Adaptive AM–FM Signal Decomposition With Application to Speech Analysis

Yannis Pantazis, Olivier Rosec, *Member, IEEE*, and Yannis Stylianou, *Member, IEEE*

***Abstract*—In this paper, we present an iterative method for the accurate estimation of amplitude and frequency modulations (AM–FM) in time-varying multi-component quasi-periodic signals such as voiced speech. Based on a deterministic plus noise representation of speech initially suggested by Laroche *et al.* ("HNM: A simple, efficient harmonic plus noise model for speech," *Proc. WASPAA*, Oct., 1993, pp. 169–172), and focusing on the deterministic representation, we reveal the properties of the model showing that such a representation is equivalent to a time-varying quasi-harmonic representation of voiced speech. Next, we show how this representation can be used for the estimation of amplitude and frequency modulations and provide the conditions under which such an estimation is valid. Finally, we suggest an adaptive algorithm for nonparametric estimation of AM–FM components in voiced speech. Based on the estimated amplitude and frequency components, a high-resolution time–frequency representation is obtained. The suggested approach was evaluated on synthetic AM–FM signals, while using the estimated AM–FM information, speech signal reconstruction was performed, resulting in a high signal-to-reconstruction error ratio (around 30 dB).**

***Index Terms*—AM–FM decomposition, AM–FM signals, sinusoidal modeling, speech analysis.**

I. Introduction

MULTI-COMPONENT amplitude modulated and frequency modulated (AM–FM) signals are present everywhere in natural sounds, including the human voice. In many applications, such as Doppler radar applications [2], medical imaging [3], and audio and speech processing [4], [5], there is a great interest in decomposing the input signal into time-varying amplitude and frequency components. The terms "demodulation" and "separation" are usually used in the same context. Focusing on speech, amplitude, and frequency modulations are strongly related with the speech production mechanism: from the glottis to the vocal tract and to the lips. Particularly, the continuous movements of the articulators as well as the time-varying glottal source excitation cause the spectral content of the signal to change smoothly, in general. It is also fact that not only slow but also fast modulations occur in speech signals (i.e., jitter and shimmer) [4].

Manuscript received June 29, 2009; revised December 07, 2009; accepted March 16, 2010. Date of publication April 08, 2010; date of current version October 27, 2010. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Gaël Richard.

Y. Pantazis and Y. Stylianou are with the Computer Science Department, University of Crete and ICS-FORTH, 71003 Heraklion, Crete, Geece (e-mail: pantazis@csd.uoc.gr; yannis@csd.uoc.gr).

O. Rosec is with the Speech Synthesis Team, Orange Labs, 22307 Lannion, France (e-mail: olivier.rosec@orange-ftgroup.com).

Digital Object Identifier 10.1109/TASL.2010.2047682

In the literature, there are nonparametric and parametric algorithms or methods for processing an AM–FM signal and estimating its components. Examples of the nonparametric methods include time–frequency representations like spectrogram or short-time Fourier transform (STFT), Wigner–Ville distribution [6], [7], and Choi–Williams distribution [4], [6]. To improve the time–frequency representation of these approaches, the reassignment method [8] can be used. However, the tradeoff between the fine-time or fine-frequency resolution, as well as the interference components present in multi-component signals, always restricts the capabilities of these analysis methods [6]. A different nonparametric approach to AM–FM demodulation is to pass the input signal through a bank of filters and then try to estimate the instantaneous components in each of the output signals. Usually, the analytic signal, estimated through Hilbert transform, is used to demodulate an AM–FM signal [9]. Also, energy operators like the Teager–Kaiser energy operator [10] were applied for AM–FM demodulation [11]–[13]. The major limitation of these methods is their sensitivity when several components are present in the input signal. This makes the filtering process very crucial. Similarly, extended Kalman filtering [14] as well as auditory filters [15] have been used to track and, then separate the instantaneous components of the input signal.

Parametric methods are strongly related to time–frequency representations and examples of these methods include the Chirplet transform [16] and Fan-Chirp transform [17], [18]. In this case, some of the necessary parameters usually need to be estimated beforehand. For example, the chirp rate parameter that controls the time-varying frequency characteristics should be given or estimated explicitly. Recently, they have been applied in the analysis of speech signals [17], [18]. Polynomial phase signals with constant [19] or time-varying amplitude [20], [21] are also well studied. Another estimation approach is to consider linear parametric models for AM and FM and estimate the unknown parameters through maximum likelihood [22], [23]. The parameter estimation process of the above methods is highly nonlinear which may cause difficulties in reaching the global maximum (or minimum) of the maximized (or minimized) quantity during the estimation procedure.

The majority of proposed AM–FM decomposition algorithms for speech signal processing are formant-based and not at a sinusoidal/harmonic level. One reason for this is that, in order to resolve the harmonics of the signal, the use of large windows is necessary, which results in a poor resolution in time. Thus, fast time variations of amplitude and/or of frequency components may not be observed or may cause spurious effects in the time–frequency representation of the signal. The sinusoidal speech model suggested by McAulay *et al.* [24] attempts

to balance the time and frequency resolution requests of speech signals by choosing the length of the analysis window and the frame rate. However, when fast transitions occur in amplitude and/or in frequency, the sinusoidal model is inadequate for tracking the AM–FM components of the signal, since this model assumes speech stationarity within an analysis frame.

In this paper, we suggest an adaptive method for estimating the instantaneous components of voiced speech signals. The suggested approach can also be applied to other time-varying multi-component AM–FM signals like music signals, signals (or calls) generated by marine mammals, birds, etc. The suggested method is based on a model initially introduced by Laroche *et al.* for audio modeling of percussive sounds [25] and for speech analysis [1]. In [26], the model was reintroduced revealing its main properties: it was shown that the model is equivalent to a time-varying quasi-harmonic representation of speech. In the following, this model will be referred to as Quasi-Harmonic Model (QHM).

In this paper, we first show the robustness of QHM as frequency estimator. QHM assumes that an initial estimate for the frequencies of the components is provided as well as that the number of components is known *a priori*. Then, the remaining parameters are estimated by minimizing the mean squared error between the speech signal and the model which leads to a least squares (LS) solution. In practice, however, a frequency mismatch between the original and the initial estimates of frequencies is inevitable. Reformulating the QHM representation, we show that QHM provides a mechanism for iteratively estimating the frequency mismatch and consequently improving the amplitude estimation in the mean-squared sense. Robustness issues of the frequency mismatch estimation algorithm are discussed and bounds on the magnitude of frequency mismatch for which the algorithm converges are provided.

Similar to sinusoidal modeling, QHM assumes speech to be locally stationary. In the case of signals exhibiting strong AM–FM, experiments show that QHM is limited in the sense that it can capture variations of frequencies and amplitudes but only up to a certain point. To overcome this limitation, we expand QHM to a novel speech representation referred to as adaptive QHM, or aQHM, where speech is not assumed to be locally stationary, taking into account the trajectories of the instantaneous frequencies. Then, the speech signal is projected in a space generated by time varying nonparametric sinusoidal basis functions. Therefore, the basis functions are adapted to the local characteristics of the input signal. The basis functions are updated by minimizing the mean squared error between the input signal and the aQHM model, at each adaptation step. This leads to a non-parametric AM–FM decomposition algorithm for speech signals. This is indeed, one key difference between aQHM and other AM–FM decomposition approaches like those suggested in [20], [22], where the amplitude and/or phase functions are represented by linear parametric models. In that sense, aQHM suggests a more flexible AM–FM decomposition algorithm. Another key difference between these approaches and aQHM, lies in the fact that aQHM takes into account the specific properties of voiced speech signals (i.e., quasi-harmonic structure). In that sense, aQHM suggests a representation that is more suitable for speech than the generic models suggested in [20] and [22]. On the other hand, the suggested representation can be applied in a straightforward way on other multicomponent signals, simply by adjusting accordingly the expected distribution of the frequency information. For instance, in the case of singing voice or music, or certain marine biologic sounds, frequencies are expected to have a quasi-harmonic structure. In other cases, however, signals may contain a sparse and uncorrelated frequency content. In any case, prior knowledge about the structure and properties of the input signal may reduce the complexity of models while increasing their modeling efficiency.

