AN EXTENSION OF THE ADAPTIVE QUASI-HARMONIC MODEL

George P. Kafentzis[1] , *Yannis Pantazis*[2] , *Olivier Rosec*[1] *and* *Yannis Stylianou*[3]

[1]Orange Labs TECH/ASAP/VOICE, Lannion, France
[2]Department of Mathematics and Statistics, University of Massachusetts, Amherst, MA, USA
[3]Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece
email: george.kafentzis@orange.com, pantazis@math.umass.edu, olivier.rosec@orange.com and styliano@ics.forth.gr

ABSTRACT

In this paper, we present an extension of a recently developed AM-FM decomposition algorithm, which will be referred to as the extended adaptive Quasi-Harmonic Model (eaQHM). It was previously shown that the adaptive Quasi-Harmonic Model (aQHM) [1] is an efficient AM-FM decomposition algorithm with applications in speech analysis. In this paper, we show that a simple extension of the aQHM algorithm to include not only frequency but also amplitude adaptation results in higher performance in terms of Signal-to-Reconstruction-Error Ratio (SRER). To support our hypothesis, eaQHM is tested both on synthetic signals and on a subset of the ARCTIC database of speech. Overall, compared with aQHM, eaQHM improves the SRER by more than 2 dB, on average.

Index Terms— adaptive Quasi-Harmonic Model, Frequency estimation, Amplitude estimation, Speech analysis

1. INTRODUCTION

Speech modeling has found one of its principal exponents in the sinusoidal model [2], which has been successfully applied in speech coding and speech modifications. Sinusoidal modeling expresses a speech signal as a sum of sinewaves, with constant amplitudes and frequencies over successive frames. The Harmonic plus Noise Model, HNM [3], is another well known model, with applications in speech synthesis and speech modification. The HNM decomposes speech into two components; the harmonic component, which represents the deterministic part of speech, and the noise component, which represents the stochastic part of speech. In this way, high quality prosodic modifications can be achieved. Moreover, time-varying amplitude and frequency component separation is of great interest in speech processing sciences due to its strong relationship with the speech production mechanisms [4].

However, frequency estimation sensitivity is a major drawback of these models. Poor estimation of frequencies yields very high modeling error and results in artifacts in the reconstructed speech signal. Recently, a time-varying sinusoidal representation referred to as the Quasi-Harmonic Model (QHM) has been proposed, which has been shown to have low sensitivity to frequency estimation errors [1]. This is due to the fact that the model contains a frequency mismatch corrector which is able to estimate and consequently correct frequency mismatches. Moreover, in [1], an adaptive QHM (aQHM) was suggested, where the model adapts to the local characteristics of the signal and provides high-quality speech reconstruction. In [5], the aQHM was used for the accurate estimation of amplitude and frequency (AM-FM) modulations in speech. The aQHM models a signal as a sum of exponentials with linearly time-varying amplitudes and non-stationary phases. Hence, the model phase parameters adapt to the local characteristics of the signal phase and thus, its frequency. However, in many cases, e.g. in speech, rapid nonlinear amplitude changes occur within short time intervals. As a consequence, phase adaptation alone is not enough and amplitude adaptation is also necessary in these cases.

In this paper, we present an extension of the aQHM, referred to as the extended adaptive Quasi-Harmonic Model (eaQHM). In the eaQHM, the adaptation process includes not only the frequencies but also the amplitudes of the basis functions onto which the signal is projected. This yields a model which can adapt to the analyzed signal better than aQHM. Experiments conducted on synthetic and real speech signals show that eaQHM improves the Signal-to-Reconstruction-Error Ratio (SRER) compared to that obtained by aQHM. The latter is demonstrated through synthetic and real speech signals.

The rest of the paper is organized as follows. In Section 2 we will briefly review the Quasi-Harmonic Model, QHM, and the adaptivity algorithm, aQHM. Section 3 presents the extension of aQHM, eaQHM. Section 4 shows a synthetic signal example and addresses the robustness of the eaQHM in white Gaussian noise, compared to aQHM. Section 5 provides real speech analysis results for both algorithms. Finally, Section 6 concludes the paper.

