Proc. 1999 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, New York, Oct. 17-20, 1999

ANALYSIS AND ENHANCEMENT OF LOCALLY HARMONIC SIGNALS USING ADAPTIVE MULTI-KERNEL METHODS

Stéphanie Dubost, Olivier Cappé

Ecole Nationale Supérieure des Télécommunications, Dpt. TSI / LTCI (CNRS URA 820)
46 rue Barrault, 75634 Paris cedex 13, France
dubost, cappe @tsi.enst.fr

ABSTRACT

This paper is concerned with the analysis/synthesis and enhancement of signals that can be efficiently modeled as "quasi-harmonic". We first provide a formal definition of quasi-harmonic signals and show that known estimation methods for such signals can be interpreted in a non-parametric local approximation framework. We then investigate the possibility of using "adaptive multi-kernel" estimation methods for which the duration of the analysis frame is tuned independently for each harmonic depending on the local behavior of the signal.

1. INTRODUCTION

Sinusoidal modeling now is a well established approach in speech and audio processing, with applications such as speech coding and enhancement or musical synthesis [1]. The basic methodology of sinusoidal analysis/synthesis can be described as follow:

Analysis Consider the signal in overlapping short-time frames (with duration typically ranging from 20 to 40ms) assuming that *the signal in each frame may be modeled as a noisy sum of sinusoids*. Estimate the parameters via least-squares or Fourier transform.

Synthesis *Reconstruct a "smooth" signal by interpolation* of the results obtained in each frame, either by direct sinusoidal parameter interpolation or by time-domain overlap-add synthesis.

Although very efficient in practice (for speech signals notably), the approach sketched above raises a number of challenging (and yet unsolved) issues:

Is the short-time model appropriate? [2, 3] When modeling speech or music signals with sinusoids, the residual signal (difference between the actual signal and the synthetic one) is far from being the stationary noise process one would expect if the model was correct (even when only considering the frame level residual).

Is frame-based processing justified? [4, 5, 3] The usual frame duration of 20-40ms is nothing more than the result of a compromise: larger durations make the sinusoidal model grossly wrong, while durations shorter than a few periods render sinusoidal parameter estimation unreliable. Various schemes have been proposed to overcome the limitations of standard-frame based processing, usually by allowing high frequency components to be updated more frequently.

Finally, the optimal use of sinusoidal modeling for enhancement of signals corrupted by wideband noise is also an open question. It is well known that the basic approach to sinusoidal analysis/synthesis described above does not have any potential for denoising (see also fig. 3). When the sinusoidal model is used for denoising, the analysis stage is thus followed by a (soft or hard) thresholding of the estimated sinusoidal components [6, 7, 4].

In this contribution, we focus on the case of pitched signal and introduce in section 2 a general non-parametric model to describe such signals. In this context, the standard frame-based sinusoidal analysis can be interpreted as an instance of the Local Polynomial Approximation (LPA) method, or more precisely of the zero order LPA or kernel method. In section 3, we motivate and describe recent techniques which make it possible to adapt the kernel (or short-time window) duration, depending on the local behavior of the signal[1]. We then discuss the application of these methods to the quasi-harmonic model and consider in particular their implications for the denoising application.

2. A MODEL FOR QUASI-HARMONIC SIGNALS

We first define an observed discrete-time locally sinusoidal signal as follows

$$s_t = a(\frac{t}{T})\cos(T\phi(\frac{t}{T})) + b(\frac{t}{T})\sin(T\phi(\frac{t}{T})) \quad \text{for } t = 1, \ldots, T \qquad (1)$$

where

- The phase and quadrature amplitudes $a(u)$, and $b(u)$ are C^2 (twice continuously differentiable) functions defined on $[0, 1]$.
- The phase $\phi(u)$ is a monotonously increasing differentiable function on $[0, 1]$, whose derivative $\omega(u) \triangleq \dot{\phi}(u)$ is C^2 and such that $\pi > \omega(u) \geq \omega_{\min} > 0$.

