University of Crete
Department of Computer Science

Decomposition of AM-FM Signals with Applications in Speech Processing

(Philosophy of Doctoral)

Yannis Pantazis

Heraklion

Summer 2010

Department of Computer Science
University of Crete

Decomposition of AM-FM Signals with Applications in Speech Processing

Submitted to the
Department of Computer Science
in partial fulfillment of the requirements for the degree of
Philosophical Doctoral

Yannis Pantazis

phone: +306973638924

web: http://www.csd.uoc.gr/~pantazis/

e-mail: pantazis@csd.uoc.gr and

yannis.pantazis@gmail.com

Summer, 2010

Board
of inquiry:

Supervisor _____

Yannis Stylianou
Associate Professor

Member _____

Olivier Rosec
Researcher

Member _____

Athanasios Mouchtaris
Assistant Professor

Member _____

Panagiotis Tsakalides
Professor

Member _____

Georgios Tziritas
Professor

Member _____

Alexandros Potamianos
Associate Professor

Member _____

Vassilis Digalakis
Professor

Heraklion, Summer, 2010

Abstract

During the last decades, sinusoidal model gained a lot of popularity since it is able to represent non-stationary signals very accurately. The estimation of the instantaneous components (i.e. instantaneous amplitude, instantaneous frequency and instantaneous phase) is an active area of research. In this thesis, we develop and test models and algorithms for the estimation of the instantaneous components of sinusoidal representation. Our goal is to reduce the estimation error due to the non-stationary character of the analyzed signals by taking advantage of time-domain information. Thus, we re-introduce a time-varying model referred to as QHM which is able to adjust its frequency values closer to the true frequency values. We further show that an iterative scheme based on QHM produce statistically efficient sinusoidal parameter estimation. Moreover, we extend QHM to chirp QHM (cQHM) which is able to capture linear evolution of instantaneous frequency quite satisfactorily.

However, neither QHM nor cQHM are not able to represent highly non-stationary signals adequately. Thus, we further extend QHM to adaptive QHM (aQHM) which uses time-domain frequency information. aQHM is able to adjust its non-parametric basis functions to the time-varying characteristics of the signal. This results to reduction of the estimation error of the instantaneous components. Moreover, an adaptive AM-FM decomposition algorithm based on aQHM is proposed. Results on synthetic signals as well in voiced speech showed that aQHM greatly reduce the reconstruction error compared to QHM or sinusoidal model of McAulay and Quatieri [1].

Concentrating on speech applications, we develop an analysis/synthesis speech system based on aQHM. Actually, aQHM is used for the representation of the quasi-periodicities of speech while the aperiodic part of speech is modeled as a time- and frequency-modulated noise. The resynthesized speech signal produced by the proposed system is indistinguishable from the original. Finally, another application of speech analysis where aQHM can be applied is the extraction of vocal tremor characteristics. Since vocal tremor is defined as modulations of the instantaneous components of speech, aQHM is the appropriate model for the representation of these modulations. Indeed, results showed that the reconstructed signals are close to the original signals which validate our method.

Περίληψη

Acknowledgements

This thesis is the result of a four-year working experience as PhD candidate at the research group for Media Informatics Lab at Computer Science Department of the University of Crete. This work was supported by a fellowship from Orange Labs through industrial contract NB• 200141932 in collaboration with the Institute of Computer Science, FORTH.

First of all I would like to thank Yannis Stylianou, my supervisor, for his guidance and encouragement during this work. Our collaboration, which started back in 2002, all these years have been very instructive, giving rise to several interesting tasks more or less related to the research. I would like to also thank Olivier Rosec for his fruitful discussions and suggestions during the fulfillment of this work.

The pleasant and creative research atmosphere at the group is of course also an outcome of all the other members of the group. Of these I want to give special thanks to A. Holzapfel, M. Makraki and M. Vasilakis for sharing with me some of their profound knowledge on digital signal processing, to N. Melikides, M. Fundulakis and Z. Haroulakis my friends, for the discussions on just everything that is interesting in this world. I would like to also thank Euaggelia Faitaki for supporting and tolerating me all this time.

Finally, I feel obliged to thank my family, for providing me with the ideal environment to grow up in, and exhorting me in changing for the better. Their constant support and encouragement have really brought me here.

I have had a lot of fun. Thank you all (I hope that I've not forgotten someone).

Contents

List of Tables

List of Figures

List of Abbreviations

AM	Amplitude Modulation
A/S	Analysis/Synthesis
CRLB	Cramer-Rao Lower Bound
DESA	Discrete Energy Separation Algorithm
FFT	Fast Fourier Transform
FM	Frequency Modulation
FT	Fourier Transform
GN	Gauss-Newton (method)
HM	Harmonic Model
HNM	Harmonic+Noise Model
LS	Least Squares
QHM	Quasi-Harmonic Model
aQHM	adaptive Quasi-Harmonic Model
cQHM	chirp Quasi-Harmonic Model
iQHM	iterative Quasi-Harmonic Model
MLE	Maximum Likelihood Estimation
MAE	Mean Absolute Error
MSE	Mean Squared Error
S-G	Savinzky- Golay (filter)
SM	Sinusoidal Model
SNR	Signal-to-Noise Ration
SRER	Signal-to-Reconstruction Error Ratio

Chapter 1

Introduction

One of the most important aspects in signal processing is the extraction of useful information from measurements obtained from physical or mechanical systems. Physical systems include speech production [2], musical instruments [3], marine mammals [4] while, mechanical systems include radars and sonars [5], [6] or digital communication systems [7]. In these systems, measurements are usually called signals which are functions of time or space or any other special domain. Typically, signals are represented by a parametric model whose parameters should be accurately estimated. The choice of the model depends crucially on the physical properties of the analyzed signal as well on the efficiency of the parameter estimation method. Indeed, a simple model usually has a straightforward and easy estimation solution but it is unable to represent accurately the signal, while, a very complex model may lead to intractable parameter estimation.

Concentrating on speech processing, an accurate representation of speech by parametric models is necessary for applications such as speech analysis/synthesis and speech modification/transformation. Actually, in such applications, the quality of the output speech is more significant, to some extend, than the computational burden. Indeed, if the signal is not accurately modeled, then the modeling error produced during the representation/analysis step will be propagated to the modification/transformation/synthesis step resulting in perceptual degradation of the quality of the resynthesized signal. Another area of speech processing where the modeling accuracy is crucial is that of voice pathology, where recorded speech is used as a non-invasive technique for the extraction of information related with the voice-production process and organs. In all these applications, high quality analysis of speech is required. There, the non-stationary, nonlinear, and non-Gaussian character of speech signals should be considered. Thus, the estimation and more generally the whole processing of speech becomes a quite demanding

task.

In this thesis, signals —most of the times speech— are represented using the sinusoidal model (SM) which adequately addresses the non-stationarity of speech signals. In sinusoidal representation, signals are assumed to consist of a superposition of sinusoids whose amplitude and frequency are time-varying. The goal is to accurately estimate the time-varying components of each sinusoid. The following Section reviews the sinusoidal representation as well approaches which have been proposed in the literature for the estimation of the unknown parameters of SM. The limitations of the estimation methods due to the multi-component and non-stationary character of the analyzed signals are also presented. Moreover, the connection between the sinusoidal representation and AM-FM signals is provided.

1.1 Review of Sinusoidal Models

Sinusoidal representation, as it was introduced by McAulay and Quatieri [1], received great popularity due to its simplicity in formulation and estimation as well its wide applicability in speech and audio synthesis [1, 8, 9, 10], coding [11, 12] and modification [13]. The estimation of the instantaneous components in sinusoidal representation engaged a lot of research work during the last decades. In the original work of McAulay and Quatieri [1], the analyzed signal is chopped into frames and the basic assumption is that locally each frame consists of sinusoids with constant amplitudes and constant frequency. Then, the sinusoidal components for one frame are determined from the maxima of the magnitude of the Fourier transform of the frame. This algorithm is known as spectral pick-peaking and it is motivated from the fact that periodogram is asymptotically an efficient frequency and amplitude estimator [14]. Quadratic interpolation is used for reducing the bias due to the discretization of the frequency-domain. Recently, further studies [15, 16, 17, 18] on the bias of the pick-peaking estimation algorithm led to improvements in the accuracy of the quadratically interpolated FFT-based pick-peaking estimation algorithm.

It is noteworthy that SM had been studied earlier by Hedelin [19] and by Almeida and Tribolet [20] but in a limited framework. Indeed, in [19] and in [20] as well in the work of Serra [9] and Serra and Smith [21] only the voiced speech was represented by SM while [1] suggested, under certain conditions that also unvoiced speech can be represented by SM. A second sinusoidal parameter estimation method has been proposed by George and Smith [22, 23, 24] which is based on an analysis-by-synthesis scheme. In order to determine the sinusoidal parameters, their method uses a successive approximation-based analysis-by-synthesis procedure rather than peak-picking.

A third approach for the estimation of sinusoidal parameters is based on the minimization of an error function which is usually the weighted sum of squared error between the model and the analyzed frame. This approach is known as the least squares (LS) method [25], [26] and for Gaussian noise is equivalent to the maximum likelihood estimator (MLE). In the context of sinusoidal parameter estimation, the minimization of the LS is highly nonlinear for the frequency parameters while it is linear for the amplitude and phase parameters. Thus, the estimation is typically split into two subproblems. The first subproblem is to compute an estimate for the frequencies using methods such as Pisarenko [27] or Yule-Walker [28]. The second subproblem is the estimation of the complex amplitudes —which merge both real amplitude and phase— by linear LS [29], [30]. Moreover, iterative schemes such as Gauss-Newton method [31] can be applied for further improving the accuracy of the frequency estimation leading to asymptotically efficient estimation. Particularly in voiced speech, Stylianou [32] assumed that frequencies are integer multiples of a fundamental frequency and proposed to estimate fundamental frequency by autocorrelation combined with spectral methods and then the complex amplitudes are computed by linear LS. In this thesis, we also use linear LS method for the estimation of the parameters of the suggested models.

Even though the sinusoidal model with FFT-based parameter estimation is widely used because of its simplicity, there are some limitations in this approach. Indeed, for multicomponent signals, such as speech, the interference between the components affects the accuracy of the FFT-based estimation methods. In order to alleviate the errors due to component interference, the duration of the analysis window should be increased. But then, the assumption of local stationarity is less valid resulting again in biased estimation, this time, due to the non-stationary character of the signal. On the other hand, LS amplitude estimation method tackles the problem of interference by canceling the interfering components, hence, windows with shorter duration may be used. Even though shorter windows can be used, the stationary assumption is not always valid within an analysis frame, thus, both amplitude and frequency modulation of the signal during the frame may produce bias at the parameter estimation, consequently, artifacts at the signal representation. For instance, one well known artifact in sinusoidal modeling is the pre-echo effect due to amplitude bursts which is tackled using for instance exponentially damped sinusoids [33, 34] or filterbank approaches as in [35].

However, the estimation of frequency modulations is more crucial compared with the estimation of amplitude modulations due to the fact that the parametric estimation of the time-varying amplitude given the time-varying frequency is always linear [36, 37] in the context of LS esti-

mation. A lot of effort has been put in signal processing community on modeling the frequency modulation of the analyzed frame. In order to remove the local stationarity assumption within a frame, the most common extension is to assume linear evolution of the frequency, thus, a chirp model replaces SM for one frame. Based on the fact that the Gaussian window has Fourier transform which is again a Gaussian function, many studies [38, 39, 40, 41] estimate the chirp rate by fitting a quadratic function to the log-magnitude spectrum. Other chirp rate estimation methods include discrete polynomial phase transform [42, 43] which is able to handle even higher-order frequency modulations, and, in the context of phase vocoder, chirp rate can be estimated from the slope of the derivative over time of the estimated instantaneous phase [44].

A different way of improving the accuracy of the sinusoidal parameter estimation is to increase the resolution of the Fourier spectrum. Reassignment method [45], [46] is a technique which refines both time and frequency resolution of the spectrogram. Moreover, variants of the Fourier transform such as Chirplet transform [47], [48] or fractional Fourier transform [49], [50] or Fan-Chirp transform [51], [52] are applied for smearing out the non-stationary sinusoids which are spread exactly due to the non-stationarity. One limitation of these approaches is that the parameters which determine the non-stationarity (for instance, the chirp rate which determines the linear evolution of the frequency) should somehow provided a priori. Other time-frequency distributions, such as Wigner-Ville [53], can be used with optimal results for some special cases. However, their use is rather limited in the case of multicomponent signals due to high amplitude interfering components.

1.1.1 AM-FM Signals and Demodulation

The definition of SM is well connected with the definition of an AM-FM signal. Actually, the mathematical definition of both models are mainly the same, although, there are differences between them. For instance, the components of an AM-FM signal may cross-over and usually the carrier frequency is of orders greater than the modulation frequency which is not typical for SM. Moreover, the number of components in AM-FM signals is usually smaller than the number of components in SM. In voiced speech, for instance, SM may be applied for modeling the harmonics, while AM-FM representation models the formants of speech (usually one formant per kHz). Since we develop an algorithm which is able to decompose both time-varying sinusoids and AM-FM signals, we will review most of the AM-FM demodulation algorithms presented in the literature.

The demodulation of an AM-FM signal depends on the number of components it contains.

For the mono-component case, analytic signal through Hilbert transform [54, 55] provides an estimate of the instantaneous amplitude and instantaneous phase. Instantaneous frequency is then computed by differentiate the unwrapped instantaneous phase. Another well-known algorithm for mono-component AF-FM demodulation is the Discrete Energy Separation Algorithm (DESA) developed by Maragos, Quatieri and Kaiser [56, 57]. DESA utilizes the nonlinear Teager-Kaiser operator which has fine time-resolution. For a comparison between Hilbert method and DESA, please refer to [58] and [59]. Phase-locked loops [60], as well extended Kalman filter [61] are also utilized for the demodulation of mono-component AM-FM signals.

However, the generalization to multi-component AM-FM signals is not a trivial task. Even the well-posiness of the definition of a multi-component AM-FM signal received great attention [62, 63, 64]. The most common solution for demodulation of a multicomponent AM-FM signal is to pass the signal from a filterbank and then apply the preferred mono-component AM-FM demodulation algorithm to the output of each filter [65, 66, 67, 35]. This approach is similar to phase vocoder algorithm [68] used in speech processing. However, the interference between the adjacent filters as well the crossing of a component between different filters add limitations to this approach. Another AM-FM component-separation approach has been proposed relatively recently by Santhanam and Maragos [69] which separates the AM-FM components algebraically based on the periodicity characteristics of the components. This algorithm is very attractive since the separation is accurate even when the AM-FM components cross each other. The weak point of this demodulation method is that the period of each AM-FM component should be correctly computed. Finally, a novel multi-component AM-FM decomposition algorithm was proposed in [70] which is highly accurate for signal representation, however, the extracted AM-FM components usually lacks physical meaning especially when the components have approximately equal strength.

1.2 Contribution of this thesis

The importance of accurate sinusoidal parameter estimation has been highlighted. Different approaches based on stationary or time-varying models which perform sinusoidal parameter estimation as well their limitations have been presented. The time-varying frequencies of the signals put a major limitation in the estimation methods since bias is introduced. Thus, a way to efficiently tackle this issue is of high interest. The main objective of this thesis is to develop time-varying models which are able to adjust locally their frequency information to the frequency of

the analyzed signal. This results in reducing the estimation bias since the model representation is more accurate. Hence, the quality of the signal representation is improved. Then, the novel models are applied in applications such as speech analysis/synthesis and voice quality assessment.

The major contributions of the work presented in this thesis are:

- A time-varying model which is referred to as Quasi-Harmonic Model (QHM) was revised in a new basis and its properties have been fully explored [71]. The estimation of the unknown parameters of QHM is performed using Least Squares (LS) method.

- The most significant property of QHM is its ability to estimate the frequency mismatch between the original frequency and the initially provided frequency for each component of the signal. This is achieved by proper decomposition of the estimated QHM parameters. Thus, an iterative algorithm called iQHM similar to Gauss-Newton (GN) method is developed for the estimation of sinusoidal parameters given an initial estimate of frequencies [72]. Statistical efficiency of iQHM is also tested.

- The region of convergence of the iterative algorithm i.e., bounds on the maximum allowed frequency mismatch is provided. It is shown that the frequency mismatch should be less than one third of the bandwidth of the squared analysis window.

- An extension of QHM referred to as chirp QHM (cQHM) [73] which is able to capture linear evolution of the frequency without the need of providing a priori the chirp rate is also presented. Basic properties of cQHM are given.

- Another even more powerful extension of QHM, referred to as adaptive QHM (aQHM) is proposed. Instead of initially estimated frequencies, aQHM uses an estimate of the instantaneous phase. Thus, time information is added to the model which results in an adaptive to the input signal and a non-stationary signal representation.

- An AM-FM decomposition algorithm is suggested [73], [74]. This algorithm is initialized by QHM, which serves as a frequency tracker, providing, thus, an initial estimate of the instantaneous components of the signal. The accuracy of the estimation is then improved by aQHM.

- An interpolation scheme for the instantaneous phase is proposed. It is based on the integration of the instantaneous frequency plus a correction term which guarantees the continuity of phase and of frequency.

- Analysis/synthesis of speech based on separation of speech into two parts, the deterministic part and the stochastic part. The deterministic part is modeled by the sinusoidal representation and the instantaneous components are estimated using a variant of the suggested AM-FM decomposition algorithm. The stochastic part is modeled as a time- and frequency-modulated noise. Time-modulation of noise is based on an estimation of the energy envelope [75] .

- Extraction of vocal tremor characteristics of sustained vowels based on the suggested AM-FM decomposition algorithm [76] is developed. The decomposition algorithm is applied for the estimation of the instantaneous components of speech signals as well for the extraction of acoustic characteristics of vocal tremor such as modulation frequency and modulation level.

1.3 Thesis Organization

The organization of this thesis is as follows. In Chapter 2, QHM, which is an extension of SM, is introduced. Parameter estimation of QHM is performed through LS. The properties of QHM both in time-domain and in frequency-domain are provided. Moreover, an iterative scheme is presented which is able of unbiased estimation of sinusoidal parameters. The convergence of the iterative algorithm are also investigated. Finally, an extension of QHM which is called chirp QHM (cQHM) is provided.

Chapter 3 shows that stationary sinusoidal analysis is inappropriate for the case where the analyzed frame is non-stationary. Thus, a novel model is introduced, namely aQHM, which is able to adaptively estimate the time-varying characteristics of the frame. We show that aQHM is fundamentally different from QHM. Furthermore, aQHM suggests an AM-FM decomposition algorithm which is also presented. The performance of the new AM-FM decomposition algorithm is tested on synthetic signals and on real voiced speech.

Chapter 4 presents an analysis/synthesis speech system based on the decomposition of speech into two parts. The deterministic part, which accounts for the quasi-periodicities of speech, is modeled by the sinusoidal representation whose parameters is estimated from the suggested AM-FM decomposition algorithm. The stochastic part, which accounts for the aperiodicities of speech, is modeled as time-modulated and frequency-modulated noise. Details on the implementation issues are given.

Chapter 5 applies the proposed AM-FM decomposition algorithm for the estimation of vocal

tremor properties. Thus, vocal tremor in sustained phonation is defined and an algorithm based on the suggested AM-FM decomposition is used for the extraction of acoustical vocal tremor characteristics such as modulation frequency and modulation level.

Chapter 6 resumes the major contributions and results of this thesis and gives some directions for further research on sinusoidal parameter estimation, on AM-FM signal decomposition methods as well on extensions of the presented speech applications.

Finally, Appendix A presents various computational tricks for faster estimation of QHM unknown parameters while Appendix B shows the equivalence between iQHM and a sequential version of GN method.

Chapter 2

Quasi-Harmonic Model

In this Chapter, we introduce the Quasi-Harmonic Model (QHM) for the representation of almost (or quasi) periodic signals. QHM is not a novel model since it has been firstly introduced by Laroche [77] back in 1989 for the representation of percussive sounds and, later, for modeling of voiced speech by Stylianou [32]. However, the main properties of QHM were not extensively explored. For instance, it was known that QHM contains frequency information ([32], pg. 83) but it was not known why and, furthermore, how this information can be extracted correctly. Recently, Valin et al. [78] defined a variant of QHM using linear approximations of trigonometric functions. In this Chapter, we derive the time-domain and, most importantly, the frequency-domain properties of QHM. We show that a proper decomposition of QHM's parameters results in estimation of the frequency mismatch between the analysis frequency and the true frequency of a sinusoid, whenever there is a frequency mismatch. Furthermore, an iterative algorithm is derived for the estimation of sinusoidal parameters using QHM and Least-Squares (LS) method. However, some approximations are performed for the estimation of the frequency mismatch, hence, we present the effects and the limitations of these approximations. Also the robustness of the estimation process under noisy conditions is explored. Finally, a variant of QHM which is able to capture not only frequency mismatches but also the chirp rate of the analyzed signal is presented.

2.1 Preliminaries

In the context of sinusoidal representation, the real signal to be analyzed is viewed as a sum of amplitude-modulated and frequency-modulated sinusoids given by

$$s(t) = A_0(t) + \sum_{k=1}^{K(t)} 2A_k(t)cos(\phi_k(t)) = \sum_{k=-K(t)}^{K(t)} A_k(t)e^{j\phi_k(t)} \tag{2.1}$$

where $A_k(t)$ is the instantaneous amplitude, while $\phi_k(t)$ is the instantaneous phase of the kth component, respectively. Instantaneous frequency is defined as the derivative of instantaneous phase with respect to time scaled by $1/(2\pi)$, i.e.,

$$f_k(t) = \frac{1}{2\pi}\frac{d\phi_k(t)}{dt} \tag{2.2}$$

Note also that the number of components is not constant over time which is necessary for the representation of non-stationary signals such as music or speech.

In a frame-by-frame sinusoidal analysis, signal $s(t)$ is chopped into pieces called frames which are denoted by

$$s_l(t) = s(t - t_l)w(t) \tag{2.3}$$

where t_l is the center of the frame and $w(t)$ be the analysis window function with support in $t \in [T_l, T_l]$. Typically, the window function vanish at the limits of its support so as to alleviate the discontinuities at the boundaries of the frame as well to eliminate the side-lobe interference between the components. Note also that the window length may depend on the particular frame and, in speech analysis for instance, it is usual to depend on the local pitch period.

Sinusoidal modeling assumes that one frame has stationary components, meaning that it consists of a superposition of sinusoids with constant amplitudes and constant frequencies, i.e., one frame is modeled as[1]

$$h_s(t) = \sum_{k=-K}^{K} a_k e^{j2\pi f_k t}w(t), \qquad t \in [-T, T] \tag{2.4}$$

where K is the local number of components while f_k and a_k are the local frequency and local complex amplitude of the kth sinusoid, respectively. As stated in Chapter 1, there is a vast

[1]In order to be consistent with (2.3), we should add a subscript to each parameter to denote the particular frame number. However, it is not necessary for this chapter and it is dropped for simplicity. The additional frame indexing is applied in Chapter 3 where the AM-FM decomposition algorithm is presented.

literature on the estimation of the unknown sinusoidal parameters.

A usual restriction of the admissible frequency values of SM which is utilized in many real signals results in the harmonic model (HM). In HM, the frequencies are not arbitrary, rather they are determined as integer multiples of a fundamental frequency. Hence, HM is given by

$$h_h(t) = \sum_{k=-K}^{K} a_k e^{j2\pi k f_0 t} w(t), \qquad t \in [-T, T] \tag{2.5}$$

where f_0 is the local fundamental frequency and a_k is again the local complex amplitude of kth sinusoid. The typical estimation approach for HM's unknown parameters is to firstly estimate the local fundamental frequency using time-domain and/or frequency-domain techniques and then estimate the complex amplitudes using linear LS [32]. However, this estimation approach suffers from the fact that it produces bias in the estimation of complex amplitudes whenever the local fundamental frequency is erroneous or whenever the frequencies of the real signal are not exactly integer multiples of fundamental frequency. In QHM, which follows, the goal is again the estimation of the complex amplitudes using LS method without the limitations due to inaccurate frequency estimation. As we will show, QHM is able to correct frequency estimation errors, thus, it produces unbiased estimates for the complex amplitudes.

2.2 Definition of Quasi-Harmonic Model, QHM

As in sinusoidal model (2.4), one frame is assumed to consist of a superposition of sinusoids with constant frequencies and constant amplitudes. Nevertheless, we suggest modeling one frame using a time-varying model referred to as QHM, which is defined by

$$h_q(t) = \sum_{k=-K}^{K} (a_k + t b_k) e^{j2\pi \hat{f}_k t} w(t), \qquad t \in [-T, T] \tag{2.6}$$

where K is the number of sinusoidal components, $\hat{f}_k$ is the analysis frequency for the kth component which are assumed to be known, a_k is the complex amplitude while b_k is the complex slope of the kth component, respectively. Please note that $a_{-k} = \bar{a}_k$ and $b_{-k} = \bar{b}_k$ as well that $a_0, b_0 \in \mathbb{R}$ when the analyzed signal is real. Hence, QHM has $4K + 2$ unknown real variables in the real signal case. Analysis window, $w(t)$, has support in $[-T, T]$. We assume that the true frequencies of the analyzed signal are not known, but an estimate of them is provided. Hence,

there is a frequency mismatch error given by

$$\eta_k = f_k - \hat{f}_k \tag{2.7}$$

where f_k are the true frequency of the kth component. It is very common in speech the a priori provided frequencies, $\hat{f}_k$, being integer multiples of an estimated fundamental frequency, i.e. $\hat{f}_k = k\hat{f}_0$.

