that the discrete-time STFT of the resulting signal using a window function $h(n)$ is given by the convolution of the response of the window function $H(f)$ by the spectrum of $X(f)$ sampled at harmonic frequencies $f_k = kf_0$, i.e.

$$Y(f) = \sum_k H(f_k - f)X(f_k) \tag{1}$$

One drawback of the TD-PSOLA approach is that $x(n)$ is itself a windowed signal of several periods length,

$$x(n) = s(n)w(n) \tag{2}$$

where $w(n)$ is the window and $s(n)$ the signal that's being analyzed. Therefore, $X(f)$ is the convolution of $W(f)$ and the signal's true spectrum $S(f)$. This means that the sampled spectrum is a smoothed version of the true signal's spectrum, with a spectral resolution determined by the width of the window function $W(f)$, actually wider than several harmonics. This happens even in the case of a pure periodical signal. Although this is an inherent problem, its effect can be minimized by inverse filtering the analyzed signal and processing the residual as in LP-PSOLA.

The mentioned PSOLA algorithms don't allow modifying individual harmonic components. However, our interest is to model those frequency components in wide-band conditions and at the same time be able to transform them independently, therefore combining the good temporal resolution of typical time-domain techniques with the flexibility of frequency-domain methods. In the following sections we present the proposed method in detail.

2. WIDE-BAND HARMONIC ANALYSIS

Our intention is to estimate the parameters of the harmonics of a periodic signal $s(n)$ in the widest possible band conditions by means of a STFT. We assume that the period of the signal has been already estimated by any appropriate technique (e.g. [5]). Let's define $s(n)$ as a stationary periodic signal sampled at a rate of f_s, composed of $T/2$ sinusoids with constant amplitude, frequency and initial phase values, and a known fundamental period of T samples

$$s(n) = \sum_{k=1}^{T/2} a_k \cos\left(2\pi \frac{f_k}{f_s} n + \theta_k\right) \quad , \quad f_k = \frac{kf_s}{T} \tag{3}$$

The discrete-time STFT of $s(n)$ using a rectangular window $w_R(n)$ is given by

$$x(n) = s(n)w_R(n) \tag{4}$$

$$X(f) = \sum_k W_R(f_k - f)S(f_k) \tag{5}$$

where f_k denotes the harmonic frequencies, and $S(f)$ and $W_R(f)$ are respectively the Discrete Time Fourier Transform (DTFT) of $s(n)$ and $w_R(n)$. Thus the value of $X(f)$ at an arbitrary frequency f is the result of the contribution of all harmonic components multiplied by the transform of the window evaluated at the frequency difference $f_k - f$.

The DTFT of a normalized rectangular window of N samples is given by

$$w_R(n) = \begin{cases} 1/N & for \quad 0 \le n \le N-1 \\ 0 & for \quad n \notin [0, N-1] \end{cases} \tag{6}$$

$$W_R(f) = \sum_{n=0}^{T-1} \frac{1}{N} e^{-j2\pi \frac{f}{f_s} n} = e^{-j\pi \frac{f}{f_s}(N-1)} \frac{\sin\left(\pi \frac{f}{f_s} N\right)}{N \sin\left(\pi \frac{f}{f_s}\right)} \tag{7}$$

Note that it has zeros at frequencies $f_g = (g f_s)/N$, $g = 1, 2, ..., N-1$.

Since the fundamental period is known, the harmonic frequencies are also known ($f_k = k f_s / T$) and therefore we would like to arrive to $X(f_k) = S(f_k)$. Observing eq. (5), this will be true if the energy contribution of a given harmonic to other harmonic frequencies is zero. In other words,

$$W_R\left(\frac{k f_s}{T}\right) = 0 \quad \forall k \in [1, T-1] \tag{8}$$

The previous condition will happen whenever the length of the rectangular window is a multiple of the signal's period ($N = gT$, $g \in \mathbb{N}$). Therefore, the maximum widest-band condition is achieved for $N = T$, when the rectangular window covers exactly one period of the signal.

In practical implementations it is inefficient to compute the DTFT. Instead, the Discrete Fourier Transform (DFT) is used, which actually samples the DTFT at frequencies equidistant by f_s/N. Denoting the DFT of $x(n)$ as $\bar{X}(k)$ we obtain

$$\bar{X}(k) = \sum_{k=0}^{N-1} x(n) e^{-j2\pi \frac{k}{N} n} = X\left(\frac{k f_s}{N}\right) \tag{9}$$

For $N = T$ we obtain

$$\bar{X}(k) = X\left(\frac{k f_s}{T}\right) = X(f_k) = S(f_k) \tag{10}$$

This means that each bin of the DFT actually corresponds to one harmonic of $s(n)$, and that from its complex value we can simply compute harmonic parameters as

$$\begin{aligned} f_k &= \frac{k f_s}{T} \\ a_k &= \left|\bar{X}(k)\right| \qquad k = 1, ..., \frac{T}{2} \\ \theta_k &= \angle \bar{X}(k) \end{aligned} \tag{11}$$

This way we can efficiently estimate the harmonic parameters from one individual signal period without spectral smoothing due to the windowing process. For computational efficiency, it is preferred to use the Fast Fourier Transform (FFT) algorithm for computing the DFT. However, if T is not a power of 2, using the FFT algorithm will often[1] require to zero-pad the signal and this will modify the frequency of the spectral bins so that they won't correspond anymore to a harmonic. Moreover, the FFT is limited to an integer number of samples but not the period T which is a real value.

