ON THE PROPERTIES OF A TIME-VARYING 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 present the properties of a parametric speech model based on a deterministic plus noise representation of speech initially suggested by Laroche et al. [1]. Aiming at a high resolution analysis of speech signals for voice quality control (transformation) and assessment, we focus on the deterministic representation and we reveal the properties of the model showing that such a representation is equivalent to a time-varying quasi-harmonic representation of speech. Results show that the model is appropriate in estimating accurately linear amplitude modulations and modeling the inharmonicity of speech.

Index Terms— Speech analysis, speech modeling, quasi-harmonic models, inharmonicity

1. INTRODUCTION

Sinusoidal and harmonic or deterministic plus noise models have successfully been applied in speech and music signals for coding, modifications and speech synthesis [2], [3]. Especially for speech signals, it has been shown that for coding/compression purposes the harmonic representation of voiced areas (and unvoiced areas when a low fundamental frequency is considered) produces good quality of speech [2] at low bit rates. The harmonic structure of speech is also supported by the simple linear source filter model for the speech production mechanism.

However, it is well known that speech is not really a periodic signal and indeed it is usually referred to as a *quasi-harmonic* signal. Although for many speech applications this may be considered as a detail, in other speech applications like high quality speech modification, speech synthesis and voice quality assessment (i.e., in pathologic voices) it is a property of the signal worth considering. Looking at the magnitude spectrum of short-term Fourier Transform it can easily be seen that the local maxima (peaks) are not exactly at the multiples of a fundamental frequency. Actually this was the main motivation on developing the sinusoidal model by Quatieri et al. [2]. In [2], the amplitude and frequency of the sinusoids were considered constant over the analysis window although it is also well known that these are time-varying parameters. The Deterministic plus Noise model suggested by Laroche et al. [1] can take into account this time-varying character of speech. It has also been shown in [4] and in [5] that this model is not a harmonic model since the phase is not a linear function of time. Indeed, although this was not mentioned explicitly in these works (i.e., [5]), sinusoids near to the corresponding harmonic frequencies (inharmonic, or 'detuning' of individual harmonics) were implicitly used into the model. In this paper, we will refer to this inharmonicity as *quasi-harmonicity*. A different time-varying approach has been suggested by Marques et al. [6] where a generalized polynomial phase function was used to increase modeling accuracy of the speech model. Time-varying quasi-harmonic models have also been suggested for long-term speech modeling [7] and enhancement [8] and in music signal processing for pitch estimation [9]. All the above time-varying quasi-harmonic approaches try to model the speech time-series by representing the time-varying amplitude and phase function as a linear combination of fixed basis functions where many parameters should be estimated in order to efficiently represent the sequence of speech samples. For example, over 1000 parameters need to be estimated for a 200 ms speech segment in [7]. When all the parameters of these models are considered unknown (like the number of components, the pitch) then a probabilistic approach should be used (e.g., based on MCMC) instead of the usual least-squares approach [9].

Although these models may be successful in representing the sequence of speech samples (however, by increasing the complexity or the order of the suggested model) they are not suitable in describing the main speech characteristics and properties of speech as these are reflected by the speech production mechanism and by the (nonlinear) interaction of the glottal airflow signal with the vocal tract filter. In this paper, we would like to revisit the lower-order speech model suggested by Laroche et al. [1] and reveal and discuss the main properties of the model. This will allow us to track important characteristics of speech with high accuracy and be able to use this model (or in the future, its extension) in applications like Voice Transformation and Voice Quality Assessment (like analysis of pathologic voices). Instead of using the term "deterministic" as in Laroche et al. [1], we will refer to this component as "quasi-harmonic" which better reflects the properties of the model as we will show shortly. Also, since we will limit our presentation to this component to represent the lower frequencies of speech, we will refer to this model to as Quasi-Harmonic Model (QHM) in the following sections.

The paper is organized as follows. Section 2 presents an short overview of QHM. Sections 3 and 4 present the time-domain and frequency-domain properties of the model, respectively. In Section 5, results from synthetic signals and real speech are presented. Finally, Section 6 concludes the paper and provides directions for future work.

2. A SHORT OVERVIEW OF QHM

Within an analysis window the deterministic (or otherwise, quasi-harmonic) component of a speech signal is modeled as (Chapter 4 in [4]):

$$s(t) = \left(\sum_{k=-L}^{L} (a_k + t b_k) e^{2\pi j k f_0 t} \right) w(t), \qquad (1)$$

where f_0 is the fundamental frequency of the harmonic signal, L specifies the order of the model i.e. the number of harmonics, a_ks are the complex amplitudes and b_ks are the complex slopes. $w(t)$ denotes the analysis window. Window is typically a rectangular or a Hamming window and it is zero outside a symmetric interval $[-t_0, t_0]$. This model is an extension to the classic harmonic model where the $t\,b_k$ term is omitted [3]. Hence, the signal in eq. (1) is projected to the complex exponential functions as in the simple harmonic case and in addition to functions of type $t\,e^{2\pi jkf_0t}$.