The estimation of QHM unknown parameters is performed by minimizing the Least Squares (LS) error. The error is defined by

$$\begin{aligned} \epsilon(\mathbf{a},\mathbf{b}) &= \sum_{n=-N}^{n=N} |s(t_n) - h_q(t_n)|^2 \\ &= (\mathbf{s} - E\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix})^H W^H W (\mathbf{s} - E\begin{bmatrix}\mathbf{a}\\ \mathbf{b}\end{bmatrix}) \end{aligned} \tag{2.8}$$

where $\mathbf{a} = [a_{-K}, ..., a_K]^T$ and $\mathbf{b} = [b_{-K}, ..., b_K]^T$ are the unknown vectors of size $(2K+1)\times 1$, $\mathbf{s} = [s(t_{-N}), ..., s(t_N)]^T$ is the samples of the analyzed frame of size $(2N+1)\times 1$, $E = [E_0|E_1]$ is the matrix with the exponentials of size $(2N+1)\times(4K+2)$. Furthermore, submatrices E_0 and E_1 have elements which are given by $(E_0)_{n,k} = e^{j2\pi\hat{f}_k t_n}$ and $(E_1)_{n,k} = t_n e^{j2\pi\hat{f}_k t_n} = t_n (E_0)_{n,k}$, respectively, while W is a diagonal $(2N+1)\times(2N+1)$ matrix with entries the window values. The superscript ^{H} denotes the Hermitian operator. It is noteworthy that while we used continuous-time in (2.6), we switch to discrete-time in (2.8) in order to perform the LS computation. Please note that in this thesis, discrete-time formulation is used only for the estimation of unknown parameters of the models. In any other case, continuous-time will be used since it is easier for mathematical manipulation.

The minimization of the error function is linear with respect to the complex unknown parameters given that the analysis frequencies, $\hat{f}_k$, are known. The solution in matrix notation is given by

$$\begin{bmatrix}\hat{\mathbf{a}}\\ \hat{\mathbf{b}}\end{bmatrix} = (E^H W^H W E)^{-1} E^H W^H W \mathbf{s} \tag{2.9}$$

Appendix A shows the fully discrete formulation and how the solution can be speed up by, firstly, using some explicit formulas and, secondly, by taking advantage of the properties of the derived matrices. Furthermore, we show in Appendix A that the time-consuming part of LS estimation is not so much the inversion of the involved matrix but mostly its construction.

Finally, for an objective evaluation of the modeling performance, we propose to reconstruct the signal and then measure the Signal-to-Reconstruction Error Ratio (SRER) which is defined as

$$SRER = 20\log_{10}\frac{\sigma_{s(t)}}{\sigma_{s(t)-\hat{s}(t)}} \tag{2.10}$$

where σ_s denotes the standard deviation of $s(t)$, and $\hat{s}(t)$ is the reconstructed signal. Please note that SRER is measured in decibel (dB). For QHM, the reconstructed signal —actually, reconstructed frame— is given by

$$\hat{s}(t) = \sum_{k=-K}^{K} (\hat{a}_k + t\hat{b}_k)e^{j2\pi\hat{f}_k t}w(t), \qquad t \in [-T, T] \tag{2.11}$$

2.2.1 Motivation Example

The advantage of QHM over HM is revealed when the analysis (or input, or a priori provided) frequencies are different from the true ones. In such cases, the estimation of the complex amplitude is biased due to the frequency mismatch and the representation of the analyzed signal is not accurate. This is depicted in Figure 2.1 where a pure sinusoid with frequency $100Hz$ (line with circles) is modeled with HM (solid line) and QHM (dashed line) with analysis frequency at $90Hz$. The duration of the signal is $40ms$ ($T = 20ms$) and Hamming window is used. Obviously, HM is incapable of modeling the original signal while QHM is able to remedy the frequency mismatch quite satisfactorily. Indeed, SRER for QHM is $20.5dB$ while SRER for HM is $8.5dB$. In the following Sections, we provide a solid theoretical analysis of this behavior and we suggest ways to exploit it for the estimation of sinusoidal parameters in the context of LS estimation.

2.3 Properties of QHM

In this Section, we study the time-domain and frequency-domain properties of QHM showing in parallel the differences between QHM and SM.

2.3.1 Time-Domain Properties

The time-domain characteristics of the model are discussed in this subsection. From (2.6), it is easily seen that the instantaneous amplitude of the kth component is a time-varying function which is given by

$$M_k(t) = |a_k + tb_k| = \sqrt{(a_k^R + tb_k^R)^2 + (a_k^I + tb_k^I)^2} \tag{2.12}$$

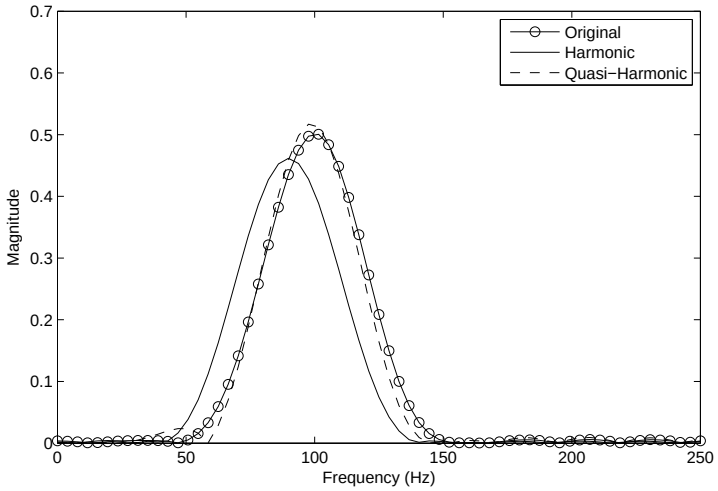

Figure 2.1: The Fourier spectra of the original signal (line with circles), of the reconstruction of HM (solid line) and of the reconstruction of QHM (dashed line). Obviously, QHM representation is closer to the original signal compared with HM.

where x^R and x^I denote the real and the imaginary part 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 for the kth component is given by

$$\Phi_k(t) = 2\pi\hat{f}_k t + \angle(a_k + tb_k) = 2\pi\hat{f}_k t + atan\frac{a_k^I + tb_k^I}{a_k^R + tb_k^R} \tag{2.13}$$

while instantaneous frequency is given by

$$F_k(t) = \frac{1}{2\pi}\Phi_k'(t) = \hat{f}_k + \frac{1}{2\pi}\frac{a_k^R b_k^I - a_k^I b_k^R}{M_k^2(t)} \tag{2.14}$$

Substituting (2.12) to (2.14), it is easily observed that the instantaneous frequency is a bell-shaped curve similar to Cauchy distribution. Figure 2.2 shows the instantaneous frequency of QHM (dashed line) as it is computed from (2.14). Obviously, it is closer to the true frequency (line with circles) especially at the middle of the analysis window even though the analysis is performed at a wrong frequency (solid line). From the same Figure, it is also obvious that the overall shape of the instantaneous frequency of QHM has no correlation with the original instantaneous frequency of the sinusoid, which is constant in this example. Finally, a feature of the model worth noting is that the 2nd term of the instantaneous frequency in (2.14) depends

on the instantaneous amplitude which means that the accuracy of frequency estimation (or, the estimation of phase function) depends on the amplitude strength. This observation is in accordance with the Cramer-Rao lower bound of frequency estimation [26].

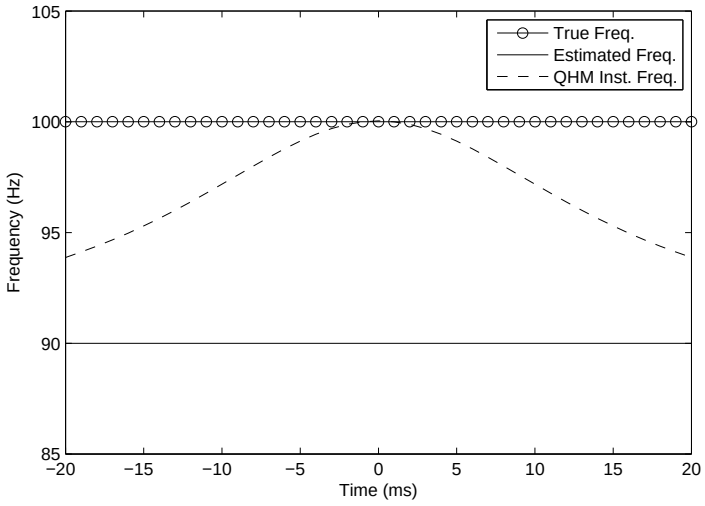

Figure 2.2: A frame of $40ms$ duration which contains a pure sinusoid with frequency $100Hz$ (line with circles) is analyzed at $90Hz$ (solid line). Instantaneous frequency of QHM (dashed line) tries to adjust to the true frequency of the sinusoid.

2.3.2 Frequency-Domain Properties

Let us consider the Fourier transform of $h_q(t)$ in (2.6) given by

$$H_q(f) = \sum_{k=1}^{K} \left(a_k W(f - \hat{f}_k) + \frac{j b_k}{2\pi} W'(f - \hat{f}_k) \right) \tag{2.15}$$

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

$$H_{q,k}(f) = a_k W(f - \hat{f}_k) + \frac{j b_k}{2\pi} W'(f - \hat{f}_k) \tag{2.16}$$

To reveal the main properties of QHM, we suggest the projection of b_k onto a_k as illustrated in Figure 2.3. Accordingly,

$$b_k = \rho_{1,k} a_k + \rho_{2,k} j a_k \tag{2.17}$$

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

$$\rho_{1,k} = \frac{a_k^R b_k^R + a_k^I b_k^I}{|a_k|^2} \tag{2.18}$$

and

$$\rho_{2,k} = \frac{a_k^R b_k^I - a_k^I b_k^R}{|a_k|^2} \tag{2.19}$$

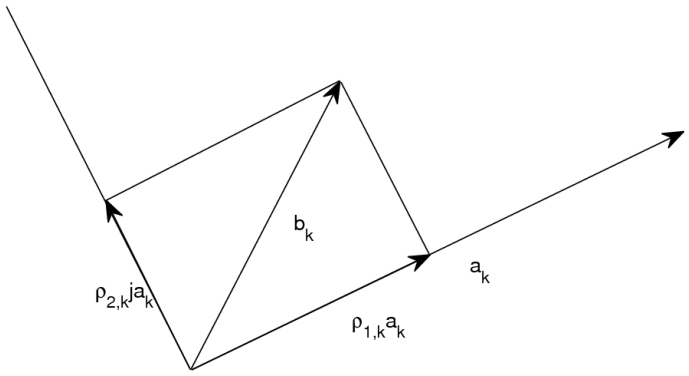

Figure 2.3: The projection of b_k into one parallel and one perpendicular to a_k component. Complex numbers are thought as vectors on the plane.

Thus, the kth component of $H_q(f)$ is rewritten as

$$H_{q,k}(f) = a_k W(f - \hat{f}_k) - \frac{a_k \rho_{2,k}}{2\pi} W'(f - \hat{f}_k) + \frac{j a_k \rho_{1,k}}{2\pi} W'(f - \hat{f}_k) \tag{2.20}$$

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

$$\begin{aligned} W(f - \hat{f}_k - \frac{\rho_{2,k}}{2\pi}) &= W(f - \hat{f}_k) - \frac{\rho_{2,k}}{2\pi} W'(f - \hat{f}_k) + O(\rho_{2,k}^2 W''(f - \hat{f}_k)) \\ &\approx W(f - \hat{f}_k) - \frac{\rho_{2,k}}{2\pi} W'(f - \hat{f}_k) \end{aligned} \tag{2.21}$$

Consequently, from (2.20) and (2.21) it follows that

$$H_{q,k}(f) \approx a_k \left[W(f - \hat{f}_k - \frac{\rho_{2,k}}{2\pi}) + j \frac{\rho_{1,k}}{2\pi} W'(f - \hat{f}_k) \right] \tag{2.22}$$

Going back in the time-domain, (2.6) (i.e., QHM) is approximated as

$$h_q(t) \approx \sum_{k=-K}^{K} a_k \left[e^{j(2\pi \hat{f}_k + \rho_{2,k})t} + \rho_{1,k} t e^{j2\pi \hat{f}_k t} \right] w(t) \tag{2.23}$$

From (2.23), it is clear that $\frac{\rho_{2,k}}{2\pi}$ can be thought as an estimate of the frequency mismatch between the actual frequency of the kth component and the provided frequency, $\hat{f}_k$, 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 (2.14) and (2.19), it follows that

$$\frac{\rho_{2,k}}{2\pi} = F_k(0) - \hat{f}_k \tag{2.24}$$

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

$$\rho_{1,k} = \frac{\frac{dM_k(t)}{dt}|_{t=0}}{M_k(0)} \tag{2.25}$$

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

Presumably, the decomposition of b_k gives a way to estimate the frequency mismatch between the true frequency and the analysis frequency. Thus, it is straightforward to construct an algorithm which performs sinusoidal parameter estimation.

2.4 Application to Sinusoidal Parameter Estimation

Previous Section suggests that an estimate of the frequency mismatch of the kth sinusoidal component is given by

$$\hat{\eta}_k = \rho_{2,k}/2\pi \tag{2.26}$$

Thus, an algorithm which is able to iteratively estimate the frequency mismatches and correct the frequencies is suggested. We name this iterative algorithm iQHM and it is given in pseudo-code by

---

Iterative Sinusoidal Parameter Estimation

---

1. Initialization

   i. Get an initial estimate of frequencies, $\{\hat{f}_k\}_{k=1}^K$

   ii. Estimate $\{a_k, b_k\}_{k=-K}^K$ given $\{\hat{f}_k\}_{k=1}^K$ using (2.9)

2. Do iterations

   i. For each kth component:

      a. Estimate $\hat{\eta}_k$ using (2.26)

      b. Update frequencies: $\hat{f}_k \leftarrow \hat{f}_k + \hat{\eta}_k$

   ii. Reestimate $\{a_k, b_k\}_{k=1}^K$ given $\{\hat{f}_k\}_{k=1}^K$ using (2.9)

The above iterative algorithm converges to the true parameters when the frequency mismatch, η_k, is adequately small. In the next Section we will provide the necessary conditions of the frequency mismatch for the convergence of iQHM. Once the frequency mismatch is within the appropriate region of convergence, the number of iterations needed for reaching a stable estimate is very low. Typically two to four iterations are enough. Alternatively, a convergence criterion can be used for stopping the iterative algorithm. For instance, a convergence criterion may be: if $\frac{|\hat{f}_k^{new} - \hat{f}_k^{old}|}{\hat{f}_k^{old}} < \epsilon$ is satisfied for all k, then stop. Please note that an estimate of the complex amplitude of the kth sinusoid is provided by a_k.

Finally, using Taylor series expansion, it can be shown that QHM is a linearization of the frequency mismatch. This linearization in conjunction with LS method make iQHM similar to other iterative sinusoidal estimation method. Indeed, iQHM can be viewed as a variant of the Gauss-Newton (GN) optimization method which is developed in Appendix B. In Appendix B, the similarities and the differences between iQHM and GN method are explored.

2.5 Effects of approximations on the frequency estimation process and noise robustness

In the previous Section, we showed that $\rho_{2,k}/(2\pi)$ is an estimator, under certain conditions, of the frequency mismatch between the true and the initially provided analysis frequencies of the

underlying sine-wave. In this Section, we explicitly refer to these conditions and investigate their effects. Namely, these are the effect of the analysis window and the effect of the approximation in (2.21). Finally, we discuss the robustness of the frequency estimator under noise.

2.5.1 Effect and Importance of Window Duration

Since QHM has $4K + 2$ unknown real parameters, the length of the analysis window should be at least $4K + 2$ (in samples) in order to obtain stable LS solutions. Moreover, low frequency components need larger windows, and an empirical choice for the analysis window length is that this should be at least $2\lfloor \frac{f_s}{\min\limits_k f_k} \rfloor$ where $\lfloor \cdot \rfloor$ denotes the floor operator while f_s is the sampling frequency. Furthermore, when the original signal is contaminated by noise, more samples (i.e. larger window) are needed in order to perform more robust and accurate estimation of the unknown parameters [26], [30]. On the other hand, when larger windows are used, the possibility of the signal being non-stationary is higher, which may introduce errors and biases in the estimation process. Additionally, we will show in the following subsection that the smaller the window length the more valid is the approximation in (2.21). From the above discussion, it should be clear that the length of the analysis window is very important and there is a trade-off between the accuracy of the proposed iterative sinusoidal parameter estimation algorithm and its robustness. As a rule of thumb, we suggest the use of as small as possible window length.

2.5.2 Estimation Error of Frequency Mismatch

Due to the approximation in (2.21), the suggested estimator for the frequency mismatch, $\rho_{2,k}/(2\pi)$, is generally not an unbiased estimator. Moreover, frequency mismatch estimator 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 non-stationary, 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 by far the most significant one. Indeed, if the correct value of the frequency of a component of the input signal is known, then, unbiased estimates of the corresponding complex amplitude is obtained through LS [26], [30]. Thus, the focus is on the frequency mismatch estimation.

Let us consider the mono-component case of a stationary signal where one frame is given by

$$\begin{aligned} s(t) &= A_1 e^{j2\pi f_1 t} w(t) = A_1 e^{j(2\pi \hat{f}_1 t + 2\pi\eta_1 t)} w(t) \\ &= A_1 (cos(2\pi\eta_1 t) + j sin(2\pi\eta_1 t)) e^{j2\pi \hat{f}_1 t} w(t), \qquad t \in [-T, T] \end{aligned} \tag{2.27}$$

where A is the complex amplitude, f_1 is the true frequency, $\hat{f}_1$ is the estimated frequency and η_1 is the frequency mismatch between them to be estimated.

In the context of QHM, the original signal is modeled as

$$h_q(t) = (a_1 + t b_1) e^{j2\pi \hat{f}_1 t} w(t), \qquad t \in [-T, T] \tag{2.28}$$

where a_1 and b_1 are the unknown complex amplitude and slope, respectively, which are estimated through LS as presented in Section 2.2. It can be shown that the LS method involves the projection of the input signal onto two orthogonal basis functions: $e^{j2\pi \hat{f}_1 t} w(t)$ and $te^{j2\pi \hat{f}_1 t} w(t)$. Thus, for a rectangular window the complex amplitude is obtained by

$$\begin{aligned} a_1 &= \frac{< w(t)s(t), w(t)e^{j2\pi \hat{f}_1 t} >}{< w(t)e^{j2\pi \hat{f}_1 t}, w(t)e^{j2\pi \hat{f}_1 t} >} \\ &= A_1 \frac{sin(2\pi\eta_1 T)}{2\pi\eta_1 T} \end{aligned} \tag{2.29}$$

where $< ., . >$ denotes the inner product between functions, defined as

$$< x_1(t), x_2(t) >= \int_{-T}^{T} x_1(t) \bar{x}_2(t) dt$$

The complex slope is obtained by

$$\begin{aligned} b_1 &= \frac{< w(t)x(t), w(t)te^{j2\pi \hat{f}_1 t} >}{< w(t)te^{j2\pi \hat{f}_1 t}, w(t)te^{j2\pi \hat{f}_1 t} >} \\ &= 3jA_1 \left(\frac{sin(2\pi\eta_1 T)}{(2\pi\eta_1)^2 T^3} - \frac{cos(2\pi\eta_1 T)}{2\pi\eta_1 T^2}\right) \end{aligned} \tag{2.30}$$

Then, the estimated value for η_1 is given by

$$\hat{\eta}_1 = \frac{1}{2\pi}\rho_{2,1} = 3\left(\frac{1}{\eta_1 (2\pi T)^2} - \frac{cot(2\pi\eta_1 T)}{2\pi T}\right) \tag{2.31}$$

To inquire the properties of this estimator, it is worth computing its error in estimating the

frequency mismatch (i.e., estimation error)

$$er(\eta_1) = \eta_1 - \hat{\eta}_1 \tag{2.32}$$

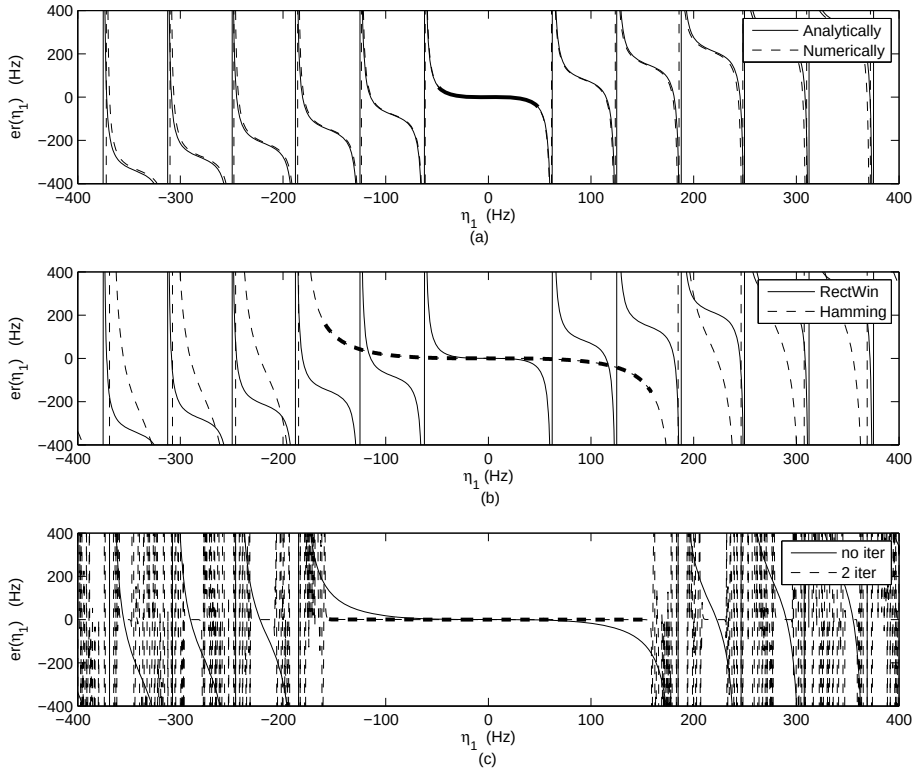

Figure 2.4: Upper panel: Estimation error of frequency mismatch for a rectangular window computed analytically (solid line) and numerically (dashed line). Middle panel: The estimation error for a rectangular (solid line) and Hamming window (dashed line). Lower panel: The estimation error using the Hamming window (as in b) without (solid line) and with two iterations (dashed line). Note that the iterative estimation fails when $|\eta_1| > B/3$.

In the case of a mono-component signal and using a rectangular window, the estimation error can be computed analytically as above. Figure 2.4(a) depicts the error for a rectangular window of $16ms$ ($T = 8ms$) obtained analytically via (2.31) (solid line), and numerically through LS computation of $\{a_1, b_1\}$ and then applying (2.19) (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 computing numerically the estimation error to infer its analytical value, whenever the latter is not computationally tractable. In Figure 2.4(a), the estimation error

is small[2] (see the bold line) if the frequency mismatch is below $50Hz$. For a Hamming window, the error is small if the frequency mismatch is below $135Hz$ as shown on Figure 2.4(b).

In order to get further insight on the role played by the analysis window, we can first notice from (2.29) and (2.30), 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 [54]) of the squared analysis window. Note also that the bandwidth of a (squared) rectangular window of length $2T$ is $B = 1/T = 125Hz$ ($T = 8ms$) while for a squared Hamming window we have $B = 3/T = 375Hz$, which may explain why the region with small estimation error is about 3 times larger for a Hamming window than for a rectangular window in Figure 2.4. 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

$$|\eta_1| < B/3 \tag{2.33}$$

where B is the bandwidth of the squared analysis window.

Applying the iterative scheme (iQHM), it is expected to reduce the estimation error of frequency mismatch to zero at least for the cases where the frequency mismatch is less than $B/3\ Hz$. Indeed, in Figure 2.4(c) the estimation error is depicted for no iteration (solid line) and after two iterations (dashed line). Again, Hamming window of duration $2T = 16ms$ is used as in Figure 2.4(b). We observe that the estimation error is considerably reduced (mainly is zero) if the initial frequency mismatch is smaller than $B/3$. It is worth noting that two iterations are adequate for reducing the estimation error of frequency mismatch, thus, the frequency error to zero.

2.5.3 Robustness in Noise

In this subsection, the performance of QHM and iQHM is assessed for the case when a signal with multiple sinusoidal components is contaminated by white Gaussian noise. Concisely, the ability of the proposed model to improve the accuracy of the frequency estimation – hence the accuracy of the complex amplitude estimation – is tested. The signal consists of 4 sinusoids and it is corrupted by noise while window's duration is $16ms$ ($T = 8ms$) and sampling frequency

[2]By small, we mean that $|er(\eta_1)| < |\eta_1|$.

$16000Hz$ (i.e, so the duration of the window duration in samples is 257). Moreover, Hamming window is used, thus, the maximum allowed frequency mismatch is $125Hz$. In Table 2.1, the frequency and the amplitude of each component are given. Two closely-spaced sinusoids and two well-separated sinusoids are considered. Monte Carlo simulations are used for the assessment of the robustness of the proposed method. For each simulation, the frequency mismatch of each sinusoid is sampled uniformly on the intervals defined in Table 2.1.

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 interval (Hz)	$[-20, 20]$	$[-20, 20]$	$[-75, 75]$	$[-75, 75]$

Table 2.1: Parameters of a synthetic sinusoidal signal with four components and intervals of allowed frequency mismatch per component.

Figures 2.5 and 2.6 respectively depict the mean squared error (MSE) of the complex amplitude and frequency of each component after 10^5 Monte Carlo simulations. Please note that MSE for a parameter θ is generally given by

$$MSE(\theta) = \frac{1}{M}\sum_{i=1}^{M} |\theta - \hat{\theta}^{(i)}|^2 \tag{2.34}$$

where M is the number of simulations while $\hat{\theta}^{(i)}$ is the estimated parameter at the ith simulation.