[1] Some implementations of the FFT allow non-power-of-2 window sizes at the cost of some increased computation

2.1. Non-integer size FFT

There are several ways for computing the spectrum of a non-integer number of samples using the FFT algorithm:

- PERIODIZATION: one period of the input signal is windowed with $w_R(n)$, and repeated several times at the rate defined by T so that the FFT buffer of length M covers in the end several periods. The repetition implies interpolating both the signal samples and the window function. Then the resulting signal $s_r(n)$ is windowed by an analysis window function $w_A(n)$, and the spectrum obtained is actually the convolution of such analysis window response $W_A(f)$ by the spectrum of $S_r(f)$ sampled at harmonic frequencies, i.e.

$$X_r(f) = \sum_k W_A(f - f_k) S_r(f_k) \tag{12}$$

  where actually $S_r(f)$ is the STFT of length T. In general, the frequencies of the spectral bins don't correspond to the harmonic frequencies but to

$$\bar{X}_r(b) = \sum_{b=0}^{M-1} x_r(n) e^{-j2\pi \frac{b}{M} n} = X_r\left(\frac{b f_s}{M}\right) \tag{13}$$

  Therefore estimating harmonic parameters (i.e. frequency, amplitude and phase) requires interpolating the spectral bins around harmonic peaks. Besides, zero-padding can help to improve the estimation accuracy. This method is depicted in Figure 1, although a rectangular window of T samples is not used but a longer one so to overlap samples at borders and therefore avoid discontinuities. In the following subsection it will be shown the need of this overlapping method.

- UPSAMPLING: Another way of computing the STFT of a non-integer number of samples is to upsample the input signal so that one period matches the closest FFT size M, i.e. $M = 2\,\hat{}\,(\lfloor \log_2(T) \rfloor + 1)$. Downsampling is not desirable in this case because some of the higher harmonics should be removed to avoid aliasing. Computing the FFT of the upsampled signal $s_u(n)$ would result into

$$\begin{aligned} \bar{X}_u(k) = X_u\left(\frac{k f_s}{T}\right) &= \sum_g W_R\left(\frac{k f_s}{T} - f_g\right) S_u(f_g)\Big|_{f_g = \frac{g f_s}{T}} \\ &= S_u(f_k) \end{aligned} \tag{14}$$

  where $S_u(f)$ is the STFT of length T and only the first bins up to $T/2$ would be relevant.

Ideally both methods would output exactly the same results. However, due to inaccuracies of the sample and spectral interpolation methods used some differences are expected, although insignificant.

2.2. Inter-harmonic energy contribution

We saw before that in order to achieve $\bar{X}(k) = S(f_k)$ the energy contribution of each harmonic to other harmonic frequencies should be zero. Thus, in order to have an initial evaluation of the goodness of the proposed approach, we explore the inter-harmonic energy contribution by computing the one-period-STFT of a sinusoid with the upsampling method. This gives us a measure of the noise present at other harmonic frequencies.

