POLYNOMIAL QUASI-HARMONIC MODELS FOR SPEECH ANALYSIS AND SYNTHESIS

Gilles Fay, Eric Moulines, Olivier Cappé and Frédéric Bimbot

Ecole Nationale Supérieure des Télécommunications, Paris, France

ABSTRACT

Harmonic plus noise models have been successfully applied to a broad range of speech processing applications, including, among others, low bit-rate speech coding, and speech restoration and transformation. In conventional methods, the frequencies, the relative phases and the amplitudes of the pitch-harmonic components are assumed to be piecewise constants over an analysis frame. This assumption is inadequate in segments where fast variations of these parameters may occur, e.g. phoneme-to-phoneme boundaries or speech onsets. In this contribution, a time-varying model of the pitch-harmonic parameter is presented. It is based on a basis expansion technique, consisting in representing the time-varying functions as a linear combination of fixed basis function. An estimation procedure for the parameters of this expansion is presented. Results are provided to demonstrate the effectiveness of this approach.

1. HARMONIC-PLUS-NOISE MODELS

Sinusoidal as well as Harmonic-plus-noise models have been applied with success to solve many speech processing problems. Harmonic-plus-noise model (HNM) assumes the speech signal to be composed of a quasi-harmonic part $s(t)$ and a noise part $n(t)$:

$$x(t) = s(t) + n(t) \tag{1}$$

The quasi-harmonic part, in voiced segments, reflects the (local) periodicity of the speech signal. The noise part $n(t)$ accounts for the cycle-to-cycle variations of the glottal airflow, the friction noise, etc...

Harmonic plus noise models have proven to be useful in many speech processing applications, including among others low-bit rate speech coding [6], speech transformation (time scaling, pitch scaling) [3], text-to-speech synthesis and co-channel speaker separation [7].

In a quasi-harmonic model, $s(t)$ is represented as a superposition of almost harmonically related sinusoidal components

$$s(t) = \sum_{k=1}^{K(t)} \rho_k(t) \cos(\phi_k(t)) \tag{2}$$

where $\rho_k(t)$ and $\phi_k(t)$ denote respectively the amplitude and phase at time t of the k-th pitch-harmonic component. Note that such decomposition is not unique unless some constraints are imposed for $\rho_k(t)$ and $\phi_k(t)$. Typically, it is assumed that $\rho_k(t)$ is a low-pass function, *i.e.* the support of the Fourier transform $\mathcal{F}(\rho_k)$ of ρ_k is *approximately* included in $[-\delta, \delta]$, where δ is a 'small' number, and the support of $\mathcal{F}(\cos(\phi_k(t)))$ lies outside this interval. Under this assumption, the instantaneous frequency of each component in (2) is defined unambiguously as $\dot{\phi}_k(t)$, where $\dot{f}(t)$ denotes the time derivative of $f(t)$. The model is *quasi*-harmonic because the instantaneous frequencies $\dot{\phi}_k(t)$ are assumed to be *approximately harmonically related*, in the sense that

$$\dot{\phi}_k(t) \approx 2\pi k / P(t)$$

where $P(t)$ is the *local* pitch-period. The time-variations of $\rho_k(t)$ and the deviations of $\dot{\phi}_k(t)$ from exact harmonicity reflects the variations of the supra-glottal cavities transfer function and of the glottal pulse shape.

The noise component $n(t)$ may be modeled as a quasi-stationary random sequence. It has been shown that this component may be adequately fitted by filtering a stationary white noise by a time-varying linear filter (see, for example, [4]).

In this contribution, we focus only on the estimation of the quasi-harmonic component. With few exceptions, the methods proposed to date to estimate $\rho_k(t)$ and $\phi_k(t)$ are frame-based: the functions $\rho_k(t)$ and $\dot{\phi}_k(t)$ are assumed to be approximately constant [1] over the analysis frame (typical values correspond to between 20 and 30 ms). The functions $\rho_k(t)$ and $\phi_k(t)$ are then reconstructed by (linear or polynomial) interpolation between the values estimated over the successive analysis frames. While this approximation is reasonable in stationary speech segments (e.g., vowel nucleus) it is not accurate when modeling sharp transitions occuring at phoneme boundaries, voicing onsets, etc. These effects are typically not annoying in applications where coarse-grained speech analysis is sufficient (e.g. low bit-rate coding), the perceptual effect of these mismatches being in general limited. For fine-grained analysis (which is required, for exam-

[1] An exception to that rule is the work by Marques and Almeida [5] who proposed to use a generalized polynomial phase for sinusoidal component to boost accuracy.

ple, for high-quality speech transformation, speech disorder diagnosis...), piecewise constant approximation is clearly inappropriate.