Moreover, Cramer-Rao lower bound (CRLB) [79, 14] for the amplitude and for the frequency are depicted at both Figures. CRLB for the amplitude of the kth component is given by

$$CRLB(a_k) = \frac{\sigma^2}{2N+1} \tag{2.35}$$

while CRLB for the frequency of the kth component is given by

$$CRLB(f_k) = \frac{f_s}{2\pi}\frac{12\sigma^2}{|a_k|^2(2N)(2N+1)(2N+2)} \tag{2.36}$$

where σ^2 is the variance of the white noise while $2N+1$ is the duration of the analysis window in samples. Figures indicate that the estimation of both complex amplitudes and frequencies asymptotically reaches the CRLB after three iterations which means that iQHM is a statistically efficient sinusoidal estimator. This result is expected since iQHM is closely related with GN method (see Appendix B) which is a statistically efficient sinusoidal parameter estimator.

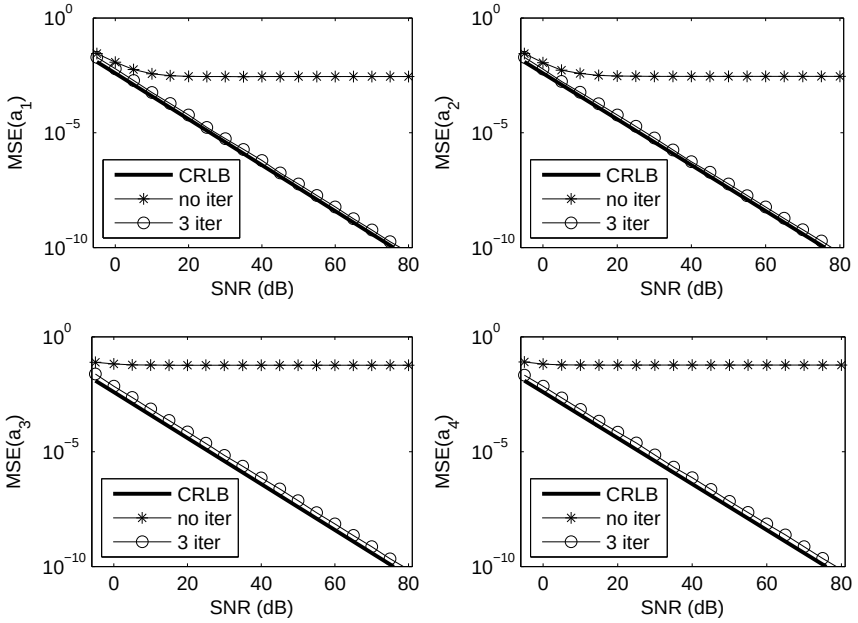

Figure 2.5: MSE of the four amplitudes as a function of SNR. Please note that no iterations refers to QHM while 3 iterations refers to iQHM.

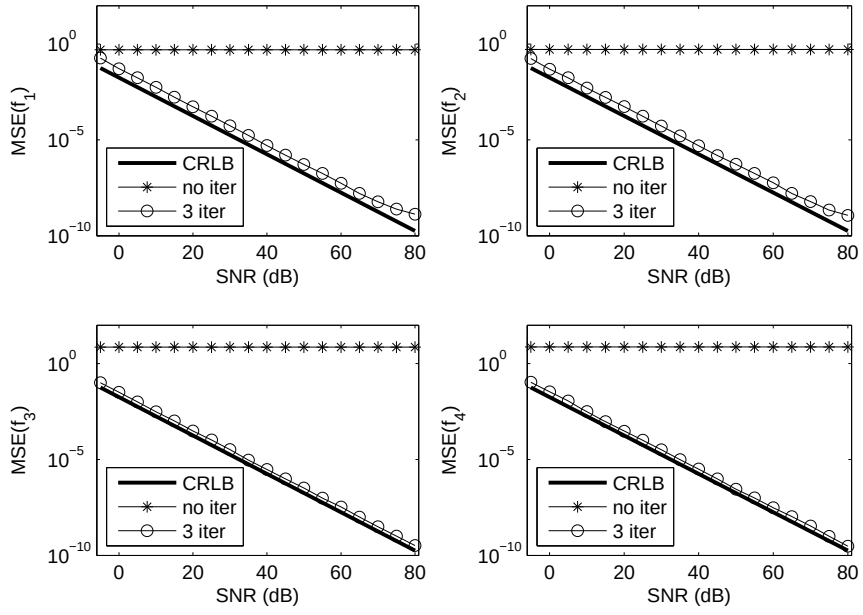

Figure 2.6: MSE of the four frequencies as a function of SNR. Please note that no iterations refers to QHM while 3 iterations refers to iQHM.

2.6 QHM and Real Signals

There is a vast literature in coding, synthesis, modification, etc. where both speech and music signals are modeled frame-by-frame as a sum of harmonically related sinusoids. However, looking at the magnitude spectrum of short-term Fourier transform it is easily seen that the local maxima (peaks) are not exactly at the integer multiples of the fundamental frequency. This inharmonicity – also called detuning – induces biased estimation of the complex amplitudes. Furthermore, even if the frequencies of the real signals were perfect harmonics, errors may occur in the estimation of the fundamental frequency, hence once again, bias is introduced in the amplitude estimation.

Previously, we showed theoretically and on synthetic signals that QHM is able to tackle with small frequency errors. Hence, we are interested to compare its performance with HM for real signals. For that purpose we select a $30ms$ frame from a reasonably stationary section of speech. The magnitude spectra computed by FFT and estimated using the classic harmonic representation as in [32] as well QHM are shown in Figure 2.7. Interestingly, the harmonics between 1.5kHz and 2kHz where the second formant takes place are greatly detuned and are missed by a purely harmonic model. By contrast, QHM provides a better spectral estimation. In terms of Signal-to-Reconstruction Error Ratio (SRER), the improvement is $3.9dB$. Moreover, the iterative scheme is not necessary in these case since the estimation of the fundamental frequency is accurate enough. However, when the estimation of fundamental frequency is inaccurate the iterative scheme is applied to correct the frequency estimation. Similarly, Figure 2.8 depicts the comparison between QHM and HM for a $30ms$ frame from a saxophone sound (i.e. musical signal). Again, QHM represents the sinusoidal components more accurately compared with HM. In terms of SRER the improvement is $3.9dB$.

Finally, these observations are consistent by testing more than 5 minutes of voiced speech from both male and female voices where the average SRER improvement is found to be $4.3dB$.

2.7 Capturing Chirp Signals: A variant of QHM

QHM, as SM, assumes that locally the analyzed frame is stationary. However, this is rarely the case. The frequencies as well the amplitudes are time-varying during the period of few ms (one frame duration) for natural sounds like speech. In order to remove the local stationarity assumption, a very common extension is to assume that frequencies are varying linearly over time, which means that one frame is modeled as a chirp signal. In this Section, we investigate the representation of chirp signals by an extension of QHM which is called chirp QHM (cQHM).

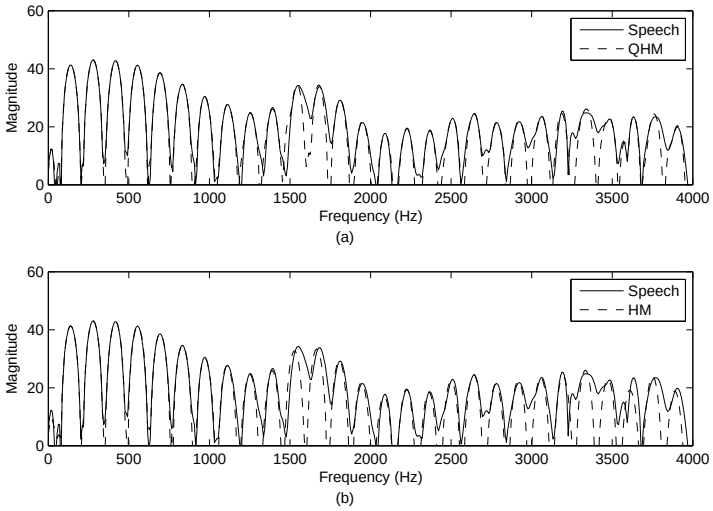

Figure 2.7: Upper panel: speech modeling using QHM. Lower panel: speech modeling using HM. The estimated fundamental frequency is $138.9Hz$.

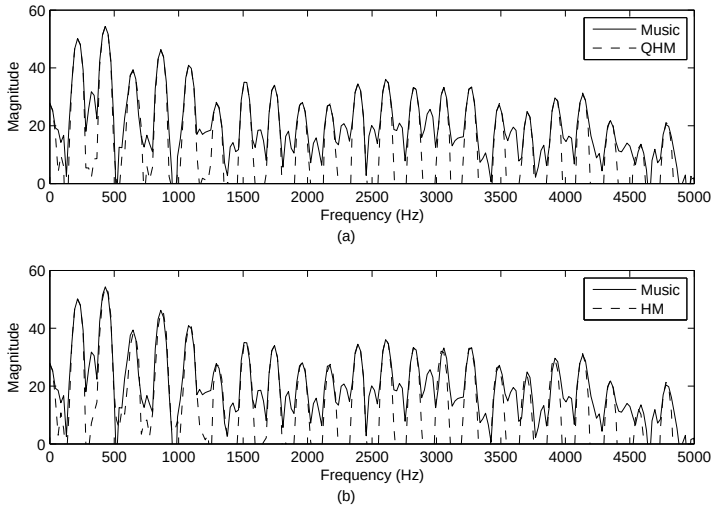

Figure 2.8: Upper panel: music modeling using QHM. Lower panel: music modeling using HM. The estimated fundamental frequency is $217.5Hz$.

To begin, the multi-component chirp signal is defined for a frame as

$$s(t) = \sum_{k=-K}^{K} A_k e^{j2\pi(\hat{f}_k t + \eta_{1,k} t + \eta_{2,k} t^2)} w(t), \qquad t \in [-T, T] \tag{2.37}$$

where K is the number of harmonics, $\hat{f}_k$ and A_k are the initially provided frequency and the complex amplitude of the kth component, respectively, while $\eta_{1,k}$ is, as in QHM, the frequency

mismatch (i.e. $f_k = \hat{f}_k + \eta_k$ is the true frequency) and $2\eta_{2,k}$ is the chirp rate of the kth component, respectively.

The estimation of the unknown parameters of the chirp signal in (2.37) is a highly nonlinear procedure. In order to obtain a linear estimation problem, a simple, yet powerful, technique is to approximate the signal in (2.37) by Taylor series expansion. Thus, for the kth component, the first order Taylor series approximation gives

$$s_k(t) \approx A_k[1 + j2\pi(\eta_{1,k}t + \eta_{2,k}t^2)]e^{j2\pi\hat{f}_k t}w(t), \qquad t \in [-T, T] \tag{2.38}$$

Motivated by the above approximation as well by QHM, we propose to model a frame of the original signal by a second order polynomial with complex coefficients given by

$$h_c(t) = \sum_{k=-K}^{K} (a_k + b_k t + c_k t^2)e^{j2\pi\hat{f}_k t}w(t), \qquad t \in [-T, T] \tag{2.39}$$

where, as before, K is the number of harmonics and $\hat{f}_k$ is the estimated frequency of the kth component, while $\{a_k, b_k, c_k\}_{k=-K}^{K}$ are complex coefficients which contain both amplitude and phase/frequency information.

The estimation of the complex unknown parameters $\{a_k, b_k, c_k\}_{k=-K}^{K}$ is performed again through linear LS. In matrix form, the solution is given by

$$\begin{bmatrix} \hat{\mathbf{a}} \\ \hat{\mathbf{b}} \\ \hat{\mathbf{c}} \end{bmatrix} = (E^H W^H W E)^{-1} E^H W^H W \mathbf{s} \tag{2.40}$$

where $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ and $\mathbf{s}$ are the vectors constructed from a_k, b_k, c_k and $s(t)$, respectively, while W is a diagonal matrix whose elements are the values of the analysis window function. Finally, $E = [E_0|E_1|E_2]$ where the elements of submatrices E_i for $i = 0, 1, 2$ are given by $(E_i)_{n,k} = (t_n)^i e^{j2\pi\hat{f}_k t_n}$. The reconstruction of the analyzed frame is then given by

$$\hat{s}(t) = \sum_{k=-K}^{K} (\hat{a}_k + \hat{b}_k t + \hat{c}_k t^2)e^{j2\pi\hat{f}_k t}w(t), \qquad t \in [-T, T] \tag{2.41}$$

Finally, please note that cQHM has $6K + 3$ unknown real parameters in the real signal case which means that cQHM requires larger analysis windows compared to QHM, for robust estimation of its parameters. Consequently, the computational load for cQHM parameter estimation is

about 8 times more compared with the corresponding QHM computational load. This may put a limitation on the use of cQHM for real-time applications. Nevertheless, cQHM combines both the linear evolution of frequency as well the frequency mismatch as it is shown in the following subsections.

2.7.1 Time-domain Properties

From (2.39), the instantaneous amplitude for the kth component is given by

$$M_k(t) = |a_k + tb_k + t^2 c_k| = \sqrt{(a_k^R + tb_k^R + t^2 c_k^R)^2 + (a_k^I + tb_k^I + t^2 c_k^I)^2} \tag{2.42}$$

while the instantaneous phase is computed as

$$\Phi_k(t) = 2\pi \hat{f}_k t + \angle(a_k + tb_k + t^2 c_k) = 2\pi \hat{f}_k t + atan \frac{a_k^I + tb_k^I + t^2 c_k^I}{a_k^R + tb_k^R + t^2 c_k^R} \tag{2.43}$$

Finally, instantaneous frequency which is the derivative of instantaneous phase over time is given by

$$F_k(t) = \frac{1}{2\pi}\Phi_k'(t) = \hat{f}_k + \frac{1}{2\pi}\frac{(a_k^R b_k^I - a_k^I b_k^R) + 2t(a_k^R c_k^I - a_k^I c_k^R) + t^2(b_k^R c_k^I - b_k^I c_k^R)}{M_k^2(t)} \tag{2.44}$$

Obviously, the instantaneous frequency of cQHM is richer compared to QHM. Figure 2.9 shows the instantaneous frequency of a chirp signal (line with circles) as well the instantaneous frequency of cQHM as it is computed by (2.44). Even though the analysis have been performed with constant frequency (solid line), the instantaneous frequency of cQHM is able to follow the original instantaneous frequency at least around the center of the analysis window.

2.7.2 Towards the target model

Following similar ideas as in Section 2.3, the decomposition of b_k and c_k into two components one collinear and one orthogonal to a_k yields

$$b_k = \rho_{1,k} a_k + \rho_{2,k} j a_k \tag{2.45}$$

and

$$c_k = \sigma_{1,k} a_k + \sigma_{2,k} j a_k, \tag{2.46}$$

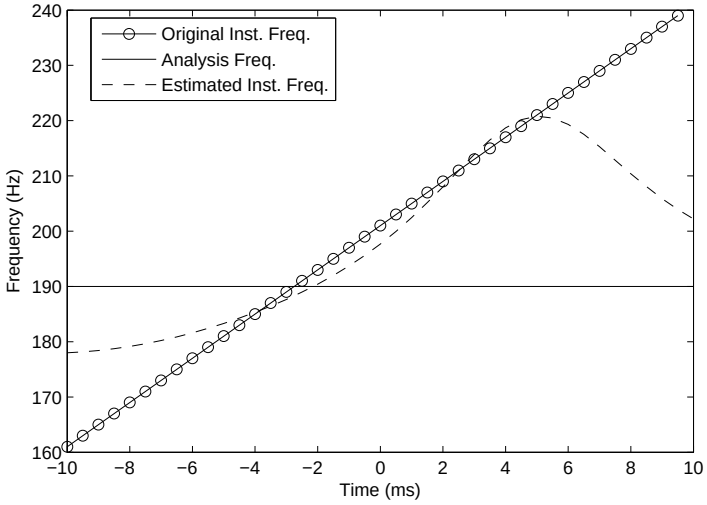

Figure 2.9: A frame of $20ms$ duration which contains a chirp sinusoid with instantaneous frequency $f_1(t) = 200 + 4000tHz$ (line with circles) is analyzed at $190Hz$ (solid line). Instantaneous frequency (dashed line) tries to adjust to the true instantaneous frequency of the sinusoid. Hamming window of $20ms$ duration was used.

where $\rho_{1,k}$, $\rho_{2,k}$, $\sigma_{1,k}$, and $\sigma_{2,k}$ are the projections of b_k and c_k onto a_k and ja_k, respectively. Mathematically, the projections are given by

$$\rho_{1,k} = \frac{a_k^R b_k^R + a_k^R b_k^R}{|a_k|^2} \qquad \text{and} \qquad \rho_{2,k} = \frac{a_k^R b_k^I - a_k^I b_k^R}{|a_k|^2} \tag{2.47}$$

while

$$\sigma_{1,k} = \frac{a_k^R c_k^R + a_k^R c_k^R}{|a_k|^2} \qquad \text{and} \qquad \sigma_{2,k} = \frac{a_k^R c_k^I - a_k^I c_k^R}{|a_k|^2} \tag{2.48}$$

With this notation, (2.39) can be rewritten as

$$h_c(t) = \sum_{k=-K}^{K} a_k \left[1 + (\rho_{1,k} + j\rho_{2,k})t + (\sigma_{1,k} + j\sigma_{2,k})t^2\right] e^{j2\pi \hat{f}_k t}. \tag{2.49}$$

Finally, from (2.38) and (2.49), an estimate of the kth frequency mismatch and kth chirp rate are obtained by

$$\hat{\eta}_{1,k} = \frac{\rho_{2,k}}{2\pi} \tag{2.50}$$

and

$$\hat{\eta}_{2,k} = \frac{\sigma_{2,k}}{2\pi} \tag{2.51}$$

Note that $\rho_{1,k}$ and $\sigma_{1,k}$ can be used for the estimation of the slope and higher-order quantities of the instantaneous amplitude.

2.7.3 Iterative Estimation

Once the instantaneous phase parameters $\{\hat{\eta}_{1,k}, \hat{\eta}_{2,k}\}$ of the frame have been computed using (2.50) and (2.51), they can be used to define a new model which takes into account the estimated parameters leading to a more accurate representation of the signal. Hence, we suggest an iterative procedure where, at each iteration, one frame is modeled as

$$h_{ic}(t) = \sum_{k=-K}^{K} (a_k + b_k t + c_k t^2) e^{j2\pi(\hat{f}_k t + \hat{\eta}_{1,k} t + \hat{\eta}_{2,k} t^2)} w(t), \qquad t \in [-T, T] \tag{2.52}$$

where a_k, b_k and c_k are again complex coefficients estimated by LS method. Technically, submatrices E_i has elements $(E_i)_{n,k} = (t_n)^i e^{j2\pi(\hat{f}_k t_n + \hat{\eta}_{1,k} t + \hat{\eta}_{2,k} t^2)}$. This procedure is repeated until convergence, i.e., once a criterion based on the evolution of the LS error or based on the relative chance of the estimated parameters is satisfied. Then, the reconstruction of the analyzed frame is provided by

$$\hat{s}(t) = \sum_{k=-K}^{K} (\hat{a}_k + \hat{b}_k t + \hat{c}_k t^2) e^{j2\pi(\hat{f}_k t + \hat{\eta}_{1,k} t + \hat{\eta}_{2,k} t^2)} w(t), \qquad t \in [-T, T] \tag{2.53}$$

Region of Convergence of iterative cQHM

The convergence of iterative cQHM is very difficult to analyze analytically because the model is changing at each iteration since the estimates of the previous iteration are used. Nevertheless, we explore numerically the estimation error of both frequency mismatch and chirp rate for a mono-component chirp signal and then some clues about the region of convergence (ROC) of iterative cQHM are provided.

Figure 2.10 shows the estimation error of frequency mismatch when cQHM and (2.50) is applied. Analysis window is a Hamming window of $16ms$ duration. Frequency mismatch takes values in the interval $[-400, 400]Hz$, while chirp rate takes values in the interval $[-100, 100]Hz/ms$. Have in mind that a chirp rate of $50Hz/ms$ means that in one millisecond, instantaneous frequency changes 50 Hertz. It is obvious from Figure 2.10 that the error is not a convex function, however, there are regions where the convexity holds and in these regions we expect the iterative cQHM to converge. Figure 2.11 shows with white color the region where the estimation error is

less than the initial frequency mismatch. In this region, iterative cQHM is expected to converge since the frequency mismatch is reduced.

Figures 2.12 and 2.13 shows the estimation error and the convergence region of chirp rate parameter, η_2, respectively. Now, the region of convergence (white color) is considerably smaller compared with the corresponding ROC for the frequency mismatch which limits the maximum allowed chirp rate value. Nevertheless, motivated by Figures 2.11 and 2.13, different iterative schemes may be suggested. For instance, one suggestion could be to iteratively reduce the frequency mismatch and then try to reduce iteratively the chirp rate error.

2.7.4 Application to speech

To test the performance of cQHM for multi-component cases, we apply cQHM on real signals. In Figure 2.14, one frame of a female voice is analyzed using cQHM. The sampling frequency of the signal is 16kHz and the number of harmonics is set to 15. In this example, after careful manual inspection of the evolution of the glottal cycle, it was observed that within the analysis window, the fundamental frequency approximately increases from 180Hz to 220Hz. It must be also pointed out that for speech signal the chirp rate is larger for higher harmonics. Actually, it is expected to be k times the chirp rate of the fundamental frequency. Consequently, there may be cases in which the Taylor approximation in (2.38) is not valid. This is noticeable in Figure 2.14 where some partials has opposite slope than the expected. To handle such cases, it is recommendable to use a single fan-chirp rate η_2 estimated from the chirp rates of the first K_0 components. We will refer to this as *restricted chirp rate estimation procedure*. As an estimate of the single chirp rate, η_2, is used a weighted average of the chirp rate of the first K_0 components given by

$$\hat{\eta}_2 = \frac{1}{K_0} \sum_{k=1}^{K_0} \hat{\eta}_{2,k}/k \tag{2.54}$$

Then, the iterative analysis is carried out using chirp rate $\hat{\eta}_{2,k} = k\hat{\eta}_2$ for the kth harmonic. Figure 2.15 shows the same frame analyzed with the restricted chirp rate estimation approach for each component. K_0 is set to 3. Now, the frequency evolution is consistent for each component which results in higher accuracy. Indeed, in this example, the SRER is improved $2dB$ when the single chirp rate parameter is used.

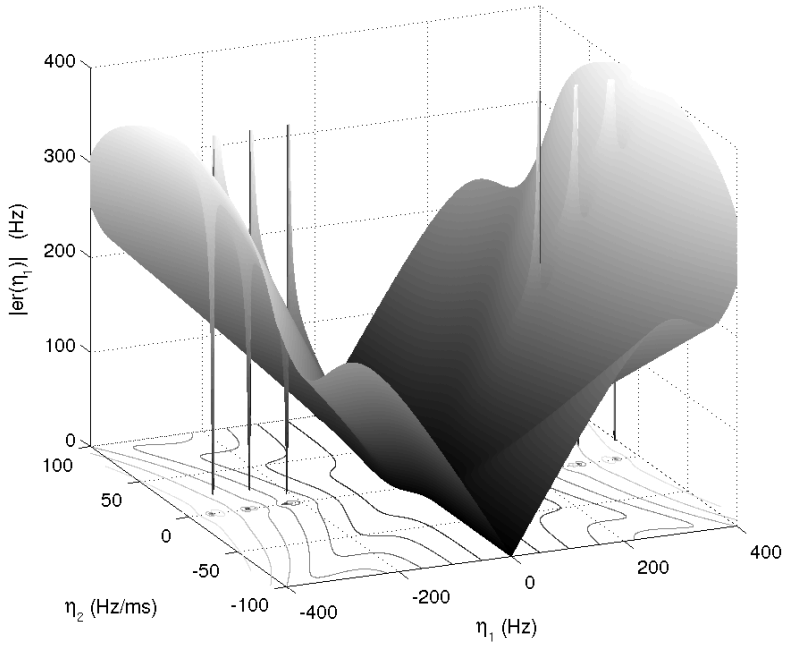

Figure 2.10: Absolute value of the frequency mismatch estimation error using cQHM. Please note that $er(\eta_1) = \eta_1 - \hat{\eta}_1$.

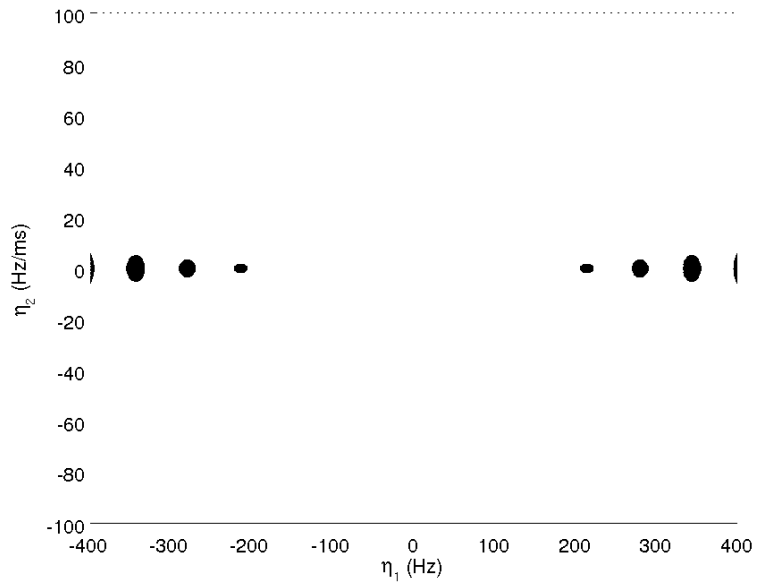

Figure 2.11: Region of convergence (white region) for the frequency mismatch using the iterative cQHM. It is worth noting that almost for any chirp signal the frequency mismatch will be corrected.

