ON THE ROBUSTNESS OF THE QUASI-HARMONIC MODEL OF SPEECH

Yannis Pantazis[1] , *Olivier Rosec*[2] *and Yannis Stylianou*[1]

[1]Institute of Computer Science, FORTH, and Multimedia Informatics Lab, CSD, UoC, Greece
[2]Orange Labs TECH/SSTP/VMI, Lannion, France
email: pantazis@csd.uoc.gr, olivier.rosec@orange-ftgroup.com and yannis@csd.uoc.gr

ABSTRACT

In this paper we discuss the robustness of the Quasi-Harmonic model, QHM, previously suggested for speech analysis [1] and AM-FM decomposition of speech [2]. Assuming a frame by frame analysis, QHM suggests an iterative estimator for the actual frequencies of the speech components at the center of analysis window. In this paper, we show that this is a biased estimator and then, we compute analytically and numerically the bias of the estimator showing its dependence on the type and length of the analysis window. Moreover, we analyze the robustness of the QHM estimator in white Gaussian noise, showing that the suggested iterative estimator asymptotically attains the corresponding Cramer-Rao lower bound even in adverse noisy conditions. Examples of synthetic signals are provided to support our analysis.

Index Terms— Quasi-Harmonic Model, Frequency estimation, Robustness, Cramer-Rao bound, Speech analysis

1. INTRODUCTION

Speech modeling is always a timely subject in speech processing. It still has applications in speech coding for wireless and VoIP communications, while with the advances in statistical parametric speech synthesis (e.g., HMM-based speech synthesis), speech modeling shows to be a critical component of these systems for high-quality speech synthesis. Speech modification and voice conversion are other areas where speech modeling is very important. Parameters from suggested speech models can be used in speech and speaker recognition. Last but not least, speech models that offer high-resolution time-frequency analysis of speech have applications in speech analysis and voice function assessment.

Among the most prominent speech models is the sinusoidal model [3] which has found applications in speech coding and speech modifications [4]. Other sinusoidal-based speech representations include the Harmonic plus Noise Model, HNM [5], which has found applications in speech modifications and speech synthesis [6]. This two component representation of speech provides a way to treat differently the harmonic and the noise part of speech, which leads to high quality prosodic modifications. A drawback of the sinusoidal models is their sensitivity to the estimation of frequencies. Good estimation of frequencies of the speech components results in good to high-quality speech modeling. To the contrary, if the frequencies are not well estimated, the quality of modeling is very low, which may produces artifacts in reconstructed and/or modified speech signal.

In our previous works, we suggested a time-varying sinusoidal representation which is not very sensitive to frequency mistakes. In [1], we revisited the model initially introduced by J. Laroche [7] showing that this model can accurately follow the time-varying characteristics of voiced speech, suggesting that voiced speech can be efficiently modeled as sum of quasi-harmonic components (e.g., Quasi-Harmonic Model, QHM). In [2], we showed how QHM can be used for the accurate estimation of amplitude and frequency (AM-FM) modulations in speech. Furthermore, in [2], an adaptive QHM (aQHM) was suggested which provides high-quality speech reconstruction. Actually, QHM *contains* a frequency estimator or corrector, which has the ability to correct frequency mistakes, when trying to minimize the mean squared error between the model and the speech signal. The performance of this QHM frequency estimator is very crucial for the effectiveness of the adaptive QHM and subsequently, for the high-quality speech reconstruction suggested by aQHM.

In this paper, we discuss the robustness of QHM. More specifically, it is shown that the QHM frequency estimator is a biased estimator and then, we study analytically the bias of the estimator. In some cases, an analytic computation of the bias is not possible, and then the bias is computed numerically. It is shown that the bias is a function of the type and length of the analysis window. Frequency mismatch intervals are obtained indicating the bandwidth of frequency mismatch that QHM can efficiently handle. Moreover, in this paper we study the robustness of QHM estimator in white Gaussian noise. It is shown that the QHM estimator asymptotically attains the corresponding Cramer-Rao lower bound (CRLB) even in adverse noisy conditions (below $0dB$). QHM may update its estimations iteratively. We show that this iterative scheme reduces the bias of the QHM estimator and increase the robustness of QHM against noise. We provide examples with synthetic signals in order to visualize and support the results from our analysis.

The rest of the paper is organized as follows. In Section 2 we will quickly review the Quasi-Harmonic Model, QHM, and provide details about the QHM frequency estimator. In Section 3, the bias of the estimator is computed and the role of the analysis window is discussed. Section 4 addresses the robustness of the QHM estimator in white Gaussian noise. In both sections, synthetic examples are provided to visualize the properties of QHM. Finally, Section 5 concludes the paper.