2. OVERVIEW OF QHM AND aQHM

In sinusoidal modeling, a signal can be represented as follows:

$$x(t) = \left(\sum_{k=1}^{K} a_k e^{j2\pi f_k t} \right) w(t), \quad t = -N, ..., N \tag{1}$$

where K is the number of components with complex amplitudes a_k at frequencies f_k, and $w(t)$ is the analysis window. Let us assume that f_k denote the correct frequencies of the signal components. In sinusoidal modeling, frequencies are estimated first (e.g., by peak-picking, by considering harmonics of a fundamental frequency, etc.), before the estimation of the complex amplitudes. The estimated frequencies will be denoted here by $\hat{f}_k$. Then, we may write:

$$f_k = \hat{f}_k + \eta_k, \quad k = 1, ..., K \tag{2}$$

If the error, η_k, is high, then the estimation of the complex amplitudes, a_k, is severely biased.

Work of Yannis Pantazis was partially supported by the National Science Foundation (USA, CMMI-0835673). Work of Yannis Stylianou was partially supported by FP7-FET-OPEN programme LISTA. The project LISTA acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission, under FET-Open grant number: 25623. This work was also supported by France Telecom R&D agreement A09141.

2.1. Quasi-Harmonic Model, QHM

To cope with this problem, in [5] and [6] the use of the Quasi-Harmonic Model (QHM) for the representation of speech was suggested:

$$x(t) = \left(\sum_{k=1}^{K} (a_k + t b_k) e^{j2\pi \hat{f}_k t} \right) w(t), \quad t = -N, ..., N \tag{3}$$

where b_k denotes the complex slope of the kth component. In the frequency domain, the kth component is written as:

$$X_k(f) = a_k W(f - \hat{f}_k) + j \frac{b_k}{2\pi} W'(f - \hat{f}_k) \tag{4}$$

where $W(f)$ is the Fourier transform of the analysis window and $W'(f)$ is the derivative of $W(f)$ over f. In [6], it was shown that QHM is able to correct frequency mismatches using the projection of b_k onto a_k. Indeed, it was shown that an estimation of η_k can be obtained by:

$$\hat{\eta}_k = \frac{\rho_{2,k}}{2\pi} = \frac{1}{2\pi} \frac{a_k^R b_k^I - a_k^I b_k^R}{|a_k|^2}, \tag{5}$$

where a_k^R, b_k^R and a_k^I, b_k^I are the real and imaginary parts of a_k and b_k, respectively. In [6], it was also shown that this correction depends on the magnitude of $\rho_{2,k}$ and the value of the term $W''(f)$ at f_k. The estimation of a_k, b_k is performed via Least Squares (LS).

2.2. Adaptive Quasi-Harmonic Model, aQHM

To better model the speech signal, especially its non stationary part, an adaptive QHM model has been suggested.

$$x(t) = \left(\sum_{k=1}^{K_l} (a_k + t b_k) e^{j(\hat{\phi}_k(t+t_l) - \hat{\phi}_k(t_l))} \right) w(t), \;\; t \in [-T_l, T_l] \tag{6}$$

where $\phi_k(t)$ denotes the phase function of the k^{th} component and t_l is the center of the analysis window. The term b_k 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. Given the samples of the input signal in vector s, the model parameters are found via LS:

$$\begin{bmatrix} \hat{\mathbf{a}} \\ \hat{\mathbf{b}} \end{bmatrix} = (E^H W^H W E)^{-1} E^H W^H W s \tag{7}$$

where $\mathbf{a} = [a_1, \cdots, a_{K_l}]$, $\mathbf{b} = [b_1, \cdots, b_{K_l}]$, W is the matrix containing the window values in the diagonal, s is the input signal vector, the matrix E is defined as $E = [E_0 | E_1]$, the submatrices E_i, $i = 0, 1$ have elements given by

$$(E_0)_{n,k} = e^{j(\phi_k(t_n + t_l) - \phi_k(t_l))} \tag{8}$$

and

$$(E_1)_{n,k} = t_n e^{j(\phi_k(t_n + t_l) - \phi_k(t_l))} = t_n (E_0)_{n,k}, \tag{9}$$