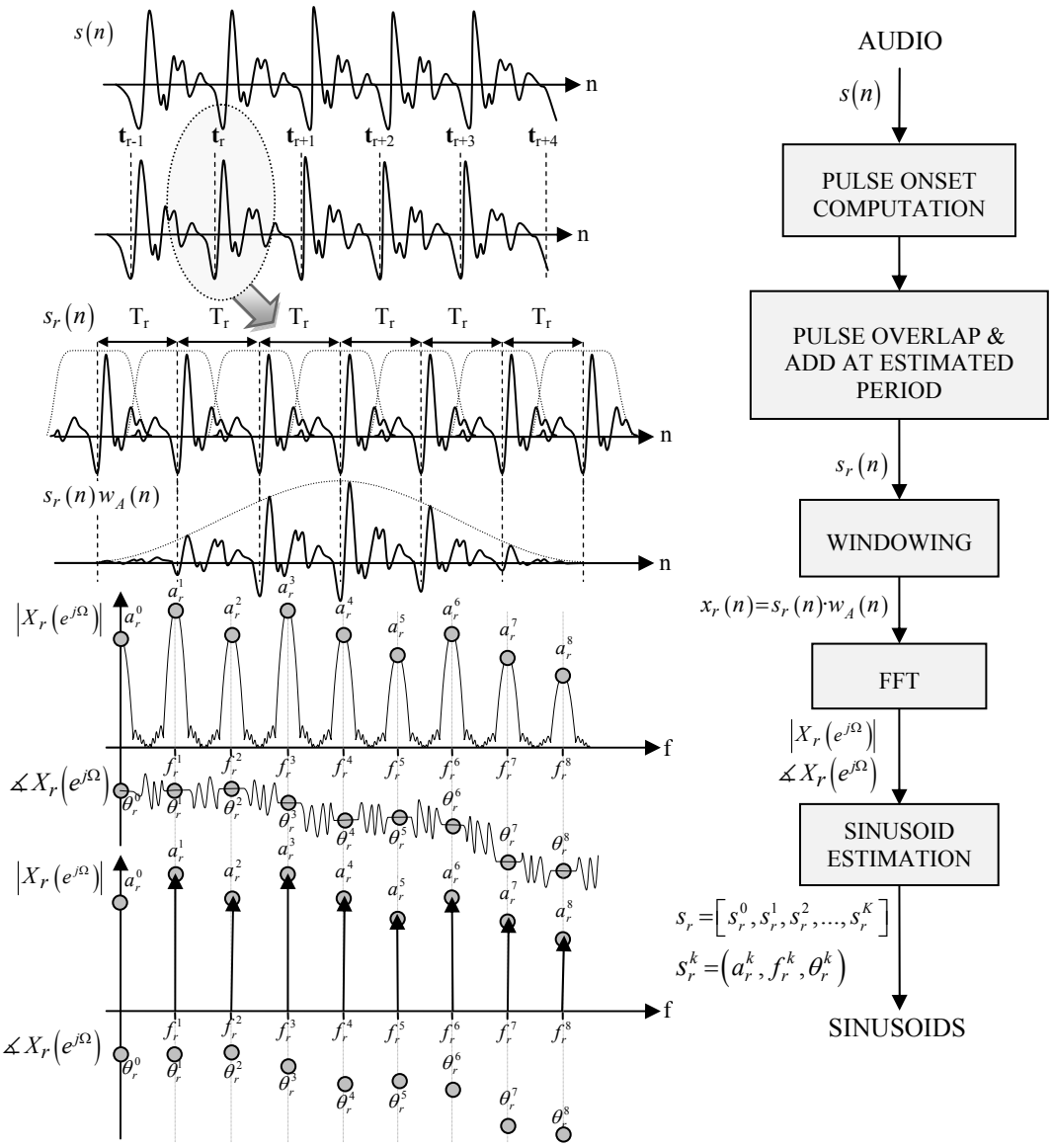

Figure 1 *Block diagram of the analysis phase using the periodization method*

Figure 7a-f show the one-period-STFT of a signal containing only one sinusoid at the fundamental frequency, whereas Figure 7g-h show the result when the sinusoid frequency corresponds to different multiples of the fundamental frequency. Inaccuracies introduced by any of the estimators or interpolation methods will degrade the analysis performance. We consider the following aspects:

- UPSAMPLING: the upsampling process is performed using a polyphase implementation. Figure 7a shows the case of integer period values where contributions are negligible since they fall below -100dB. Periods between 65 and 127 are upsampled to have a length of 128 samples.
- NON-INTEGER PERIODS: In Figure 7b we observe negligible contributions for several real-valued periods between 64 and 128.
- PITCH ESTIMATION ERRORS: Figure 7c shows the contributions for pitch estimation errors up to 20 cents. The contribution to the adjacent harmonic goes from -65.45dB/1cent to -36.73dB/20cents. The reason for this bias relies both in the discontinuity between borders of the STFT input signal and the fact that bins frequencies depart from harmonic frequencies. In (d) we see how the numbers can be greatly improved by interpolating the values around borders in the way shown in Figure 1.
- NON-STATIONARY SIGNALS: Figure 7e shows the results in the case of both non-stationary sinusoids and pitch estimation errors. The sinusoid frequency shifts approximately from 125 to 133Hz along the analysis window, and the estimation errors go from 0 to 20 cents. Obviously the best case is when the fundamental frequency is well detected, with contributions around -50 and -60dB for the two closest partials, and slowly decaying to -90dB for the 15th harmonic. These values are good enough for real world signals. However, the worst case of 20 cents is not that good. The contributions range from -38dB for the 2nd harmonic to -45dB for

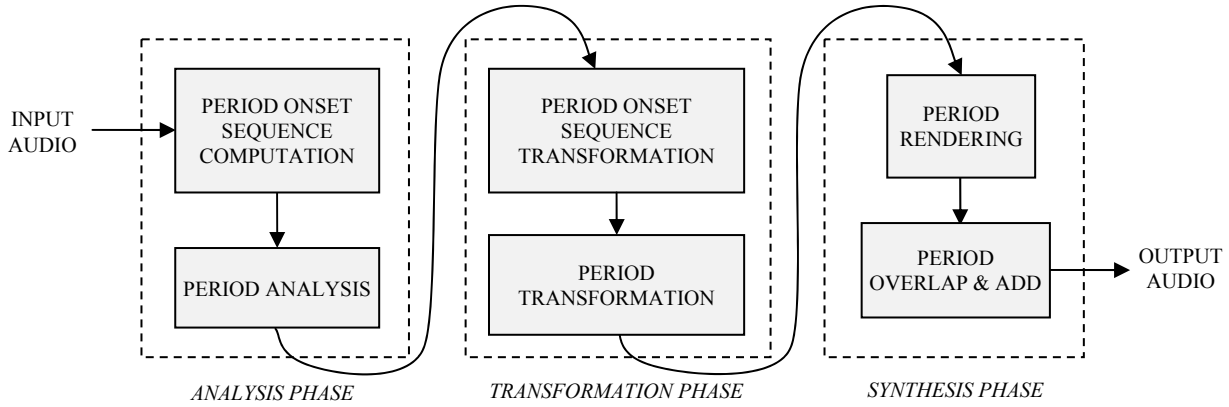

Figure 2 *Processing framework*

the 15th harmonic. Applying interpolation around borders as previously exposed the results can be greatly improved, as shown in Figure 7f, with values falling from -39 to -80 dB, good enough for practical uses.