2. OVERVIEW OF QHM

In the sinusoidal context, speech is assumed to be:

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

where there are K components with complex amplitude a_k at frequencies f_k. The analysis window is denoted by $w(t)$. Let us as-

This is Reproducible Research. Matlab code for generating the figures of this paper can be downloaded from
http://www.csd.uoc.gr/~pantazis/ICASSP2010/RobustnessOfQHM_FiguresCode.zip

sume that f_k denote the correct frequencies of the components of the signal. In sinusoidal modeling, frequencies are estimated (e.g., by peak-picking, by considering harmonics of a fundamental frequency, etc.), which 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 which will create artifacts in the reconstruction or modification stage of speech using the sinusoidal models. To cope with this problem, in [1] and [2] we suggested the use of the Quasi-Harmonic Model, QHM, for the representation of speech:

$$x(t) = \left(\sum_{k=1}^{K} (a_k + t b_k) e^{j 2\pi \hat{f}_k t} \right) w(t), \quad t = -N, ..., N \tag{3}$$

where b_k denotes the complex slope of the kth component. In 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 [1] it was suggested to project b_k to a_k:

$$b_k = \rho_{1,k} a_k + \rho_{2,k} j a_k \tag{5}$$

where ja_k denotes the perpendicular (vector) to a_k. Then, the kth component is written as:

$$\begin{aligned} X_k(f) = \quad a_k \Big[& W(f - \hat{f}_k) - \frac{\rho_{2,k}}{2\pi} W'(f - \hat{f}_k) \\ & + j \frac{\rho_{1,k}}{2\pi} W'(f - \hat{f}_k) \Big] \end{aligned} \tag{6}$$

Let us consider the Taylor series expansion of $W(f - \hat{f}_k - \frac{\rho_{2,k}}{2\pi})$:

$$\begin{aligned} W(f - \hat{f}_k - \frac{\rho_{2,k}}{2\pi}) = \quad & W(f - \hat{f}_k) - \frac{\rho_{2,k}}{2\pi} W'(f - \hat{f}_k) + \\ & O(\rho_{2,k}^2 W''(f - \hat{f}_k)) \end{aligned} \tag{7}$$

Notice that if the value of term $W''(f - f_k)$ at f_k is small, then for small values of $\rho_{2,k}$ we can approximate (7) as

$$W(f - \hat{f}_k - \frac{\rho_{2,k}}{2\pi}) \approx W(f - \hat{f}_k) - \frac{\rho_{2,k}}{2\pi} W'(f - \hat{f}_k) \tag{8}$$

Consequently, from (6) it follows that

$$X_k(f) \approx a_k \left[W(f - \hat{f}_k - \frac{\rho_{2,k}}{2\pi}) + j \frac{\rho_{1,k}}{2\pi} W'(f - \hat{f}_k) \right] \tag{9}$$

which is written in the time domain as

$$x_k(t) \approx a_k \left[e^{j(2\pi \hat{f}_k + \rho_{2,k})t} + \rho_{1,k} t e^{j 2\pi \hat{f}_k t} \right] w(t) \tag{10}$$

From (10) and (2), we see that $\rho_{2,k}/2\pi$ can be an estimator of the frequency error η_k:

$$\hat{\eta}_k = \rho_{2,k}/2\pi \tag{11}$$

while $\rho_{1,k}$ accounts for the normalized amplitude slope of the kth component. In other words, QHM suggests a frequency correction to the input frequencies $\hat{f}_k$ (or a frequency estimator). This suggestion is however conditional on the magnitude of $\rho_{2,k}$ and the value of term $W''(f)$ at f_k as it was mentioned above (going from (7) to (8)).

In the rest of the paper, we will discuss the validity of the QHM frequency estimator and its robustness against noise.

3. VALIDITY OF THE QHM FREQUENCY ESTIMATOR

The QHM frequency estimator suggested in the previous section depends on the window and on the amount of frequency mismatch, $\rho_{2,k}$. Let us first address the issue of the analysis window.

3.1. Influence of analysis window

Like in any frequency estimation problem, the analysis window length should be large enough to achieve high frequency resolution and robust estimation of the unknown parameters. On the other hand, since the model suggested in (1) is a stationary model, the analysis of natural signals like speech, will require the window length to be small enough in order to accommodate the non-stationary characteristics of the analyzed natural signal. Considering only windows that satisfy this trade-off, we should select from them those that offer small value for $W''(f - f_k)$ at f_k. 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 $W''(f - f_k)$ at f_k. Thus, considering analysis windows that fulfill the tradeoff mentioned above, we will prefer the one that has the largest bandwidth (e.g., we will prefer the hamming over the rectangular window).