Assuming that we know signal $s(t)$ at time instants $t_1, t_2, ..., t_N \in [-t_0, t_0]$, then the estimation of model parameters $\{f_0, L, a_{-L}, ..., a_L, b_{-L}, ..., b_L\}$ is performed into two steps. At first, the fundamental frequency, f_0 and the number of harmonic components, L, are estimated using spectral and autocorrelation information as described in [4]. Then, the computation of a_k and b_k, $k \in \{-L, ..., L\}$ is performed by minimizing a mean squared error which naturally leads to Least Squares.

3. TIME-DOMAIN PROPERTIES OF QHM

The time-domain characteristics of the model are discussed in this section. From eq.(1), it is easily seen that the instantaneous amplitude is a time-varying function and it is given for each harmonic by:

$$\begin{aligned} m_k(t) &= |a_k + tb_k| \\ &= \sqrt{(a_k^R + tb_k^R)^2 + (a_k^I + tb_k^I)^2} \quad , \end{aligned} \tag{2}$$

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

Since both amplitudes and slopes (a_k, b_k) are complex variables, instantaneous phase and instantaneous frequency are not constant functions over time. Indeed, instantaneous phase is given for each harmonic by:

$$\begin{aligned} \phi_k(t) &= 2\pi k f_0 t + \angle(a_k + tb_k) \\ &= 2\pi k f_0 t + atan\frac{a_k^I + tb_k^I}{a_k^R + tb_k^R}, \end{aligned} \tag{3}$$

while instantaneous frequency is given by:

$$\begin{aligned} f_k(t) &= \frac{1}{2\pi}\phi_k'(t) \\ &= kf_0 + \frac{1}{2\pi}\frac{a_k^R b_k^I - a_k^I b_k^R}{m_k^2(t)}. \end{aligned} \tag{4}$$

From eq.(4) we can easily see that the instantaneous frequency is a bell-shaped curve similar to Cauchy distribution. A feature of the model worth noting is that the 2nd term of the instantaneous frequency in eq.(4) depends on the instantaneous amplitude. This means that the accuracy of frequency estimation (or, the estimation of phase function) depends on the amplitude information [10].

4. FREQUENCY-DOMAIN PROPERTIES

In this section, we provide an in-depth analysis of the properties of QHM in the frequency domain and show that this model can be used to get an accurate estimation of harmonic frequencies and/or to track amplitude variations.

Using standard relations from Fourier analysis eq. (1) is written in frequency domain as:

$$S(f) = \sum_{k=-L}^{L} \big(a_k W(f - kf_0) + jb_k W'(f - kf_0)\big) \tag{5}$$

where $W(f)$ is the Fourier transform of the analysis window, $w(t)$.

Let $\vec{a}_k$ and $\vec{b}_k$ denote the vectors corresponding respectively to the complex a_k and b_k. In order to get further insight on the properties of QHM, we decompose $\vec{b}_k$ into two components: one collinear to $\vec{a}_k$ and one perpendicular to $\vec{a}_k$. Thus, $\vec{b}_k$ is given by

$$\vec{b}_k = \rho_{1,k}\vec{a}_k + \rho_{2,k}\vec{a}_k^{\perp}, \tag{6}$$

where $\vec{a}_k^{\perp} = (-a_k^I, a_k^R)^T$,

$$\rho_{1,k} = \frac{\langle \vec{a}_k, \vec{b}_k \rangle}{|\vec{a}_k|^2}$$

and

$$\rho_{2,k} = \frac{\langle \vec{a}_k^{\perp}, \vec{b}_k \rangle}{|\vec{a}_k|^2}.$$

Note that $\langle .,. \rangle$ is the inner product between two vectors. Then, the k^{th} component in eq. (5) can be written as:

$$S_k(f) = a_k[W(f - kf_0) - \rho_{2,k}W'(f - kf_0) + j\rho_{1,k}W'(f - kf_0)]. \tag{7}$$

For small values of $\rho_{2,k}$, using a first order approximation of the Taylor series of $W(f)$, we have

$$W(f - kf_0) - \rho_{2,k}W'(f - kf_0) \approx W(f - kf_0 - \rho_{2,k}) \tag{8}$$

and finally eq. (7) can be approximated as follows:

$$S_k(f) \approx a_k[W(f - kf_0 - \rho_{2,k}) + j\rho_{1,k}W'(f - kf_0)]. \tag{9}$$

Figure 1, depicts the effect of this approximation using a Hamming window. For reasonable values of the frequency shift, it can be observed that the approximation is very good.