In (1), the variations of the instantaneous amplitude and frequency are modeled as regularly sampled C^2 functions which formalizes our intuition that for real-life narrow-band signals, the variations of the instantaneous amplitude and frequency should be smooth. Compared to similar constructions such as that of [8], the advantages of (1) are twofold: *i)* it is a discrete-time definition, *ii)* it is asymptotically (ie. for large values of T) coherent with usual properties of discrete-time sinusoidal signal, as eg.

$$\left| s_t - 2\cos(\omega(\frac{t}{T}))s_{t-1} + s_{t-2} \right| \leq \frac{C}{T} \qquad (2)$$

[1] In non-parametric estimation, the size of the kernel is usually referred to as "the bandwidth". We will not use this term which is clearly misleading in a signal processing context.

where the constant C depends only upon the functions $a(u), b(u)$ and $\omega(u)$ (and not on t). (2) justifies that $\omega(\frac{t}{T})$ may indeed be considered as the instantaneous frequency of s_t when the number of observations is large. It is important to note however that (1) does not correspond to the sampling of a fixed continuous-time signal (see also [9] on this point).

The natural extension of (1) to the case of quasi-harmonic signals is given by

$$s_t = \sum_{k=1}^{K} \left[a_k(\frac{t}{T}) \cos(kT\phi(\frac{t}{T})) + b_k(\frac{t}{T}) \sin(kT\phi(\frac{t}{T})) \right] \quad (3)$$

where the amplitude functions a_k and b_k (for $k = 1, \dots, K$) are C^1 on the interval $[0, 1]$. (3) correspond to a so-called *non-parametric model* because the quantities that define s_t are functions and thus cannot be described by a set of finite dimensional parameters.

In the sequel, we consider that the observed signal is a noisy version of (3)

$$X_t = s_t + B_t \quad (4)$$

where $\{B_t\}$ is a stationary random noise sequence. (4) typically correspond to the case of a voiced sequence of speech observed in measurement noise. In order to analyze complete sections of speech or music, the model should be extended to allow for an additional random locally stationary signal component. We omit this issue for the sake of simplicity and furthermore assume that the psd. of B_t is known.

3. LOCAL POLYNOMIAL APPROXIMATION

The rationale underlying the Local Polynomial Approximation (LPA) approach [10] is that an unknown function may be approximated locally using Taylor expansions

$$a(\frac{t}{T}) \approx a(\frac{t_0}{T}) + \frac{(t-t_0)}{T} \dot{a}(\frac{t_0}{T}) + \dots \quad (5)$$

The quality of this approximation is related to the magnitude of the terms of higher order that are omitted in the development (and thus to higher-order derivatives of $a(u)$ in $\frac{t_0}{T}$). The LPA estimate of $a(\frac{t_0}{T})$ is then obtained by computing the tapered least-squares solution corresponding to the pseudo-model

$$X_t = a(\frac{t_0}{T}) + \frac{(t-t_0)}{T} \dot{a}(\frac{t_0}{T}) + B_t \quad \text{for } t_0 - h \leq t \leq t_0 + h$$

with kernel (vector of diagonal least-squares weights) $w_{t_0-h}, \dots, w_{t_0+h}$. The LPA or order 0 is usually referred to as *kernel smoothing*. Application of the kernel method to the amplitudes and frequencies in (1) yields the following approximation

$$X_t = \sum_{k=1}^{K} [\alpha_k \cos((t-t_0)k\Omega) + \beta_k \sin((t-t_0)k\Omega)] + B_t \quad \text{for } t_0 - h \leq t \leq t_0 + h \quad (6)$$

where

$$\begin{cases} \Omega = \omega(\frac{t_0}{T}) \\ \alpha_k = a_k(\frac{t_0}{T}) \cos(T\phi(\frac{t_0}{T})) + b_k(\frac{t_0}{T}) \sin(T\phi(\frac{t_0}{T})) \\ \beta_k = -a_k(\frac{t_0}{T}) \sin(T\phi(\frac{t_0}{T})) + b_k(\frac{t_0}{T}) \cos(T\phi(\frac{t_0}{T})) \end{cases}$$