and the instantaneous phase of the k^{th} component can be computed as

$$\hat{\phi}_k(t) = \int_{t_l}^{t_l + t} 2\pi f_k(u) du, \;\; t \in [-T_l, T_l], \tag{10}$$

where $f_k(t)$ is the frequency trajectory of the k^{th} component. In contrast to QHM, where the argument of the basis functions is parametric and stationary, in aQHM the argument of the basis functions is neither parametric nor necessarily stationary. Moreover, the aQHM basis functions use the instantaneous phases which have been estimated from the input signal. In that sense these are also adaptive to the estimates of the current characteristics of the signal.

3. EXTENSION OF aQHM, eaQHM

The extension of aQHM to include amplitude adaptation is straightforward:

$$x(t) = \left(\sum_{k=1}^{K_l} (a_k + t b_k) \frac{A_k(t + t_l)}{A_k(t_l)} e^{j(\hat{\phi}_k(t+t_l) - \hat{\phi}_k(t_l))} \right) w(t), \quad t \in [-T_l, T_l] \tag{11}$$

where t_l is still the center of the analysis window and $A_k(t)$ is the instantaneous amplitude of the k^{th} component. The estimation of the unknown parameters of eaQHM is similar to that of QHM:

$$\begin{bmatrix} \hat{\mathbf{a}} \\ \hat{\mathbf{b}} \end{bmatrix} = (E_e^H W^H W E_e)^{-1} E_e^H W^H W s \tag{12}$$

where $\mathbf{a} = [a_1, \cdots, a_{K_l}]$, $\mathbf{b} = [b_1, \cdots, b_{K_l}]$, W is the matrix containing the window values in the diagonal, s is the input signal vector, the matrix E_e is defined as $E_e = [E_{e0} | E_{e1}]$, and the submatrices E_{ei}, $i = 0, 1$ have elements given by

$$(E_{e0})_{n,k} = \frac{A_k(t_n + t_l)}{A_k(t_l)} e^{j(\phi_k(t_n + t_l) - \phi_k(t_l))} \tag{13}$$

and

$$(E_{e1})_{n,k} = \frac{A_k(t_n + t_l)}{A_k(t_l)} t_n e^{j(\phi_k(t_n + t_l) - \phi_k(t_l))} = t_n (E_{e0})_{n,k}, \tag{14}$$

It is clear that the basis functions are adapted to the local amplitude characteristics of the signal. Note that the instantaneous amplitude $A_k(t)$ is divided by $A_k(t_l)$, so as to have unit value at the center of the analysis window.

Like aQHM, eaQHM requires an initialization step, so QHM will be used for this purpose, although any frequency estimation algorithm can be used. Thus, the initials steps consist of the following:

$$\hat{f}_k^0(t_l) = \hat{f}_k^0(t_{l-1}) + \hat{\eta}_k \tag{15}$$

$$\hat{A}_k^0(t_l) = |a_k^l|, \;\; \hat{\phi}_k^0(t_l) = \angle a_k^l \tag{16}$$

where t_l is the center of the l^{th} analysis frame. The AM-FM decomposition algorithm using eaQHM is provided below:

1. **Initialization:**
   Provide initial frequency estimate $f_k^0(t_1)$
   FOR frame $l = 1, 2, \cdots, L$
   (a) Compute a_k^l, b_k^l using LS
   (b) Update $\hat{f}_k^0(t_l)$ using (15)
   (c) Compute $\hat{A}_k^0(t_l)$ and $\hat{\phi}_k^0(t_l)$ using (16)
   (d) $f_k^0(t_{l+1}) = \hat{f}_k^0(t_l)$
   END
   Interpolation of the parameters $\{\hat{A}_k^0(t), \hat{f}_k^0(t), \hat{\phi}_k^0(t)\}$
2. **Adaptation of amplitudes and phases:**
   FOR adaptation $i = 1, 2, \cdots$
   FOR frame $l = 1, 2, \cdots, L$
   (a) Compute a_k^l, b_k^l using $\hat{\phi}_k^{i-1}(t)$ and (11)
   (b) Update $\hat{f}_k^i(t_l)$ using (5)
   (c) Compute $\hat{A}_k^i(t_l)$ and $\hat{\phi}_k^i(t_l)$ using (16)
   END
   Interpolation of the parameters $\{\hat{A}_k^i(t), \hat{f}_k^i(t), \hat{\phi}_k^i(t)\}$
   END