The Sinusoidal Model (SM) suggests a maximum-likelihood frequency estimator, assuming a harmonic structure for the speech waveform. This leads to a simple peak-picking algorithm for estimating, at the analysis stage, the parameters of the model from the short-time Fourier transform of the signal [24]. Then at the synthesis stage, parametric models for the instantaneous amplitude (linear model) and phase (cubic model) are used for the signal reconstruction. Comparisons between SM, QHM, and aQHM were conducted for validation purposes, on synthetic signals under various noise conditions, while for evaluation we used voiced speech signals. Results show that the suggested approaches outperform SM in terms of accuracy in parameters estimation (for synthetic signals) and in signal-to-reconstruction error ratio (for speech signals).

The organization of the paper is as follows. Section II presents an overview of QHM and reveals its main time and frequency properties showing that under certain conditions, QHM may be used to enhance frequency estimation. In Section III, these conditions are further examined and bounds where such an enhancement is possible in the context of QHM, are provided. The iterative adaptive AM–FM decomposition algorithm based on aQHM is presented in Section IV. The validity of the proposed approach, its robustness under noisy conditions, as well as its comparison with the SM on synthetic AM–FM signals are shown in Section V. To further support our suggestions, results of signal reconstruction on voiced speech signals from many speakers using SM, QHM, and aQHM are provided in Section VI. Section VII concludes the paper and discusses future directions of this work.

II. Quasi-Harmonic Model

A. Overview QHM

Within an analysis window, the deterministic (i.e., quasi-harmonic) component of a speech signal is modeled as ([27, Ch. 4])

$$s(t) = \left(\sum_{k=-K}^{K} (a_k + t b_k) e^{j2\pi k f_0 t} \right) w(t) \tag{1}$$

where f_0 is the fundamental frequency of the harmonic signal, K specifies the order of the model, i.e., the number of harmonics, a_k are the complex amplitudes, b_k are the complex slopes, and $w(t)$ denotes the analysis window which is typically a Hamming window and zero outside a symmetric interval $[-T, T]$. Thus, $t = 0$ will always denote the center of the analysis window. Note that for real signals such as speech, audio,

etc., $a_{-k} = a_k^*$ and $b_{-k} = b_k^*$, where * is the conjugate operator. This model is an extension to the classic harmonic model in which the $t\,b_k$ term is omitted [27]. As a result, it follows that the signal in (1) is projected onto the complex exponential functions, as in the simple harmonic case, and to functions of type $t\,e^{j2\pi k f_0 t}$. It is worth noting that the model in (1) can also be written for nonharmonically related frequency components. Therefore, a more general model can be expressed as

$$s(t) = \left(\sum_{k=-K}^{K} (a_k + t b_k) e^{j2\pi f_k t} \right) w(t) \tag{2}$$

where f_k will be referred to as initial estimates of the frequencies that will be considered to be known. Initial frequencies f_k (or $k f_0$) are considered to be known but not necessary optimal in representing the input signal in the mean squared error sense. For speech as well as for music signals, this is quite often the case.

Assuming that the speech signal $x(t)$ is defined on $[-T, T]$, the estimation of the model parameters $\{f_0, K, a_k, b_k\}$ is performed into two steps. At first, the fundamental frequency, f_0 and the number of harmonic components, K, are estimated using spectral and autocorrelation information as described in [27]. Then, the computation of $\{a_k, b_k\}_{k=-K}^{K}$ is performed by minimizing a mean squared error which leads to a simple least squares solution [27]. The same procedure is applied if the initial estimates of frequencies f_k are not restricted to be multiples of a fundamental frequency. In this case, frequencies f_k may be obtained by peak picking the magnitude spectrum of the Fourier transform of the input signal as suggested in [24]. In the following, we will not restrict the analysis to $f_k = k f_0$, unless otherwise mentioned.

From (2), it is easily seen that the instantaneous amplitude of each component is a time-varying function given by

$$\begin{aligned} M_k(t) &= |a_k + t b_k| \\ &= \sqrt{(a_k^R + t b_k^R)^2 + (a_k^I + t b_k^I)^2} \end{aligned} \tag{3}$$

where x^R and x^I denote the real and the imaginary parts of x, respectively.

Since both the amplitudes a_k and the slopes b_k are complex variables, the instantaneous frequency of each component is not a constant function over time but varies according to

$$\begin{aligned} F_k(t) &= \frac{1}{2\pi} \frac{d\Phi_k(t)}{dt} \\ &= f_k + \frac{1}{2\pi} \frac{a_k^R b_k^I - a_k^I b_k^R}{M_k^2(t)} \end{aligned} \tag{4}$$

while the instantaneous phase is given by

$$\begin{aligned} \Phi_k(t) &= 2\pi f_k t + \angle(a_k + t b_k) \\ &= 2\pi f_k t + atan \frac{a_k^I + t b_k^I}{a_k^R + t b_k^R}. \end{aligned} \tag{5}$$

A feature of the model worth noting is that the second term of the instantaneous frequency in (4) depends on the instantaneous amplitude. In other words, the accuracy of the frequency estimation (or, the estimation of phase function) depends on the amplitude information [28].

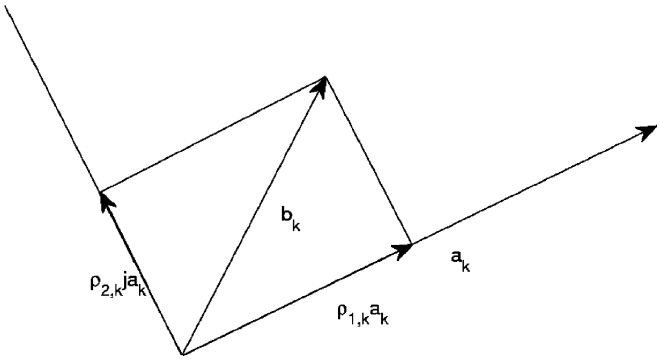

Fig. 1. Projection of b_k into one parallel and one perpendicular to a_k component.

B. Frequency-Domain Properties of QHM

To understand the important features of QHM, we suggest discussing its frequency domain properties. To this end, let us consider the Fourier transform of $s(t)$ in (2)

$$S(f) = \sum_{k=1}^{K} \left(a_k W(f - f_k) + \frac{j b_k}{2\pi} W'(f - f_k) \right) \tag{6}$$

where $W(f)$ is the Fourier transform of the analysis window, $w(t)$, and $W'(f)$ is the derivative of $W(f)$ over f. For simplicity, we will only consider the kth component of $S(f)$

$$S_k(f) = a_k W(f - f_k) + \frac{j b_k}{2\pi} W'(f - f_k). \tag{7}$$

To reveal the main properties of QHM, we suggest the projection of b_k onto a_k as illustrated in Fig. 1. Accordingly,

$$b_k = \rho_{1,k} a_k + \rho_{2,k} j a_k \tag{8}$$

where $j a_k$ denotes the perpendicular (vector) to a_k, while $\rho_{1,k}$ and $\rho_{2,k}$ are computed as

$$\rho_{1,k} = \frac{a_k^R b_k^R + a_k^I b_k^I}{|a_k|^2} \tag{9}$$

and

$$\rho_{2,k} = \frac{a_k^R b_k^I - a_k^I b_k^R}{|a_k|^2} \tag{10}$$

Thus, the kth component of $S_k(f)$ can be written as

$$\begin{aligned} S_k(f) = a_k W(f - f_k) - \frac{a_k \rho_{2,k}}{2\pi} W'(f - f_k) \\ + \frac{j a_k \rho_{1,k}}{2\pi} W'(f - f_k). \end{aligned} \tag{11}$$

Considering the Taylor series expansion of $W(f - f_k - \rho_{2,k}/2\pi)$ we obtain

$$\begin{aligned} &W(f - f_k - \frac{\rho_{2,k}}{2\pi}) = \\ &= W(f - f_k) - \frac{\rho_{2,k}}{2\pi} W'(f - f_k) + O(\rho_{2,k}^2 W''(f - f_k)). \end{aligned} \tag{12}$$

For a rectangular window it holds that $W''(f) \propto T^3$, where T is the duration of the analysis window, $w(t)$. Since the duration of the analysis window determines its bandwidth, it turns out that the larger the bandwidth the smaller the value of the term $W''(f)$ at f_k. Thus, assuming short analysis windows and low values for $\rho_{2,k}$ we can approximate (12) as

$$W(f - f_k - \frac{\rho_{2,k}}{2\pi}) \approx W(f - f_k) - \frac{\rho_{2,k}}{2\pi} W'(f - f_k). \quad (13)$$