(6) of course correspond to the familiar sinusoidal analysis but is now considered only as *an approximation* of the real signal given by (3). The important feature of the non-parametric model (3) is that the estimation will not systematically improve when increasing the kernel size h: When h decrease, the variance of the estimates do increase because the number of observations taken into account is $2h + 1$, however the systematic bias due to the approximation decrease because the higher order terms in (5) get smaller and smaller. The kernel size h thus control the *bias-variance trade-off.*

In practice, computation of the least-squares solution of (6) requires some numerical optimization since the problem is non-linear in Ω. The results of Katkovnic [11] (although obtained under slightly different hypotheses) however suggest that even in the non-parametric context, the optimal rate of estimation for the pitch is much faster than that associated with the amplitude of the harmonics[2]. Thus, there is asymptotically no loss in estimating the time-varying pitch beforehand (with an optimal estimation procedure!) and using (3) as if the phase component $\phi(\frac{t}{T})$ was known exactly (which was also the conclusion of the example considered in [3]). In the present work, we do not consider optimal adaptive pitch estimation and simply proceed as follow

1. Estimate the pitch using fixed-duration 30ms frames with 5ms hop and interpolate it to sample precision
2. Compute an estimate of
$$\phi(\frac{t}{T}) = \int_0^{\frac{t}{T}} \omega(u) du + \phi(0)$$
using a cumulative sum approximation ($\phi(0)$ is arbitrary and can be set to 0).
3. Consider the model of (3) with **known phases** $\phi(\frac{t}{T})$.

The kernel method applied to the amplitudes of (3) now yields the following approximation

$$X_t = \sum_{k=1}^{K} \left[\alpha_k \cos(T\phi(\frac{t}{T})) + \beta_k \sin(T\phi(\frac{t}{T})) \right] + B_t \quad \text{for } t_0 - h \leq t \leq t_0 + h \quad (7)$$

where $\alpha_k = a_k(\frac{t_0}{T})$, $\beta_k = b_k(\frac{t_0}{T})$.

Although the regressors are not anymore exact sinusoids (but sinusoids with slowly drifting frequencies), it can be shown that *if $\phi(u)$ verifies the assumptions mentioned under (1)*, the least-squares estimate $\hat{\alpha}_k$ (or $\hat{\beta}_k$) satisfies

$$\begin{cases} Var(\hat{\alpha}_k) = \frac{C_w}{h} \gamma_B(k\omega(\frac{t_0}{T})) + O(\frac{1}{h^2}) \\ Cov(\hat{\alpha}_k, \hat{\beta}_j) = O(\frac{1}{h^2}) \\ Cov(\hat{\alpha}_k, \hat{\alpha}_j) = O(\frac{1}{h^2}) \quad \text{for } k \neq j \\ Bias(\hat{\alpha}_k) \triangleq a_k(\frac{t_0}{T}) - E[\hat{\alpha}_k] = O\left(\left(\frac{h}{T}\right)^2 \left| \ddot{a}_k(\frac{t_0}{T}) \right| \right) \end{cases} \quad (8)$$

where C_w is a constant that depend on the type of the kernel (equals to 1 for a rectangular kernel) and $\gamma_B(\lambda)$ is the psd of the noise B_t at frequency λ. As in the case of stationary harmonic signals, (8) suggests that when the kernel duration is large, the phase and quadrature components corresponding to all the harmonics form a set of $2K$ quasi-orthogonal waveforms.

[2] See [12] for a proof of this result in the parametric context (assuming that the model in (6) is exact).