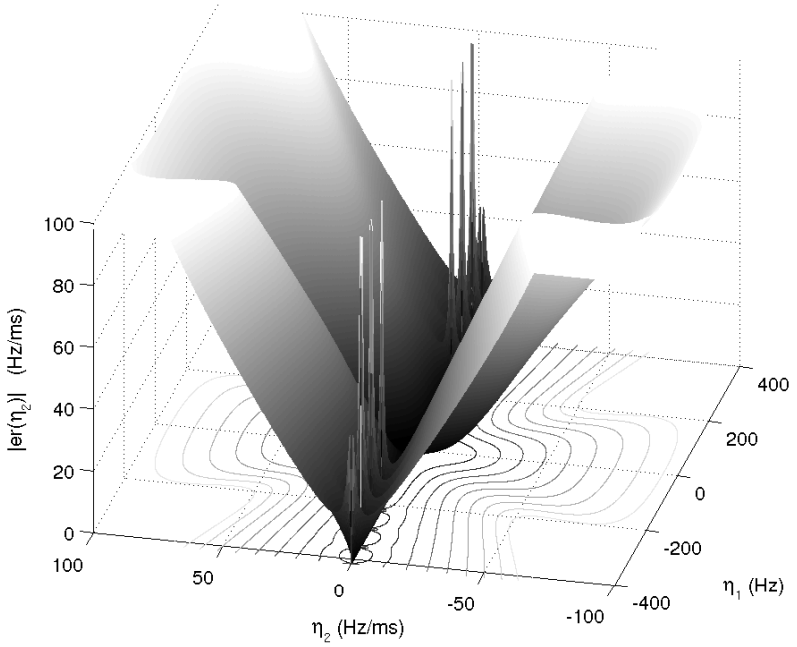

Figure 2.12: Absolute value of the chirp rate estimation error using cQHM. Please note that $er(\eta_2) = \eta_2 - \hat{\eta}_2$.

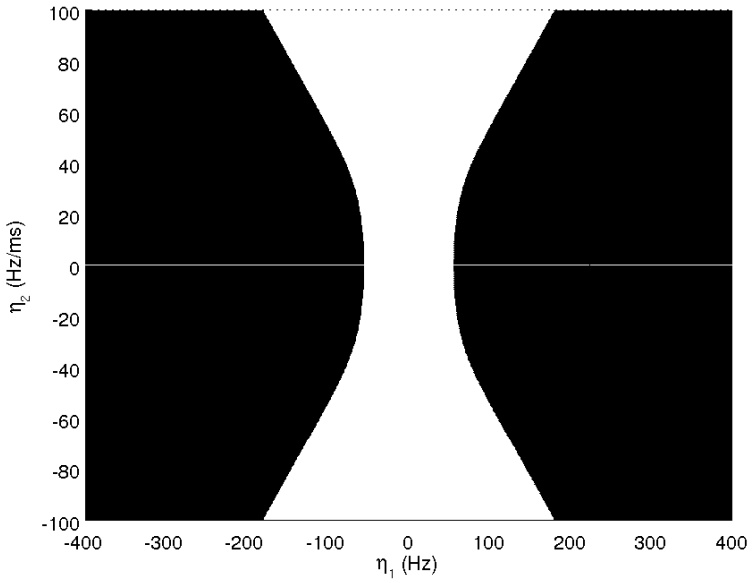

Figure 2.13: Region of convergence (white region) for the chirp rate using the iterative cQHM.

2.8 Conclusion

In this Chapter, we re-introduced a time-varying model which is referred to as QHM. The estimation of the unknown parameters was performed through linear LS. Then, the main properties

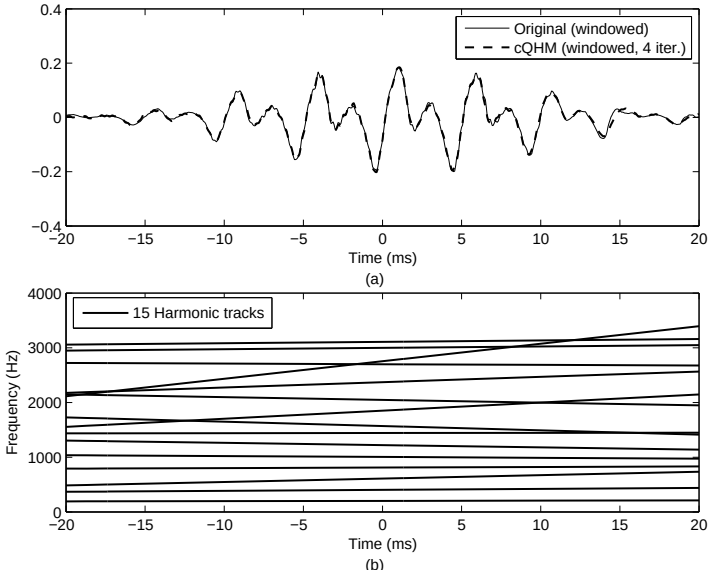

Figure 2.14: $40ms$ of female speech. Upper panel: Original (solid) and reconstructed (dashed) signals (SRER = $11.1dB$). Sinusoidal components may have arbitrary chirp rates. Lower panel: The estimated frequency evolution of the 15 first harmonics.

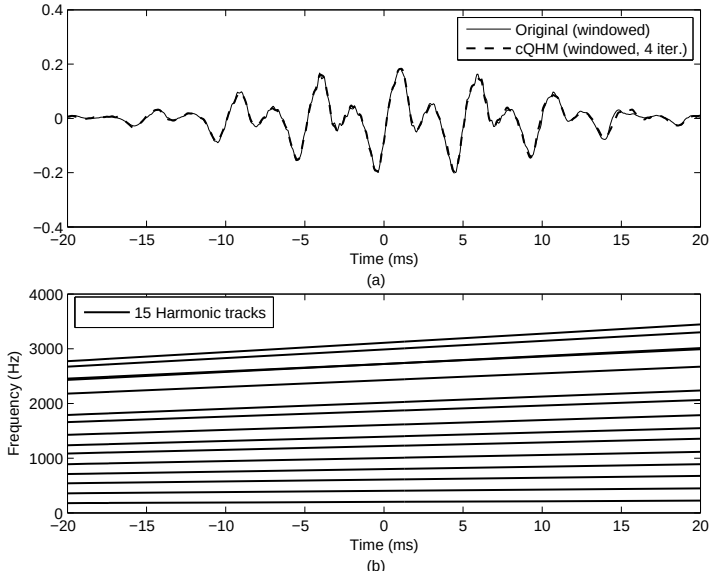

Figure 2.15: $40ms$ of female speech. Upper panel: Original (solid) and reconstructed (dashed) signals (SRER = $13.1dB$). Sinusoidal components have chirp rates which are integer multiples of a fundamental chirp rate. Lower panel: The estimated frequency evolution of the 15 first harmonics.

of QHM were presented. We showed that an important property of QHM is its ability to detect frequency mismatch errors and then correct them. Thus, an iterative algorithm (iQHM),

which efficiently estimates the sinusoidal parameters, was proposed. The region of convergence of the iterative algorithm was provided and the importance of the window type and duration was highlighted. Moreover, the robustness of QHM and iQHM under additive noisy conditions was demonstrated. Furthermore, QHM was tested on real signals such as voiced speech and music signals showing its superiority over HM. Finally, an extension of QHM, namely chirp QHM (cQHM), was presented which is able to model not only the frequency mismatch but also the linear evolution of the frequency. Iterative parameter estimation was also applied for this model in order to reduce the estimation error.

Chapter 3

Adaptive QHM

In previous Chapter, we showed that QHM is able, under certain conditions, to correct frequency mismatch errors efficiently when the analyzed frame is locally stationary. In practice however, even for frames of duration of few ms, the natural signals like speech are non-stationary. The use of cQHM, which is a more complex model than simple QHM or SM, is not satisfactory due to the fact that computation cost becomes very high as well the convergence of iterative cQHM is not guaranteed in the multi-component case. In this Chapter, we propose a non-parametric and adaptive model referred to as adaptive QHM (aQHM) which is able to model efficiently the non-stationarity of the analyzed frame. Furthermore, an algorithm for the decomposition of AM-FM signals based on aQHM is developed. Then, the AM-FM decomposition algorithm is tested on synthetic signals with highly non-stationary characteristics as well under noise. Finally, the application of aQHM to voiced speech reveals its superiority in terms of SRER against QHM and FFT-based SM.

3.1 Limitations of QHM

We showed that if the analyzed frame is a sum of stationary sinusoids, QHM is able to correct the frequency mismatch[1] between the initially provided frequencies and the true frequencies. Moreover, applying iQHM, we showed that both frequency and amplitude estimation errors approach the CRLB, which means that iQHM is a statistically efficient estimator of sinusoidal parameters. However, even locally the sinusoids have variations both in amplitude and in frequency. In Section 2.7, we further expanded QHM to cQHM in order to represent chirp signals, thus, we

[1] As presented in previous Chapter, the maximum allowed frequency mismatch depends on the bandwidth of the analysis window.

manage to model linear evolution of frequency within a frame with the additional cost of more parameters, larger analysis windows, and higher computational cost. As concerns the ability of QHM and cQHM in capturing amplitude modulations, QHM is able to model linear evolution of amplitude while cQHM is able to capture quadratic amplitude evolution. Nevertheless, the frequency estimation of non-stationary signals is by far more important and the performance of the models presented in the previous Chapter is not satisfactory. This is shown in the following example where a more complex mono-component signal is considered. One frame of the signal is given by

$$x(t) = (1 + \alpha_1 t + \alpha_2 t^2 + \alpha_3 t^3) e^{j2\pi(\hat{f}_1 t + \eta_1 t + \eta_2 t^2 + \eta_3 t^3)} w(t), \qquad t \in [-T, T] \tag{3.1}$$

where the amplitude coefficients, $\{\alpha_i\}_{i=1}^3$, as well as the phase coefficients, $\{\eta_i\}_{i=1}^3$, are real numbers. Again, $\hat{f}_1$ is an estimate of the signal frequency which is used for the computation of the unknown complex parameters of QHM (or cQHM). In order to test whether QHM is able to correctly estimate the frequency mismatch parameter, η_1, we resort to numerical computations and Monte Carlo simulations. Thus, each parameter in (3.1) takes values uniformly distributed on the intervals provided in Table 3.1. The analysis window is a Hamming window of duration

	min	max
α_1	$-2/T$	$2/T$
α_2	$-2/T^2$	$2/T^2$
α_3	$-2/T^3$	$2/T^3$
η_1	$-16/T$	$16/T$
η_2	$-2/T^2$	$2/T^2$
η_3	$-2/T^3$	$2/T^3$

Table 3.1: Intervals for each parameter in (3.1).

$2T = 16ms$. Note that the synthetic signal under consideration changes its characteristics very fast. For example, if all the coefficients in (3.1) are set to zero except for α_1, which is set to $\frac{1}{T}$, then the instantaneous amplitude starts from 0 at the beginning of the frame and ends (after 16ms) at the value of $2A$. Figure 3.1(a) depicts the estimation error of frequency mismatch for 10^5 Monte-Carlo runs. It can be seen that a reasonable estimate of frequency mismatch is obtained if the frequency mismatch is smaller than $100Hz$, which is less than in the stationary case ($125Hz$) that corresponds to the specific window type and length. Hence, the region of convergence of iQHM is smaller for non-stationary signals. More importantly, even for very low

frequency mismatch, a persistent error is present. This is why further updates of the frequencies (depicted in Figure 3.1(b)) provide only marginal refinements but do not systematically decrease the estimation error at each iteration as is the case for the mono-component stationary signal.

Similar results are obtained when cQHM is used instead of QHM. Figure 3.2(a) shows the estimation error of frequency mismatch, η_1, using cQHM. Even though the region of convergence using cQHM is almost doubled (actually, the maximum allowed frequency mismatch is $175Hz$), a pertinent error in the estimation of frequency mismatch remains, even if the iterative scheme is applied (Figure 3.2(b)). As a conclusion, neither QHM nor cQHM are able to model adequately the non-stationarity of the analyzed frame. Thus, a different approach should be utilized for highly non-stationary signals. In the following Section, an adaptive model, which extends QHM, is suggested. We show that the new model uses time-dependent frequency information at the estimation level and it is able to represent non-stationary signals with higher accuracy using the same number of parameters as in SM and using .

3.2 Definition of adaptive QHM, aQHM

In this Section, we suggest a different approach where the basis functions of the model are not restricted to be chirp or exponential functions but can adapt to the locally estimated instantaneous frequency/phase components. More specifically, one frame is projected in a space generated by time varying non-parametric sinusoidal basis functions. We will refer to this modeling approach as adaptive QHM (aQHM).

Lets assume for the moment that an estimate of the instantaneous components of the signal, $\{\hat{A}_k(t), \hat{f}_k(t), \hat{\phi}_k(t)\}_{k=-K}^{k}$, are given. Then, one frame, $s_l(t)$, of the signal centered at time instant t_l is modeled as[2]

$$h_a^l(t) = \sum_{k=-K_l}^{K_l} (a_k^l + t b_k^l) e^{j(\hat{\phi}_k(t+t_l) - \hat{\phi}_k(t_l))} w(t), \qquad t \in [-T_l, T_l] \tag{3.2}$$

where a_k^l and b_k^l are again complex numbers. The term b_k^l 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. Note also that the old phase value at t_l (i.e., $\hat{\phi}_k(t_l)$) is subtracted from the instantaneous phase, so as the argument of the basis

[2]The frame indexing is necessary for aQHM, so, it reappears here.

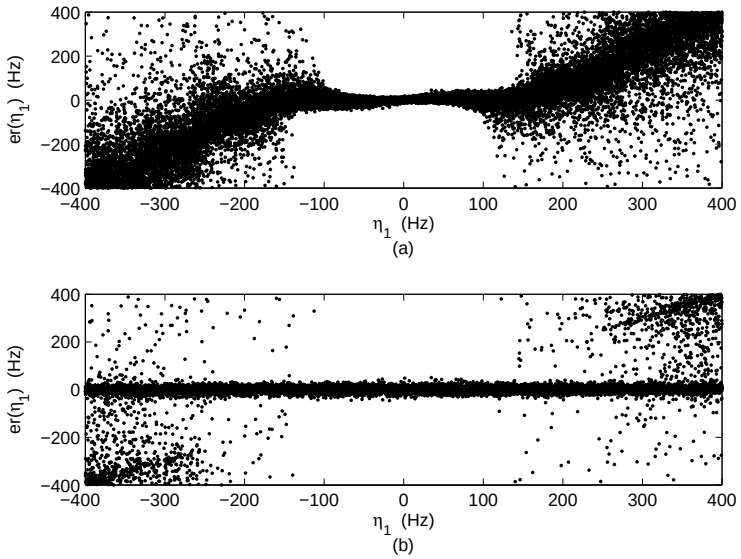

Figure 3.1: Upper panel: The estimation error of η_1 using QHM and a Hamming window of 16ms length, after 10^5 Monte-Carlo simulations of (3.1). Lower panel: Same as above, but with two iterations for the estimation of η_1.

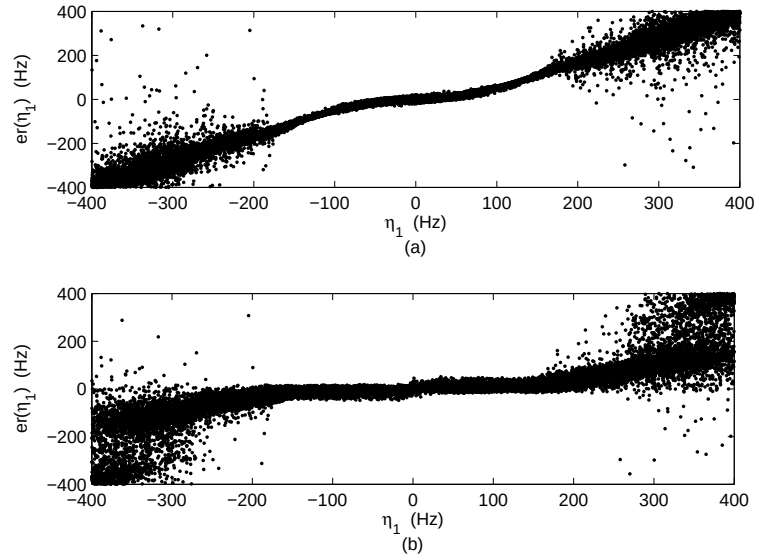

Figure 3.2: Upper panel: The estimation error of η_1 using cQHM and a Hamming window of 16ms length, after 10^5 Monte-Carlo simulations of (3.1). Lower panel: Same as above, but with two iterations for the estimation of η_1.

function has zero value at the center of the analysis. Thus, a new phase estimate at time-instant t_l is obtained from the argument of a_k^l.

The estimation of the unknown parameters of aQHM is similar to the QHM's parameter

estimation. The mean squared error between the signal and the model is minimized. The solution is straightforward and it is provided by

$$\begin{bmatrix} \hat{\mathbf{a}} \\ \hat{\mathbf{b}} \end{bmatrix} = (E^H W^H W E)^{-1} E^H W^H W \mathbf{s} \tag{3.3}$$

where now submatrices $E_i,\ i = 0, 1$ of matrix $E = [E_0|E_1]$ have elements given by $(E_0)_{n,k} = e^{j2\pi(\hat{\phi}_k(t_n+t_l)-\hat{\phi}_k(t_l))}$ and $(E_1)_{n,k} = t_n e^{j2\pi(\hat{\phi}_k(t_n+t_l)-\hat{\phi}_k(t_l))} = t_n(E_0)_{n,k}$. Unfortunately, most of the improvements for the computation of the above linear equation system presented in the Appendix A are not applicable to aQHM because the basis functions are non-parametric. The only applicable improvement is to diagonalize the submatrices of $E^H W^H W E$ which results in speed-ups of its construction and its inversion. The reconstruction of the frame is given by

$$\hat{s}_l(t) = \sum_{k=-K_l}^{K_l} (\hat{a}_k^l + t\hat{b}_k^l) e^{j(\hat{\phi}_k(t+t_l)-\hat{\phi}_k(t_l))} w(t), \qquad t \in [-T_l, T_l] \tag{3.4}$$

3.2.1 Difference between aQHM and QHM or cQHM

In contrast to QHM or cQHM, where the argument of the basis functions is parametric and/or stationary, in aQHM the argument of the basis functions is non-parametric neither necessarily stationary. Moreover, since the aQHM basis functions use the instantaneous phases which have been estimated from the input signal, these are also adaptive to the current characteristics of the signal. In other words, they are adaptive to the analyzed signal. This is depicted in Figure 3.3 where the original instantaneous frequency (line with circles) is shown for the frame centered at time t_l, along with the frequency track used by QHM (solid line) as well the frequency track used by aQHM (dashed line). It is obvious from Figure 3.3 that aQHM will produce less error in the estimation of a_k^l and b_k^l compared to QHM because the signal is projected to basis functions which are closer to the original instantaneous frequency.

Indeed, an interpretation of LS estimation method is that the instantaneous frequency (or more correctly instantaneous phase) of the basis function is subtracted from the original instantaneous frequency and then an averaging is performed in order to provide the estimates of the unknown parameters. Looking at Figure 3.3 the difference between the original and the instantaneous frequency of the models reveals that it is smaller for aQHM rather than for QHM. Finally, note that aQHM needs an initial estimation for the instantaneous phase. This is provided by QHM which acts as a frequency tracker as it will be shown next.

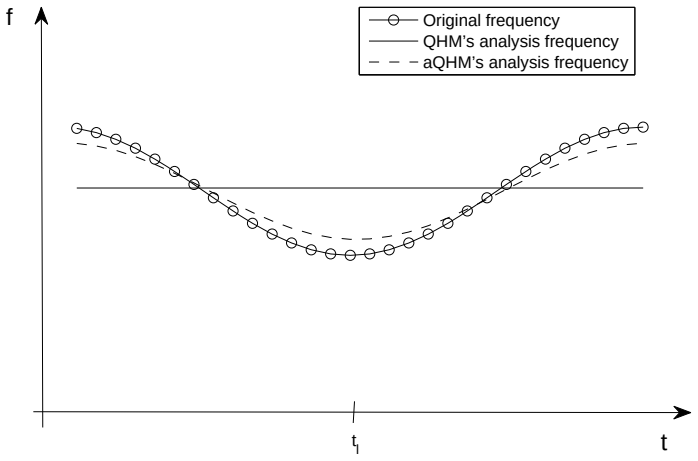

Figure 3.3: QHM vs aQHM. The instantaneous frequency of the mono-component signal (line with circles) is assumed to be constant for QHM (solid line) while aQHM (dashed line) does not make any assumption about the shape of instantaneous frequency.

3.2.2 Initialization of aQHM

In order to apply aQHM, we actually need an estimate for the instantaneous phase for each sinusoidal component. Any algorithm that produces such an estimate can be used as an initialization for aQHM. In this thesis, we suggest initial estimate of the instantaneous phase (and amplitude/frequency) to be provided by QHM. Since QHM is able to correct small frequency mismatch errors, it can also be used as a frequency tracker. Indeed, assuming that the analysis is moved from time-instant t_{l-1} to time-instant t_l, the estimated frequency at time-instant t_{l-1} can be used as an initial estimate of the frequency for QHM. Thus, let $\hat{f}_k(t_l)$, $\hat{A}_k(t_l)$, and $\hat{\phi}_k(t_l)$, denote the frequency, the corresponding amplitude and phase at time-instant t_l (center of analysis window) of the kth component, with $l = 1, \ldots, L$ where L be the number of frames. These parameters are estimated using QHM as

$$\hat{f}_k(t_l) = \hat{f}_k(t_{l-1}) + \frac{\rho^l_{2,k}}{2\pi} \tag{3.5a}$$

$$\hat{A}_k(t_l) = |a^l_k| \tag{3.5b}$$

$$\hat{\phi}_k(t_l) = \angle a^l_k \tag{3.5c}$$

Considering now all the estimations made at t_l, with $l = 1, \cdots, L$, we may construct the corresponding time series for the instantaneous amplitude, frequency and phase, for each of the

components of the signal. Then, we should consider the effect of the step size, or otherwise the distance between the centers of the analysis frames, t_l, in the performance of aQHM.

Effect and importance of step size

For applications in speech analysis such as voice function assessment (i.e., voice disorders, analysis of vocal tremor) or voice modification, one sample time-step is accepted. In this case, the instantaneous values of frequency, amplitude, and phase are just provided, since at each sample an estimate of these parameters are computed. In other applications however, such as speech synthesis, larger time-steps are required. In SM, between two consecutive synthesis instants, linear interpolation for the amplitudes and cubic interpolation for the phases were suggested [1]. In aQHM, many interpolation schemes can be considered for the estimation of the intermediate samples. For instantaneous amplitude we suggest linear interpolation because it guarantees that the instantaneous amplitude will be always positive which is a necessary condition for the well-posiness of the instantaneous amplitude. Cubic or spline interpolation unfortunately do not guarantee positiveness of the instantaneous amplitude. For the instantaneous frequency, we suggest using spline interpolation because it provides smooth estimates of the frequency trajectories (which is considered to be representative of the typical voiced speech; in other types of sounds different approaches may be applied). However, such simple solutions are not possible for the interpolation of instantaneous phase. For this purpose, we will describe in the following a non-parametric approach based on the integration of instantaneous frequency, as an alternative to the cubic phase interpolation method suggested in [1].

Phase interpolation

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 of the kth component can be computed as

$$\check{\phi}_k(t) = \hat{\phi}_k(t_{l-1}) + \int_{t_{l-1}}^{t} 2\pi \hat{f}_k(u)du \tag{3.6}$$

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 (3.6) in order to guarantee phase continuation over

frame boundaries as following

$$\hat{\phi}_k(t) = \hat{\phi}_k(t_{l-1}) + \int_{t_{l-1}}^{t} 2\pi \hat{f}_k(u) + r_k^l \sin\left(\frac{\pi(u - t_{l-1})}{t_l - t_{l-1}}\right) du \tag{3.7}$$

In (3.7), 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). Please note that the instantaneous frequency is re-estimated as the derivative of the modified instantaneous phase with respect to time. Moreover, it can be easily shown that the instantaneous phase of (3.7) at t_l will be equal to $\hat{\phi}_k(t_l) + 2\pi M$ if r_k^l is selected to be

$$r_k^l = \frac{\pi(\hat{\phi}_k(t_l) + 2\pi M - \check{\phi}_k(t_l))}{2(t_l - t_{l-1})} \tag{3.8}$$

where M is computed as before. Moreover, r_k^l is not just a correction factor and it can be thought as a measure of how valid is the assumption that the analyzed signal is a superposition of time-varying sinusoids. To be more specific, r_k^l is used for correcting small errors due to discretization of the instantaneous components as well estimation errors of frequency or phase. Whenever the signal is indeed an AM-FM signal, then the correction factor should be small. On the other hand, when the signal is not an AM-FM signal and contains wide-band information, then the correction factor should be high.

In Figure 3.4, (3.6) and (3.7) are compared on a synthetic example. The signal was analyzed frame-by-frame using QHM at time-instants $t_1, ..., t_l, ..., t_L$ with time-step $4ms$ and estimates of the instantaneous components at these instants are obtained. Figure 3.4(a) shows the true instantaneous frequency contour (dashed line) of the AM-FM signal along with the estimated instantaneous frequency which is computed as the derivative of the instantaneous phase computed by (3.6). Figure 3.4(b) shows the same but, now, (3.7) has been used for the instantaneous phase computation. 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 these spikes.