Consequently, from (11) and (13) it follows that [26]

$$S_k(f) \approx a_k \left[W(f - f_k - \frac{\rho_{2,k}}{2\pi}) + j\frac{\rho_{1,k}}{2\pi} W'(f - f_k) \right] \quad (14)$$

which is written in the time domain as

$$s_k(t) \approx a_k \left[e^{j(2\pi f_k + \rho_{2,k})t} + \rho_{1,k} t e^{j2\pi f_k t} \right] w(t). \quad (15)$$

From (15), it is clear that $\rho_{2,k}/2\pi$ accounts for the mismatch between the frequency of the kth component and the initial estimate of the frequency, f_k, while $\rho_{1,k}$ accounts for the normalized amplitude slope of the kth component. Another way to see this relationship, is to associate the time domain and the frequency domain properties of QHM. From (4) and (10) it follows that

$$\frac{\rho_{2,k}}{2\pi} = F_k(0) - f_k. \quad (16)$$

Therefore, $\rho_{2,k}/2\pi$ accounts for a frequency deviation between the initially estimated frequency, f_k, and the value of the instantaneous frequency at the center of the analysis window ($t = 0$). Similarly, for $\rho_{1,k}$, we have

$$\rho_{1,k} = \frac{\frac{dM_k(t)}{dt}|_{t=0}}{M_k(0)} \quad (17)$$

which shows that $\rho_{1,k}$ provides the normalized slope of the amplitude for the kth component, considering the instantaneous amplitude at the center of the analysis window.

III. Effects of Assumptions and Approximations on the Frequency Estimation Process

In the previous section, we showed that $\rho_{2,k}$ can be an estimator, under certain conditions, of the frequency mismatch between the original and the initially estimated frequencies of the underlying sine waves. In this section, we shall treat the effects of these conditions.

A. Effect and Importance of Window Length

From (1) and assuming a real signal, we see that there are $2K$ complex unknown parameters (K for a_k and K for b_k). Thus, the length of the analysis window should be at least $4K$ (in samples) in order to obtain stable least squares solutions. Moreover, low-frequency components need larger windows, and an empirical choice for the analysis window length is that this should be at least $2\lfloor f_s/\min_k f_k \rfloor$, where $\lfloor \cdot \rfloor$ denotes the floor operator while f_s is the sampling frequency. Furthermore, when the AM–FM signal is contaminated by noise, more samples (larger window) are needed in order to perform more robust and accurate estimation of the unknown parameters [28], [29]. On the other hand, when larger windows are used, the possibility the signal is nonstationary is higher, which may introduce errors and bias in the estimation procedure. Additionally, the analysis presented in the following section shows that the smaller the window length the more valid is the approximation in (13). From the above discussion, it should be clear that the length of the analysis window is very important and there is a tradeoff between the accuracy of the proposed algorithm and its robustness. As a general rule, we suggest the use of as small as possible window length.

B. Frequency Mismatch Computation

Because of the approximation in (13) and of the expected nonstationary character of the input signal, the suggested estimator $\rho_{2,k}$ for the frequency mismatch is generally not an unbiased estimator. The frequency mismatch (i.e., the error between the actual and the estimated values), cannot be, in the general case, computed analytically. Nevertheless, it is important to examine the adequacy and the validity of the proposed algorithm. In the case where the signal has multiple components and/or is characterized as nonstationary, the estimation of frequency mismatch will be analyzed numerically. However, in the case where the input signal is mono component and stationary, the estimation of the frequency mismatch can be derived analytically. Note also that the frequency parameter is the most significant since if the correct value of a frequency is known, then unbiased estimates of the amplitude and the phase for that frequency can be obtained through least-squares (LS) [28], [29]. Thus, the focus is on the frequency mismatch estimation.

Let us consider the mono component case ($K = 1$) of a stationary signal given by

$$\begin{aligned} x(t) &= Ae^{j(2\pi\zeta_1 t + \phi_1 t + \varphi)} \\ &= Ae^{j\varphi}(\cos(\phi_1 t) + j\sin(\phi_1 t))e^{j2\pi\zeta_1 t} \end{aligned} \quad (18)$$

where A is the amplitude, φ is the phase offset, ζ_1 is the carrier (or analysis) frequency and ϕ_1 is the (angular) frequency mismatch to be estimated. In the context of QHM, the input signal is modeled in one analysis frame as

$$s(t) = (a_1 + tb_1)e^{j2\pi f_1 t} w(t) \quad -T \le t \le T \quad (19)$$

where a_1 and b_1 are the unknown complex amplitude and slope, respectively, which are estimated through LS if f_1 is known. At the moment, we assume that $f_1 = \zeta_1$. It can be shown that the LS method involves the projection of the input signal onto two orthogonal basis functions: $e^{j2\pi f_1 t}$ and $te^{j2\pi f_1 t}$. Thus, for a rectangular window the complex amplitude is obtained by

$$\begin{aligned} a_1 &= \frac{\langle w(t)x(t), w(t)e^{j2\pi f_1 t}\rangle}{\langle w(t)e^{j2\pi f_1 t}, w(t)e^{j2\pi f_1 t}\rangle} \\ &= Ae^{j\phi_0}\frac{\sin(\phi_1 T)}{\phi_1 T} \end{aligned} \quad (20)$$

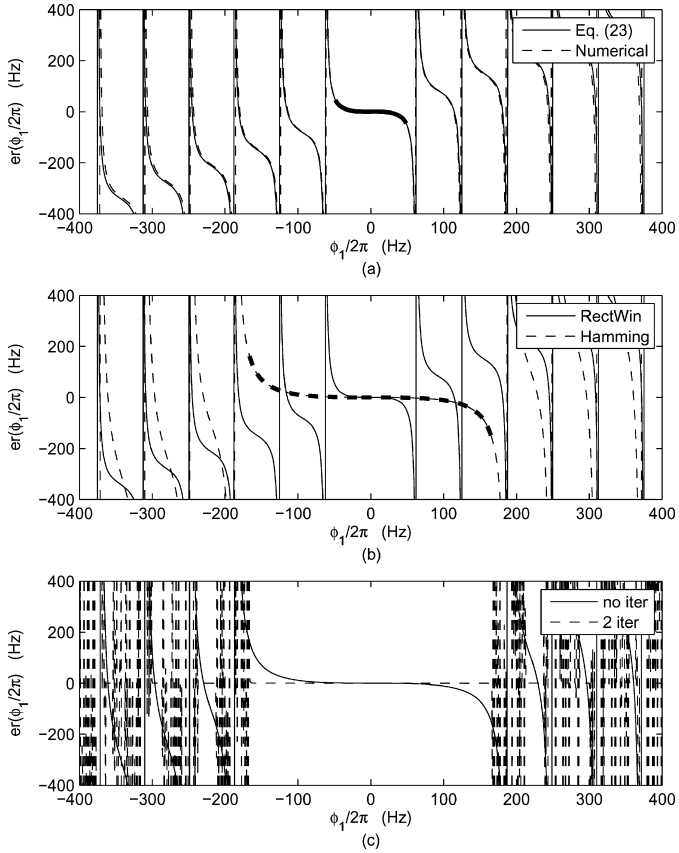

Fig. 2. Upper panel: the estimation error for a rectangular window computed analytically (solid line) and numerically (dashed line). Middle panel: the estimation error for a rectangular (solid line) and Hamming window (dashed line). Lower panel: error using the Hamming window (as in b) without (solid line) and with two iterations (dashed line). Note that the iterative estimation fails when $|\phi_1|/2\pi > B/3$.

where $\langle .,. \rangle$ denotes the inner product between functions, defined as

$$\langle x_1(t), x_2(t) \rangle = \int_{-T}^{T} x_1(t) x_2^*(t) dt.$$

The complex slope is obtained by

$$\begin{aligned} b_1 &= \frac{\langle w(t)x(t), w(t)te^{j2\pi f_1 t} \rangle}{\langle w(t)te^{j2\pi f_1 t}, w(t)te^{j2\pi f_1 t} \rangle} \\ &= 3jAe^{j\phi_0} \left(\frac{\sin(\phi_1 T)}{\phi_1^2 T^3} - \frac{\cos(\phi_1 T)}{\phi_1 T^2} \right). \end{aligned} \tag{21}$$