3.2. Bias computation

Let us next compute the bias of the QHM frequency estimator. For this, let us assume a mono-component signal:

$$x(t) = \alpha e^{j(2\pi \hat{f} t + \eta t)} \tag{12}$$

and the corresponding QHM model:

$$s(t) = (a + tb) e^{j(2\pi \hat{f} t)} \quad -T \leq t \leq T \tag{13}$$

Please note that η in (12) denotes an angular frequency and not a linear frequency as in (2). Assuming a rectangular window, $w(t)$, of length $2T$, the least squares solution for a and b is given by:

$$\begin{aligned} a &= \alpha \frac{sin(\eta T)}{\eta T} \\ b &= \alpha \, 3j \left(\frac{sin(\eta T)}{\eta^2 T^3} - \frac{cos(\eta T)}{\eta T^2} \right) \end{aligned} \tag{14}$$

Then, the coefficient ρ_2 in (5) can be shown to be:

$$\rho_2 = 3 \left(\frac{1}{\eta T^2} - \frac{cot(\eta T)}{T} \right) \tag{15}$$

or, in other words, QHM suggests that:

$$\hat{\eta} = 3 \left(\frac{1}{\eta T^2} - \frac{cot(\eta T)}{T} \right) \tag{16}$$

It is therefore worth studying the bias of this estimator:

$$Bias(\eta) = \eta - \hat{\eta} \tag{17}$$

In Fig. 1(a) the bias is plotted for a rectangular window of length $16ms$ ($T = 8ms$) with solid line. For the readability of the paper we show frequency mismatch (η) in Hz (e.g., $\eta/2\pi$). Although the computation of bias for the rectangular window is simple, it is a bit more complicated for other windows like the Hamming window. Alternatively, the bias can be computed numerically. To confirm that

the numerical method provides about the same bias as the analytic formula, the bias for the rectangular window was also computed numerically and is shown in Fig. 1(a) by a dashed line. Notice the similarity between the two approaches.
The bandwidth where the bias is considered small (e.g., $|Bias(\eta)| < |\eta|$) is shown by a bold line. From this figure, it turns out that in the case of analysis using a rectangular window, QHM can correct frequency mismatches if they are below $45Hz$.
The bias for a Hamming window of the same size ($16ms$) is computed numerically and it is shown in Fig. 1(b). Again, the bandwidth where the bias is considered small is shown as solid line. It is worth noting that in the case of the Hamming window, QHM can correct frequency mismatches that are below $135Hz$. This gain factor of 3, can be explained by the ratio of the bandwidth of the main lobe, B, of the squared Hamming window ($375Hz$) over the corresponding bandwidth of the (squared) rectangular window ($125Hz$)[1]. Testing with a variety of window types it was found that the bias is small when the frequency mismatch is smaller than one third of the bandwidth of the squared analysis window.

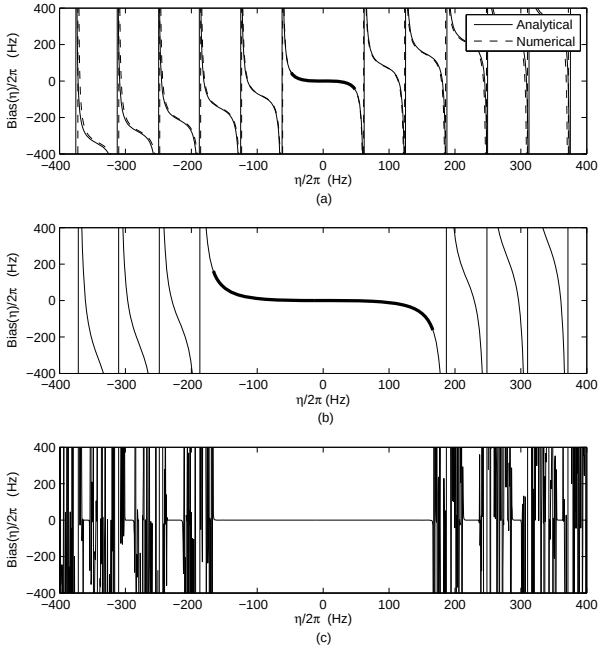

Fig. 1. Upper panel: The bias for a rectangular window computed analytically (solid line) and numerically (dashed line). Middle panel: The bias for a Hamming window computed numerically. Lower panel: Bias using the Hamming window (as in (b)) with two iterations.