4. ADAPTIVE ESTIMATION

To make the analysis really "adaptive" (dependent on the local behavior of the signal around t_0) we need to tune the value of h in t_0. For this purpose, we use the method of Lepski [13] which is asymptotically minimax with respect to the mean squared error risk

$$E|\hat{a}(\frac{t_0}{T}) - a(\frac{t_0}{T})|^2 = Var[\hat{a}(\frac{t_0}{T})] + Bias^2[\hat{a}(\frac{t_0}{T})]$$

Compared to more traditional solutions detailed in [10], Lepski's procedure is advantageous because it does not involve estimating (even indirectly) the bias due to the local approximation: The idea rather consist in computing $\hat{a}(\frac{t_0}{T})$ for different increasing values of h on a logarithmic scale in order to detect the first value of h where the variations of $\hat{a}(\frac{t_0}{T})$ with h become large compared to the variance. The point of this construction is that contrary to the bias, the variance does not depend on the unknown signal and can be readily computed according to (8). The reader is referred to [13] or [8] for detail concerning the kernel selection procedure.

For the locally harmonic model (3), there are several options in applying this kernel selection rule. the experiments that we have done so far indicate that

1. Global kernel selection, ie. selection of a single kernel duration used for estimating all the harmonic components based on the signal estimate

$$\hat{s}_{t_0} = \sum_{k=1}^{K} \hat{\alpha}_k \cos(T\phi(\frac{t_0}{T})) + \hat{\beta}_k \sin(T\phi(\frac{t_0}{T}))$$

   is not effective for speech. This is due to the large difference in level between the different harmonics: While all harmonics contribute equally to the variance of $\hat{s}_{t_0}$, only the first harmonics (or more generally, a few highest harmonics) significantly contribute to the bias. As a consequence, global kernel selection appears as a compromise between incompatible requirements (compare the durations selected for H1 and H5 on figs. 1 and 2).

2. Harmonic-dependent kernel selection is faced with the problem that the phase $a_k(\frac{t_0}{T})$ and quadrature $b_k(\frac{t_0}{T})$ components may significantly differ in magnitude for some values of t_0. For this reason, the estimated contribution of the k-th harmonic in t_0 ($\hat{\alpha}_k \cos(T\phi(\frac{t}{T})) + \hat{\beta}_k \sin(T\phi(\frac{t}{T}))$) is not a pertinent choice for selecting the kernel duration. The best results where obtained using separate kernel selection procedures for both the estimated phase $\hat{\alpha}_k$ and quadrature $\hat{\beta}_k$ components in t_0. In cases where these two selections do not agree, the smallest of the two durations is kept for estimating the contribution of the k-th harmonic component in t_0.

For speech processing, it is not realistic to repeat the above estimation procedure at each time index t_0. The adaptive estimation scheme is thus applied to time indexes t_0 spaced of 5ms and the approximation given by (7) is used to reconstruct the signal on 10ms frames, the signal being interpolated via a standard overlap-add procedure. Note however that the duration of the synthesis frame is always much shorter than that of the analysis frame and that the former is not varied with h.

5. RESULTS AND DISCUSSION

The method previously described is applied to a section of voiced speech corrupted by Gaussian white noise. The signal as well as its 10dB noisy version are displayed on fig. 3. For kernel selection, Lepski's procedure is applied with 6 different possible kernel durations ranging from 16ms to 320ms on a logarithmic scale.

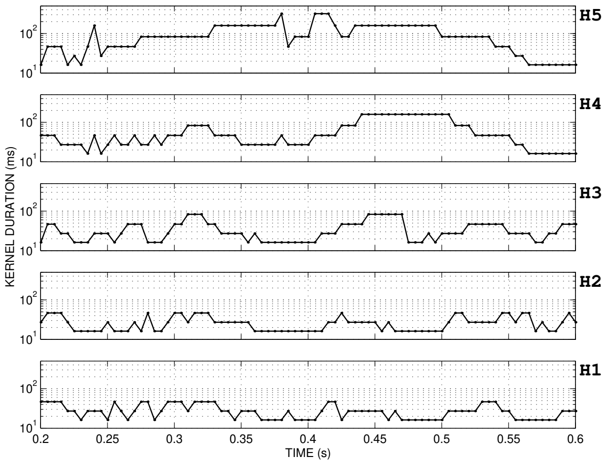

Figure 1: *Kernel durations selected for each of the first five harmonics (30 dB SNR in the [0,1kHz] band).*