Then, the estimated value for ϕ_1 is given by

$$\begin{aligned} \hat{\phi}_1 &= \rho_{2,1} \\ &= 3 \left(\frac{1}{\phi_1 T^2} - \frac{\cot(\phi_1 T)}{T} \right). \end{aligned} \tag{22}$$

To inquire as to the properties of this estimator, it is worth computing its error in estimating the frequency mismatch (i.e., estimation error)

$$er(\phi_1) = \phi_1 - \hat{\phi}_1. \tag{23}$$

In the case of a mono-component signal and using a rectangular window, the estimation error can be computed analytically as above. Please note that ϕ_1 is an angular frequency expressed in rad/s. For the readability of the paper, we will present in Figs. 2 and 3 the estimation error in Hz, i.e., $er(\phi_1/2\pi)$ as a function of $\phi_1/2\pi$. Fig. 2(a) depicts the error for a rectangular window of 16 ms ($T = 8$ ms) obtained analytically via (22) (solid line), and numerically through LS computation of $\{a_1, b_1\}$ and then applying (10) (dashed line). Both ways to compute the estimation error provide the same result. Although there is no guarantee that this will be true in the general case, we suggest to compute numerically the estimation error to infer its analytical value, whenever the latter is not computationally tractable. In Fig. 2(a), the estimation error is small[1] (see the bold line) if the frequency mismatch is below 50 Hz. For a Hamming window, the error is small if the frequency mismatch is below 135 Hz as shown in Fig. 2(b).

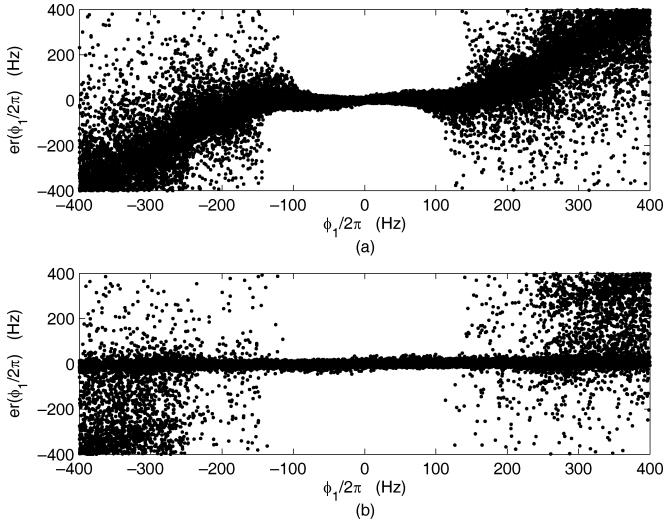

Fig. 3. Upper panel: the estimation error of ϕ_1 using a Hamming window of 16 ms length, after 10^5 Monte-Carlo simulations of (25). Lower panel: same as above, but with two iterations for the estimation of ϕ_1.

In order to get further insight on the role played by the analysis window, we can first notice from (20) and (21) that the Fourier Transform of the square of the analysis window appears in the LS estimates of a_1 and b_1 and consequently in the denominator of $\rho_{2,k}$. Thus, the frequency mismatch must be smaller than the bandwidth (i.e., the width of the main lobe [30]) of the squared analysis window. Note also that the bandwidth of a (squared) rectangular window of length $2T$ is $B = 1/T = 125$ Hz ($T = 8$ ms) while for a squared Hamming window we have $B = 3/T = 375$ Hz, which may explain why the region with small estimation error is about three times larger for a Hamming window than for a rectangular window. After testing a variety of window types and window lengths, we found that for mono-component stationary signals the estimation error is small when the frequency mismatch is smaller than one third of the bandwidth of the squared analysis window, i.e., when

$$\frac{|\phi_1|}{2\pi} < \frac{B}{3}. \tag{24}$$

Once an initial estimation of ϕ_1 is obtained through $\rho_{2,1}$, the initial estimation for frequency f_1 in (19) can be updated and then the input signal can be modeled again by QHM using now the updated frequency value, i.e., $\hat{f}_1 = f_1 + (1/2\pi)\rho_{2,1}$. Thus,

[1]By small, we mean that $|er(\phi_1)| < |\phi_1|$.

TABLE I
INTERVALS FOR EACH PARAMETER IN (25)

	min	max
α_1	$-2/T$	$2/T$
α_2	$-2/T^2$	$2/T^2$
α_3	$-2/T^3$	$2/T^3$
ϕ_1	$-16/T$	$16/T$
ϕ_2	$-2/T^2$	$2/T^2$
ϕ_3	$-2/T^3$	$2/T^3$

new estimations of ϕ_1 can be obtained iteratively. In Fig. 2(c), the estimation error is depicted for no iteration (solid line) and after two iterations (dashed line). We observe that the estimation error is considerably reduced (mainly is zero) if the initial frequency mismatch is smaller than $B/3$.

We now consider a more complicated case where the input signal has time-varying components (AM and FM)

$$x(t) = A(1+\alpha_1 t+\alpha_2 t^2+\alpha_3 t^3)e^{j(2\pi\zeta_1 t+\phi_1 t+\phi_2 t^2+\phi_3 t^3)} \quad (25)$$

where $t \in [-T, T]$ and the amplitude coefficients as well as the phase coefficients are real numbers. Based on the QHM model [i.e., on (19)], and assuming $f_1 = \zeta_1$, we would like to estimate ϕ_1. Since—even for such a mono-component signal—the estimation error cannot be computed analytically, we resort to evaluating the error by numerically computing $\rho_{2,k}$ and Monte-Carlo simulations. Each parameter in (25) takes values uniformly distributed on the intervals provided in Table I. The analysis window is as before a Hamming window of duration $2T = 16$ ms. Note that the synthetic signal under consideration changes its characteristics very fast. For example, if all the coefficients in (25) are set to zero except for α_1, which is set to $1/T$, then the instantaneous amplitude starts from 0 at the beginning of the frame and ends (after 16 ms) at the value of $2A$. Fig. 3(a) depicts the results of this test for 10^5 Monte-Carlo runs. It can be seen that a reasonable estimate can be obtained if the frequency mismatch is smaller than 100 Hz, which is less than in the stationary case (125 Hz). More importantly, even for very low-frequency mismatch a persistent error is present. This is why further updates of the frequencies [depicted in Fig. 3(b)] only provides marginal refinements but do not systematically decrease the estimation error at each iteration as is the case for a mono-component stationary complex exponential. In the following section, an adaptive scheme based on QHM is suggested which is able to model non stationary signals such as in (25).

IV. ADAPTIVE AM–FM DECOMPOSITION

In Section III-B, we showed that the iterative process suggested by QHM successfully adjusts the frequencies when the frequency components evolves slowly. In the case of fast variations, refinement of the frequencies can also be obtained, but only up to a certain point, since Fig. 3(b) clearly shows a persistent error even for a small frequency mismatch. This error is due to the fact that in such cases, stationary basis functions are used which are not adequate to model the input signal. Stated differently, iterating the update process by projecting onto a basis that does not fit the characteristics of the signal is not pertinent. In [31], a variant of this iterative method was proposed, using a basis of chirp functions in order to track linear variations of the frequencies.

In this section, we suggest a different approach where the basis functions are not restricted to be chirp or exponential functions but can adapt to the locally estimated instantaneous frequency/phase components. More specifically, an input signal is projected in a space generated by time varying nonparametric sinusoidal basis functions. The nonparametric basis functions are updated iteratively, minimizing the mean squared error at each iteration. We will refer to this modeling approach as adaptive QHM, or aQHM.

Initialization of aQHM is provided by QHM. Let $\hat{f}_k(t_l)$, $\hat{A}_k(t_l)$, and $\hat{\phi}_k(t_l)$, denote the updated frequencies, the corresponding amplitudes and phases at time instant t_l (center of analysis window), with $l = 1, \ldots, L$, where L is the number of frames. We recall that these parameters are estimated using QHM as follows:

$$\hat{f}_k(t_l) = F_k(0) = f_k(t_l) + \frac{\rho_{2,k}}{2\pi} \quad (26a)$$

$$\hat{A}_k(t_l) = M_k(0) = |a_k| \quad (26b)$$

$$\hat{\phi}_k(t_l) = \Phi_k(0) = \angle a_k. \quad (26c)$$

In case the distance between the consecutive analysis time instants correspond to one sample then, QHM provides an estimation of the instantaneous amplitude, $\hat{A}_k(t)$ and instantaneous phase $\hat{\phi}_k(t)$. Then, in aQHM the signal model is given as

$$s(t) = \left(\sum_{k=-K}^{K} (a_k + t b_k) e^{j(\hat{\phi}_k(t-t_l) - \hat{\phi}_k(t_l))} \right) w(t) \quad (27)$$

with $|t| \leq T$, where $2T$ denotes the duration of the analysis window. In contrast to (1) or (2), where the basis functions are stationary and parametric, in (27) these are not parametric neither necessarily stationary. Moreover, since the time-varying characteristics of the basis functions are based on measurements from the input signal, these are also adaptive to the current characteristics of the signal. In other words, they are adaptive to the input signal. Also, note that the old phase value at t_l (i.e., $\hat{\phi}_k(t_l)$) is subtracted from the instantaneous phase, in order to obtain a new phase estimate from (26c).