3.3 AM-FM decomposition algorithm

Summarizing, aQHM suggests a non-parametric AM-FM decomposition algorithm which proceeds by successive adaptations of the basis functions of the model to the characteristics of the underlying sine-waves of the input signal. Initial estimate of the instantaneous phase necessary for aQHM is provided by QHM. A pseudo-code of the algorithm is presented below.

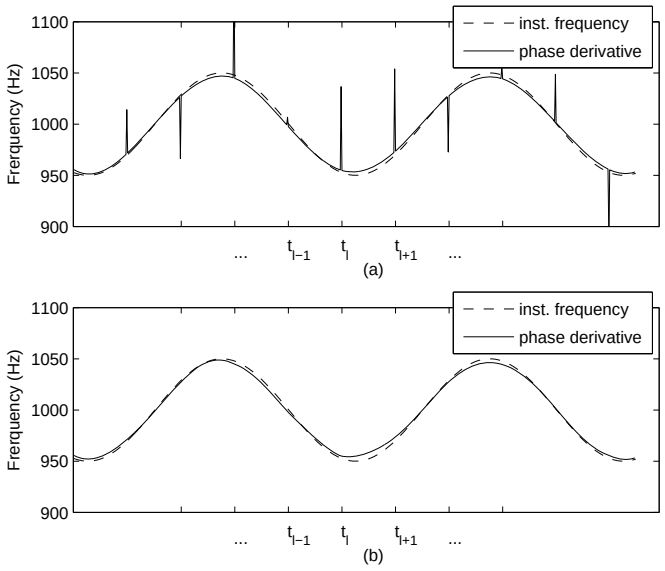

Figure 3.4: Actual instantaneous frequency (dashed line) and estimated instantaneous frequency (solid line) as the derivative of the instantaneous phase computed from (3.6) (upper panel) and (3.7) (lower panel).

Adaptive AM-FM Decomposition Algorithm

---

1. Initialization step:

   Provide initial frequency estimate $f_k^0(t_1)$

   For $l = 1, 2, ..., L$

   (a) Compute a_k^l, b_k^l using $f_k^0(t_l)$ as initial frequency estimates in (2.6)

   (b) Update $\hat{f}_k^0(t_l)$ using (3.5a) and (2.19)

   (c) Compute $\hat{A}_k^0(t_l)$ and $\hat{\phi}_k^0(t_l)$ using (3.5b) and (3.5c), 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. Adaptation step:

   For $i = 1, 2, ...$

   For $l = 1, 2, ..., L$

(a) Compute a_k^l, b_k^l using $\hat{\phi}_k^{i-1}(t)$ in (3.2)

(b) Update $\hat{f}_k^i(t_l)$ using (3.5a) and (2.19)

(c) Compute $\hat{A}_k^i(t_l)$ and $\hat{\phi}_k^i(t_l)$ using (3.5b) and (3.5c), respectively

end

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

end

The aQHM-based AM-FM decomposition algorithm is intuitively simple, and, as concerns its complexity, the most time-consuming part is the computation of a_k^l and b_k^l via LS at each time-step. For comparison purposes, when there is only one component, the complexity of each time-step is $O(N)$ where $2N+1$ is the duration of the analysis window. This order of complexity is comparable to AM-FM decomposition algorithms with very low complexity such as the DESA algorithm [56]. For multi-component signals, the complexity of each step is $O((N+K)K^2)$ where K is the number of components. Please note also that the window duration may be frame-dependent and usually it depends on the smallest frequency.

The signal is reconstructed by summing the time-varying components, i.e.

$$\hat{s}(t) = \sum_{k=-K}^{K} \hat{A}_k(t) e^{j\hat{\phi}_k(t)} \tag{3.9}$$

An objective measure on how close the reconstructed signal is to the original signal is given by SRER at it is defined by (2.10). Now, SRER measures the overall performance of the AM-FM decomposition algorithm, hence, it is considered as a global measure[3]. Thus, the adaptation step can be iterated until changes in the SRER are not significant. As we will show, the number of adaptations depends on the amount of non-stationarity of the signal.

3.4 Validation on Synthetic Signals

In this Section, the performance of the suggested adaptive AM-FM decomposition algorithm 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

[3]Up to now, SRER measured the modeling error for one frame, hence, it was considered as a local measure. Nevertheless, SRER is able to measure the total performance of a method/model.

the robustness of the proposed algorithm. For all synthetic examples, please note that we consider a sampling frequency of $f_s = 8000Hz$ and the time-step will be fixed to one sample ($t_l - t_{l-1} = 1$). Thus, interpolation of the instantaneous components is not necessary.

For comparison purposes, we suggest comparing the AM-FM decomposition algorithm which we denote it by aQHM with QHM (i.e. only the initialization step of the AM-FM decomposition algorithm) and the estimation procedure used in the sinusoidal modeling of [1]. 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.

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. MAE for a time-varying parameter $\theta(t)$ with support in $[0, T]$ is defined as

$$MAE(\theta) = \frac{1}{M} \sum_{i=1}^{M} \int_0^T |\theta(t) - \hat{\theta}^{(i)}(t)| dt \tag{3.10}$$

where M is the number of simulations while $\hat{\theta}^{(i)}(t)$ is the estimated time-varying parameter at the ith simulation.

3.4.1 Mono-component AM-FM signal

Firstly, let us consider the following mono-component chirp signal with a 2nd order amplitude modulation given by

$$x(t) = (11 - 340t + 4000t^2) e^{j2\pi(100t + 19500t^2)}, \qquad t \in [0, 0.1] \tag{3.11}$$

whose instantaneous frequency is $f_1(t) = 100 + 39000t$ (Hz). Note that the chirp rate is significant and starting from $100Hz$, it reaches $4000Hz$ in $0.1s$, which is the maximum allowed frequency since the sampling frequency is $8000Hz$. Figure 3.5(a) shows the real part of the chirp signal while Figure 3.5(b) shows its spectrogram.

Based on the analysis presented before (Section 2.5), 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 use a $8ms$ ($T = 4ms$)

Hamming window, so the squared window has bandwidth $B = 3/T = 750Hz$. Therefore the maximum frequency mismatch is $\pm 250Hz$. The center of the first analysis window is located at $4ms$, where the actual instantaneous frequency is $256Hz$. We set the initial frequency estimate to $200Hz$ which means that there is frequency mismatch of $56Hz$. The upper plots of Figure 3.6 show the original (line with circles) and the estimated using aQHM (bold dashed line) instantaneous components of the chirp signal. Three adaptation passes are enough for aQHM to converge. The lower plots of Figure 3.6 show the estimation error of the instantaneous components not only for aQHM (dashed line) but also for SM (solid line) and QHM (dotted line). The performance of QHM and SM is similar which is expected since both methods use stationary basis functions. On the other hand, aQHM adapts to the characteristics of the analyzed signal, thus, its estimation error is greatly reduced.

We now consider the case of complex additive white Gaussian noise of $30dB$ and $10dB$ local SNR[4]. Then, the average performance of each algorithm was measured based on 10^4 simulations of noise realization. Table 3.2 reports the MAE between the estimated and the actual AM and FM component, for QHM, aQHM, and SM. Please note that two or three 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 non-stationary 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 non-stationarity characteristics of the input signal while the other stems from the additive noise. The former source seems to be more important for the case of QHM and SM, while the latter affects more aQHM. However, even for $10dB$ local SNR, aQHM is more than 200% and 60% better than SM (in terms of MAE) in estimating the AM and FM components, respectively. Finally, in the case of additive noise, the reported SRER suggests that aQHM is not an overdetermined method (i.e. aQHM does not model the noise) since SRER is approximately equal to the local SNR.

[4]By local SNR, we mean that SNR is constant at any time instant, i.e. SNR is independent from the instantaneous amplitude of the analyzed signal. This is achieved by multiplying the additive noise with the instantaneous amplitude of the signal.

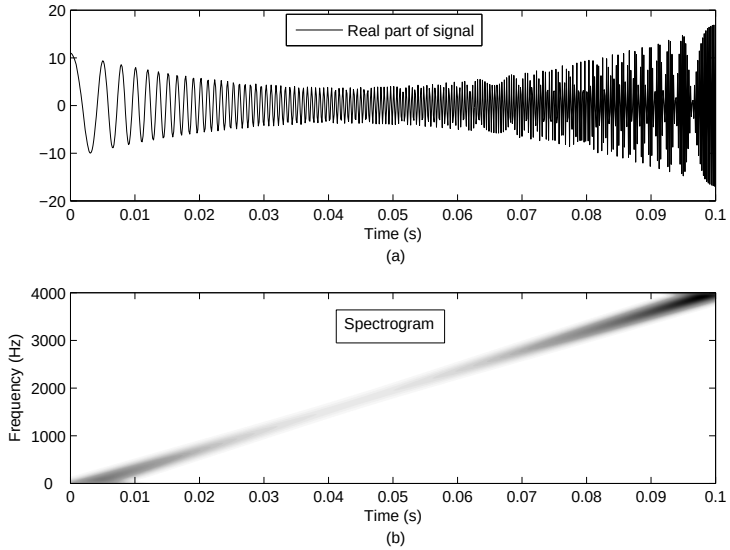

Figure 3.5: Upper panel: The real part of the mono-component AM-FM signal. Lower panel: Its STFT with squared Hamming window of $8ms$ as analysis window and the time-step is set to 1 sample.

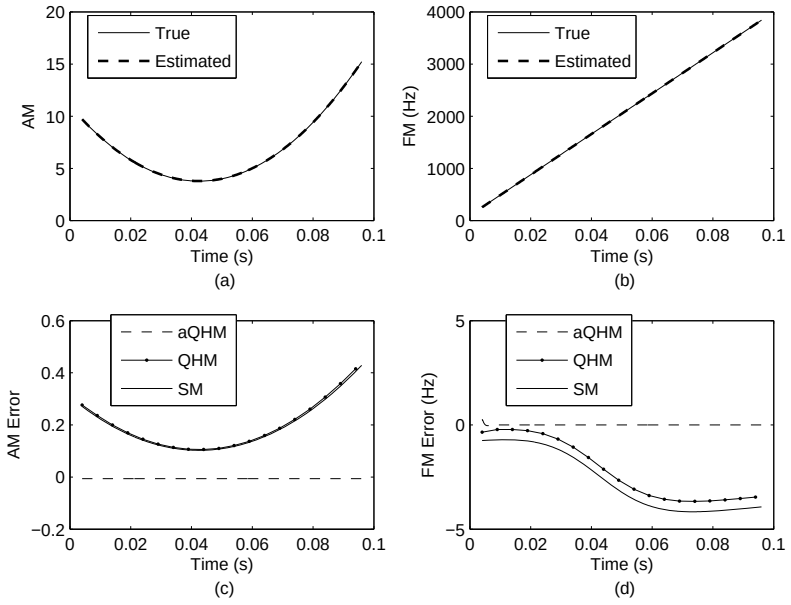

Figure 3.6: Upper panels: The true and the estimated by aQHM instantaneous components. Lower panels: The error between the true and the estimated components by aQHM (dashed line), by SM (solid line) and QHM (dotted line). Note that the estimation error for aQHM is mainly zero for both AM and FM components.

SNR	Method	AM	FM (Hz)	SRER (dB)
∞dB	QHM	0.19	2.21	15.0
	aQHM	0.006	0.002	61.8
	SM	0.19	2.69	13.1
$30dB$	QHM	0.20	2.30	14.8
	aQHM	0.02	0.66	29.5
	SM	0.19	2.73	13.0
$10dB$	QHM	0.23	5.04	10.1
	aQHM	0.22	6.23	10.1
	SM	0.23	5.23	8.2

Table 3.2: MAE of AM and FM components for QHM, aQHM and SM without noise, and with complex additive white Gaussian noise at $30dB$ and $10dB$ local SNR. SRER is also reported.

3.4.2 Two-component AM-FM signal

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

$$\begin{aligned} s(t) &= 2(1+0.4cos(2\pi 30t))e^{j(2\pi 700t+cos(2\pi 130t))} \\ &+ 2(1+0.3cos(2\pi 50t))e^{j(2\pi 1000t+cos(2\pi 130t))} \end{aligned} \tag{3.12}$$

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 high: 130 cycles per second. Figure 3.7(a) shows the real part of the two-component AM-FM signal while Figure 3.7(a) shows its spectrogram. It is worth-noting that the two components cannot be distinguished from the spectrogram.

For the proposed AM-FM decomposition algorithm, a Hamming window of length $16ms$ ($T = 8ms$) is used. In case of QHM, an initial frequency mismatch of $32Hz$ is assumed for both components, which is below the maximum allowable mismatch (namely $B/3 = 125Hz$ in this example). The performance of the proposed AM-FM decomposition algorithm is shown in Figure 3.8 where the original (solid line) as well the estimated by aQHM (bold dashed line) instantaneous components after 14 adaptations are presented. Figure 3.9 shows the modeling error of each instantaneous component estimated by SM (solid line), by QHM (dotted line) and by aQHM (dashed line). Even though the estimation error of aQHM is not fully eliminated, it is greatly reduced compared to the estimation error of the other two methods.

Moreover, the performance of the algorithms is tested with complex additive white Gaussian

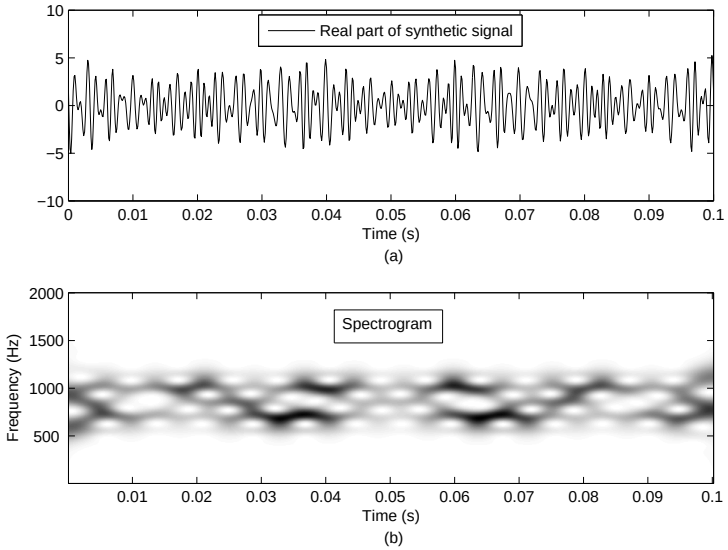

Figure 3.7: Upper panel: The real part of the two-component AM-FM signal. Lower panel: Its STFT with squared Hamming window of 16ms as analysis window and the time-step is set to 1 sample. It is noteworthy that the two components are not well-separated.

noise of $10dB$ local SNR. As previously, in case of additive noise, the average performance of each algorithm was measured based on 10^4 simulations of noise realization. In Table 3.3, the performance of QHM, aQHM, and SM is shown in terms of MAE as well in terms of SRER. Indeed, aQHM is about 500% better than SM or QHM in terms of MAE for the noiseless case and more than 300% for the 10dB noise level. 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 non-stationary signal. Specifically, in $16ms$, about 2 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 in order to reduce MAE for each of these components. Indeed, aQHM required 14 adaptations to converge (meaning that no significant changes in SRER were observed) in case of clean data while 8 adaptations are required in case of additive noise.

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 due to the interference between the components. Also, aQHM outperforms both QHM and SM for all the parameters and under all conditions. Furthermore, in contrast to the mono-component case, aQHM is not so sensitive to the additive noise. In this case, the source of the estimation error due to the highly non-stationary

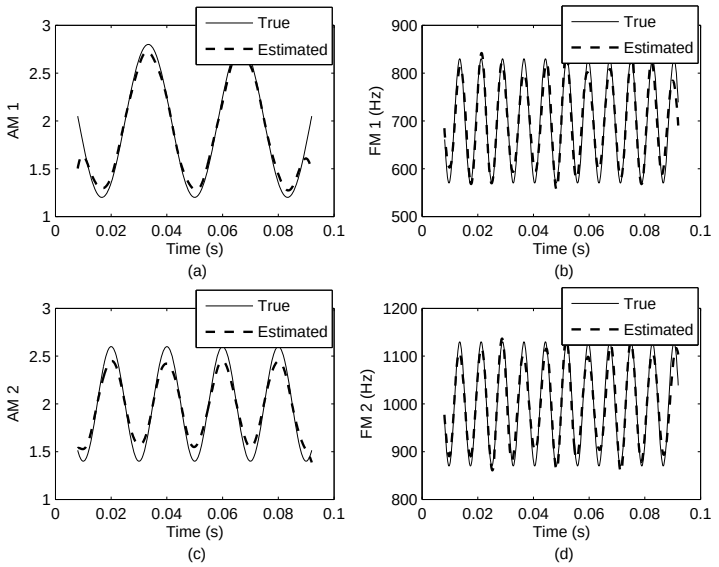

Figure 3.8: Upper panels: The true and the estimated by aQHM instantaneous amplitude and frequency for the first AM-FM component. Lower panels: The same but for the second AM-FM component.

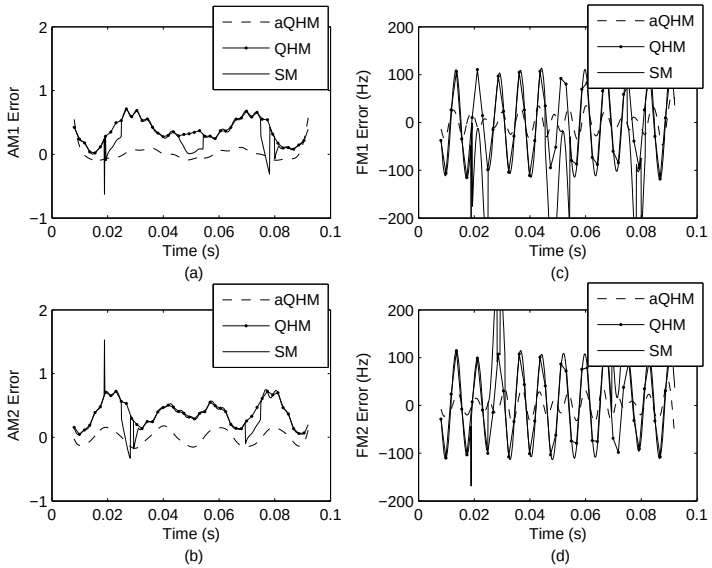

Figure 3.9: Upper panels: The error between the true and the estimated by SM (solid line), by QHM (dotted line) and by aQHM (dashed line) instantaneous amplitude and frequency for the first AM-FM component. Lower panels: The same but for the second AM-FM component.

character of the input signal is more important than the corresponding error source due to the presence of noise. Therefore, decreasing the SNR, does not significantly affect the performance of aQHM.

SNR	Method	AM1	AM2	FM1 (Hz)	FM2 (Hz)	SRER (dB)
∞dB	QHM	0.36	0.38	70.02	69.81	5.4
	aQHM	0.07	0.10	17.11	16.54	20.1
	SM	0.30	0.34	88.02	78.84	5.4
$10dB$	*QHM*	0.36	0.08	71.02	69.81	4.6
	aQHM	0.10	0.12	22.80	20.55	10.0
	SM	0.31	0.34	88.39	79.46	4.2

Table 3.3: 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 $10dB$ local SNR.

3.5 Application to Voiced Speech

The suggested adaptive 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. The only modification in the previously presented the AM-FM decomposition algorithm in order to work, is that instead of tracking each frequency, a fundamental frequency is tracked and then the analysis frequencies for QHM are provided as integer multiples of the estimated fundamental frequency, i.e. $f_k^0(t_l) = kf_0(t_l)$ for each k. The reason is that voiced speech could be highly non-stationary and sinusoidal components are born or die making the tracking of each frequency extremely difficult while fundamental frequency is a quantity which is always present in voiced speech. Thus, providing just the fundamental frequency for the first frame of the voiced segment and the number of components, the whole voiced segment is analyzed by the suggested AM-FM decomposition algorithm. It is worth noting that the accuracy of the fundamental frequency estimator is not crucial for QHM, since frequency mismatches are easily corrected (of course, we exclude cases of fundamental frequency doubling or halving).

In this Section, we compare aQHM with QHM and SM in terms of SRER for 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.

In Figure 3.10(a), a segment from a voiced speech signal generated by a male speaker is shown

(sampling frequency 16kHz). The analysis was performed using a Hamming window of $24ms$ and with *one sample as time-step*. For QHM, we set $f_0(t_1) = 140Hz$ (the average fundamental frequency of the segment) and $K = 40$. The results from QHM were used as an initialization for aQHM, where three adaptations were 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 (3.9). The reconstruction error for each method is depicted in Figure 3.10(b), (c) and (d), for QHM, aQHM, and SM, respectively. Again, aQHM provides the best reconstruction compared to the other two alternatives even if only one iteration is applied. The SRER is $19.5dB$ for SM, $24.1dB$ for QHM, and $30.5dB$ for aQHM.

3.5.1 Large-scale Objective Test

In case the time-step is bigger than one sample, then the instantaneous amplitudes and phases should be computed from the estimated parameters at the analysis time-instances. The instantaneous phase of QHM and aQHM is computed from (3.7). For SM, instantaneous amplitude is computed with linear interpolation while for the instantaneous phase, cubic interpolation is used [1]. Using three different step sizes, namely $1ms$, $2ms$, and $4ms$, we analyze 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 of the speech signals is $16000Hz$. Assuming an average pitch of $100Hz$ and $160Hz$ for male and female speakers, respectively, we use Hamming windows of fixed length; 2.5 times the average pitch period. Thus, we used a fixed length analysis window: $25ms$ for male and $15ms$ for female speakers. The same windows is used for all the algorithms. The number of components is set to $K = 40$ for male voices and to $K = 30$ for female voices. The average and standard deviation of the SRER (in dB) is provided in Table 3.4 along with various time-steps. Table 3.4 also presents the mean number of adaptations (NoA) needed for aQHM to converge. Since only aQHM suggests an adaptive algorithm, this column of the table is considered only for aQHM.

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, time-step is a crucial parameter in QHM and in aQHM. Results show that there is a minor decrease in the performance of these two algorithms when the time-step is increasing.

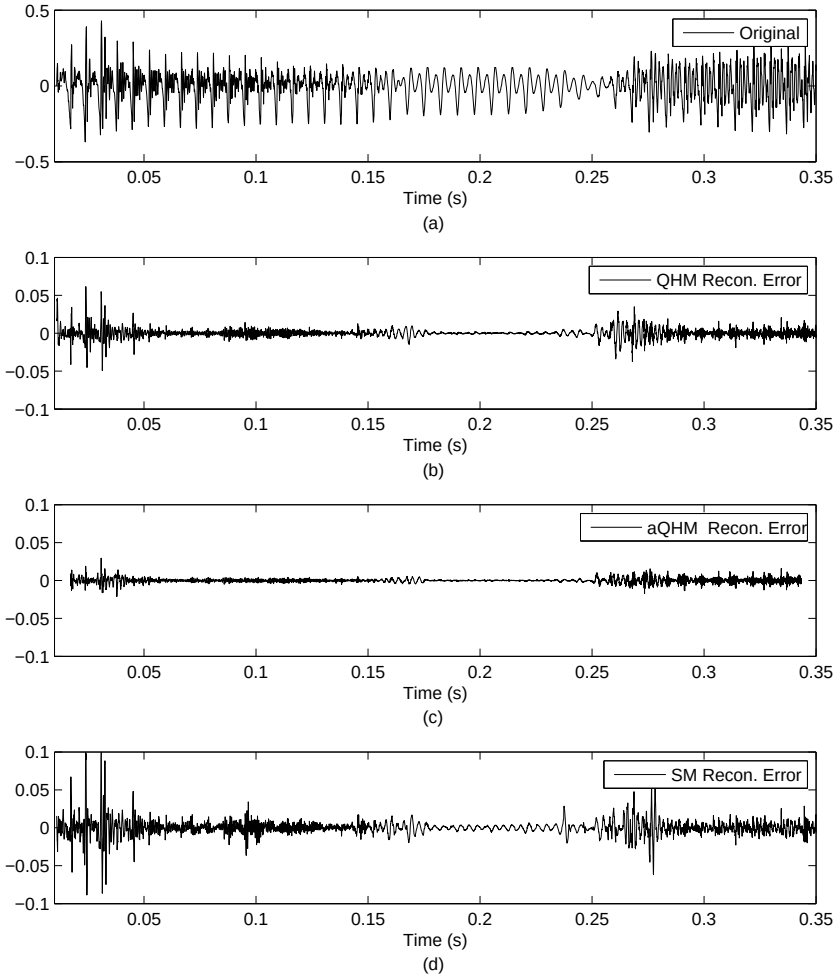

Figure 3.10: (a) Original speech signal and reconstruction error for (b) QHM, (c) aQHM after three adaptations, and (d) SM, using $K = 40$ components. Obviously, aQHM has the smallest reconstruction error.

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.

		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	–

Table 3.4: Mean and Standard Deviation of SRER (in dB) for approximately 200 minutes of voiced speech from TIMIT.

3.6 Conclusion

In this Chapter, we showed that QHM (or cQHM) is not appropriate for modeling highly non-stationary signals. Thus, we proposed an extension of QHM, which is referred to as adaptive QHM (aQHM), for the modeling of locally non-stationary signals. In this case, the basis functions of the model are non-parametric and they are able to adjust to the time-varying characteristics of the signal.

Moreover, an AM-FM decomposition algorithm based on aQHM was developed. Since aQHM requires an initial estimate of the instantaneous phase, aQHM is initialized by QHM. Results on synthetic signals showed that aQHM estimates efficiently the instantaneous components of the signals. Comparisons with QHM and SM on synthetic AM-FM signals showed that aQHM outperforms both of them. Finally, similar results were obtained when aQHM was compared to QHM and SM on voiced speech signals.