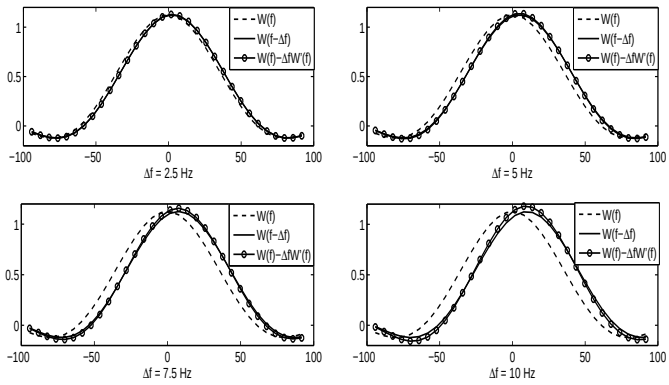

Fig. 1. Approximation of $W(f) - \Delta f W'(f)$ by $W(f - \Delta f)$ when W is the Hamming window.

From the above developments, it appears that the angle between vectors $\vec{a}_k$ and $\vec{b}_k$ plays an important role in the frequency-domain characteristics of QHM. We now analyze two particular settings where $\vec{a}_k$ and $\vec{b}_k$ are respectively collinear and orthogonal, before providing results on a more general case.

4.1. $\vec{b}_k$ collinear to $\vec{a}_k$

If $\vec{b}_k$ is collinear to $\vec{a}_k$, $\rho_{2,k} = 0$ and eq. (7) becomes

$$S_k(f) = a_k[W(f - kf_0) + j\rho_{1,k}W'(f - kf_0)], \qquad (10)$$

or, equivalently, the k^{th} component in time is

$$s_k(t) = 2|a_k|(1 + \rho_{1,k}t)cos(k\omega_0 t + \angle a_k)w(t). \qquad (11)$$

In this case, $\rho_{1,k}$ is the slope of the k^{th} instantaneous amplitude of the model, while the k^{th} instantaneous frequency is constant and equal to kf_0 since $a_k^R b_k^I - a_k^I b_k^R$ of eq. (4) equals to zero.

To illustrate the behavior of the model in this case, we present an example of a single sinusoid at 350 Hz whose amplitude varies linearly in time with a slope of 50 per second. In Figure 2, the original signal and the instantaneous amplitude as estimated by eq. (2) are shown on the upper plot. The lower plot depicts the true frequency and the instantaneous frequency as estimated by eq. (4). From the estimates of a_k and b_k, the angle between vectors $\vec{a}_k$ and $\vec{b}_k$ was found to be 0. Moreover, we obtain $\rho_{1,k} = 50.0$ and $\rho_{2,k} = 10^{-13}$, which shows that the model provides accurate estimates of both the amplitude slope and the instantaneous frequency.

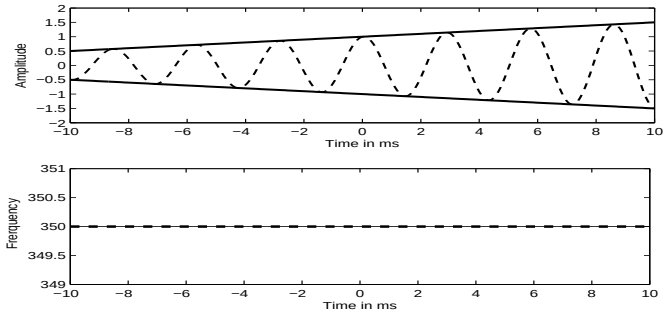

Fig. 2. Amplitude modulation. Upper panel: Original signal (dashed line) and estimated inst. amplitude (solid line). Lower panel: True (dashed line) and estimated inst. frequency (solid line).