The term b_k in (27) plays the same role as in QHM; it provides a means to update the frequency of the underlying sine wave at the center of the analysis window t_l. The suggestions regarding the type and size of the analysis window made for QHM, are also valid for aQHM, since the same update mechanism is used. Therefore, an iterative analysis procedure using (27) is possible. In fact, using the initial estimates from QHM, new values of a_k and b_k are then computed using; however, in case of aQHM, the basis functions described in (27). Similar to QHM, the mean squared error between the signal and the model is minimized. The solution is straightforward and it is provided by least squares, as for QHM. Then, new instantaneous values are computed using (26). The procedure can be iterated until changes in the mean-squared error are not significant. At the

last step of aQHM, the signal can be finally approximated as the sum of its AM–FM components

$$\hat{x}(t) = \sum_{k=-K}^{K} \hat{A}_k(t) e^{j\hat{\phi}_k(t)}. \tag{28}$$

Therefore, aQHM suggests an algorithm for the adaptive AM–FM decomposition of a signal.

For applications such as speech analysis for the purpose of voice function assessment (i.e., voice disorders, analysis of tremor), and voice modification, the one sample time step is accepted. In other applications however, such as speech synthesis, larger steps are required. In this case, the instantaneous values of frequency, amplitude, and phase should be estimated from the set of parameters computed at every analysis time instant t_l. In SM, between two consecutive synthesis instants, linear interpolation for the amplitudes and cubic interpolation for phases were suggested [24]. In aQHM, many analysis time instants can be considered under the analysis window. For instantaneous amplitude estimation, we used splines, although other choices, such as linear interpolation, are possible. Such a simple solution is not, however, possible for the estimation of instantaneous phase. For this purpose, we will now describe a nonparametric approach as an alternative to the cubic interpolation method suggested in [24].

Based on the definition of phase, the instantaneous phase for the kth component can be computed as the integral of the computed instantaneous frequency. For instance, between two consecutive analysis time instants t_{l-1} and t_l, the instantaneous phase can be computed as

$$\check{\phi}_k(t) = \hat{\phi}_k(t_{l-1}) + \int_{t_{l-1}}^{t} 2\pi \hat{f}_k(u) du. \tag{29}$$

This solution however does not take into account the frame boundary conditions at t_l, which means that there is no guarantee that $\check{\phi}_k(t_l) = \hat{\phi}_k(t_l) + 2\pi M$, where M is the closet integer to $|\hat{\phi}_k(t_l) - \check{\phi}_k(t_l)|/(2\pi)$. We suggest modifying (29) in order to guarantee phase continuation over frame boundaries as follows:

$$\hat{\phi}_k(t) = \hat{\phi}_k(t_{l-1}) + \int_{t_{l-1}}^{t} 2\pi \hat{f}_k(u) + a \sin\left(\frac{\pi(u - t_{l-1})}{t_l - t_{l-1}}\right) du. \tag{30}$$

Note that the derivative of the instantaneous phase over time in both formulas provide the instantaneous frequency computed from t_{l-1} to t_l. In (30), the continuation of instantaneous frequency at the frame boundaries is also guaranteed by the use of the sine function (although other choices may be used as well). Moreover, it can be easily shown that using (30) the instantaneous phase at t_l will be equal to $\hat{\phi}_k(t_l) + 2\pi M$ if a is selected to be

$$a = \frac{\pi(\hat{\phi}_k(t_l) + 2\pi M - \check{\phi}_k(t_l))}{2(t_l - t_{l-1})} \tag{31}$$

where M is computed as before.

In Fig. 4, the two above formulas are compared on a synthetic example. The signal was analyzed frame-by-frame using QHM to compute the phase values at the frame boundaries. The actual instantaneous frequency contour (shown by a dashed line) of the signal is used in (29) and in (30) for the instantaneous phase computation. Then an estimation of the instantaneous frequency is obtained as the derivative over time of the computed instantaneous phases. Fig. 4(a) and (b) show the estimated instantaneous frequency (solid line) when (29) and (30) are used, respectively. It is obvious that in the former case there are spikes at the frame boundaries while in the latter case, the estimated instantaneous frequency is free of artefact and very close to the actual one.

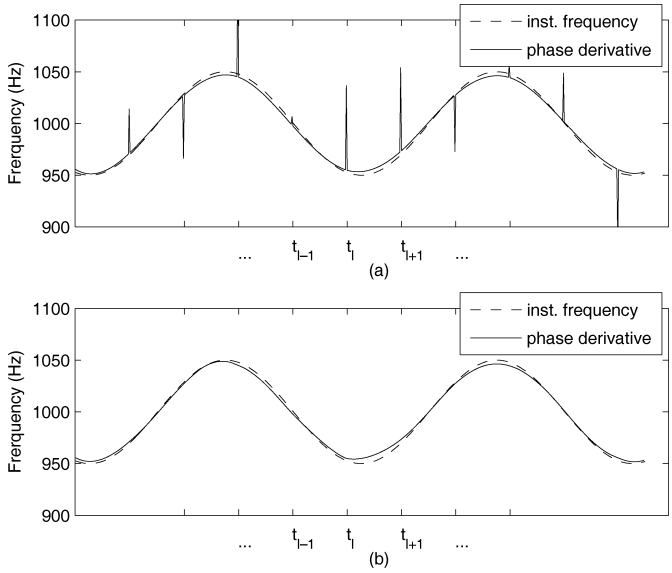

Fig. 4. Actual instantaneous frequency (dashed line) and estimated instantaneous frequency (solid line) as the derivative of the instantaneous phase computed from (29) (upper panel) and (30) (lower panel).

Having computing the instantaneous amplitude and phase information, the signal reconstruction is obtained by (28), as previously.

Summarizing, aQHM suggests a non-parametric AM–FM decomposition algorithm which proceeds by successive adaptations of the decomposition basis functions to the characteristics of the underlying sine waves of the input signal. A pseudocode of the algorithm is presented next.

Adaptive AM–FM decomposition alg. (aQHM)

1) Initialization (*QHM*):
Provide initial estimate $f_k^0(t_1)$
For $l = 1, 2, \ldots, L$
a) Compute a_k, b_k using $f_k^0(t_l)$ as initial frequency estimates in **(2)**
b) Update $\hat{f}_k^0(t_l)$ using **(26a)** and **(10)**
c) Compute $\hat{A}_k^0(t_l)$ and $\hat{\phi}_k^0(t_l)$ using **(26b) and (26c)**, respectively
d) $f_k^0(t_{l+1}) = \hat{f}_k^0(t_l)$
end
Interpolate $\hat{f}_k^0(t)$, $\hat{A}_k^0(t)$, $\hat{\phi}_k^0(t)$ as described
2) Adaptations:
For $i = 1, 2, \ldots$
For $l = 1, 2, \ldots, L$

a) Compute a_k, b_k using $\hat{\phi}_k^{i-1}(t)$ and **(27)**

b) Update $\hat{f}_k^i(t_l)$ using **(26a)** and **(10)**

c) Compute $\hat{A}_k^i(t_l)$ and $\hat{\phi}_k^i(t_l)$ using **(26b) and (26c)**, respectively

end

Interpolate $\hat{f}_k^i(t)$, $\hat{A}_k^i(t)$, $\hat{\phi}_k^i(t)$ as described

end

The aQHM algorithm is intuitively simple, and, as concerns its complexity, the most time-consuming part is the computation of a_k and b_k via LS at each time step. For each time step, the cost is $O((2K)^3+2KN)$, where N is the window length is samples. For comparison purposes, when there is only one component, the complexity of each step is $O(2N)$. This order of complexity is comparable to AM–FM decomposition algorithms with very low complexity such as the DESA algorithm [11].