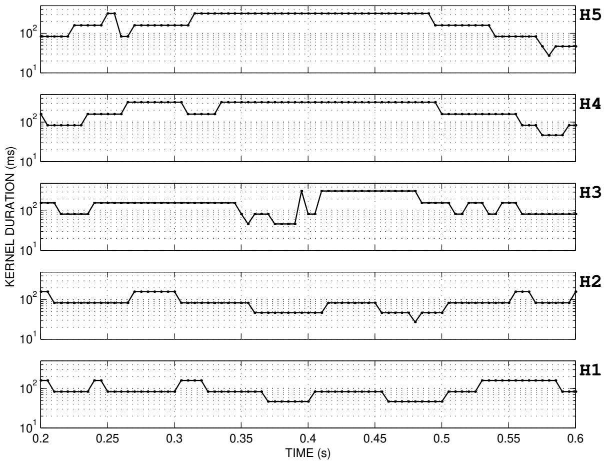

Figure 2: *Kernel durations selected for each of the first five harmonics (10dB SNR in the [0,1kHz] band).*

Figs. 1 and 2 display the selected kernel durations for each harmonics (with the first harmonic H1 at the bottom of the plot as in a usual spectrograms) for two value of the SNR. A first interesting remark is that for both noise conditions, higher harmonics tend to be analyzed with longer frame durations. This results which is in complete contradiction with the analysis scheme used in [4] can be interpreted as follow: (8) show that multiplying the kernel duration by 2 reduce the variance by only 3dB whereas the bias term is reduced by 12dB when dividing the kernel duration by a factor 2. In noisy situations, the fact that higher harmonics have smaller magnitude and hence lower SNR is thus predominant, even if there are good reasons to think that for speech and music signal terms such as $\left|\ddot{a}_k(\frac{t}{T})\right|$ may take larger values for higher harmonics. Closer inspection of fig. 2 together with fig. 3 shows that the variations of the selected kernel duration for each harmonic are clearly related to the amplitude of the harmonic. Another point of interest is that for noisy signals, the selected kernel durations are larger than is common in speech processing: For the first harmonic, selected durations which lie around $20ms$ for the low noise condition (fig. 1) are typically of the order of $100ms$ when the noise level is high (fig. 2).

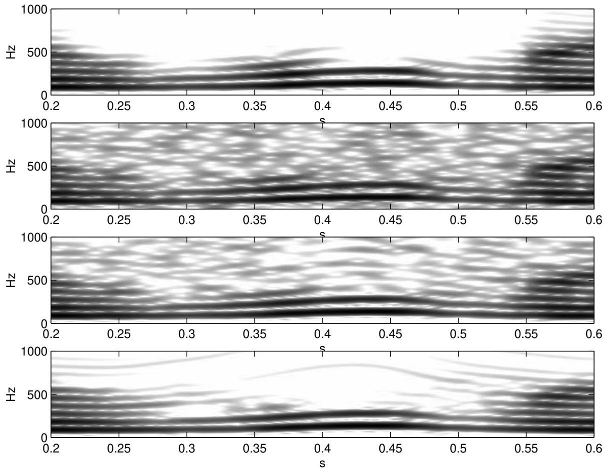

Figure 3: *From top to bottom:* **A** *Original signal;* **B** *Noisy signal (10dB SNR in the [0,1kHz] band);* **C** *Sinusoidal estimate with fixed 25ms frames;* **D** *Adaptive sinusoidal estimate with multi-kernel selection.*