4.2. $\vec{b}_k$ orthogonal to $\vec{a}_k$

When $\vec{b}_k$ is orthogonal to $\vec{a}_k$, then $\rho_{1,k} = 0$ and for small values of $\rho_{2,k}$, the approximation in eq. (9) falls down to:

$$S_k(f) \approx a_k W(f - kf_0 - \rho_{2,k}). \qquad (12)$$

Going back in time domain, the k^{th} component can be written as:

$$s_k(t) \approx 2|a_k|cos\left(2\pi(kf_0 + \rho_{2,k})t + \angle a_k\right) w(t) \qquad (13)$$

In this case, the instantaneous amplitude is constant while $kf_0 + \rho_{2,k}$ is its instantaneous frequency. Thus, it is worth noting that QHM enables the estimation of a frequency shift for each harmonic component which is very important when the frequencies of the sinusoids are not exactly at integer multiples of f_0 but slightly vary from this position. To illustrate this property, we consider a single sine wave whose frequency is 350 Hz. In this example, the analysis is carried out with a frequency of 356 Hz. Figure 3 shows that the estimated instantaneous frequency can be recovered with a good accuracy. The estimated angle between $\vec{a}_k$ and $\vec{b}_k$ is 89.99 degrees which mainly corresponds to the case where $\vec{a}_k$ and $\vec{b}_k$ are orthogonal. Finally, $\rho_{1,k} = 10^{-13}$, which means that there is not amplitude modulation while $\rho_{2,k} = -6.005$, which is a good estimate of the frequency mismatch.

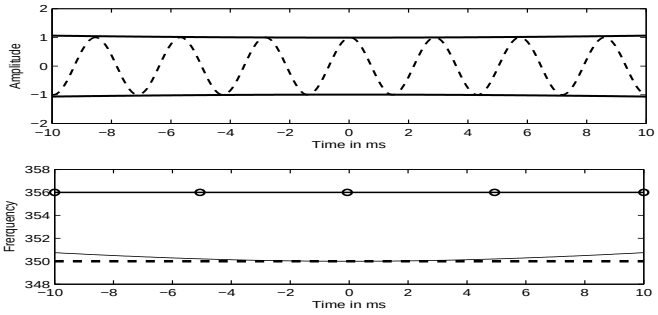

Fig. 3. Frequency mismatch. The analysis is performed at 356 Hz (6 Hz away from the correct frequency; solid line with circles). Upper panel: Original signal (dashed line) and estimated inst. amplitude (solid line). Lower panel: True (dashed line) and estimated inst. frequency (solid line). Note that the inst. frequency is very close to the true value.

4.3. Random angle between $\vec{b}_k$ and $\vec{a}_k$

We now address the case when both linearly time-varying amplitude and frequency mismatch is present. We consider the synthetic signal presented in Figure 4 which is a sinusoid at 350 Hz with an amplitude slope of 50 per period. When the analysis is carried out with a frequency of 356 Hz,the angle between $\vec{a}_k$ and $\vec{b}_k$ is 24.286 degrees and the following estimates were obtained : $\rho_{1,k} = 47.75$ and $\rho_{2,k} = -5.926$. Thus, QHM enables the estimation of the amplitude slope and provides a refinement of the frequency. However, it is worth noting that the refined frequency is less accurate than in the example in Figure 3. Moreover, the curve of the instantaneous frequency (Figure 4) exhibits larger variations than that in Figure 3. In order to further refine this frequency estimate, we suggest to use an iterative procedure which simply consists in updating the analysis frequency with the estimated frequency. With only two iterations of this procedure the frequency can be obtained with a very good accuracy ($\rho_{2,k} = -6.0$), while the estimated slope is $\rho_{1,k} = 49.97$.

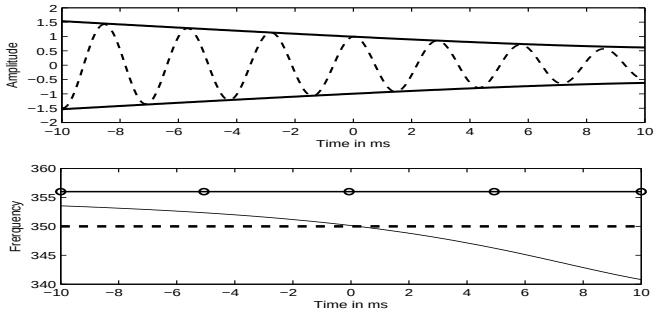

Fig. 4. Amplitude modulation and frequency mismatch. Upper panel: Original signal (dashed line) and estimated inst. amplitude (solid line). Lower panel: True (dashed line) and estimated inst. frequency (solid line). Initial analysis is performed at 356 Hz (solid line with circles)

5. RESULTS

In this section, we illustrate the abilities of QHM on various signals with more than one components (synthetic and speech signals).

5.1. Multi-component synthetic signals

Two quasi-harmonic signals with 10 components and fundamental frequencies of 120 Hz, and 200 Hz were considered as our synthetic signals. The frequencies of the components were set to be not at exactly integer multiplies of the fundamental frequency while each component has a linear time varying inst. amplitude. Hamming window of four pitch periods was used for the estimation of complex amplitudes and phases. Analysis was performed using the true fundamental frequency as well as wrong frequency ($5Hz$ of mismatch).