Finally, having provided the algorithm for the iterative AM–FM decomposition, we would like to make the following comment that may be found useful during the implementation of the algorithm. As it was mentioned previously, frequencies are updated by adding a corrective term $\rho_{2,k}$ which depends on the amplitude as follows:

$$\rho_{2,k} = \frac{1}{2\pi}\frac{a_k^R b_k^I - a_k^I b_k^R}{|a_k|^2}. \tag{32}$$

Therefore, if the value of the kth amplitude a_k is close to zero, the estimated value for the kth instantaneous frequency will be erroneously high. This may create serious accuracy problems in the estimation of the other components of the signal and subsequently during the adaptation procedure. To overcome this, we suggest to exclude that component from the updating process and keep the initial value of the component. Alternatively, one could consider that this is not a component of the input signal. Such exclusion occurs if, at iteration i, $\hat{f}_k^i(t_l) \notin [f_{k-1}^i(t_l), f_{k+1}^i(t_l)]$. Since the corresponding amplitude has a low value, any solution will not have a serious effect in the signal reconstruction process.

V. Validation on Synthetic Signals

In this section, the performance of the suggested adaptive AM–FM decomposition algorithm (aQHM) will be validated on two AM–FM synthetic signals. The first signal is a chirp signal with a second-order polynomial for AM, while the second signal has two sinusoidally time-varying AM–FM components. Moreover, we will consider the case with additive noise in order to further validate the robustness of the proposed algorithm. Since the synthetic signals are parametric in AM–FM components, we will use as a validation metric the mean absolute error (MAE) between the true and the estimated AM–FM components. For comparison purposes, we suggest comparing aQHM with QHM and the estimation procedure used in the SM [24]. Regarding SM, at each analysis frame, we compute the Fourier transform of the windowed signal and determine the frequency and amplitude of each component of the signal by performing peak picking in the magnitude spectrum. To improve the frequency resolution of this standard approach, parabolic interpolation in the magnitude spectrum is used. The Fourier transform of the signal is computed at 2048 frequency bins (from 0 to 2π). Since we use a peak picking approach for the frequency estimation in the case of SM, the notion of frequency mismatch is irrelevant in this case. Therefore, frequency mismatch is only considered for the QHM and aQHM cases. Thus, it is worth noting that QHM and aQHM have to cope with both the initial frequency mismatch and the additive noise, while SM has to cope only with the additive noise.

TABLE II
Mean Absolute Error (MAE) of AM and FM Components for QHM, aQHM, and SM Without Noise, and With Complex Additive White Gaussian Noise at 30-dB and 10-dB Local SNR

SNR	Method	AM	FM (Hz)
∞dB	*QHM*	0.42	2.43
	aQHM	0.009	0.001
	SM	0.42	2.43
$30dB$	*QHM*	0.42	2.43
	aQHM	0.02	0.22
	SM	0.42	2.44
$10dB$	*QHM*	0.43	3.13
	aQHM	0.13	2.23
	SM	0.42	3.57

For all synthetic examples, we will consider a sampling frequency $f_s = 16\,000$ Hz and for all methods the same fixed-length Hamming window will be used. The time step (hop size) will be fixed to one sample ($t_{l+1} - t_l = 1$).

First let us consider the following monocomponent ($K = 1$) chirp signal with a second-order amplitude modulation

$$x(t) = (11 - 340t + 4000t^2)e^{j2\pi(280t+19500t^2)},\ t \in [0, 0.1] \tag{33}$$

whose instantaneous frequency is $f(t) = 280 + 39\,000t$ (Hz). Based on the analysis presented in Section III, the maximum frequency mismatch between the initial estimate and the actual frequency of the signal is defined as one third of the bandwidth of the squared analysis window. In this experiment, we will use a 10-ms Hamming window, for which the squared window has bandwidth $B = 3/T = 300$ Hz. Therefore, the maximum frequency mismatch is ± 100 Hz. The center of the first analysis window is located at 5 ms, where the actual instantaneous frequency is 475 Hz. We set the initial frequency estimate to 400 Hz which means that there is frequency mismatch of 75 Hz. We consider the case without noise as well as with complex additive white Gaussian noise of 30-dB and 10-dB local SNR. In case of additive noise, the average performance of each algorithm was measured based on 10^4 simulations.

Table II reports the MAE between the estimated and the actual AM and FM component, for QHM, aQHM, and SM. Two iterations were used in the case of QHM, and two iterations (or adaptations) were used for aQHM. First, we observe that aQHM outperforms all the other approaches, while QHM and SM present about the same performance. When there is no additive noise, aQHM efficiently resolves the nonstationary character of the signal in contrast to the other two approaches. As the local SNR decreases, the performance of aQHM decreases too, while the performance of QHM and SM remains about the same. In this experiment, estimation error has mainly two sources. One stems from the nonstationarity characteristics of the input signal. The other stems from the additive noise. The former source seems

TABLE III
MEAN ABSOLUTE ERROR FOR QHM, AQHM AND SM FOR THE TWO-COMPONENT SYNTHETIC AM–FM SIGNAL, WITHOUT NOISE, AND WITH COMPLEX ADDITIVE WHITE GAUSSIAN NOISE AT 10-dB LOCAL SNR.

SNR	Method	AM1	AM2	FM1 (Hz)	FM2 (Hz)
∞dB	*QHM*	0.36	0.38	69.99	69.74
	aQHM	0.06	0.11	21.43	20.14
	SM	0.32	0.36	83.50	81.99
$10dB$	*QHM*	0.38	0.40	70.70	70.48
	aQHM	0.07	0.11	23.13	22.27
	SM	0.37	0.38	84.48	82.84

to be more important for the case of QHM and SM, while the latter affects more aQHM. However, even for 10-dB local SNR, aQHM is more than 200% and 60% better than SM (in terms of MAE) in estimating the AM, and FM components, respectively.

Let us consider a two component AM–FM signal of the form

$$s(t) = 2(1+0.4\cos(2\pi 30t))e^{j(2\pi 700t+\cos(2\pi 130t))} + 2(1+0.3\cos(2\pi 50t))e^{j(2\pi 1000t+\cos(2\pi 130t))} \quad (34)$$

where instantaneous amplitudes and frequencies present sinusoidally time-varying characteristics. Note that the AM of the second component (AM2) varies faster than the corresponding AM of the first component (AM1), and that frequency modulation for both components is important: 130 cycles per second. A Hamming window of length 16 ms is used. In case of QHM, an initial frequency mismatch of 32 Hz is assumed for both components, which is a bit below the maximum allowable mismatch (namely $B/3 = 41$ Hz in this example) The performance of the algorithms is tested without additive noise and with complex additive white Gaussian noise of 10-dB local SNR. As previously, in case of additive noise, the average performance of each algorithm was measured based on 10^4 simulations. In Table III, the performance of QHM, aQHM, and SM is shown in terms of MAE. It is worth noting here, that over the duration of the window length, the signal components change quickly; therefore, it may be seen as a highly nonstationary signal. Specifically, in 16 ms, about two periods of the FM components are observed. Regarding amplitude modulation, this is about half of one period for AM1 and about one period for AM2. Therefore, more iterations in *aQHM* are expected to reduce the MAE for each of these components. Indeed, aQHM required 11 iterations (or adaptations) to converge (meaning that no significant changes in MAE were observed) in case of clean data and eight adaptations in case of additive noise. QHM required two iterations.

As in the mono component signal, QHM and SM have similar performance regarding the AM components, while for the FM components, QHM performs better than SM. It seems that the presence of two components affects more SM than QHM because of the interference between the components. Also, aQHM outperforms both QHM and SM for all the parameters and under all conditions. In contrast to the mono component case, however, aQHM is not so sensitive to the additive noise. In this case, the source of the estimation error, because of the highly nonstationary character of the input signal, is more important than the corresponding error source because of the presence of noise. Therefore, decreasing the SNR, does not significantly affect the performance of aQHM.