Fig. 3 displays the signals on a gray-level narrow-band (25ms analysis frames) spectrographic representation with 35dB of depth. Fig. 3-C clearly shows that standard sinusoidal estimation enhances the signal in a very limited way because the harmonic components that fall below the noise level are estimated with an estimation variance that is larger than their power: The visual and auditive impression is that the noise is still present at a lower level with slightly different characteristics. In comparison, on fig. 3-D there is a marked noise reduction effect. Compared to fig. 3-C, the additional noise reduction can be evaluated more precisely by noting that in the middle section of the signal the kernel durations selected for the upper harmonics (H3 to H5 on fig. 2) are approximately 10 times larger than 25ms which correspond to a 10dB attenuation (assuming that the harmonics are indeed well below the noise level in this section). Note that there is a price to pay for this noise reduction which is a slight augmentation of the bias ("over-smoothing") which is visible for the third harmonic (fig. 3-D, around time 0.35s).

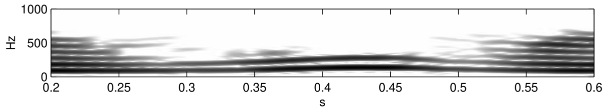

Figure 4: *Adaptive sinusoidal estimate with multi-kernel selection and hard thresholding.*

As is clear from the previous discussion, complete denoising of the signal would require testing for kernel durations much larger than 320ms. Note however that the spurious low level sinusoids visible on fig. 3 have very smooth variations because there are estimated from long kernels. In these conditions, thresholding can be done reliably and, as an example, fig. 4 shows the result of zeroing all harmonics for which the longest kernel duration was selected.

6. REFERENCES

[1] T. F. Quatieri and R. J. McAulay, "Audio signal processing based on sinusoidal analysis/synthesis," in *Applications of Signal Processing to Audio and Acoustics*, M. Kahrs and K. Brandenburg, Eds. Kluwer Academic Publishers, 1998.

[2] J. Laroche, Y. Stylianou, and E. Moulines, "HNS: Speech modification based on a harmonic + noise model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, 1993.

[3] G. Fay, E. Moulines, O. Cappé, and F. Bimbot, "Polynomial quasi-harmonic models for speech analysis and synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Seattle, May 1998, pp. II–865–II–868.

[4] D. V. Anderson and M. A. Clements, "Audio signal noise reduction using multi-resolution sinusoidal modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, 1999.

[5] S. N. Levine, T. S. Verma, and J. O. Smith, III, "Alias-free multiresolution sinusoidal modeling for polyphonic, wideband audio," in *IEEE Workshop on App. of Sig. Proc. to Audio and Acoust.*, New Paltz, NY, 1997.

[6] T.F. Quatieri and R.J McAulay, "Noise reduction using a soft-decision sine-wave vector quantizer," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Albuquerque, New Mexico, 1990.

[7] J. Hardwick, C. D. Yoo, and J. S. Lim, "Speech enhancement using the dual excitation speech model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, 1993, pp. II–367–II–370.

[8] V. Katkovnik, "On adaptive local polynomial approximation with varying bandwith," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, 1998.

[9] R. Dahlhaus, "Asymptotic statistical inference for nonstationary process with evolutionary spectra," in *Lecture notes in Statistics 115*. Springer-Verlag, 1996.

[10] J. Fan and I. Gijbels, *Local Polynomial Modelling and Its Applications*, Chapman & Hall, 1996.

[11] V. Katkovnik, "Local polynomial periodograms for signals with time-varying frequency and amplitude," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, 1996.

[12] B. G. Quinn and P. J. Thomson, "Estimating the frequency of a periodic function," *Biometrika*, vol. 78, no. 1, pp. 65–74, 1991.

[13] O. V. Lepski and V. G. Spokoiny, "Optimal pointwise adaptive methods in nonparametric estimation," *Annals of Statistics*, vol. 25, no. 6, pp. 2512–2546, 1997.