Chapter 4

Analysis/Synthesis Speech System based on aQHM

This Chapter develops an analysis/synthesis (A/S) speech system which is able to produce indistinguishable resynthesized speech. Taking into account the different sources that constitute speech, we choose to follow a hybrid representation of speech. Hybrid models separate speech into a deterministic component and a stochastic component [80, 9, 81]. The deterministic component models the quasi-periodic features of speech while the stochastic component models the non-periodic characteristics of speech. Voiced speech usually contains both components. The source separation results in better manipulation of the different components leading to more flexible and efficient speech modification algorithms.

One well known hybrid model for speech is the Harmonic+Noise model (HNM) developed by Stylianou [32] and Stylianou et al. [81] and it was used for high quality time-scale/pitch-scale modification of speech and for voice transformation. HNM decomposes speech into two bands: the lower band (deterministic part) where the speech signal is modeled as a sum of harmonically related sinusoids and the upper band (stochastic part) where the speech signal is modeled as modulated noise. In the literature, the separation of periodic and aperiodic components of speech has gained a lot of research interest [82, 83]. In our separation scheme, the deterministic part captures the speech signal up to a maximum voiced frequency and the residual signal between the speech signal and the reconstructed deterministic component defines the stochastic component.

We suggest modeling the deterministic part using aQHM initialized by QHM as in the AM-FM decomposition algorithm presented in the previous Chapter. Taking advantage of the time-varying characteristics of the analyzed signal, aQHM is able to address efficiently the local non-

stationarity of the speech signal. Compared to HNM or SM, this new approach reduces further the bias in the estimation of the sinusoidal parameters, yielding a more accurate, compared to these models, signal representation. However, the AM-FM decomposition algorithm cannot be applied directly. One reason is that speech is a non-stationary process and some components may be "born" or "die" and the AM-FM decomposition algorithm is not able to cope with such cases. Actually, the AM-FM decomposition algorithm assumes that the number of AM-FM components is known (and constant). Thus, the tracking of the frequency trajectories is very difficult. Another reason is that not only the frequency tracking is difficult but also the tracking of the fundamental frequency is not always robust. Actually, it has been observed that sometimes the tracking of the fundamental frequency is lost which results in deterioration of the quality of the reconstruction. Thus, we suggest adding a module to the A/S system which performs fundamental frequency estimation. Having an estimation of the fundamental frequency, both the initialization step and the definition of the frequency tracks are simplified. Indeed, in the initialization step, QHM uses as initial frequencies integer multiples of the estimated fundamental frequency up to a maximum voiced frequency, while the frequency tracks are defined by the number of harmonics.

The stochastic component is modeled as a time-modulated and frequency-modulated Gaussian noise. Frequency modulation is achieved by AR modeling and LPC analysis while the time modulation is achieved by a time-domain envelope. The time-domain envelope is very important for correct fusion of the two components and it was shown in [71] that an energy-based envelope gives the best perceptual result. Moreover, the analysis of stochastic part can be performed synchronous or asynchronous to the deterministic part. Our choice is to use an asynchronous analysis for the stochastic part.

In the synthesis step, the deterministic part is synthesized as a time-varying sum of amplitude-modulated and frequency-modulated sinusoids. Indeed, in aQHM, frame-by-frame interpolation of the parameters is more natural than overlap-add (OLA) method. On the other hand, the stochastic part is synthesized frame-by-frame using the OLA method. Listening tests show that the reconstructed signal is indistinguishable from the original which validates the high-quality speech representation.

4.1 Analysis

The separation of speech signal, $s(t)$, into two additive parts is given by

$$s(t) = s_d(t) + s_s(t) \tag{4.1}$$

where $s_d(t)$ denotes the deterministic part while $s_s(t)$ denotes the stochastic part. Voiced segments contain both parts while deterministic part is zero in unvoiced segments.

Deterministic part which models the periodicities of voiced speech segments as a sum of time-varying sinusoidal components (i.e. SM) is written as

$$s_d(t) = \sum_{k=-K(t)}^{K(t)} A_k(t) e^{j\phi_k(t)} \tag{4.2}$$

where $K(t)$ is the time-varying number of components, $A_k(t)$ and $\phi_k(t)$ are the instantaneous amplitude and instantaneous phase of the kth component, respectively. Instantaneous frequency is once again given by $f_k(t) = \frac{1}{2\pi}\frac{d\phi_k(t)}{dt}$.

Stochastic part models the aperiodicities of speech signal as a time- and frequency-modulated Gaussian noise. As stated above, stochastic part models all the information of unvoiced segments. For voiced segments, stochastic part is defined as the residual between the speech signal and the reconstructed deterministic part. However, deterministic part cannot fully represent the periodicities especially at the extremely non-stationary regions of voiced segment, thus, the residual signal is highpass filtered. In other words, this processing step asserts that below a frequency, voiced speech contains only quasi-periodic information. To sum up, stochastic part is given by

$$s_s(t) = (s(t) - \hat{s}_d(t)) \star p(t) \tag{4.3}$$

where $\hat{s}_d(t)$ is the reconstructed deterministic part while $p(t)$ is the impulse response of a zero-phase highpass filter with cutoff frequency F_m and $\star$ denotes convolution.

4.1.1 Deterministic Part

Preliminary Analysis

Recorded speech contains various types of sounds, hence, it is usual to separate a speech file into speech and nonspeech regions and for the speech regions a further discrimination is performed

between voiced and unvoiced segments. Then, fundamental frequency estimation is performed for voiced segments. The detection of speech/non-speech and voiced/unvoiced segments is performed in a frame-by-frame procedure. The energy of each frame is first computed and if this is above a threshold B_e, then, it is assigned as speech, otherwise it is assigned as nonspeech (silence, in our case). In Table 4.1.1, parameter values used in our implementation of the A/S speech system are provided. For the voiced/unvoiced decision, speech signal is lowpass filtered with cutoff frequency F_v and the following condition is tested. If the energy (measured by the standard deviation of the speech samples) of the speech frame minus the energy of the smoothed speech frame is below B_d and if the energy of the smoothed signal is above B_s, then, the frame is assigned as voiced, otherwise, it is assigned as unvoiced. The frame duration was set to $30ms$ while the time-step was set to $5ms$. Finally, in order to eliminate isolated decisions, a median filter is applied to the estimated decisions. An adequate order for the median filter was found to be 5.

Parameter	Value (dB)	Parameter	Value (Hz)	Parameter	Value (#)
B_e	-60	F_v	1000	K_f	3
B_d	10	F_m	1500	K_e	4
B_s	-50	F_M	$5500Hz$		
B_m	-55				

Table 4.1: Various parameter values used in the implementation of the analysis step.

Pitch Estimation

A novel fundamental frequency estimator based on time-domain information is derived. It is inspired from the visual inspection of voiced signals and how human eye (not ear) understands and "measures" the pitch period. Indeed, speech can be viewed as the output of a filter, which represents the vocal tract, excited by the glottal flow derivative. Thus, the proposed pitch estimator searches for the local minima of speech which are related with the minima of the glottal flow derivative waveform. As it will be shown, the suggested pitch estimation algorithm eliminates doubling or halving problems especially at the beginning or ending of the voiced segment. Note also that the accuracy of the estimated pitch period is not crucial in our A/S speech system since QHM is able to correct small frequency mismatch errors.

The description of the algorithm is as follows[1]. As a first step, and for each voiced segment the minimum value of the smoothed signal is found. The assumption here is that around the

[1]Note that the estimation of pitch period is performed on the smoothed speech signal.

minimum value the signal is more stationary and, thus, the estimation of pitch period is more robust. Next, using the autocorrelation function around the minimum value an estimate of the pitch period is found. Moving forward (or backward) using as a step the locally estimated pitch period, the next (or previous) local minimum is searched. The search is performed in a region of $5ms$ and $3.5ms$ for male and female voices, respectively. Finally, we move to the next expected minimum value of the signal and continue in this way until the end (forward) or the beginning (backward) of the voiced segment is reached. Figure 4.1 shows a particular instant of the pitch estimation algorithm. Fundamental frequency at a point is computed as a weighted sum of the reciprocals of the two closest to the point pitch periods. Finally, fundamental frequency is passed through a median filter to smooth out perturbations of the computed fundamental frequency.

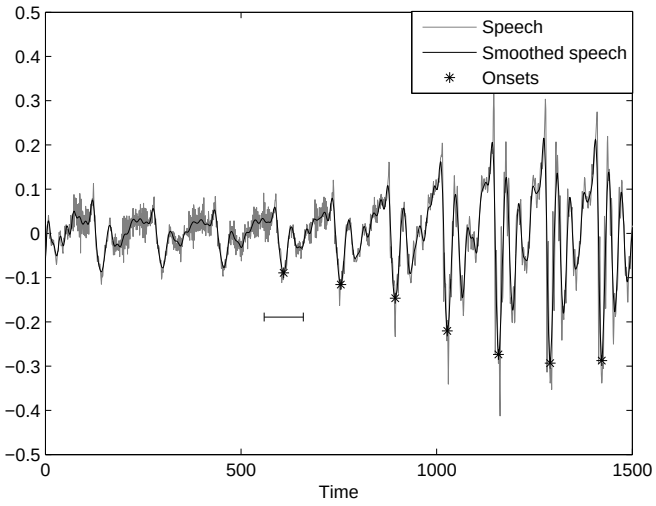

Figure 4.1: The pitch estimation algorithm take advantage of speech production mechanism and tries to find local minima around a defined area. These local minima are attributed to local minima of the glottal flow derivative waveform.

Initialization Step: QHM

Within the lth frame which is centered at time instant t_l, the deterministic component is modeled by QHM as

$$h_q^l(t) = \sum_{k=-K_l}^{K_l} (a_k^l + t b_k^l) e^{j2\pi k f_0^l t} w(t), \qquad t \in [-T_l, T_l] \tag{4.4}$$

where f_0^l is the fundamental frequency of lth frame estimated at the previous step. K_l specifies the order of the model, i.e., the number of harmonics which is given by $K_l = \lfloor \frac{F_M}{f_0^l} \rfloor$ where F_M is the maximum voiced frequency and $\lfloor \cdot \rfloor$ denotes the floor operator. The window is typically a Hamming window with support in the symmetric interval $[-T_l, T_l]$. The window length depends on the local pitch period and it is equal to three pitch periods. We found that this is a good compromise between the necessary samples for robust estimation of the sinusoidal components and the non-stationary character of speech signals.

Since mistakes may take place in the estimation of the fundamental frequency, the kth harmonic may have a frequency error which is k times the estimation error of the fundamental frequency. This may lead to problems in the parameter estimation as well in the determination of the frequency tracks. Thus, once an initial estimation of the frequency mismatch is obtained for the kth harmonic from $\rho_{2,k}^l$, then the local fundamental frequency, f_0^l, can be updated using the first K_f harmonics by

$$f_0^l = f_0^l + \Delta f_0^l = f_0^l + \frac{1}{K_f} \sum_{k=1}^{K_f} \frac{\rho_{2,k}^l}{k} \tag{4.5}$$

where K_f is a small integer value. In our implementation, we set K_f to be 3 as Table 4.1.1 reports. Then, the input signal can be modeled again by QHM using now the updated fundamental frequency. Figure 4.2 shows a frame of speech in time-domain and in frequency-domain as well the analysis frequencies before and after applying the correction of fundamental frequency. For the particular frame shown in Figure 4.2, the initial estimate of the fundamental frequency (circles) is not accurate, but, after correcting fundamental frequency (stars) the frequency values are correct and meaningful. However, there are frames (especially at the boundaries of voiced segments) where the update of the fundamental frequency results in lower accuracy in terms of reconstruction error. Thus, we suggest keeping the updated fundamental frequency only if the reconstruction error is improved.

The instantaneous components are estimated at time-instant t_l from the parameters of QHM as in the AM-FM decomposition algorithm. Hence,

$$\hat{f}_k(t_l) = k f_0^l + \frac{\rho_{2,k}^l}{2\pi} \tag{4.6a}$$

$$\hat{A}_k(t_l) = |a_k^l| \tag{4.6b}$$

$$\hat{\phi}_k(t_l) = \angle a_k^l \tag{4.6c}$$

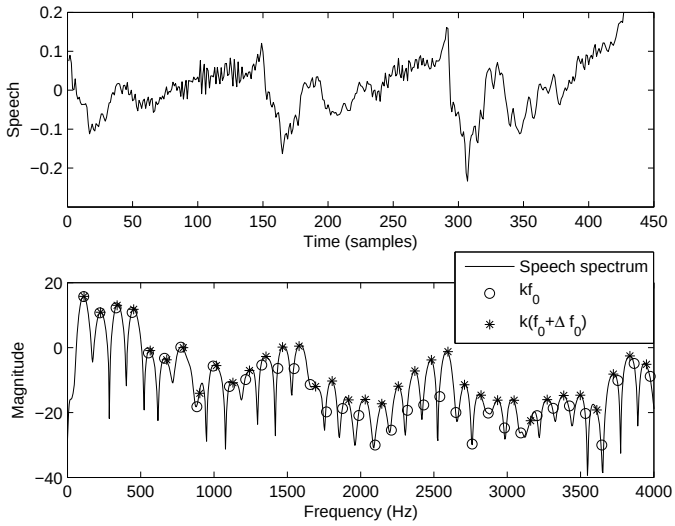

Figure 4.2: Upper plot: A speech frame of three pitch periods. Lower plot: Spectrum of the upper frame, the analysis frequencies (circles) and the refined analysis frequencies (stars). The refinement is performed using one iteration of QHM.

One major problem with QHM is the existence of components with very low energy and in particular when this is combined with high noise level because of the speech production mechanism (frictions etc.). For instance, nasal phonemes have antiformants which result in frequency bands with low amplitude. Also there are phonemes with high friction and high noise levels at some frequency bands. In these cases, the assumption of existence of time-varying sinusoids is very weak and causes problems in the QHM estimation procedure such as incorrect frequency mismatch estimation as well matrix ill-conditioning during the LS estimation. To cope with these problems, we check for two conditions for each harmonic before applying (4.6). First, the amplitude of kth harmonic should be at most B_m (dB) less than the highest amplitude of the frame and second, the frequency correction term for each harmonic (i.e. $\frac{\rho^l_{2,k}}{2\pi}$) should be at most $\frac{f^l_0}{2}$. If these two conditions are not satisfied for a sinusoidal component, then we assume that it does not exists. Finally, the interpolation of the instantaneous components (amplitudes, frequencies, phases) is exactly the same as it is described in the development of the AM-FM decomposition algorithm.

Adaptation Step: aQHM

In the adaptation step, the analysis is performed on the time-varying basis functions which use the estimated instantaneous phase. In this way, the signal is projected in functions that are

adapted to the signal.

$$h_a^l(t) = \sum_{k \in \mathcal{A}_l} (a_k^l + t b_k^l) e^{j(\hat{\phi}_k(t_l - t) - \hat{\phi}_k(t_l))} w(t), \qquad t \in [-T_l, T_l] \tag{4.7}$$

where $\mathcal{A}_l$ is a set which contains the index of the time-varying components that exists at time-instant t_l. Note that the instantaneous components have been determined at the initialization step and their duration cannot be changed at the adaptive step. This means that the conditions used for the amplitudes are frequencies at the initialization step are adequate for robust estimation of the instantaneous components. Finally, caution should be put for frames where some components are born or die. In such cases, the instantaneous frequency is expanded with a constant value which equals to the boundary frequency value. This is shown in Figure 4.3 where the estimated frequency trajectories (lines with circles) are depicted and if a component is born or die within the frame then it is continuously expanded (dashed lines).

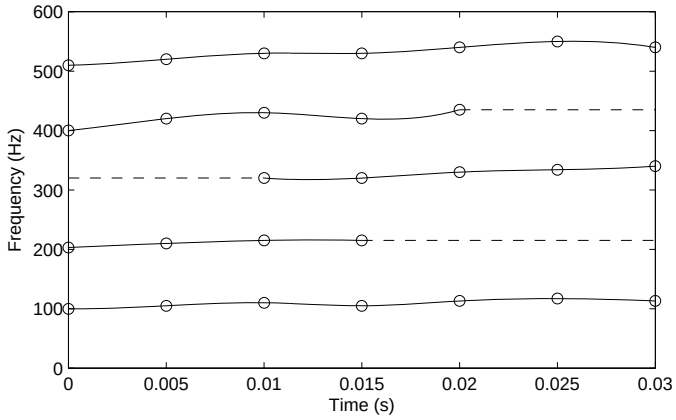

Figure 4.3: Five frequency tracks within a voiced frame. Second and forth trajectories are dying during the frame while third frequency trajectory is born.

4.1.2 Stochastic Part

Unvoiced segments are modeled frame-by-frame as frequency-modulated Gaussian noise. The frequency modulation is modeled by an AR filter whose parameters are estimated from linear prediction (LP) analysis. A Hamming window of duration $30ms$ and time-step of $5ms$ is used for the analysis of both unvoiced and voiced frames. For voiced segments, whatever is not modeled by the deterministic part, it belongs to the stochastic part. Also remember that the residual signal between speech and reconstructed deterministic part is highpass filtered at cutoff frequency F_m.

Stochastic part for one frame is then modeled as

$$s_s^l(t) = e^l(t)[u^l(t) \star q^l(t)] \tag{4.8}$$

where $u^l(t)$ denotes Gaussian noise process filtered by a time-varying AR filter with impulse response $q^l(t)$ while $e^l(t)$ is the time-domain envelope. As concerns the frequency modulation, the estimation of the AR filter is performed by LP analysis as in unvoiced segments, while the time-domain envelope, which is very important for the fusion of the two components, is an energy-based envelope represented as a sum of sinusoids. In [75], various time-domain envelopes such as triangular envelope [32], Hilbert-based envelope [84] and energy-based envelope were tested. Listening tests showed that the energy-based envelope outperformed all the other considered envelopes.

The idea behind the energy envelope is to compute the energy variation of the stochastic component and model it as a low-order sum of sinusoids. The energy envelope of the stochastic part is computed by a local mean average of the absolute stochastic part. Mathematically, energy envelope is given by

$$e(t) = \int_{t-T_o}^{t+T_o} |s_s(\tau)| d\tau \tag{4.9}$$

where T_o is $1ms$. Time-domain envelope for frame l is then approximated by a sum of sinusoids as

$$\hat{e}^l(t) = \sum_{k=-K_e}^{K_e} d_k^l e^{j2\pi\zeta_k^l t} \tag{4.10}$$

where K_e is the number of harmonics which is a small integer, while frequencies, ζ_k^l, and complex amplitudes, d_k^l, are computed by peak picking the spectrum of the time envelope as in sinusoidal model. An instance of an estimated energy envelope is depicted in Figure 4.4 for the stochastic part of a voiced frame. Figure 4.4(a) shows the energy envelope (solid line) computed by (4.9) as well the reconstructed energy envelope (dashed line) from (4.10). While, in Figure 4.4(b), the frequency contents of the frame as well the estimated frequency envelope are depicted.

4.2 Synthesis

In the synthesis step, the deterministic part is resynthesized as a sum of time-varying sinusoids. Note that this synthesis method is preferred from overlap-add (OLA) method because the time-varying frequency trajectories were already used by aQHM in the analysis step. In the case when

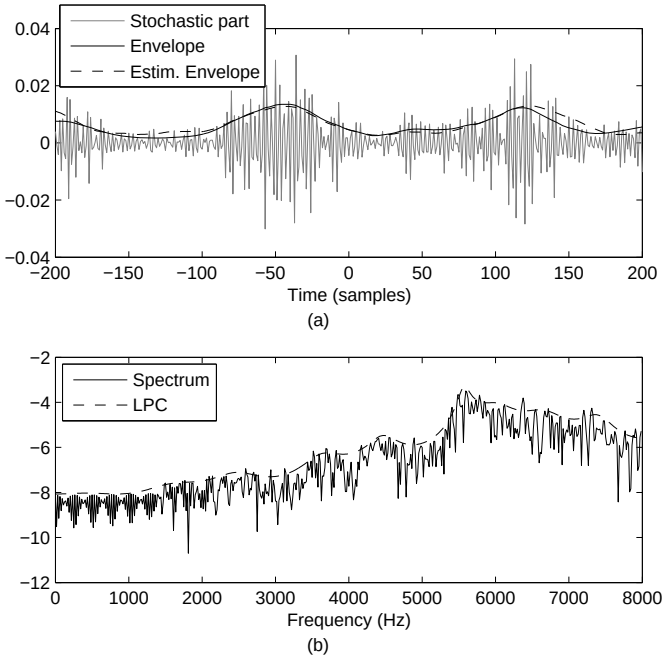

Figure 4.4: Upper plot: A frame of the stochastic part, its energy time-envelope and the estimated time-envelope. The envelope has pitch synchronous behavior. Lower plot: Frequency representation for the upper frame and its AR modeling

a sinusoidal component is born or dying, the instantaneous amplitude vanishes linearly until the next analysis time-instant while instantaneous frequency remains constant until the component vanishes.

The stochastic part is resynthesized using OLA method. For each frame, white noise is passed through the AR filter to obtain the frequency modulation of the stochastic part. Then, the energy envelope is computed from (4.10) and its multiplication with the frequency-modulated noise provides the reconstructed stochastic frame.

4.3 Evaluation

The overall performance of the A/S speech system is shown in Figures 4.5–4.9. The original speech sentence uttered by a male speaker (Figure 4.5), the reconstructed speech (Figure 4.6), the reconstruction of the deterministic part (Figure 4.7) as well the stochastic part and its reconstruction (Figures 4.8 and 4.9, respectively) are shown in both time and frequency domains. The time-step used in the analysis of the deterministic part is $5ms$. Evidently, the reconstruction of

speech is very close to the original speech in both domains.

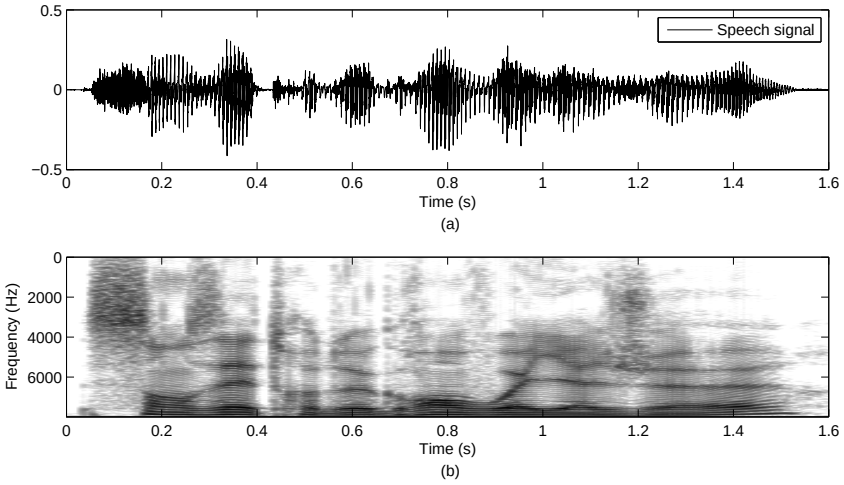

Figure 4.5: A speech sentence uttered by a male speaker in both time (a) and frequency (b).

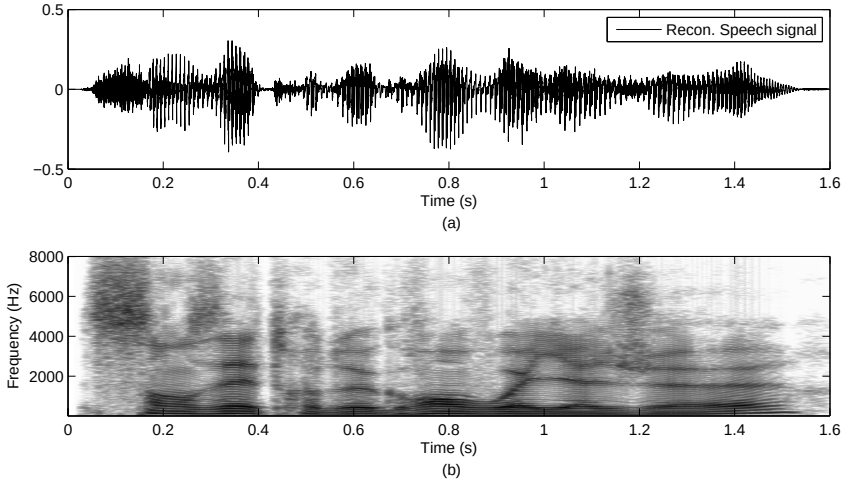

Figure 4.6: The reconstruction of the speech signal shown in Figure 4.5 in both domains.

4.3.1 Listening Examples

The best way to evaluate the performance of an A/S speech system is to listen to the reconstructed signals. Table 4.2 presents speech examples from various databases of both male and female speakers. The proposed A/S speech system denoted by aQHM is compared with the SM of McAulay and Quatieri [1], the HNM of Stylianou [32] and the STRAIGHT of Kuwahara [85]. Further examples and possibly updates can be found in `www.csd.uoc.gr/~pantazis/source/thesis/variousModels.html`.

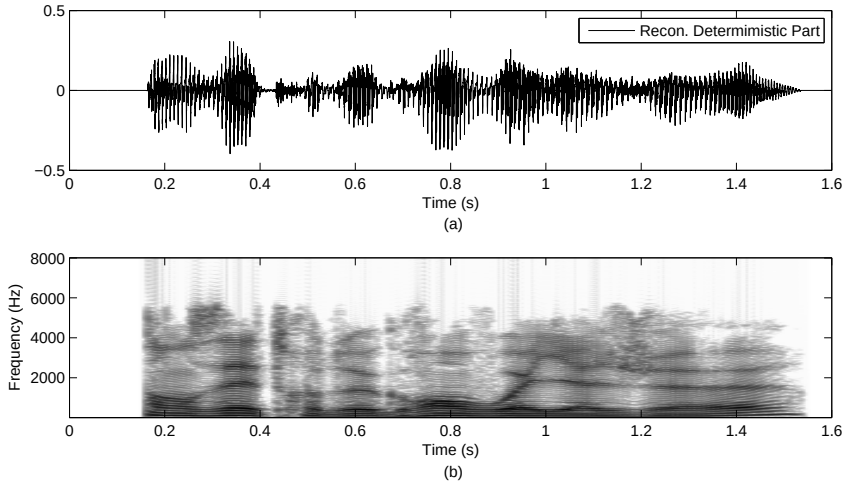

Figure 4.7: The reconstruction of the deterministic part of the signal shown in Figure 4.5 in both domains.