3.3. Iterations on QHM

Once an initial estimation of frequency mismatch (η) is obtained through ρ_2, the analysis frequency $\hat{f}$ can be updated and then the input signal can be analyzed again by QHM using now the updated frequency value, i.e., $\hat{f} = \hat{f} + \frac{1}{2\pi}\rho_2$. Thus, new estimations of η can be obtained iteratively. In Fig. 1(c) the bias is depicted for the same Hamming window as in Fig. 1(b), after two iterations. We observe that the bias is considerably reduced (mainly to zero) if the initial frequency mismatch is smaller than $B/3$, where we recall that B denotes the bandwidth of the squared analysis window.

[1]The squared of the analysis window appears in the least-squares solution for the estimation of complex amplitudes and slopes in QHM.

4. ROBUSTNESS AGAINST NOISE

In this section, the performance of QHM in the case of a signal, $x(t)$, contaminated by additive noise is assessed. As an example, we assume a signal with four components contaminated by white Gaussian noise $v(t)$, as follows:

$$y(t) = \sum_{k=1}^{4} a_k e^{j2\pi f_k} + v(t) \tag{18}$$

Table 1, provides information about the frequency and the amplitude of each component. Two closed-space sinusoids and two well-separated sinusoids are considered. For the analysis of this signal, a Hamming window of $17ms$ ($T = 8.5ms$) length is used, and a sampling frequency of $8000Hz$ is considered. The last row of this table contains the interval of allowed frequency mismatch per component. The frequency mismatch should be relative to the frequency distance between the components. For the low-frequency and closely-spaced sinusoids the frequency mismatch is therefore smaller ($\pm 10Hz$) than for the high-frequency and well-separated sinusoids ($\pm 100Hz$).

Sinusoid	1st	2nd	3rd	4th
Frequency (Hz)	100	200	1000	2000
Amplitude	$e^{j\pi/10}$	$e^{j\pi/4}$	$e^{j\pi/3}$	$e^{j\pi/5}$
Freq. Mismatch (Hz)	± 10	± 10	± 100	± 100

Table 1. The parameters of the synthetic signal and frequency mismatch intervals.

Monte Carlo simulations are used for the assessment of the robustness of the QHM frequency estimator. For each simulation, the frequency mismatch for each component is sampled from a uniform distribution on the corresponding interval defined in the last row of Table 1.

Assuming that the power spectral density of the noise is $V(f)$, then, the *local* Signal to Noise Ratio, SNR, for the k^{th} sinusoid is given by [8]

$$SNR_k \approx 10log_{10}\frac{(2N+1)|a_k|^2}{V(f_k)} \tag{19}$$

where $2N+1$ is the length of the analysis window in samples. Please note that in order to obtain the SNR value widely used, one must subtract from the *local* SNR the quantity $10log_{10}(2N+1)$. Since $N = 68$ samples ($8.5ms$), then a *local* SNR of $20dB$ corresponds to $-1.36dB$ SNR, while a *local* SNR of $30dB$, corresponds to a bit less than $8.63dB$ ordinary SNR.

The Cramer-Rao lower bound (CRLB) for the frequency of the kth component is given by [8]

$$CRLB\{f_k\} = \frac{3V(f_k)}{|a_k|^2 N(N+1)(2N+1)} \tag{20}$$

while, the CRLB for the amplitude of the kth component is given by

$$CRLB\{a_k\} = \frac{V(f_k)}{2N+1} \tag{21}$$

The performance of QHM is measured through the mean squared error for frequencies:

$$MSE\{\hat{f}_k\} = \frac{1}{\mathrm{M}} \sum_{i=1}^{\mathrm{M}} |\hat{f}_k(i) - f_k|^2 \quad (22)$$

and for amplitudes:

$$MSE\{\hat{a}_k\} = \frac{1}{\mathrm{M}} \sum_{i=1}^{\mathrm{M}} |\hat{a}_k(i) - a_k|^2 \quad (23)$$