VI. AM–FM DECOMPOSITION OF VOICED SPEECH

The suggested iterative AM–FM decomposition algorithm based on aQHM can be applied on voiced speech signals in a straightforward way. Actually, the aQHM algorithm can be applied on large voiced speech segment. Indeed, assuming that voiced speech is quasi-periodic and that the frequency content of voiced speech signals does not change very fast, then we only need to provide the fundamental frequency of the first voiced frame at the beginning of the voiced segment, $f_0(t_1)$ and then assume $f_k^0(t_1) = kf_0(t_1)$. After the QHM analysis of the first voiced frame, an updated set of $\hat{f}_k$ will be obtained for that frame. The updated set of frequencies can then be used as initial estimates for the next voiced frame. Continuing in this way, the whole voiced region will be analyzed by providing just the fundamental frequency for the first frame of the voiced segment. Another option could be to use the average fundamental frequency of the voiced segment as an initial frequency estimation $f_0(t_1)$. It is worth noting that the accuracy of the fundamental frequency estimator is not crucial for aQHM, since frequency mismatches are easily corrected (of course, we exclude cases of fundamental frequency doubling or halving). For evaluation of the adaptive AM–FM decomposition algorithm we propose to reconstruct the original signal by using the estimated AM–FM components and measure then the signal-to-reconstruction-error ratio (SRER) defined as

$$SRER = 20\log_{10}\frac{\sigma_{x(t)}}{\sigma_{x(t)-\hat{x}(t)}} \quad (35)$$

where σ_x denotes the standard deviation of x, and $\hat{x}(t)$ is the reconstructed signal computed from (28).

In this section, we will compare aQHM with QHM and SM in terms of quality of voiced speech signal reconstruction. If time step is one sample, then all algorithms have an estimation of the instantaneous amplitude and phase as these are estimated at the center of their analysis windows. For SM, parabolic interpolation in the magnitude spectrum is used in order to improve frequency resolution. Phases are then computed from the phase spectrum by considering the phase at the point nearest the interpolated frequency. As previously, the Fourier transform of the signal is computed at 2048 frequency bins (from 0 to 2π).

In Fig. 5(a), a segment from a voiced speech signal generated by a male speaker is shown (sampling frequency 16 kHz). The analysis was performed using a Hamming window of 16 ms and with one sample as step size. For QHM, we set $f_0(t_1) = 140$ Hz (the average fundamental frequency of the segment) and $K = 40$. Only one iteration was used for QHM. The results from QHM were used as an initialization for aQHM, where only one adaptation was performed. Regarding SM, the most prominent 40 components in the magnitude spectrum were selected after peak picking and parabolic interpolation. We verified that the frequency of the selected peaks were closely related to the updated frequencies, $\hat{f}_k$ of QHM. The estimated instantaneous amplitude and phase information for all the methods (QHM, aQHM, and SM) were then used to reconstruct the speech signal as in (28). The reconstruction error for each method is depicted in Fig. 5(b)–(d), for QHM, aQHM, and SM, respectively. Again, aQHM provides the best reconstruction compared to the other

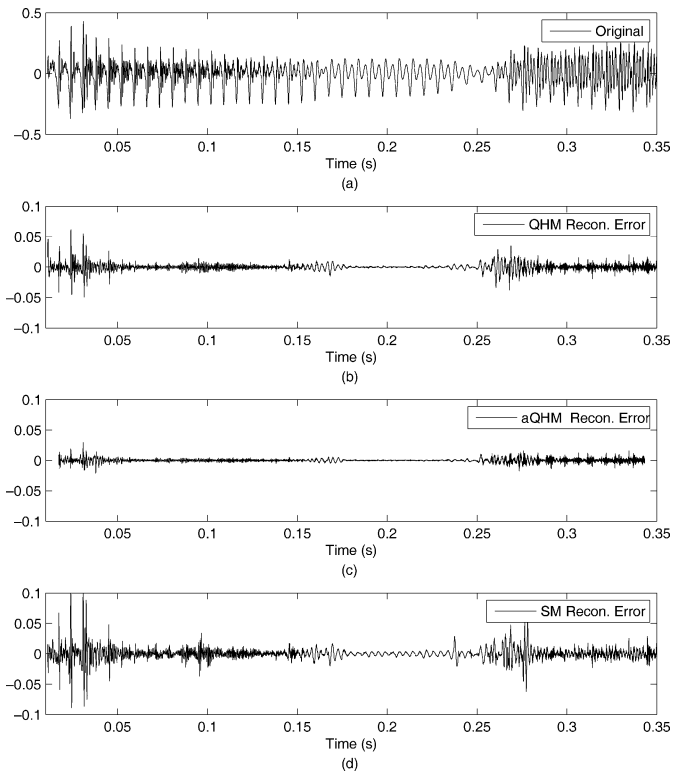

Fig. 5. (a) Original speech signal and reconstruction error for (b) QHM after one iteration, (c) aQHM after one adaptation, and (d) SM, using $K = 40$ components.

two alternatives even if only one iteration is applied. The SRER is 19.5 dB for SM, 24.1 dB for QHM, and 30.5 dB for aQHM.

In case the step size is bigger than one sample, then the instantaneous amplitudes and phases should be computed from the estimated parameters at the analysis time instances. For QHM and aQHM, the instantaneous amplitude and instantaneous frequency are computed using splines. The instantaneous phase is computed from (30). For SM, instantaneous amplitude is computed with linear interpolation while for the instantaneous phase, cubic interpolation is used [24]. Using three different step sizes, namely 1, 2, and 4 ms, we analyzed and reconstruct about 200 minutes of voiced speech from 20 male and 20 female speakers (about 5 minutes per speaker) from the TIMIT database. The sampling frequency is 16 000 Hz. Assuming an average pitch of 100 and 160 Hz for male and female speakers, respectively, we used Hamming windows of fixed length; 2.5 times the average pitch period. Thus, we used a fixed length analysis window: 25 ms for male and 15 ms for female speakers. The same windows was used for all the algorithms. The number of components was set to $K = 40$ for male voices and to $K = 30$ for female voices. The average and standard deviation of the SRER (dB) is provided in Table IV along with various step sizes. For QHM, only one iteration was used. In the case of aQHM, the relative change of SRER is observed. When this is below a threshold, convergence is assumed. Table IV presents the mean number of adaptations needed for aQHM to converge. Since only aQHM suggests an adaptive algorithm, this column of the table is considered only for aQHM.

TABLE IV
MEAN AND STANDARD DEVIATION OF SIGNAL-TO-RECONSTRUCTION-ERROR RATIO (IN dB) FOR APPROXIMATELY FIVE MINUTES OF VOICED SPEECH FOR 20 MALE AND 20 FEMALE SPEAKERS FROM TIMIT.

		Male		Female		
Step	Method	Mean	Std	Mean	Std	NoA
$1ms$	*QHM*	23.9	4.9	29.1	4.7	–
	aQHM	29.1	4.4	34.1	4.3	2.4
	SM	17.5	5.2	21.1	6.0	–
$2ms$	*QHM*	22.4	5.5	28.3	4.9	–
	aQHM	28.3	4.3	33.6	4.4	2.6
	SM	17.8	5.1	21.4	5.9	–
$4ms$	*QHM*	19.9	6.1	25.7	5.7	–
	aQHM	26.2	4.9	30.9	4.5	2.8
	SM	18.2	4.9	20.9	5.5	–

We observe that the reconstruction error has lower power for the female voices than for the male voices. This is expected as the duration of analysis window is shorter in this case. As already mentioned, step size is a crucial parameter, in QHM, and aQHM. Results show that there is a minor decrease in the performance of these two algorithms when the time step is increasing. Comparing aQHM with SM, we see that the improvement in SRER is between 56% (for males) and 55% (for females), thus providing an average improvement of over 55%. Compared to QHM, aQHM provides an average improvement of 22% in SRER.

VII. CONCLUSION

In this paper, we showed the robustness of the QHM as a frequency estimator and we expanded QHM to a novel speech representation referred to as adaptive QHM, or aQHM. The resulting representation suggests an iterative nonparametric AM–FM decomposition of speech. The algorithm was validated on synthetic mono and multi component AM–FM signals with or without the presence of additive noise. The proposed approach was also applied to estimate the AM–FM components of voiced speech. Signal reconstruction using the estimated instantaneous components of the signal leads to a high-quality reconstruction of voiced speech. An average signal to reconstruction error ratio of about 30 dB was obtained, which shows the accuracy of the suggested estimator. Furthermore, based on the proposed decomposition algorithm, high resolution time–frequency representations of voiced speech can be obtained revealing details in the structure of speech signals. Based on these results, the proposed method is expected to be useful in many speech applications including speech analysis, speech synthesis and modifications, and objective voice function assessment.