The convergence criterion for both algorithms was selected to be the following:

$$\frac{SRER^{i-1} - SRER^{i}}{SRER^{i-1}} < \epsilon$$

where $SRER^i$ is the Signal-to-Reconstruction-Error Ratio of the resynthesized signal in the i^{th} adaptation, defined as

$$SRER = 20log_{10}\frac{\sigma_{x(t)}}{\sigma_{x(t)-\hat{x}(t)}} \quad (17)$$

where σ_x denotes the standard deviation of $x(t)$, $x(t)$ is the actual signal and $\hat{x}(t)$ is the reconstructed signal, and where ϵ is a threshold set to 0.02 in our experiments. As a last step of the algorithm, the signal can finally be approximated as the sum of its AM-FM components:

$$\hat{x}(t) = \sum_{k=1}^{K} \hat{A}_k(t) e^{j\hat{\phi}_k(t)}$$

4. VALIDATION ON SYNTHETIC SIGNALS

For the purpose of demonstrating the performance of eaQHM, we consider a two-component signal with modulated amplitudes and frequencies, defined as:

$$x(t) = a_1(t)e^{j(2\pi f_1 t + \phi_1(t))} + a_2(t)e^{j(2\pi f_2 t + \phi_2(t))} \quad (18)$$

where the above parameters are given in Table 1, and the sampling

Sinusoid	1st	2nd
f_i	700	1000
$\phi_i(t)$	$\frac{\pi}{10} + \cos(2\pi 80t)$	$\frac{\pi}{3} + \cos(2\pi 50t)$
$a_i(t)$	$2 + 0.8\cos(2\pi 100)t$	$2 + 0.6\cos(2\pi 100t)$

Table 1. The parameters of the synthetic signal.

frequency is $F_s = 8$ kHz, while the window length is 10 msec. It should be noted that the amplitudes of the signal components are high-frequency modulated and thus, the local amplitude linearity is violated. The time-varying amplitudes $a_i(t)$ and the time-varying frequencies $F_i = f_i + \frac{1}{2\pi}\frac{d}{dt}\phi_i(t)$, for $i = 1, 2$, are to be estimated. In Figure 1, the parameters as they are estimated by aQHM are depicted, whereas in Figure 2, the same information is depicted for the eaQHM algorithm. As it can be seen in Figures 1 and 2, eaQHM performs better than aQHM in the estimation of the time varying frequencies and, especially, of the time varying amplitudes.

To test the robustness of the estimations provided by eaQHM, additive white Gaussian noise of 20 and 10 dB SNR was added to the synthetic signal $x(t)$ described above. For comparison purposes, results for the aQHM are also provided. The performance of the algorithms is measured through the Mean Absolute Error (MAE) for amplitudes and frequencies. The MAE of a parameter θ is defined as:

$$MAE\{\hat{\theta}\} = \frac{1}{M}\sum_{i=1}^{M} |\hat{\theta}^{(i)} - \theta| \quad (19)$$

where $\theta^{(i)}$ is the estimated parameter at the i^{th} simulation, and M is the number of Monte Carlo simulations. The results shown in this section are based on $M = 10000$ Monte Carlo simulations and the length of a Hamming analysis window for both models was 10 msec. The analysis step size was set to 1 sample. Table 2 presents the MAE and SRER scores for the aforementioned levels of noise.