	Original	aQHM	HNM	SM	STRAIGHT
Male 1					
Male 2					
Female 1					
Female 2					

Table 4.2: Analysis/Synthesis of speech signals using various methods.

4.4 Conclusion

In this Chapter, we developed an A/S speech system based on a hybrid representation of speech. Thus, speech was separated into a deterministic part and into a stochastic part. The deterministic part was modeled as a sum of time-varying sinusoids whose instantaneous components were estimated using aQHM. Initialization of aQHM was provided by QHM whose initial frequency estimates were obtained from a novel fundamental frequency estimator. The stochastic part was modeled as time- and frequency-modulated noise. Time-modulation is achieved using an energy-based envelope. Listening tests showed that the resynthesized speech was indistinguishable from the the original signal for both male and female speakers.

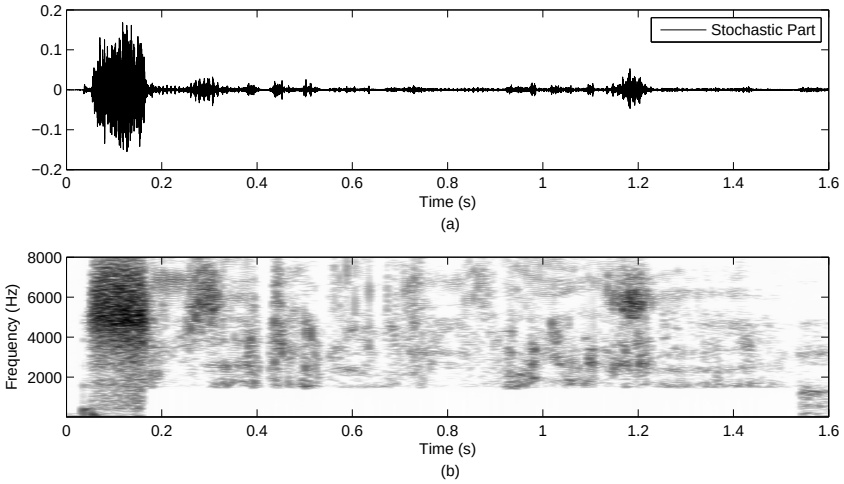

Figure 4.8: The stochastic part (i.e. the residual signal) of the signal shown in Figure 4.5 in both domains.

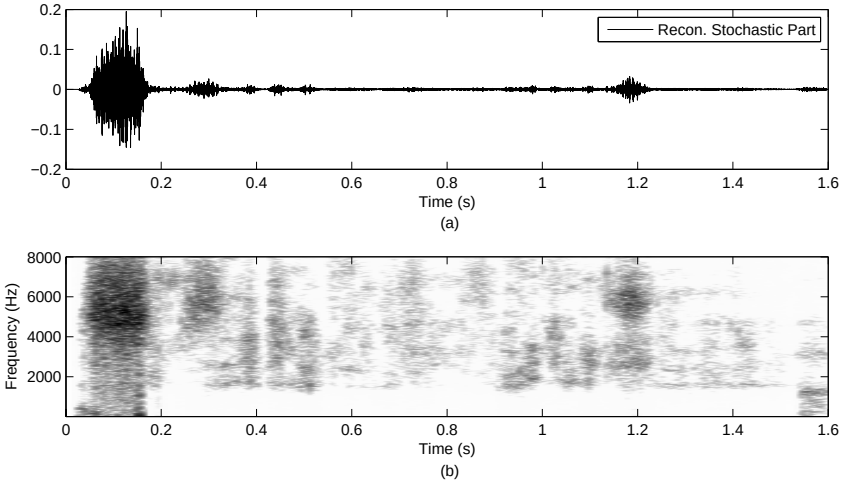

Figure 4.9: The reconstruction of the stochastic part of the above Figure in both domains.

Chapter 5

Vocal Tremor Estimation

Voice quality assessment is another area of speech processing where the accurate representation of speech is of high importance. Indeed, high-resolution speech analysis in both time and frequency may reveal interesting properties about the physiology of vocal organs such as vocal folds. In this Chapter, we suggest applying the developed AM-FM decomposition algorithm presented in Chapter 3 for voice quality assessment. More specifically, we apply the decomposition algorithm for the extraction of acoustic vocal tremor characteristics. In the following, the definition of vocal tremor as well its prominent acoustic characteristics are provided. Then, a three step algorithm based primarily on the suggested AM-FM decomposition algorithm which estimates the vocal tremor attributes is constructed. Results indicate that the proposed method is able to accurately extract the time-varying attributes of vocal tremor.

5.1 Introduction

Typically, tremor in phonation is defined as modulations of the fundamental frequency and modulations of the amplitude due to the inability of humans to keep constant the tension of their vocal folds [86]. This phenomenon affects the glottal cycle in voiced speech making the fundamental frequency and the amplitude to vary stochastically. Vocal tremor should not be confused with jitter or shimmer which are also defined as modulations of the fundamental frequency and of the amplitude, respectively. Vocal tremor refers to modulations whose modulation frequencies are slow (i.e. below $20Hz$) while jitter and shimmer refer to cycle-to-cycle modulations which are faster (i.e. modulation frequency is around the half of the local fundamental frequency).

Vocal tremor is usually categorized into the physiological tremor which is a slow natural mod-

ulation of glottal cycle and the pathological tremor which is attributed to neurological diseases such as Parkinson or tremor of the limbs [87], [88]. Most importantly, while physiological tremor makes speech sound more natural and possibly more individual, pathological tremor may influence the quality of patients voice, hence, may influence the ability of patient's communication. Moreover, while pathological tremor is characterized by stronger periodical patterns —a property that vibrato singing style has, too—, physiological tremor is more stochastic [87]. The analysis of physiological tremor is of great importance since vocal tremor in normophonic speakers may be an early sign of a neurological disease [89], [90]. Thus, it is useful to develop an estimation algorithm that is able to measure or extract features of vocal tremor even for normal voices. In the literature, acoustic analysis of tremor is usually based on the accurate estimation of fundamental frequency and then the characterization of the variations of fundamental frequency is performed [91], [88]. Modulation frequency which models the periodicity of vocal tremor and modulation level which models the strength of vocal tremor are the prominent acoustic attributes that are extracted from the instantaneous fundamental frequency [91], [88].

The objective of this Chapter is to present and validate a novel method for the estimation of the vocal tremor on sustained vowels uttered by normophonic subjects. The proposed method assumes speech as a sum of time-varying sinusoids (i.e. SM) whose instantaneous amplitude and instantaneous frequency are estimated using the suggested AM-FM decomposition algorithm based on aQHM. Then, the second step of the algorithm is to reveal the higher frequency modulations, hence, we subtract from the analyzed instantaneous component the very slow modulations ($< 2Hz$), which are attributed to sources like the cardiac rhythm. This is achieved by filtering the instantaneous component using a Savinzky-Golay smoothing filter [92]. The final step is to estimate the modulation frequency and the modulation level which are assumed time-varying attributes due to the fact that modulations are primarily non-stationary. Thus, the estimation of these vocal tremor attributes is performed using again the proposed AM-FM decomposition algorithm.

5.2 Extraction of Vocal Tremor Characteristics

In the following subsections, each step of the vocal tremor extraction algorithm is described in detail.

5.2.1 Step 1: Estimation of Instantaneous Components of Speech

Considering speech as a superposition of time-varying sinusoids, the algorithm presented in Section 3.5 is applied for the estimation of the instantaneous components of speech. We remind that the AM-FM decomposition algorithm is initialized by QHM. QHM also needs a rough estimate of the analysis frequencies. These are assumed as integer multiples of an estimated fundamental frequency. The fundamental frequency of the first frame can be computed using the autocorrelation function. Then, in the adaptation step, the use of aQHM refines the estimation of the instantaneous components. Note that time resolution of aQHM is primarily determined by the time-step of the algorithm while frequency resolution is determined by the window type and window length. For vocal tremor analysis, we choose a time-step of $5ms$ and Hamming window as window function. The window duration has been chosen to be three times the pitch period. Please note that for vocal tremor, all the speech material comes from sustained vowel phonations.

Illustratively, Figure 5.1 shows the first five harmonics extracted from sustained vowel $/a/$ using the suggested AM-FM decomposition algorithm. The signal which is reconstructed from the instantaneous components has SRER of about $32dB$ which proves that the analysis is very accurate. Figure 5.1 also shows that the modulations of higher harmonics are more evident.

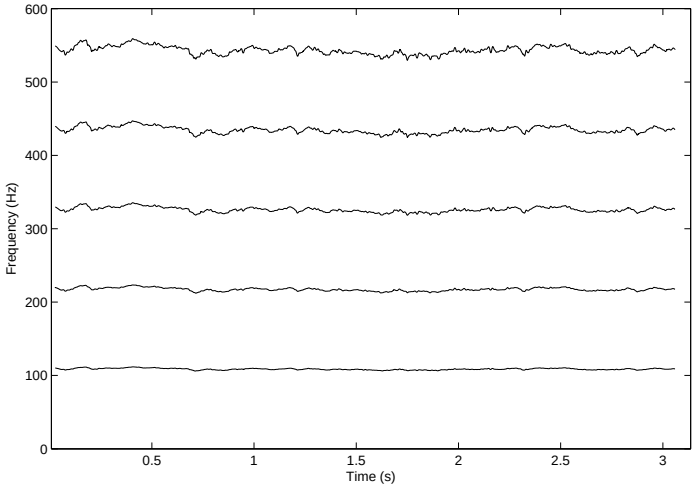

Figure 5.1: First five instantaneous frequencies of a normophonic male speaker uttered the sustained vowel /a/.

5.2.2 Step 2: Removal of Very Slow Modulations

In the literature of vocal tremor analysis, the component associated to the fundamental frequency is mainly processed. However, in our case, we have the opportunity to choose between any of the estimated instantaneous components. Based on the voice production theory, vocal tremor should have similar acoustic attributes for the different instantaneous components. Continuing the analysis process, one instantaneous component is selected, its mean value is subtracted and then the zero-mean instantaneous component is downsampled to 1000 Hz since only low frequency modulations are of our interest. Then, very slow modulations, which refer to as trend and are less than 2Hz, should be eliminated.

To remove the trend, we apply a smoothing operator. The smoothed instantaneous component is computed using the Savinzky-Golay (S-G) filter [92], [93]. S-G smoothing filter essentially performs a local polynomial regression on a distribution of equally spaced points to determine the smoothed value for each point. The main advantage of this approach is that it tends to preserve features of the distribution such as relative maxima, minima and width, which are usually "flattened" by other adjacent averaging techniques like moving averages. The order of the local polynomial used is 4 while the frame size is set to $1s$ (1000 samples). Figure 5.2(a) shows the instantaneous component (solid line) as well its smoothed version (dashed line) for a sustained vowel. Figure 5.2(b) implies that S-G filter captures the frequencies that are less than $2Hz$. Then, the smoothed instantaneous component is subtracted from the unsmoothed in order to reveal the remaining modulations of the component (see Figure 5.3(a)). Note that using different parameters for the S-G filter the smoothed signal will capture more or less of the signal's frequencies.

5.2.3 Step 3: Extracting Vocal Tremor Characteristics

The final step of the vocal tremor extraction algorithm is the modeling and estimation of the remaining modulations. As already stated, these modulations are non-stationary, hence, FFT-based approaches are not appropriate for this task. We suggest modeling the remaining non-stationary modulations as a mono-component AM-FM signal. Mathematically, it is given by

$$x(t) = m(t)cos(\psi(t)) \tag{5.1}$$

where $x(t)$ are the remaining modulations of the instantaneous component, $m(t)$ is the instantaneous amplitude while $\psi(t)$ corresponds to the instantaneous phase. Please note that an appro-

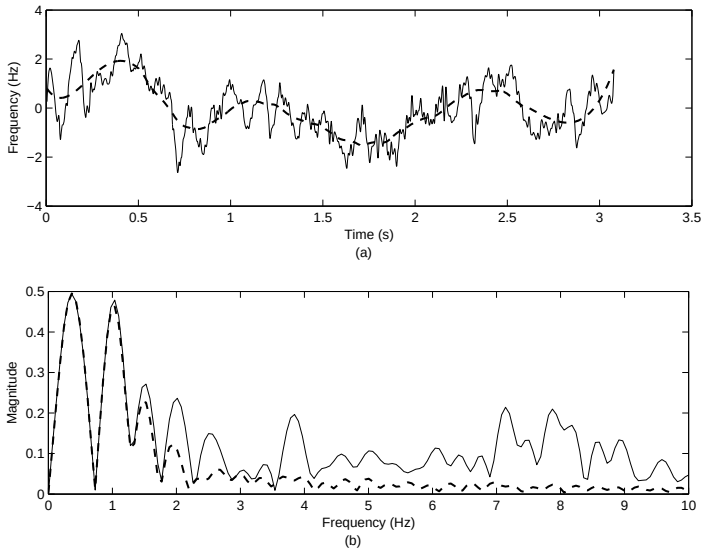

Figure 5.2: (a) First harmonic of Figure 5.1 without its mean value (continuous line) and its smoothed version (dashed line) are shown. (b) Fourier transform of signals in (a). S-G smoothing filter captures the frequencies that are below $2Hz$.

priate scaling of $m(t)$ will correspond to the modulation level of the vocal tremor. The scaling factor of the modulation level is equal to the mean value of the analyzed instantaneous component which has been subtracted at Step 2. Once again, instantaneous frequency is given by $\zeta(t) = \frac{1}{2\pi}\frac{d\psi(t)}{dt}$ and corresponds to the modulation frequency of the vocal tremor.

The proposed AM-FM decomposition algorithm is again applied for the estimation of the instantaneous components, $m(t)$ and $\zeta(t)$. The initial frequency of the first frame was computed by the largest peak of the FFT of the first frame while time-step is set to $1ms$ (i.e. one sample). Hamming window is used as analysis window and its duration is set to $0.6s$. Figure 5.3 shows the reconstructed signal (dashed line) obtained from the AM-FM decomposition algorithm in both time and frequency domains. The SRER for this particular example is $7.3dB$ while the number of adaptations of the AM-FM decomposition algorithm is 3. Figure 5.3 indicates that the decomposition algorithm adapts to the non-stationary modulations of the signal. The estimated modulation frequency as well the estimated modulation level are shown in Figure 5.4. In this example, modulation frequency takes values between $2Hz$ and $13Hz$ while modulation level is between 0.15 and 0.55 which are typical values for normophonic speakers.

Finally, an important feature of the proposed method is that any of the instantaneous components can be analyzed. Figure 5.5 shows the instantaneous amplitude of the 4th harmonic

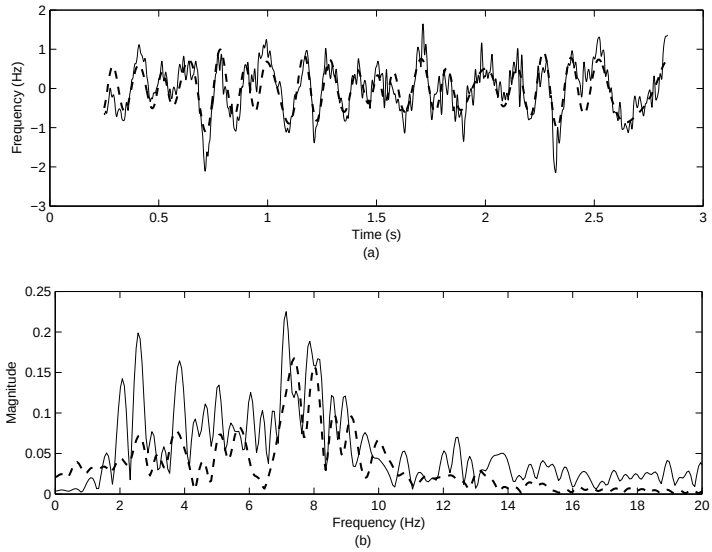

Figure 5.3: (a) Instantaneous component after subtracting its smoothed version (continuous line) and the reconstruction after applying the AM-FM decomposition algorithm (dashed line). (b) Fourier transforms of the components in (a).

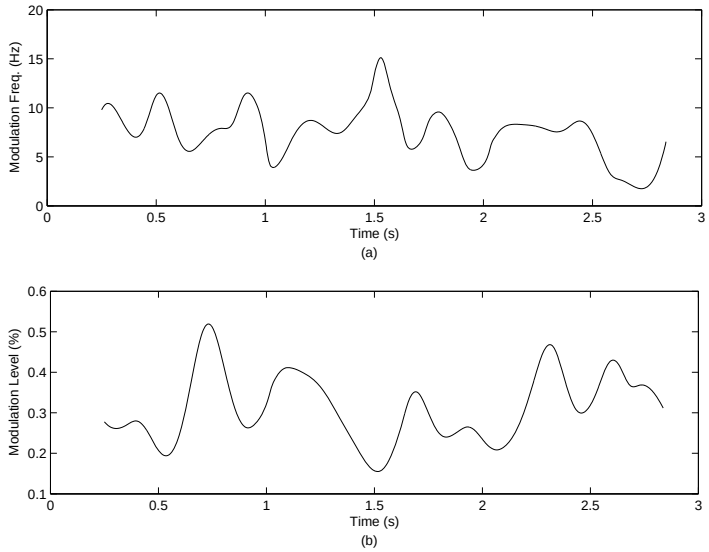

Figure 5.4: (a) Modulation frequency of the signal in Figure 5.3. (b) Modulation level of the same signal. Note that neither modulation frequency nor modulation level have constant values during the phonation.

(solid line) for the same sustained vowel as well its reconstruction (dashed line) while Figure 5.6 shows the estimated modulation frequency and modulation level. Interestingly, both modulation frequency and modulation level differs from that of Figure 5.4 which indicate that different at-

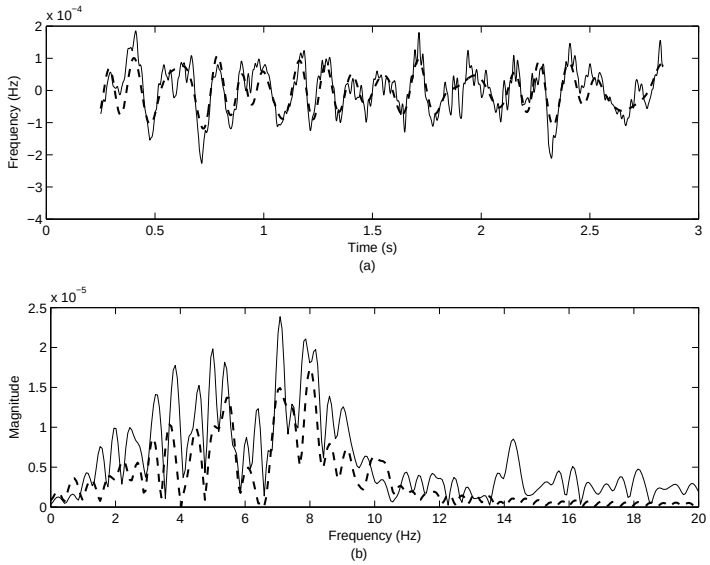

Figure 5.5: Similar to Figure 5.3 but for the instantaneous amplitude of the 4th harmonic. Note that the proposed vocal tremor extraction algorithm can be applied to any of the instantaneous component.

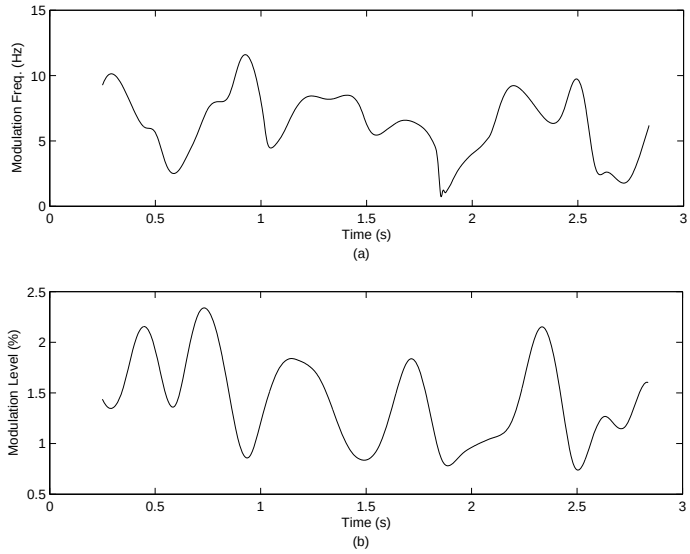

Figure 5.6: Similar to Figure 5.4 but for the instantaneous component of Figure 5.5. Similarities and differences can be found between the modulation frequency and modulation level of instantaneous components.

tributes of vocal tremor are obtained whenever different instantaneous components are analyzed. This is in contrary to the basic (and simplistic however) voice production theory and therefore requires further investigation.

5.3 Large-scale Results

The method is validated on a database of normal voices developed in our recording lab. 11 male and 5 female healthy subjects whose age varies between 23 and 45 were participated. Sustained vowels /a/, /e/, /i/, /o/ and /ou/ have been recorded at $48kHz$ and then downsampled at $16kHz$. The duration of sustained vowels varies from $2s$ to $8s$ depending primarily on the speaker. Extended tests on the database confirmed the ability of AM-FM decomposition algorithm to extract the instantaneous components of the signals accurately. This was quantified by measuring the modeling error through SRER. Indeed, average SRER for the total database is computed to be $30.7dB$. Moreover, visual inspection of the analyzed instantaneous components shows that the suggested AM-FM decomposition algorithm at Step 3 captures adequately the remaining modulations. The average SRER is $4.8dB$ for the reconstruction of the instantaneous components (Step 3).

Finally, Table 5.1 reports the averages of fundamental frequency, $\mu(f_0)$, of mean value, $\mu(mf)$, and standard deviation, $\sigma(mf)$, of modulation frequency and of mean value, $\mu(ml)$, and standard deviation, $\sigma(ml)$, of modulation level for male and female speakers uttering various vowels. It is evident that the standard deviation of modulation frequency is higher for the male speakers while the mean value of modulation frequency shows no tendency between the genders.

		$\mu(f_0)$ (Hz)	$\mu(mf)$ (Hz)	$\sigma(mf)$ (Hz)	$\mu(ml)$ (%)	$\sigma(ml)$ (%)
Male	/a/	113	4.4	1.4	0.25	0.11
	/e/	116	4.3	1.2	0.28	0.13
	/i/	119	4.1	1.3	0.25	0.11
	/o/	121	6.2	1.7	0.22	0.09
	/ou/	122	8.0	2.0	0.20	0.08
Female	/a/	233	6.6	0.9	0.36	0.14
	/e/	228	9.3	0.9	0.33	0.14
	/i/	239	3.1	0.8	0.29	0.12
	/o/	235	4.7	0.9	0.27	0.10
	/ou/	236	3.4	0.8	0.27	0.10

Table 5.1: Summary of modulation features for five vowels and both genders.

5.4 Conclusion

In this Chapter, we developed a method which is able to extract acoustic vocal tremor characteristics such as modulation frequency and modulation level from sustained vowels. Due to the fact that the analyzed signals have time-varying components, we applied the AM-FM decomposition algorithm devepoled in Chapter 3. Thus, both the instantaneous components of speech and the acoustic characteristics of vocal tremor were computed using the proposed AM-FM decomposition algorithm. Results indicated that the proposed method extracts in a robust and efficient way the vocal tremor characteristics of speech.

Chapter 6

Summary and Future Research Directions

6.1 Summary

In this thesis, we developed and tested models and algorithms for the representation of time-varying sinusoidal signals. For the estimation of the sinusoidal parameters, we reintroduced a time-varying model with complex parameters referred to as QHM. Parameters of QHM contained not only amplitude information but also frequency information. A proper decomposition of QHM parameters revealed the frequency information which could be used for the estimation of the frequency mismatch between the true frequencies and the provided frequencies. Thus, an iterative algorithm referred to as iQHM for the estimation of sinusoidal parameters were proposed. The region of convergence for iQHM is also provided. The performance of iQHM was tested under noisy conditions and its statistical efficiency was shown. Furthermore, iQHM was tested in voiced speech and its supremacy over HM was proved. An extension of QHM referred to as chirp QHM (cQHM) which is able to capture linear evolution of the frequencies was also presented. cQHM is a second-order polynomial of time with complex parameters. A similar to QHM decomposition of the cQHM parameters revealed not only the frequency mismatch but also the chirp rate of the analyzed signal.

However, in the case of fast frequency variations, we showed that QHM and cQHM refinement of the frequencies could be obtained, but only up to a certain point. This estimation error was due to the fact that the applied stationary or even chirp basis functions were not adequate to model the input signal. In order to tackle with the non-stationary character of the analyzed

signals, we suggested another extension of QHM referred to as adaptive QHM (aQHM). aQHM is a model which uses non-parametric basis functions, thus, it could take into account any arbitrary time-domain information. We showed that aQHM is able to adjust by successive adaptations its basis functions to the local characteristics of the analyzed signal. Thus, the non-stationarity of the signal was adequately modeled. Consequently, frequency estimation becomes less biased and the signal representation more accurate.