REFERENCES

[1] J. Laroche, Y. Stylianou, and E. Moulines, "HNM: A simple, efficient harmonic plus noise model for speech," in *Proc. Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, New Paltz, NY, Oct. 1993, pp. 169–172.

[2] A. W. Rihaczek, *Principles of High-Resolution Radar*. Norwood, MA: Artech House, 1985.

[3] M. S. Pattichis, C. S. Pattichis, M. Avraam, A. Bovik, and K. Kyriakou, "AM–FM texture segmentation in electron microscopic muscle imaging," *IEEE Trans. Med. Imag.*, vol. 19, no. 12, pp. 1253–1257, Dec. 2000.

[4] T. F. Quatieri, *Speech Signal Processing*. Upper Saddle River, NJ: Prentice-Hall, 2002, Signal Processing Series.
[5] A. Rao and R. Kumaresan, "On decomposing speech into modulated components," *IEEE Trans. Speech Audio Process.*, vol. 8, no. 3, pp. 240–254, May 2000.
[6] L. Cohen, *Time–Frequency Analysis*. New York: Prentice-Hall, 1995.
[7] P. J. Loughlin and B. Tacer, "On the amplitude- and frequency-modulation decomposition of signals," *J. Acoust. Soc. Amer.*, vol. 100, pp. 1594–1601, Sep. 1996.
[8] F. Auger and P. Flandrin, "Improving the readability of time–frequency and time-scale representations by the reassignment method," *IEEE Trans. Signal Process.*, vol. 43, no. 5, pp. 1068–1089, May 1995.
[9] D. Vakman, "On the analytic signal, the Teager-Kaiser energy algorithm, and other methods for defining amplitude and frequency," *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 791–797, Apr. 1996.
[10] J. F. Kaiser, "On a simple algorithm to calculate the 'energy' of a signal," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Albuquerque, NM, Apr. 1990, pp. 381–384.
[11] P. Maragos, J. Kaiser, and T. Quatieri, "On separating amplitude from frequency modulations using energy operators," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, San Francisco, CA, Mar. 1992, pp. 1–4.
[12] B. Santhanam, "Multicomponent AM–FM energy demodulation with applications to signal processing and communications," Ph.D., Georgia Inst. of Technol., Atlanta, 1997.
[13] J. H. L. Hansen, L. Gavidia-Ceballos, and J. F. Kaiser, "A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment," *IEEE Trans. Biomed. Eng.*, vol. 45, no. 3, pp. 300–312, Mar. 1998.
[14] W. C. Pai and P. C. Doerschuk, "Statistical AM–FM models, extended Kalman filter demodulation, Cramer–Rao bounds and speech analysis," *IEEE Trans. Signal Process.*, vol. 48, no. 8, pp. 2300–2313, Aug. 2000.
[15] T. F. Quatieri, T. E. Hanna, and G. C. O'Leary, "AM–FM separation using auditory-motivated filters," *IEEE Trans. Speech Audio Process.*, vol. 5, no. 5, pp. 465–480, Sep. 1997.
[16] S. Mann and S. Haykin, "The Chirplet transform: A generalization of Gabor's logon transform," in *Proc. Vision Interface*.
[17] L. Weruaga and M. Kepesi, "The Fan-Chirp transform for non-stationary harmonic signals," *Signal Process.*, vol. 87, pp. 1504–1522, 2007.
[18] R. Dunn and T. F. Quatieri, "Sinewave analysis/synthesis based on the Fan-Chirp transform," in *Proc. Workshop Applicat. Signal Process. Audio Acoust. (WASPAA)*, Oct. 2007, pp. 16–19.
[19] S. Peleg and B. Friedlander, "The discrete polynomial-phase transform," *IEEE Trans. Signal Process.*, vol. 43, no. 8, pp. 1901–1914, Aug. 1995.
[20] G. Zhou, G. B. Giannakis, and A. Swami, "On polynomial phase signals with time-varying amplitudes," *IEEE Trans. Signal Process.*, vol. 44, no. 4, pp. 848–861, Apr. 1996.
[21] M. Betser, P. Collen, G. Richard, and B. David, "Estimation of frequency for AM/FM models using the phase vocoder framework," *IEEE Trans. Signal Process.*, vol. 56, no. 2, pp. 505–517, Feb. 2008.
[22] B. Friedlander and J. M. Francos, "Estimation of amplitude and phase parameters of multicomponent signals," *IEEE Trans. Signal Process.*, vol. 43, no. 4, pp. 917–926, Apr. 1995.
[23] S. Gazor and R. R. Far, "Adaptive maximum windowed likelihood multicomponent AM–FM signal decomposition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 2, pp. 497–491, Mar. 2006.
[24] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-34, no. 4, pp. 744–754, Aug. 1986.
[25] J. Laroche, "A new analysis/synthesis system of musical signals using Prony's method. Application to heavily damped percussive sounds," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, Glasgow, U.K., May 1989, pp. 2053–2056.
[26] Y. Pantazis, O. Rosec, and Y. Stylianou, "On the properties of a time-varying quasi-harmonic model of speech," in *Proc. Interspeech*, Brisbane, Australia, Sep. 2008, pp. 1044–1047.
[27] Y. Stylianou, "Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification," Ph.D. dissertation, Ecole Nationale Supérieure des Télécomm., Paris, France, 1996.
[28] S. M. Kay, *Funtamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1993.
[29] P. Stoica, H. Li, and J. Li, "Amplitude estimation of sinusoidal signals: Survey, new results, and an application," *IEEE Trans. Signal Process.*, vol. 48, no. 2, pp. 338–352, Feb. 2000.
[30] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Englwood Cliffs, NJ: Prentice-Hall, 1989, Signal Processing Series.
[31] Y. Pantazis, O. Rosec, and Y. Stylianou, "Chirp rate estimation of speech based on a time-varying quasi-harmonic model," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, 2009, pp. 3985–3988.

Yannis Pantazis received the B.A. and M.S. degrees in computer science from the University of Crete, Heraklion, Greece, in 2004 and 2006, respectively. He is currently pursuing the Ph.D. degree in signal processing from the University of Crete.

Since 2003, he has been a member of Institute of Computer Science, FORTH, Heraklion, Greece. His research interests include speech analysis and synthesis, voice pathology, and signal modeling and estimation.

Olivier Rosec (M'07) received the M.Sc. degree in electronics and the Ph.D. degree in signal processing both from the Université de Bretagne Occidentale, Brest, France, in 1995 and 2000, respectively.

He is a Research Engineer at Orange Labs, Lannion, France. From 1996 until 1999, he was with the Acoustics and Seismics Department of IFREMER, Brest, as a Research Engineer. Since 2000, he has been a Research Engineer in signal processing within the Speech Synthesis Team, Orange Labs. His current research interests include speech modeling and analysis, speech modification and coding for concatenative synthesis, voice transformation, and voice conversion.

Yannis Stylianou (M'95) received the Diploma of Electrical Engineering from the National Technical University of Athens, Athens, Greece, in 1991 and the M.Sc. and Ph.D. degrees in signal processing from the Ecole National Superieure des Telecommunications, ENST, Paris, France, in 1992 and 1996, respectively.

He is an Associate Professor in the Department of Computer Science, University of Crete, and an Associate Researcher in the Networks and Telecommunications Laboratory, Institute of Computer Science at FORTH. From 1996 until 2001, he was with AT&T Labs Research, Murray Hill and Florham Park, NJ, as a Senior Technical Staff Member. In 2001, he joined Bell-Labs Lucent Technologies, Murray Hill (now Alcatel-Lucent). Since 2002, he has been with the Computer Science Department, University of Crete and the Institute of Computer Science at FORTH. He has over 100 peer-reviewed papers, and holds nine U.S. patents. He is on the Editorial Board of the *Journal of Electrical and Computer Engineering*, Hindawi, Associate Editor of the *EURASIP Journal on Speech, Audio, and Music Processing*, and of the *EURASIP Research Letters in Signal Processing*. He is member of ISCA and the Technical Chamber of Greece.

Prof. Stylianou is on the Board of the International Speech Communication Association (ISCA), member of the IEEE Speech and Language Technical Committee, and of the IEEE Multimedia Communications Technical Committee. He was Associate Editor for the IEEE SIGNAL PROCESSING LETTERS.