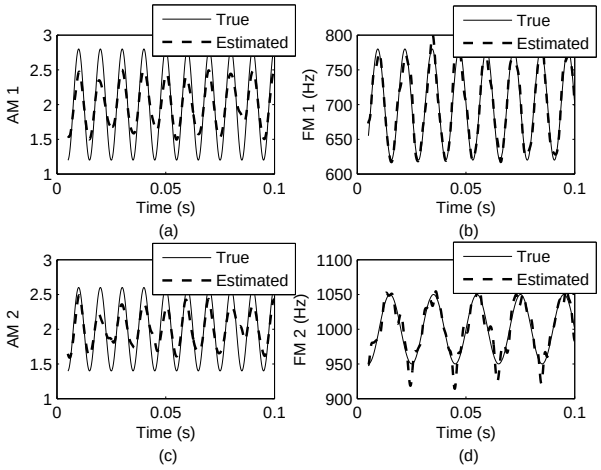

Fig. 1. Parameter estimation for aQHM. Upper panel: Amplitude (left) and Frequency (right) estimation for first component. Lower panel: Amplitude (left) and Frequency (right) estimation for second component.

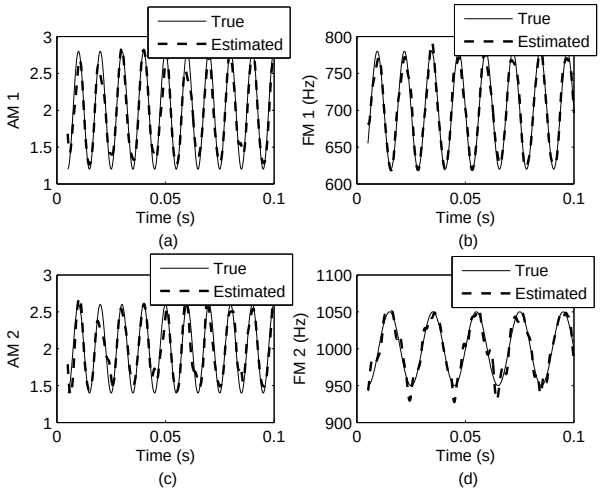

Fig. 2. Parameter estimation for eaQHM. Upper panel: Amplitude (left) and Frequency (right) estimation for first component. Lower panel: Amplitude (left) and Frequency (right) estimation for second component.

MAE scores and SRER						
SNR	Model	$a_1(t)$	$a_2(t)$	$F_1(t)$	$F_2(t)$	SRER(dB)
∞	aQHM	0.2380	0.1842	7.6105	9.1731	22.6
	eaQHM	0.0889	0.0949	5.9217	7.0505	42.0
20 dB	aQHM	0.2235	0.1735	7.2704	7.8563	18.2
	eaQHM	0.1036	0.1079	6.1682	7.1241	20.3
10 dB	aQHM	0.2317	0.1860	8.6071	9.0302	10.7
	eaQHM	0.1490	0.1476	8.0513	8.1022	10.9

Table 2. MAE scores and SRER for aQHM and eaQHM for 10^4 Monte Carlo simulations.

5. VALIDATION ON VOICED SPEECH

The next step is to test the proposed model on real speech, and in particular, on voiced speech signals. The suggested iterative AM-FM decomposition algorithm based on aQHM/eaQHM can be applied on voiced speech signals in a straightforward way. Actually, the aQHM/eaQHM algorithm can be applied on a 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 $\hat{f}_k^0(t_1) = kf_0(t_1)$. Applying QHM analysis on the first voiced frame, an updated set of $\hat{f}_k$ can 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. It is worth noting that the accuracy of the fundamental frequency estimator is not crucial for aQHM, since frequency mismatches are easily corrected. For our purpose, we consider a voiced speech signal from the CMU-ARCTIC database with sampling frequency $F_s = 16$ kHz and duration of about 0.35 sec. For both algorithms, the number of harmonics was set to $K = 40$ and an estimate of the fundamental frequency of the beginning of the segment was given to the algorithm. At most 10 adaptations were allowed to the models. The analysis window size was 2.5 pitch periods and the analysis step size was 1 sample. In the following, the signals are considered up to a fixed maximum voiced frequency (5500 Hz). The original signal, along with the aQHM/eaQHM reconstructed signals and corresponding reconstruction errors, are shown in Figure 3.