- OTHER HARMONICS THAN FUNDAMENTAL: Figure 7g shows the comparison between the contribution from fundamental and higher harmonics (2nd, 4th and 8th), with and without pitch estimation errors of 20 cents. Clearly, the contribution increases significantly for higher harmonics. For instance, the 8th harmonic estimated with a bias of 20 cents contributes to the surrounding ten harmonics with more than -40dB. Overlapping around the borders improve the results, as shown in Figure 7h, increasing significantly the contribution decay. It is important to mention that most common musical sounds and human voice tend to present spectra with energy decaying along frequency. Therefore, the observed increase of inter-harmonic contribution along frequency is not that relevant for achieving good results.

2.3. Sinusoidal modeling

Figure 3 shows the spectra obtained from a synthetic signal using those methods and a regular narrow-band STFT. The input signal consists on ten sinusoids whose frequencies are multiples of the lower one. The fundamental frequency increases along time, as can be seen in the top view (a) where periods on the left are longer than those on the right. (b) shows the waveform resulting of repeating the period in the center, whereas in (c) we see the upsampled period. Finally, (d) and (e) show respectively the amplitude and phase spectra of the previous signals, where (a) is drawn with dashed lines, (b) with solid lines, and (c) with thick solid lines. The STFT of (a) presents clear amplitude peaks only at the lower frequency harmonics, getting noisy for higher frequencies due to the non-stationary nature of the analyzed signal. Instead, the STFT of (b) presents clear peaks at expected harmonic frequencies, but also above 5Khz where no harmonics are present. This is explained by the inter-harmonic energy contributions previously discussed. On its turn, the STFT of (c) has one bin per harmonic with values matching those of (b). The exact harmonic parameters values are displayed as circles. Clearly, (b) and (c) STFTs approximate much better the input signal than (a). For instance, (a) shows bias of up to -10 dB and 0.5 radians for harmonics above 3Khz.

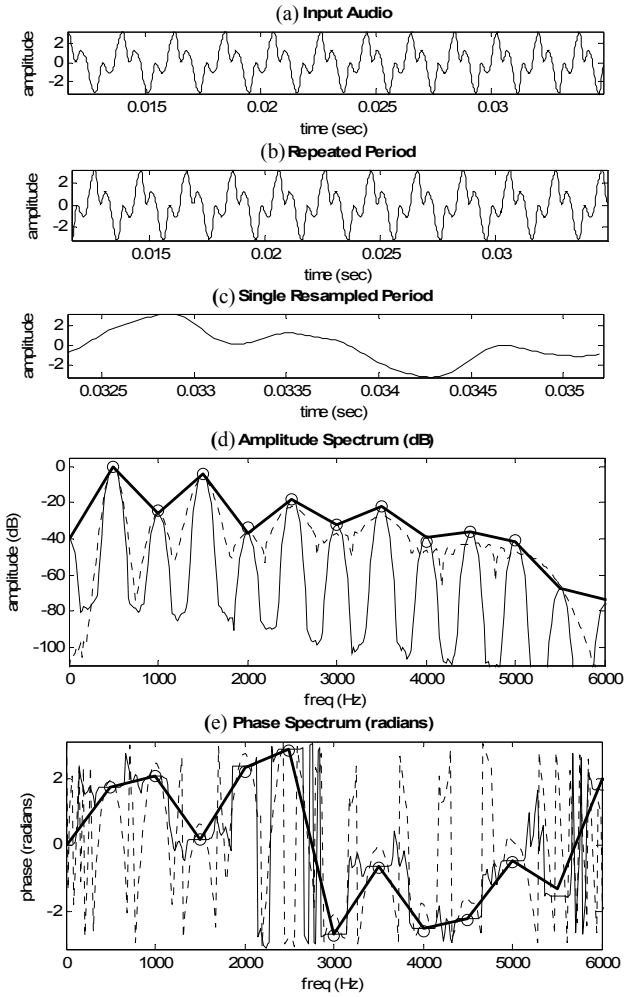

Figure 3 *Wide-band versus narrow-band analysis of a synthetic signal*

With the two methods presented in section 2.1 the resulting audio signals are purely periodic. Therefore, their spectra can be perfectly represented by a set of stationary sinusoids. It is then