In this paper, it is proposed to model $\rho_k(t)$ and $\phi_k(t)$ as time-varying functions. For that purpose, we use a *basis expansion approach*, consisting in representing the functions $\rho_k(t)$ and $\phi_k(t)$ as a linear combination of known functions of time. As shown below, this technique allows to obtain better fit in transient segments. As another application, the basis expansion technique makes possible to model syllable long speech segments (100 to 200 ms) with a single set of parameters. This distinctive property offers new perspectives for speech processing (e.g., restoration and transformation for example), that still need to be investigated.

2. A LONG-TERM PARAMETRIC MODEL FOR QUASI-HARMONIC SPEECH

The quasi-harmonic model (2) can be equivalently rewritten as

$$s(t) \simeq \sum_{k=1}^{K(t)} a_k(t)\cos(k\Phi(t)) + \sum_{k=1}^{K(t)} b_k(t)\sin(k\Phi(t))$$

The time-varying amplitude and phase in model 2 are related to $a_k(t)$ and $b_k(t)$ through the relations

$$\rho_k(t) = \sqrt{a_k(t)^2 + b_k(t)^2},$$
$$\phi_k(t) = \begin{cases} k\phi(t) - \operatorname{arctg}(b_k(t)/a_k(t)) & \text{if } a_k(t) > 0 \\ k\phi(t) - \operatorname{arctg}(b_k(t)/a_k(t)) + \pi & \text{if } a_k(t) < 0 \end{cases}$$

The derivative of the function $\dot{\Phi}(t)$ models the pitch-contour, the evolution of the fundamental frequency. Because we want to model a quasi-harmonic signal, it is assumed that the instantaneous frequency of the k-th harmonic component is (approximately) $k\dot{\Phi}(t)$: this property is 'built-in' in our model. The in-phase $a_k(t)$ and the in-quadrature $b_k(t)$ components account for both **(i)** pitch-harmonic amplitude variations, and **(ii)** deviations from the exact harmonicity, reflecting the slow variations of the phase distribution of the pitch-harmonics. Note that the amplitude and phase variations are in fact dependent, since they are both related to changes of the vocal tract transfer function and of the glottal pulse shape.

As outlined in the introduction, we model the functions $a_k(t)$ and $b_k(t)$ using a basis expansion approach. The basic idea behind this approach consists in representing the functions $a_k(t)$ and $b_k(t)$ and $\phi(t)$ as linear combinations of known functions of time.

More specifically, let $g_j(t)$, $j = 0, 1, \cdots$ a known family of functions. We wish to estimate a given (unknown) function $f(t)$ (e.g., $f(t) = a_k(t)$, $f(t) = b_k(t)$ or $f(t) = \phi(t)$). Assume that $f(t)$ can be expressed in the form of an infinite expansion

$$f(t) = \sum_{j=0}^{\infty} \theta_j^* g_j(t)$$

To limit the *effective* number of parameters in the expansion, some *smoothness* or *regularity* assumptions have to be stated. In this context, the smoothness condition about f is that the coefficients in the expansion decrease at a certain rate as $j \to \infty$. The projection approach consists in estimating a truncated expansion, $f_p(t) = \sum_{j=0}^{p} \theta_j g_j(t)$ This way, we reduce the problem of function estimation to that of parametric estimation, though the number of parameters we have to estimate is not bounded a priori and can be large. In our approach, the parameters θ_j are estimated using linear least-squares, as detailed below. Of course, the type of functions as well as the number of functions needed in the expansion is not known a priori, and should be adapted to the speech segment. These aspects are shortly discussed below.

2.1. Model for the phase function $\Phi(t)$

It is known that pitch variations in voiced segments are very smooth. In most cases a low-order polynomial function (say, of order 2 to 3) is sufficient to fit the track over the analysis frame (see section 4). As the result, the phase function $\phi(t)$, as a integral of the pitch track also is a polynomial,

$$\phi(t) = \sum_{m=1}^{M} \phi_m t^m$$

2.2. Model for the amplitude functions $a_k(t)$ and $b_k(t)$

To estimate $a_k(t)$ and $b_k(t)$, we use a projection type estimate on a B-spline (box-spline) basis [2].

Box-Spline B-splines basis has been introduced by de Boor [1] (see also [2]). A B-spline consists of polynomial pieces, connected in a special way: a B-spline of degree q, consists of $(q+1)$ polynomial pieces each of degree q; each polynomial pieces join at q inner knots: at the joining points, the derivatives up to order $q-1$ are continuous. The B-spline is positive on a domain spanned by $q+2$ knots: everywhere else, it is zero. Except at the boundaries, it overlaps with $2q$ polynomial pieces of its neighbors. At a given t, $(q+1)$ B-spline are non zeros. In practice, we use B-spline of degrees three, with regularly spaced knots.