To objectively compare the performance of both algorithms, the

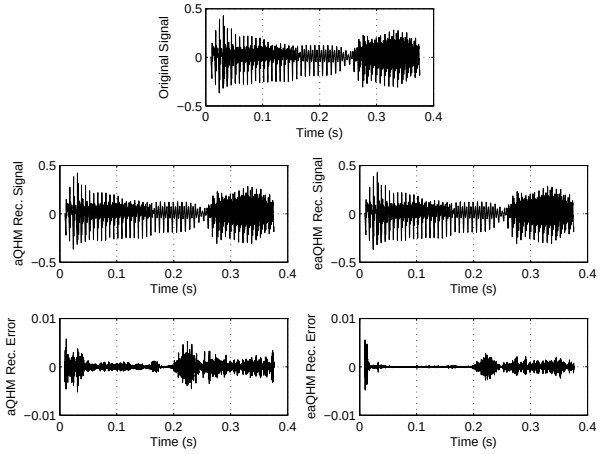

Fig. 3. Upper Panel: Original signal. Middle panel: aQHM (left) and eaQHM (right) reconstructed signal. Lower panel: aQHM (left) and eaQHM (right) reconstruction error.

SRER defined in (17) was used. The SRER was 41.2607 dB for aQHM and 45.2166 dB for eaQHM. Two adaptations for aQHM and three adaptations for eaQHM were necessary for the models to converge.

To confirm these results, a large-scale objective test was performed. Using three different step sizes, namely 1ms, 2ms, and 4ms, we analyzed and reconstructed about 50 minutes of voiced speech from 3 speakers in the ARCTIC database. The sampling frequency of the speech signals was downsampled to 16kHz. A Hamming window of fixed length was used; 3 times the average pitch period of the speaker. The same window was used for both aQHM and eaQHM. The number of components was set to $K = 30$. The average and standard deviation of the SRER (in dB) is provided in Table 3, along with various time-steps. Table 3 also presents the average number of adaptations (NoA) needed for the algorithms to converge. It is evident that, on average, eaQHM scores higher in terms of SRER, requiring, however, slightly more iterations than aQHM.

ARCTIC database evaluation				
Step	Method	Mean (dB)	Std (dB)	NoA
1 msec	aQHM	34.5	4.6	2.9
	eaQHM	35.8	5.7	3.8
2 msec	aQHM	31.0	4.0	3.5
	eaQHM	33.2	5.0	3.9
4 msec	aQHM	30.8	3.4	3.6
	eaQHM	32.8	4.6	6.1

Table 3. Mean and Standard Deviation of SRER (in dB) for approximately 50 minutes of voiced speech from the ARCTIC database.

6. CONCLUSIONS

In this paper, an extension to the recently developed aQHM algorithm was presented, called eaQHM. In eaQHM, the amplitude, along with the frequency of the signal, was included in the adaptation process in a straightforward way. Experiments on synthetic signals showed that eaQHM performs better than aQHM in terms of MAE and SRER. Its robustness in the presence of white Gaussian noise was demonstrated. Experiments on voiced speech using the ARCTIC database showed that eaQHM outperforms aQHM in terms of signal reconstruction.

7. REFERENCES

[1] Y. Pantazis, O. Rosec, and Y. Stylianou. Adaptive AMFM signal decomposition with application to speech analysis. *IEEE Trans. on Audio, Speech, and Language Processing*, 19:290–300, 2011.

[2] R. J. McAulay and T. F. Quatieri. Speech Analysis/Synthesis based on a Sinusoidal Representation. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 34:744–754, 1986.

[3] Y. Stylianou. *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*. PhD thesis, Ecole Nationale Supèrieure des Télécommunications, 1996.

[4] T. F. Quatieri. *Discrete-Time Speech Signal Processing*. Prentice Hall, Engewood Cliffs, NJ, 2002.

[5] Y. Pantazis, O. Rosec, and Y. Stylianou. AM-FM estimation for speech based on a time-varying sinusoidal model. In *Interspeech*, Brighton, Sep 2009.

[6] Y. Pantazis, O. Rosec, and Y. Stylianou. On the Properties of a Time-Varying Quasi-Harmonic Model of Speech. In *Interspeech*, Brisbane, Sep 2008.