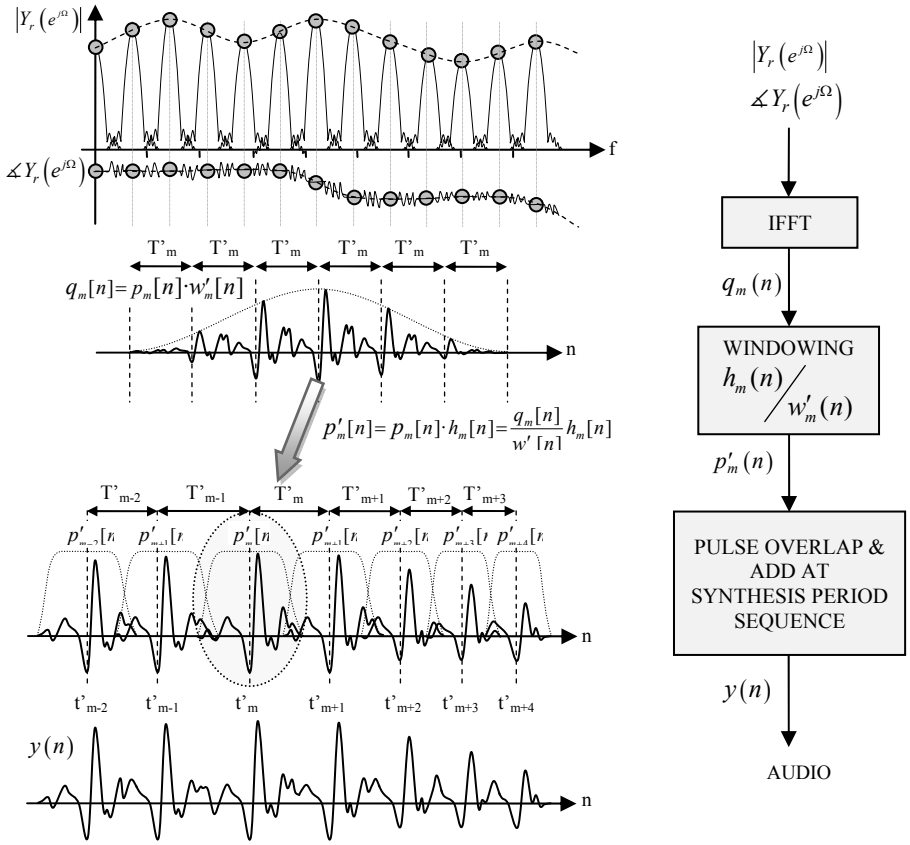

Figure 4 *Sinusoidal synthesis using the periodization method*

straightforward to use a sinusoidal model for the proposed wide-band analysis. Moreover, since the harmonic frequencies depend only on the estimated fundamental period, there is no need to use any complex method for building the harmonic trajectories along consecutive periods, but simply to connect the harmonics with the same index.

3. PROCESSING FRAMEWORK

The proposed method can be divided in three main phases, namely analysis, transformation and synthesis, as shown in Figure 2. In the analysis phase, the input signal is segmented into consecutive periods which are modeled with a set of sinusoids as already explained in the previous section. In the following subsections we detail both transformations and synthesis phases, and then discuss how the proposed method is adapted to the case of the human voice and to unvoiced signals.

3.1. Synthesis

Figure 4 and Figure 5 show the steps involved in the synthesis phase using the periodization method. For each m^{th} period to synthesize, its spectrum $Y_r\left(e^{j\Omega}\right)$ is rendered by convolving the synthesis window transform $W'_m(f)$ by each of the harmonics. It is sufficient to use a small number of coefficients per harmonic, as proposed in [6]. Next, an IFFT is applied to obtain the time domain signal $y_m(n)$, consisting of a windowed sequence of identical periods at the synthesis pitch rate T'_m. Then this signal is windowed by $h_m(n)/w'_m(n)$ obtaining $p'_m(n)$, where $w'_m(n)$ is the window whose transform was used in the sinusoidal rendering process, and $h_m(n)$ is the synthesis overlapping window. All the synthesis periods are then overlapped according to the synthesis period onset sequence and the signal $y(n)$ is obtained.

It is also possible to use an analogous synthesis method to the upsampling process used in the analysis. In that case each spectral bin of $Y_r\left(e^{j\Omega}\right)$ corresponds uniquely to one harmonic, and the IFFT computes the upsampled version of the synthesis period, $y_m(n)$. Therefore, $y_m(n)$ has to be downsampled to the analysis sampling rate f_s, and then overlapped with the other synthesized periods following the synthesis period onset sequence.

Both methods are equivalent and generate almost the same signal, with insignificant differences introduced by the downsampling and sinusoidal rendering steps. Although the second method is usually more efficient in terms of computation, whenever inharmonic components are being synthesized the first method is the most efficient one. On the other hand, it's important to point out that the input signal cannot be perfectly reconstructed when no transformations are applied due to the overlapping applied at the borders of the analysis window (see Figure 1). However, informal listening tests have shown that in most cases the synthesized signal is indistinguishable from the original one.