Denote p the order of the projection estimate and $B_j(t)$, $1 \leq j \leq p$ the B-splines; the p-th order approximation of

$a_k(t)$ and $b_k(t)$ writes

$$a_k^{(p)}(t;\mathbf{a}) = \sum_{j=1}^{p} \alpha_{k,j} B_j(t)$$
$$b_k^{(p)}(t;\mathbf{a}) = \sum_{j=1}^{p} \beta_{k,j} B_j(t)$$

where $\mathbf{a}$ denote the amplitude parameters vector

$$\mathbf{a} = [\alpha_{1,1}, \beta_{1,1}, \alpha_{1,2}, \dots, \beta_{K,p}]$$

3. PARAMETER ESTIMATION

The pth order truncation approximation of $s^{(p)}(t)$ of $s(t)$ writes

$$s^{(p)}(t;\mathbf{a},\Phi) = \sum_{k=1}^{K} a_k^{(p)}(t;\mathbf{a})\cos(k\phi(t)) + b_k^{(p)}(t;\mathbf{a})\sin(k\phi(t))$$

This approximation $s^{(p)}(t;\mathbf{a},\Phi)$ depends *linearly* on $\mathbf{a}$ and *non-linearly* on Φ.

Denote $X_1, X_2, \cdots, X_n$ the samples of the observed signal. Under the model (1), $X_k = s(kT_e) + N_k$, with $N_k = n(kT_e)$ and T_e is the sampling period (we set: $T_e = 1$ for simplicity). We estimate the parameters using least-square, *i.e.* minimize the following criterion :

$$J(X;\Phi,\mathbf{a}) = \sum_{j=1}^{n} (X_k - s(j;\mathbf{a},\Phi))^2 \qquad (3)$$

This is a non-trivial optimization problem, because the number of samples and the number of parameters involved in the minimization procedure can be very large (typical values are: $n \in [1000, 5000]$ samples, while the number of parameters can be as large as one thousand !). Hopefully, several approximations can be done that reduce dramatically the computational requirements.

Given the parameter vector Φ, linear least-squares may be used to estimate $\widehat{\mathbf{a}(\Phi)}$; Optimal values of Φ may be obtained by minimizing the reduced least-square criterion

$$\tilde{J}(X;\Phi) = J(X;\Phi,\widehat{\mathbf{a}(\Phi)})$$

This can be done using standard optimization procedure. In practice, it is possible to obtain rather accurate initialization for the parameter Φ by fitting the estimated pitch contour (obtained in a preliminary step using a standard pitch estimation device), by low-order polynomials. Only few iterations are thus needed to converge. In a simplified yet effective version, the non-linear estimation procedure needed to estimate the parameter of the phase function is simply by-passed.

Note finally that accurate approximation of the linear least-squares solution can be obtained. Denote $H_{k,n}(\Phi)$ the $(2p \times n)$ regression matrix defined as

$$H_{k,n}(\Phi) = [B_1(1:n)\cos(k\phi(1:n)) \; B_1(1:n)\sin(k\phi(1:n)), \dots, B_p(1:n)\sin(k\phi(1:n))]$$

where, for a given function $f(t)$, $f(1:n)$ denotes the column-vector $[f(1), \cdots, f(n)]^T$. For a given number p of basis functions, it may be shown that, for $\phi_1 \neq 0$, we have

$$n^{-1} H_{k,n}(\Phi)^T H_{k,n}(\Phi) = \Gamma_n + O(n^{-1}) \qquad (4)$$
$$n^{-1} H_{k,n}(\Phi)^T H_{l,n}(\Phi) = O(n^{-1}) \;\; k \neq l \qquad (5)$$

where Γ_n is a matrix which does not depend on k and Φ. In other words, given the phase function $\Phi(t)$, the B-spline modulated by $\cos(k\phi(t))$ and $\sin(k\phi(t))$ form approximately an orthogonal family of vectors. As a result, pseudo-inversion of the complete regression matrix can be by-passed, and approximate solutions of the linear least-squares estimation step can be evaluated by **(i)** computing scalar products, **(ii)** multiplying the results by a fixed $(2p \times 2p)$ Γ_n^{-1} matrix (of course, this matrix inverse can be computed once for all).

4. RESULTS

In this section preliminary results are presented to support our findings. Here, we focus on the extraction of the stochastic component $n(t)$, a problem which proves to be difficult using conventional method. Extraction of this component is useful, for example, for high-quality speech transformation (see [8]).

In this experiment, we analyze a speech segment that has been uttered by a male speaker (sampling freq.= 16 kHz). In figure 1, we display a 180 ms segment (transition between a voiced fricative /z/ and the vowel /i/). and the corresponding pitch contour. On the same plot, the estimated instantaneous frequency (obtained by fitting a third order polynomial) is displayed. It is seen on this example that a low-order polynomial is sufficient to capture the variation (whereas it seems inappropriate to assume that the pitch is constant even on a very short-time scale).