where M is the number of Monte Carlo simulations. The results shown in this section are based on $M = 10000$ Monte Carlo simulations. For comparison purposes, we include the frequency and amplitude estimation using the peak-picking after parabolic interpolation between peaks approach, which is used in the classic sinusoidal model [3]. This method will be referred to as FFT.
Fig. 2 and Fig. 3 show the MSE for amplitude and frequency for each component, respectively. For comparison purposes, FFT-based frequency and amplitude estimation is shown. The corresponding Cramer-Rao lower bounds are depicted by solid lines. We would like to mention that in this experiment there are two sources of "noise" for QHM: (i) the additive white Gaussian noise, and (ii) the frequency mismatch. For the FFT-based approach, however, only the first type of noise is applied. When no iteration is used, the frequencies are not updated, thus, the estimation errors for the frequencies correspond to the initial frequency mismatches. In that case, the FFT approach outperforms QHM for the two high and well separated frequency components (two lower panels in Fig. 2 and Fig. 3) where we allowed a maximum mismatch of $\pm 100Hz$. However, when this mismatch is lower (e.g., $\pm 10Hz$) as is the case for the other two of lower frequency components, then QHM outperforms the FFT-based approach. By contrast, when iterations are used, the iterative QHM outperforms the FFT-based approach in any case; only 3 iterations are needed for QHM to asymptotically attain the CRLB.

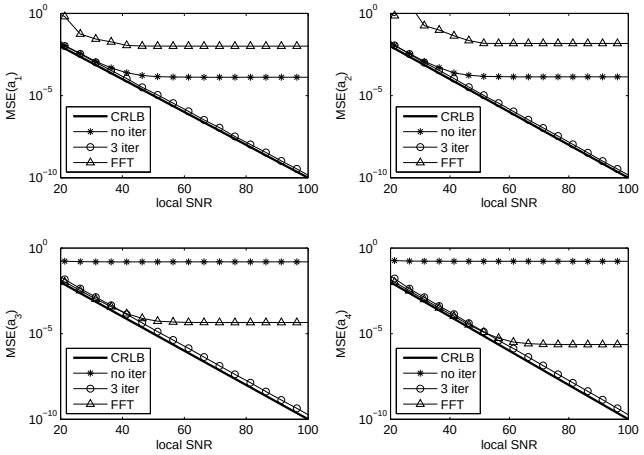

Fig. 2. MSE of amplitudes as a function of *local* SNR.

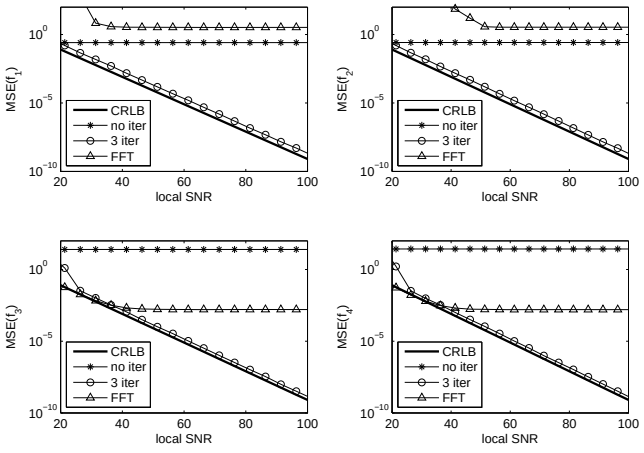

Fig. 3. MSE of frequencies as a function of *local* SNR.

5. CONCLUSIONS

In this paper we discussed the robustness of the Quasi-Harmonic Model, QHM, which was previously suggested for speech analysis and AM-FM decomposition of speech. The robustness was checked against frequency mismatches and additive white Gaussian Noise. It was shown that the iterative QHM can handle large frequency mismatches while it is robust against noise. Only few iterations were required for the QHM-based estimators (frequency and amplitude) to attain the Cramer-Rao Lower Bound.

6. REFERENCES

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

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

[3] 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.

[4] T.F. Quatieri and R.J. McAulay. Shape-Invariant Time-Scale and Pitch Modifications of Speech. 40:497–510, 1992.

[5] J. Laroche Y. Stylianou and E. Moulines. HNM: A Simple, Effecient Harmonic plus Noise Model for Speech. In *Workshop on Appl. of Signal Proc. to Audio and Acoustics (WASPAA)*, pages 169–172, New Paltz, NY, USA, Oct 1993.

[6] Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Trans. on Speech and Audio Proc.*, 9:21–29, 2001.

[7] J. Laroche. A new analysis/synthesis system of musical signals using Prony's method. Application to heavily damped percussive sounds. In *Proc. IEEE ICASSP*, pages 2053–2056, Glasgow, UK, May 1989.

[8] S. M. Kay. *Modern Spectral Estimation: Theory and Applications*. Prentise-Hall, Englewood Cliffs, NJ, 1988.