3.2. Transformations

There are two main types of transformations, the ones related to the period onset sequence and the ones related to each individual period, as depicted in Figure 2. Thinking of the traditional

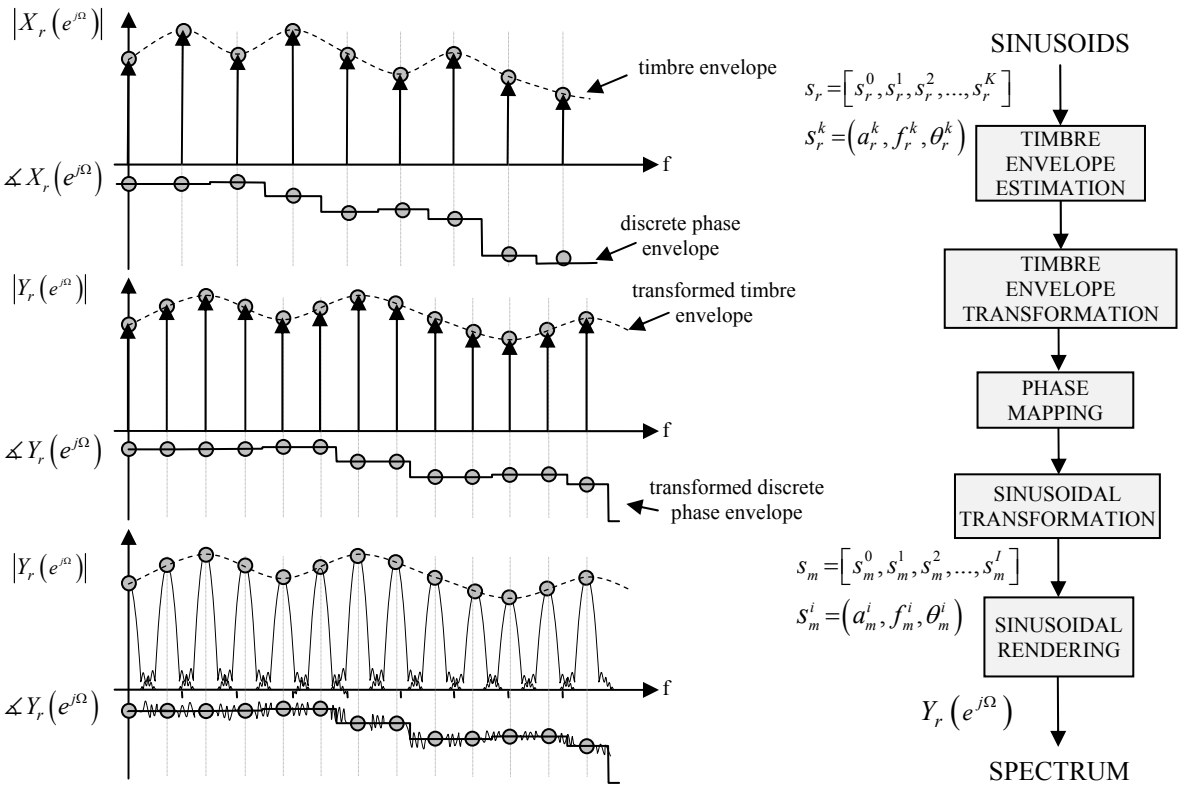

Figure 5 *Period transformations*

source-filter voice model, we could say that the former group of transformations are related to the voice source whereas the latter to the vocal tract. Traditional transformations such as time-scaling and pitch transposition involve scaling the period onset sequence, and repeating, removing or interpolating periods, in the same way as done in typical time-domain PSOLA techniques. However, pitch transposition also requires modifying the harmonic components of each period in order to match the target fundamental frequency, although phase continuation is not needed since consecutive period onsets are distant by one period.

Conversely, timbre transformations work as in typical frequency-domain techniques, by modifying the individual frequency components as depicted in Figure 5. Initially, the spectral envelope is computed by interpolation of the estimated sinusoids and properly modified. Preferably both spectral and phase envelopes should be modified by the same scaling function, with the aim of preserving the resonance-to-phase relationship. If the phase envelope is interpolated then the inherent phase wrapping has to be considered. Finally, synthesis sinusoidal components are computed out of the target fundamental frequency and both timbre and phase envelopes. Inharmonic components can be synthesized as well, although require to propagate phase so to avoid discontinuities in the synthesized signal. Figure 5 shows an example of transposition to a lower pitch and timbre stretching. Note that the phase envelope is not interpolated but a mapping function is used to determine which input harmonic's phase is used for each output harmonic.

3.3. Voice signals.

In a simplified model of voice production, a train of impulses (i.e. glottal pulses) at the pitch rate excites a resonant filter (i.e. the vocal tract). According to this model, a speaker or singer changes the pitch of his voice by modifying the rate at which these impulses occur. An interesting observation is that the shape of the time-domain waveform signal around the impulse onsets is roughly independent of the pitch, but it is dependant mainly on the impulse response of the vocal tract. This characteristic is called shape invariance. In terms of frequency domain, this shape is related to the amplitude, frequency and phase values of the harmonics at the impulse onset times. Thus, if a given transformation method is able to preserve the phase relation at analysis frame times, then it is desirable (in order to obtain the best processing quality) that analysis times match those impulse onsets mentioned before. In order to illustrate this issue, one representative example is shown in Figure 6. Left and middle figures correspond to spectra obtained when the analysis window is centered at the voice pulse onset and between two pulse onsets. In the middle figure new harmonics (in gray) are added to perform one octave down transposition. In the right figure, it is shown the spectrum of the transformed signal with the window centered at the voice pulse onset. The resulting doubled phase alignment adds an undesired roughness characteristic to the voice signal. Besides, the waveform doesn't have one voice pulse per period as expected, but two with strong amplitude modulation