We compare our estimation procedure with a conventional method (see for example [3]), consisting in **(i)** estimating the amplitude and phase of the pitch-harmonics from the short-term Fourier transform (window length 30 ms, overlap 20 ms) **(ii)** re-synthesizing the speech signal using an overlap-add procedure. In this experiment, we use 3rd order B-spline, and the number of knots has been set to 20. The number of parameters fitted to the data are equivalent in the two methods. To fit the parameter of our model, we use the simplified method (without re-estimation of the phase) (figure 3). We display, on the same scale, the residual signals, defined as the difference between the speech signal and

the fitted harmonic models. As seen on this plot, the time varying model efficiently removes the harmonic component, even at the phoneme boundary. The residual obtained with the conventional method (see figure 2) still contains harmonic components (especially at the boundary phoneme).

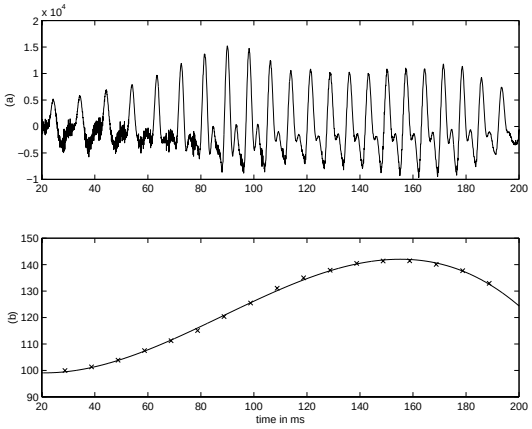

Figure 1: (a) a 180 ms voiced pitch segment, (b) pitch approximation by a 3rd order polynomial

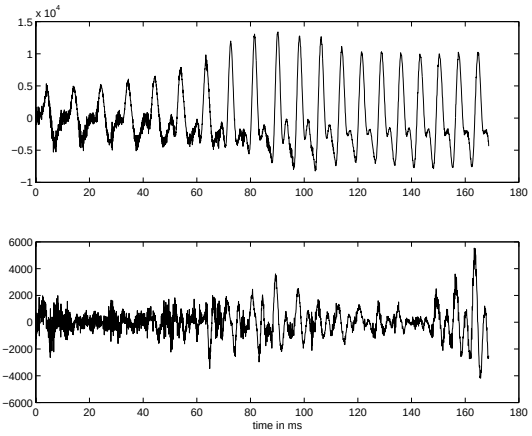

Figure 2: Synthetic signal and its corresponding residual with the conventional method

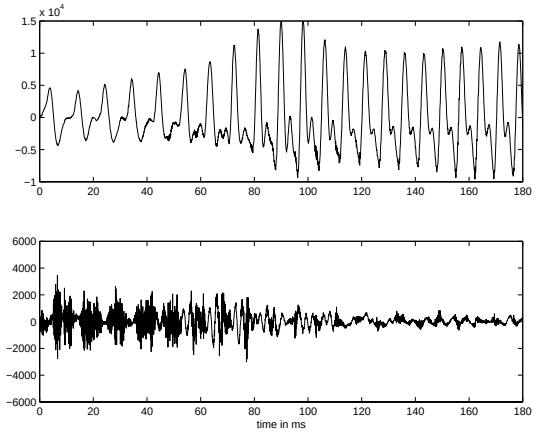

Figure 3: Synthetic signal and its corresponding residual with the proposed method

5. REFERENCES

[1] De Boor. *A Practicle Guide to Splines*. Springer, 1978.

[2] P.H.C. Eilers and B.D. Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2):89–121, 1996.

[3] E.B. George and M.J.T. Smith. Speech analysis/synthesis and modification using an analysis-by-synthesis/overlap-add sinusoidal model. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 5(5), 1997.

[4] Y. Grenier. Time-dependent ARMA modeling of nonstationary signals. *IEEE Trans. on Acoustics, Speech and Signal Processing*, pages 899–911, 1983.

[5] J.S. Marques and L.B. Almeida. A background for sinusoid based representation of voiced speech. In *Proc. Int. Conf. on Acoust. Speech and Signal. Proc.*, pages 1233–1236, 1986.

[6] R.J. McAulay and T.F. Quatieri. Low-rate speech coding based on the sinusoidal model. In Furui and Sondhi, editors, *Advances in Speech Signal Processing*, chapter 6. 1992.

[7] D. Morgan, E.B. George, L.T. Lee, and S. Kay. Cochannel speaker separation by harmonic enhancement and suprresion. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 5(5):407–425, 1997.

[8] Yannis Stylianou. *Modèles Harmoniques Plus Bruit Combinés Avec Des Méthodes Statistiques, Pour La Modification De La Parole Et Du Locuteur*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, 1996.