In each case, the iterative approach described above was applied Thus, at each iteration step the new analysis was performed using as frequencies $kf_0 + \rho_{2,k}$. Figure 5 shows the SNR at each iteration for all the synthetic signals. It is worth to note the difference in convergence in terms of SNR between the two synthetic signals.

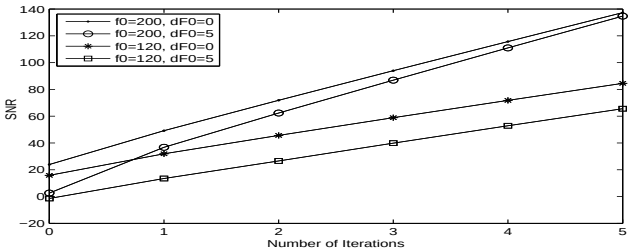

Fig. 5. SNR per iteration for a signal with fundamental frequency 120 Hz (lines with stars and squares) and for a signal with fundamental frequency 200 Hz (lines with dots and circles). $dF0$ denotes the fundamental frequency mismatch in Hertz.

5.2. Real speech example

An example of applying QHM in a speech signal generated by a male speaker with fundamental frequency to vary between 120 Hz and 160 Hz is also provided. The analysis was performed using a Hamming window of four local pitch periods long and step size of one local pitch period. The average segmental SNR between the original speech signal and the residual signal (as this is obtained by simply subtracting the reconstructed speech frame generated by QHM from the corresponding original speech segment) was 25 dB. Figure 6 shows the initial harmonic frequencies used by the model, and the estimated inharmonics as a function of time.

6. CONCLUSIONS AND FUTURE WORK

In this paper we focus on a deterministic model suggested for the analysis of speech signals 15 years ago. By re-writing the main equation which describes the model we were able to reveal the properties of the model showing that it is a time-varying quasi-harmonic model of speech. Moreover, we were able to connect the parameters of the model to time and frequency characteristics of the speech signal. This will allow us to apply the model for high-quality speech transformations and for voice quality assessment. Further extensions of the model seem possible in order to include other time varying characteristics of speech (i.e., frequency).

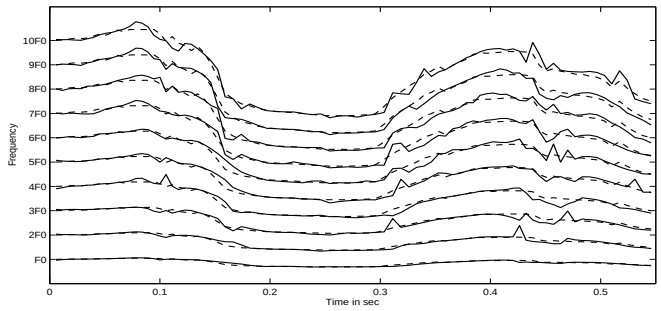

Fig. 6. Harmonic frequencies tracks (dashed lines) and estimated frequency tracks (solid lines).

7. REFERENCES

[1] 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.

[2] T. F. Quatieri and R. J. McAulay. Audio signal processing based on sinusoidal analysis/synthesis. In M. Kahrs and K. Brandenburg, editors, *Applications of Signal Processing to Audio and Acoustics*. Klwer Academic Publishers, 1998.

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

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

[5] Yannis Stylianou. *Modeling Speech based on Harmonic Plus Noise Models*, pages 244–260. Springer Berlin / Heidelberg.

[6] J. S. Marques and L. B. Almeida. A Background for Sinusoid-based Representation of Voiced Speech. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages 1233–1236, Tokyo, Japan, Apl 1992.

[7] G. Fay E. Moulines O. Cappé and F. Bimbot. Polynomial Quasi-Harmonic Models for Speech Analysis and Synthesis. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages II–865–868, Seattle.

[8] S. Dubost and O. Cappé. Analysis and Enhancement of Locally Harmonic Signals using Adaptive Multi-kernel Methods. In *Workshop on Appl. of Signal Proc. to Audio and Acoustics (WASPAA)*, pages 17–20, New York, 1999.

[9] S. Godsill and M. Davy. Bayesian Harmonic Models for Musical Pitch Estimation and Analysis. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, pages 1769–1772, Orlando, USA.

[10] Steven M. Kay. *Funtamentals of statistical signal processing: Estimation theory*. Prentice Hall, 1993.