Therefore, as depicted in Figure 1, voice pulse onsets are determined before performing the wide-band analysis, so that the analyzed periods are centered on them. Different algorithms detect voice pulse onsets relying on the minimal phase characteristics of the voice source (i.e. glottal signal) (e.g. [7]). Following this same idea we proposed in [8] a method to estimate the voice pulse onsets out of the harmonic phases based on the property that when the analysis window is properly centered, the un-

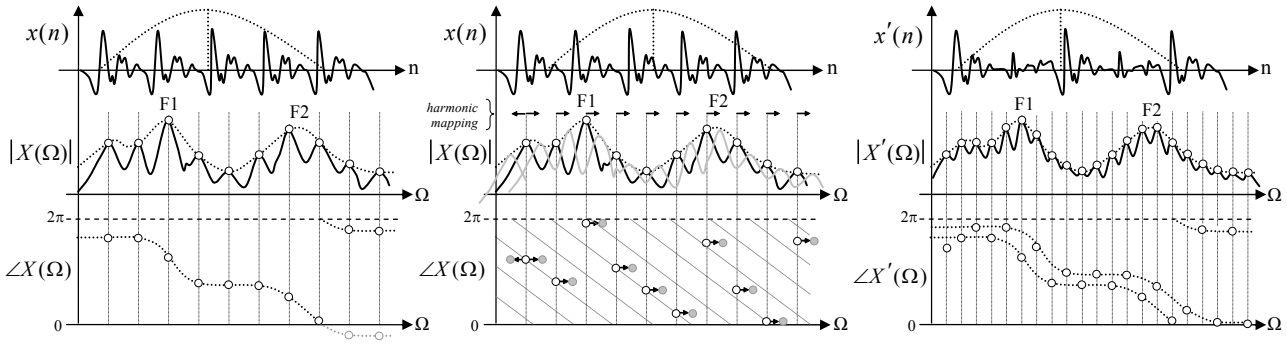

Figure 6 *Relation between analysis times and voice pulse onsets*

wrapped phase envelope defined by the harmonics is nearly flat with shifts under each formant, thus being close to a maximally flat phase alignment (MFPA) condition. This is the method we have used in our experiments.

On the other hand it is interesting to point out that both harmonic and aspirated noise components present in voiced utterances are represented exclusively by sinusoids. Actually, since the analysis is performed pitch-synchronously, the noise produces differences between consecutive periods that result into amplitude and phase modulations of the detected harmonics.

3.4. Unvoiced signals

Unvoiced signals can be processed as if they were voiced by assigning an arbitrary fundamental frequency. However, even when no transformations are applied, the analysis fundamental frequency can be slightly perceived in the synthetic signal. This can be avoided and a perfect reconstruction achieved by using a shorter period value for the period onset sequence than for the period analysis, so that from the signal obtained by the IFFT only the section not affected by the border overlapping is used to compute the output signal.

4. CONCLUSIONS

We have presented in this paper a method for wide-band sinusoidal-based modifications of harmonic signals, which combines the control of the period sequence typical of time-domain techniques with the flexibility of transformation found in frequency-domain techniques. The proposed method has been implemented as a real-time VST plug-in. Informal listening test show that the sound quality of the algorithm is at least as good as that of state-of-the-art PSOLA methods. Audio examples can be obtained from [9]. As future work we plan to perform a perceptual test to compare the sound quality of the proposed algorithm with other techniques. Besides, we plan to compare the estimation of non-stationary harmonic sinusoidal components with state-of-the-art methods such as the one in [10]. Finally, another interesting idea to explore is to separate and independently transform harmonics and surrounding noise by considering the former as slow varying signals compared to the latter.

5. ACKNOWLEDGMENTS

I would like to thank the MTG colleagues, and especially Esteban Maestre for the many fruitful discussions.

6. REFERENCES

[1] J. Laroche, "Frequency-Domain Techniques for High-Quality Voice Modification", *Proc. of the 6th Int. Conference on Digital Audio Effects.* London, UK, september, 2003.

[2] R. J. McAulay and T. F. Quatieri, “Speech analysis/synthesis based on a sinusoidal representation,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP–34, pp. 744-754, Aug. 1986.

[3] C. Hamon, E. Moulines, and F. Charpentier, “A Diphone Synthesis System Based on Time-Domain Prosodic Modifications of Speech”, in *Acoustics, Speech, and Signal Processing ICASSP*, Glasgow, UK, May 23-26, 1989, pp. 238-241.

[4] E. Moulines, C. Hamon, and F. Charpentier, “High-quality prosodic modifications of speech using time-domain overlap-add synthesis”, *Twelfth GRETSI Colloquium. Juan-les-Pins*, France, 1989.

[5] A. Cheveigné, H. Kawahara, "Comparative evaluation of F0 estimation algorithms", *7th European Conference on Speech Communication and Technology*, *EUROSPEECH-2001*, 2451-2454, Denmark, 2001.

[6] X. Rodet and P. Depalle, “A New Additive Synthesis Method using Inverse Fourier Transform and Spectral Envelope.” *Proc. Int. Computer Music Conference*, pp. 410-411, 1992.

[7] R. Smits and B. Yegnanarayana, "Determination of Instants of Significant Excitation in Speech using Group Delay Function." *IEEE Transactions on Speech and Audio Processing*, 1995.

[8] J. Bonada, "High Quality Voice Transformations based on Modeling Radiated Voice Pulses in Frequency Domain." *Proc. of the 7th Int. Conference on Digital Audio Effects.* Naples, Italy, October, 2004.

[9] http://mtg.upf.edu/~jbonada/wbhsm

[10] A. Röbel, “Frequency Slope Estimation and its Application for Non-Stationary Sinusoidal Parameter Estimation.” *Proc. of the 10th Int. Conference on Digital Audio Effects.* Bordeaux, France, September, 2007.

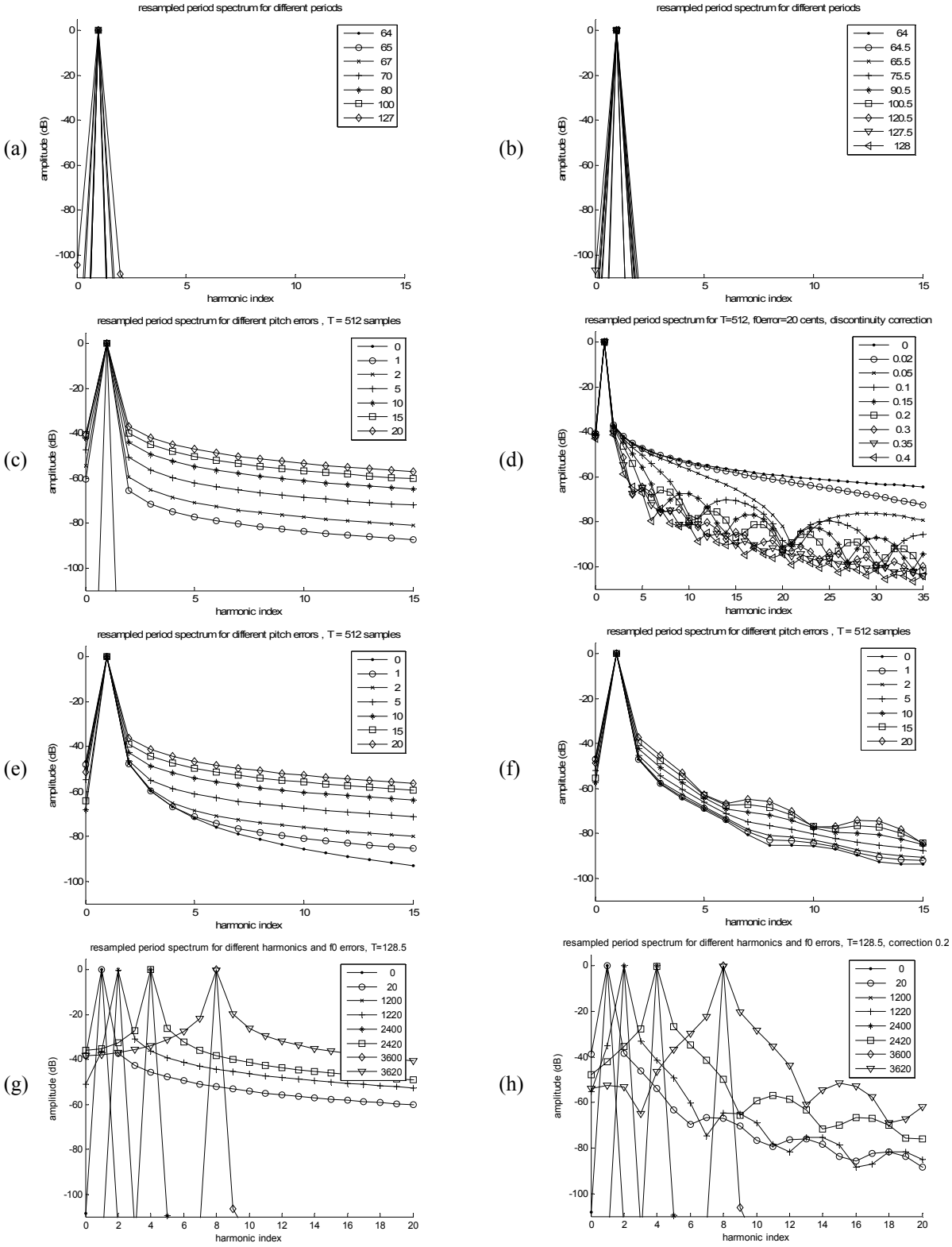

Figure 7: *Inter-harmonic contribution*