Furthermore, aQHM suggests an adaptive algorithm for the decomposition of AM-FM signals. Initialization of aQHM was provided by QHM which acted as a frequency tracker (i.e. QHM acted as a Kalman filter). Results on synthetic AM-FM signals showed that the use of aQHM greatly reduce the estimation error of the AM-FM components of the signals. We showed that very fast frequency modulations can be estimated. Extended tests on voiced speech further validate the use of aQHM as a model which addresses efficiently the non-stationary character of the analyzed signals.

Since the application of aQHM in voiced speech analysis resulted in improvements at least in terms of SRER, we developed an A/S speech system based on aQHM. The A/S speech system decomposed speech into two parts; the deterministic part which accounted for the periodicities of the signal and the stochastic part which accounted for the aperiodicities of the signal. The deterministic part was then modeled by aQHM. Minor changes from the AM-FM decomposition algorithm were needed for the estimation of the deterministic part. The stochastic part was modeled as a time-modulated and frequency-modulated white noise. Results on resynthsis showed that the reconstructed speech signals were indistinguishable from the original signals for both male and female voices.

Finally, another application presented in this thesis concerns the voice quality assessment. We developed a method which is able to extract vocal tremor attributes from sustained vowels. Since vocal tremor estimation involves the estimation of the time-varying characteristics of speech signals, the proposed adaptive AM-FM decomposition algorithm was applied for the estimation of the instantaneous components of speech. The AM-FM decomposition algorithm was further applied for the extraction of the time-varying characteristics of vocal tremor, namely, modulation frequency and modulation level, from the estimated instantaneous components. Results on normophonic speakers, which have more stochastic vocal tremor characteristics compared to pathological subjects, showed that the suggested method is able to accurately estimate the modulations attributed to vocal tremor.

6.2 Future Research Directions

By no means this thesis is a complete presentation of the addressed problems. Presumably, there are issues that should be further investigated. One such issue is the better understanding of cQHM. Indeed, the convergence properties of cQHM should be further investigated as well the behavior of cQHM on multi-component chirp signals should be thoroughly tested. Moreover, not only QHM or cQHM but also any complex time-varying amplitude model contains frequency information, and obviously, the frequency information depends on the parametric model. Thus, a new class of models can be defined, investigate their properties and possibly used in applications.

Another issue for further research is to restate aQHM to a more general framework which for the author is as a post-processing step of any AM-FM demodulation algorithm. Indeed, in order to use aQHM, an estimate of the instantaneous phase is needed and not only QHM but also any AM-FM demodulation algorithm can be used. Thus, aQHM can be considered as a refinement of any demodulation algorithm used in other areas of signal processing. Furthermore, aQHM uses only the time-varying frequency information but it does not take into account any time-varying amplitude information. Hence, a further study on the possibility of adding time-varying amplitude information in aQHM should be performed.

As concerns the speech applications, the most interesting and challenging open question in speech synthesis is how to perform speech modifications using the developed A/S speech system based on aQHM. Since aQHM is able to represent voiced speech very accurately, we expect the quality of the modifications to be high. Yet, the modification algorithm has to be developed.

Finally, in voice quality assessment, the application of the developed vocal tremor extraction method to pathological cases should be investigated. Classification of the signals into physiological and pathological using the attributes estimated by the proposed method may be a further application. Moreover, vibrato signing could be analyzed by the proposed method, thus, another application may be to measure the modulations due to vibrato and use them as a learning tool.

Appendix A

Fast LS Computations

In sinusoidal representation, LS method for the estimation of complex amplitudes is slower compared to FFT-based approaches putting a limitation to its use in real-time applications. Moreover, improvements on the computation burden of QHM is crucial for its wide acceptance. Thus, in any case, improving the computation cost of LS method is very important.

To proceed, discrete-time SM is given for a frame by

$$h_s[n] = \sum_{k=-K}^{K} a_k e^{j2\pi f_k n/f_s} w[n], \qquad n = -N, ..., N \tag{A.1}$$

while the discrete form of QHM is given for a frame by

$$h_q[n] = \sum_{k=-K}^{K} (a_k + nb_k) e^{j2\pi f_k n/f_s} w[n], \qquad n = -N, ..., N \tag{A.2}$$

LS solution for the unknown complex amplitudes of SM is given by

$$\hat{\mathbf{a}} = (E_0^H W^H W E_0)^{-1} E_0^H W^H W \mathbf{s} = R_0^{-1} \mathbf{s_0} \tag{A.3}$$

while the LS solution for the complex amplitudes and complex slopes of QHM is given by

$$\begin{bmatrix} \hat{\mathbf{a}} \\ \hat{\mathbf{b}} \end{bmatrix} = ([E_0|E_1]^H W^H W [E_0|E_1])^{-1} [E_0|E_1]^H W^H W \mathbf{s} = \begin{pmatrix} R_0 & R_1 \\ R_1^H & R_2 \end{pmatrix}^{-1} \begin{bmatrix} \mathbf{s_0} \\ \mathbf{s_1} \end{bmatrix} \tag{A.4}$$

where $R_0 = E_0^H W^H W E_0$, $R_1 = E_0^H W^H W E_1$, $R_2 = E_1^H W^H W E_1$, $\mathbf{s_0} = E_0^H W^H W \mathbf{s}$ and $\mathbf{s_1} = E_1^H W^H W \mathbf{s}$. Obviously, SM is a sub-case of QHM, hence, we concentrate mostly on the

computation of the unknown parameters of QHM. Moreover, an important sub-case of both SM and QHM is when the frequencies are integer multiples of a fundamental frequency. In such a case, sub-matrices $R_m,\ m = 0, 1, 2$ are Toeplitz which results in faster element computation and faster matrix inversion through Levinson-type algorithms.

A.1 Computations

The improvements in the LS computation of QHM parameters is threefold. Firstly, the computation of the elements of $R_m,\ m = 0, 1, 2$ is done manually, thus, there is no need for matrix multiplication. Secondly, the computation of the elements of E_0 is speed-up using trigonometric identities. And, thirdly, fast inversion of the matrix is considered.

A.1.1 Faster Computation of $R_m,\ m = 0, 1, 2$

The elements of the submatrices R_0, R_1 and R_2 are given by

$$(R_m)_{ik} = \sum_{n=-N}^{N} w^2[n] n^m e^{j2\pi(f_k - f_i)n}, \qquad m = 0, 1, 2 \tag{A.5}$$

In order to do the summation of the above equations, we have to consider what kind of window is used. Typically, Hamming window is used but other options like rectangular or Hanning windows can be applied. These three windows are parametrized into a general class of windows given by

$$w_a[n] = (1 - a) + a cos(\pi n/N) \quad n = -N, ..., N - 1, N \tag{A.6}$$

Table A.1 shows the relationship between various windows and parameter a. As (A.5) asserts, the squared window is also necessary, thus,

$$w_a^2[n] = ((1 - a) + a cos(\pi n/N))^2 = a_0 + a_1(e^{j\pi n/N} + e^{-j\pi n/N}) + a_2(e^{j2\pi n/N} + e^{-j2\pi n/N}) \tag{A.7}$$

where $a_0 = (1 - a)^2 + a^2/2$, $a_1 = a(1 - a)$ and $a_2 = a^2/4$.

$a = 0$	Rectangular
$a = 0.5$	Hanning
$a = 0.46$	Hamming

Table A.1: Different values of a provides various windows.

Applying the squared window (A.7) to (A.5), we obtain

$$\begin{aligned}(R_m)_{ik} = a_0 \sum_{n=-N}^{N} n^m \left[e^{j2\pi(f_k-f_i)}\right]^n + a_1 \sum_{n=-N}^{N} n^m \left[e^{j2\pi(f_k-f_i+\frac{1}{2N})}\right]^n + a_1 \sum_{n=-N}^{N} n^m \left[e^{j2\pi(f_k-f_i-\frac{1}{2N})}\right]^n \\ + a_2 \sum_{n=-N}^{N} n^m \left[e^{j2\pi(f_k-f_i+\frac{1}{N})}\right]^n + a_2 \sum_{n=-N}^{N} n^m \left[e^{j2\pi(f_k-f_i-\frac{1}{N})}\right]^n\end{aligned} \tag{A.8}$$

thus, the elements of matrices R_i have values that are sums of special series. Standard mathematical identity about the sum of geometric series gives that $\sum_{n=0}^{N} \alpha^{\lambda n} = \frac{1-\alpha^{\lambda(N+1)}}{1-\alpha^{\lambda}}$. Taking the derivative with respect of λ, the elements of R_1 show up, thus, they can be computed without performing the summation. Similarly, taking one more derivative over *lambda*, the elements of R_2 can be computed. Hence, the elements of the matrices R_i are given by

$$\begin{aligned}(R_m)_{ik} = a_0 g_m\left(2\pi(f_k - f_i)\right) + a_1 g_m\left(2\pi(f_k - f_i + \frac{1}{2N})\right) + a_1 g_m\left(2\pi(f_k - f_i - \frac{1}{2N})\right) \\ + a_2 g_m\left(2\pi(f_k - f_i + \frac{1}{N})\right) + a_2 g_m\left(2\pi(f_k - f_i - \frac{1}{N})\right)\end{aligned} \tag{A.9}$$

where the auxiliary functions $g_0(x)$, $g_1(x)$ and $g_2(x)$ are given by

$$g_0(x) = \begin{cases} \frac{\sin((2N+1)x/2)}{\sin(x/2)}, & x \neq 0 \\ 2N+1, & x = 0 \end{cases} \tag{A.10}$$

$$g_1(x) = \begin{cases} j\frac{\sin(Nx)}{2\sin^2(x/2)} - jN\frac{\cos((2N+1)x/2)}{\sin(x/2)}, & x \neq 0 \\ 0, & x = 0 \end{cases} \tag{A.11}$$

$$g_2(x) = \begin{cases} \frac{N^2\cos((N+1)x)+(N+1)^2\cos(Nx)}{2\sin^2(x/2)} - \frac{\sin((2N+1)x/2)}{2\sin^3(x/2)}, & x \neq 0 \\ N(N+1)(2N+1)/3, & x = 0 \end{cases} \tag{A.12}$$

Finally, due to the fact that the computations of trigonometric functions are expensive, the computation of (A.9) can be speed up by considering the following identities

$$\cos(\theta+\delta) = \cos(\theta) - [\alpha\cos(\theta) - \beta\sin(\theta)] \tag{A.13}$$

$$\sin(\theta+\delta) = \sin(\theta) - [\alpha\sin(\theta) - \beta\cos(\theta)] \tag{A.14}$$

where α and β are precomputed coefficients given by $\alpha = 2\sin^2(\delta/2)$ and $\beta = \sin(\delta)$. Hence, the sines and cosines of one of the five terms in (A.9) are required and the remaining terms are

computed using (A.13) and (A.14).

A.1.2 Faster computation of E_0

Once again, the most time-consuming part for the computation of the elements of matrix E_0 is the computation of sines and cosines. Indeed, the elements of E_0 equals to $(E_0)_{kn} = e^{j2\pi f_k n/f_s} = \cos(2\pi f_k n/f_s) + j\sin(2\pi f_k n/f_s)$. Obviously, having computed matrix E_0, the elements of E_1 are given by $(E_1)_{nk} = n(E_0)_{nk}$. In this case, the computational speed up stems from the fact that the solution $z_k[n]$ of the following second-order difference equation

$$z_k[n] - 2\cos(2\pi f_k/f_s)z_k[n-1] + z_k[n-2] = 0, \qquad n = 3, 4, \ldots \tag{A.15}$$

with initial conditions $z_k[1] = \cos(2\pi f_k/f_s)$ ans $z_k[2] = \cos(4\pi f_k/f_s)$ is given by

$$z_k[n] = \cos(2\pi f_k/f_s n), \qquad n = 1, 2, \ldots \tag{A.16}$$

Similarly, using $z_k[1] = \sin(2\pi f_k/f_s)$ and $z_k[2] = \sin(4\pi f_k/f_s)$ as initial conditions, the difference equation has solution given by

$$z_k[n] = \sin(2\pi f_k/f_s n), \qquad n = 1, 2, \ldots \tag{A.17}$$

Thus, using the above iterative equation, the computation of each trigonometric function is replaced by one multiplication and two additions.

A.1.3 Step 3: Matrix Inversion

Up to now, no approximation or discretization was performed and the LS solution has no additional error. However, if we allow small errors, we can achieve faster inversion for the matrix by discarding elements away from the diagonal. This approximation is valid because sinusoids which are away from each other has little or no interference. Thus, if we keep K_0 diagonals, the inversion is speed up significantly.

A.2 Evaluation

In this Section, the performance of the improvements is tested in both complexity and execution time. Complexity is important to understand how the computation scales as the number of

components or the window duration is increased. On the other hand, execution time is important because it says exactly the time needed for the parameter estimation.

A.2.1 Complexity

Without any improvement, the total computation cost of LS estimation is $O(NK^2 + K^3 + NK)$, considering that W is diagonal. The term $O(NK^2)$ stems from the computation of R_m, $m = 0, 1, 2$, while the term $O(K^3)$ stems from the inversion of the matrix. The term $O(NK)$ stems from the computation of $\mathbf{s_0}$ and $\mathbf{s_1}$. After the first improvement, each element of R_m is computed in constant time, thus, reducing the overall complexity to $O(K^2 + K^3 + NK)$. Adding the second improvement, we achieve to compute the sequences $\sin(2\pi f_k/f_s n)$ and $\cos(2\pi f_k/f_s n)$ using approximately $8NK$ multiplications (2 multiplications per element of E_0). This operation does not improve the complexity, but as we will see shortly, it greatly reduce the execution time.

As concerns the matrix inversion, in the case of SM where the frequencies are harmonically-related, matrix R_0 has Toeplitz structure. Thus, its inversion through Levinson-type algorithms has complexity $O(K^2)$. In any other case, the matrix inversion remains of order $O(K^3)$. However, if we allowed the "diagonalization" of the matrix to be inverted, then, the complexity is reduced to $O(K_0^2 K)$. Totally, the complexity was reduced from $O((NK+K^2+N)K)$ to $O((K+K_0^2+N)K)$.

A.2.2 Execution Time

We test the performance of the computational improvements using synthetic signals. Synthetic signals are given by

$$s[n] = \sum_{k=1}^{K} \cos(2\pi f_k/f_s n), \qquad n = -N, ..., N \tag{A.18}$$

where $f_s = 16000$ which is a typical value for the sampling frequency. Parameters K and N take values $K = 10, 20, \ldots, 50$ and $N = 150, 175, \ldots, 250$, respectively. For the special case of harmonic frequencies, $f_k = kf_0, k = 0, \ldots K$, fundamental frequency, f_0, is chosen uniformly from the interval $[80, 280]Hz$ while for the general case, we choose the frequencies uniformly in $[80, f_s/2 - 80]Hz$, under the conditions that every two frequencies should be at least $80Hz$ apart and that $f_{k-1} < f_k$.

The computer used for the experiments was equipped with: Intel Core 2 6600 CPU @ 2.4 GHz and 2GB RAM. Note that only one CPU was used to ensure accuracy of the results. The operating systems was Windows XP Professional 32 bit. In Tables A.2 and A.3, the average execution time for each improvement is shown. Note that the average was taken over 1000 runs.

In Table A.3 the Signal-to-Noise Ratio (SNR) is also reported since approximations take place. Obviously, the computational gain up to 2nd improvement is 77% for the harmonic case of SM while it is 66% for the general case of SM. The respective computational gain for QHM is 34% for the harmonic case while it is 30% for the general case. The difference in the performance between the harmonic and the general case stems from the fact that matrices $R_m,\ m = 0, 1, 2$ are Toeplitz in the former case. When the third improvement takes place, the computational gain is further improved without significant loss of SNR from turning $R_m,\ m = 0, 1, 2$ into diagonal matrices.

	SM		QHM	
	kf_0	f_k	kf_0	f_k
No improvement	4.205 ms	4.257 ms	10.871 ms	10.929 ms
1st improvement	2.467 ms	4.227 ms	4.930 ms	8.618 ms
2nd improvement	0.967 ms	2.792 ms	3.653 ms	7.676 ms

Table A.2: Average CPU time of the first and second improvement.

	SM			
	CPU time		SNR	
	kf_0	f_k	kf_0	f_k
No improvement	4.205 ms	4.257	278 dB	272 dB
3rd improvement (3)	0.810 ms	0.923 ms	88 dB	83 dB
3rd improvement (5)	0.813 ms	0.955 ms	105 dB	109 dB
3rd improvement (7)	0.839 ms	1.024 ms	117 dB	124 dB

Table A.3: Average CPU time and SNR of the third improvement. The number in the parentheses denotes how many diagonals have been used.

Appendix B

Relation of iQHM with Gauss-Newton method

This Appendix shows the relation between iQHM and Gauss-Newton (GN) method for a mono-component signal. To remind, a mono-component stationary signal is written in discrete domain as

$$s[n] = A_1 e^{j2\pi f_1 n/f_s} w[n], \qquad n = -N, ..., N \tag{B.1}$$

where A_1 is the complex amplitude, f_1 is the frequency, f_s is the sampling frequency while $w[n]$ is a symmetric window. We will show that the estimation provided by iQHM is equivalent with "sequential GN method.

B.1 iQHM Method

In iQHM, an initial estimate of the frequency (denoted $f_1^{(0)}$) is provided. Then, at the ith step ($i = 0, 1, ...$), complex amplitude, $a_1^{(i)}$, and complex slope, $b_1^{(i)}$ are computed by $a_1^{(i)} = \frac{\sum_{n=-N}^{N} w^2[n]s[n]e^{-j2\pi f_1^{(i)} n}}{\sum_{n=-N}^{N} w^2[n]}$ and $b_1^{(i)} = \frac{\sum_{n=-N}^{N} nw^2[n]s[n]e^{-j2\pi f_1^{(i)} n}}{\sum_{n=-N}^{N} n^2 w^2[n]}$. The amplitude of the signal is given by

$$A_1^{(i)} = a_1^{(i)} \tag{B.2}$$

while frequency is updated as

$$f_1^{(i+1)} = f_1^{(i)} + \frac{\rho_{2,1}^{(i)}}{2\pi} \tag{B.3}$$

where $\rho_{2,1}^{(i)}$ can be written as

$$\rho_{2,1}^{(i)} = \frac{\mathcal{R}\{a_1^{(i)}\}\mathcal{I}\{b_1^{(i)}\} - \mathcal{I}\{a_1^{(i)}\}\mathcal{R}\{b_1^{(i)}\}}{|a_1^{(i)}|^2} = \mathcal{R}\{\frac{-j\bar{a}_1^{(i)}b_1^{(i)}}{|a_1^{(i)}|^2}\} = \mathcal{R}\{\frac{-jb_1^{(i)}}{a_1^{(i)}}\} \tag{B.4}$$

where $\mathcal{R}\{\cdot\}$ and $\mathcal{I}\{\cdot\}$ denote the real and imaginary parts of a complex number while $\bar{\cdot}$ denote conjugation. Thus, the update equation for the frequency becomes

$$f_1^{(i+1)} = f_1^{(i)} - \mathcal{R}\{\frac{jb_1^{(i)}}{2\pi a_1^{(i)}}\} \tag{B.5}$$

B.2 GN Method

On the other hand, given an initial estimate of the amplitude, $A_1^{(0)}$, and of the frequency, $f_1^{(0)}$, GN method has the following updating step ($i = 0, 1, ...$)

$$\begin{pmatrix} A_1^{(i+1)} \\ f_1^{(i+1)} \end{pmatrix} = \begin{pmatrix} A_1^{(i)} \\ f_1^{(i)} \end{pmatrix} + (J^H J)^{-1} J^H r \tag{B.6}$$

where J is a $2N+1 \times 2$ matrix given by

$$J = \begin{pmatrix} w[-N]e^{j2\pi f^{(i)}(-N)} & w[-N]A_1^{(i)}j2\pi(-N)e^{j2\pi f^{(i)}(-N)} \\ \vdots & \vdots \\ w[N]e^{j2\pi f^{(i)}N} & w[N]A_1^{(i)}j2\pi N e^{j2\pi f^{(i)}N} \end{pmatrix} \tag{B.7}$$

and r is a $2N+1 \times 1$ vector given by

$$r = \begin{pmatrix} w[-N]\left(s[-N] - A_1^{(i)}e^{j2\pi f^{i}(-N)}\right) \\ \vdots \\ w[N]\left(s[N] - A_1^{(i)}e^{j2\pi f^{(i)}N}\right) \end{pmatrix} \tag{B.8}$$

Then, (B.6) equals to

$$\begin{pmatrix} A_1^{(i+1)} \\ f_1^{(i+1)} \end{pmatrix} = \begin{pmatrix} A_1^{(i)} \\ f_1^{(i)} \end{pmatrix} + \begin{pmatrix} \sum_{n=-N}^{N} w^2[n] & 0 \\ 0 & |A_1^{(i)}|^2(2\pi)^2\sum_{n=-N}^{N} n^2 w^2[n] \end{pmatrix}^{-1} \begin{pmatrix} \sum_{n=-N}^{N} w^2[n]s[n]e^{-j2\pi f_1^{(i)}n} - A_1^{(i)}\sum_{n=-N}^{N} w^2[n] \\ \mathcal{R}\left\{-j2\pi\bar{A}_1^{(i)}\sum_{n=-N}^{N} nw^2[n]s[n]e^{-j2\pi f_1^{(i)}n}\right\} \end{pmatrix} \tag{B.9}$$

which leads to

$$\begin{pmatrix} A_1^{(i+1)} \\ f_1^{(i+1)} \end{pmatrix} = \begin{pmatrix} \frac{\sum_{n=-N}^{N} w^2[n]s[n]e^{-j2\pi f_1^{(i)} n}}{\sum_{n=-N}^{N} w^2[n]} \\ f_1^{(i)} - \mathcal{R}\left\{ \frac{j\sum_{n=-N}^{N} n w^2[n]s[n]e^{-j2\pi f_1^{(i)} n}}{2\pi A_1^{(i)} \sum_{n=-N}^{N} n^2 w^2[n]} \right\} \end{pmatrix} \tag{B.10}$$

Note that the real operator is applied on the frequency update equation because frequency is a real parameter.

B.3 Relation Between the Two Methods

Using the parameters from iQHM, GN iteration becomes

$$\begin{pmatrix} A^{(i+1)} \\ f_1^{(i+1)} \end{pmatrix} = \begin{pmatrix} a_1^{(i)} \\ f_1^{(i)} - \mathcal{R}\left\{ \frac{j b_1^{(i)}}{2\pi a_1^{(i-1)}} \right\} \end{pmatrix} \tag{B.11}$$

which shows that there is a delay of one step on the estimation of the complex amplitude of the signal. However, if the estimation of the sinusoidal parameters in GN method is performed sequentially, i.e. firstly update the complex amplitude given the frequency of the previous step and then update the frequency given the updated complex amplitude, then, iQHM and GN method are equivalent.

Bibliography

[1] 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.

[2] T. F. Quatieri. *Speech Signal Processing*. Prentice Hall, Signal Processing Series, 2002.

[3] P. Guillaume. *Music And Acoustics: From Instrument to Computer*. John Wiley & Sons, 2006.

[4] W. W. L. Au and M. C. Hastings. *Principles of Marine Bioacoustics*. Springer New York, 2008.

[5] H. L. van Trees. *Detection, Estimation, and Modulation Theory: Part I*. John Wiley and Sons, 2nd edition, 2001.

[6] A. W. Rihaczek. *Principles of High-Resolution Radar*. Artech House Radar Library, 1985.

[7] J. Proakis and M. Salehi. *Digital Communications*. McGraw-Hill, 5th edition, 2007.

[8] 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.

[9] X. Serra. *A System for Sound Analysis/Transformation/Synthesis based on a Deterministic Plus Stochastic Decomposition*. PhD thesis, Stanford University, 1989.

[10] M. Goodwin and M. Vetterli. Time-Frequency Signal Models for Music Analysis, Transformation, and Synthesis. In *Proc. of the IEEE-SP Inter. Sym. on Time-Frequency and Time-Scale Analysis*, 1996.

[11] R. J. McAulay and T. F. Quatieri. Sinusoidal Coding. In W. B. Kleijn and K. K. Paliwal, editors, *Speech Coding and Synthesis*. Elsevier, 1995.

[12] Sassan Ahmadi and A. S. Spanias. Low bit-rate speech coding based on an improved sinusoidal model. *Speech Communication*, 34:369–390, 2001.

[13] T.F. Quatieri and R.J. McAulay. Shape-Invariant Time-Scale and Pitch Modifications of Speech. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 40:497–510, 1992.

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

[15] M. Abe and J.O. Smith III. Design Criteria for the Quadratically Interpolated FFT Method (I): Bias due to Interpolations. Technical Report STAN-M-114, Stanford University, California, Oct 2004.

[16] M. Abe and J.O. Smith III. Design Criteria for the Quadratically Interpolated FFT Method (II): Bias due to Interfering Components. Technical Report STAN-M-115, Stanford University, California, Oct 2004.

[17] M. Abe and J.O. Smith III. Design Criteria for the Quadratically Interpolated FFT Method (III): Bias due to Amplitude and Frequency Modulation. Technical Report STAN-M-116, Stanford University, California, Oct 2004.

[18] M. Abe and J.O. Smith III. CQIFFT: Correcting Bias in a Sinusoidal Parameter Estimator based on Quadratic Interpolation of FFT Magnitude Peaks. Technical Report STAN-M-117, Stanford University, California, Oct 2004.

[19] P. Hedelin. A Tone-oriented Voice-excited Vocoder. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, pages 205–208, 1981.

[20] L.B. Almeida and J.M. Tribolet. Nonstationary Spectral Modeling of Voiced Speech. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 31:374–390, 1983.

[21] X. Serra and J. Smith III. Spectral Modeling Synthesis: A Sound Analysis/Synthesis System based on a Deterministic plus Stochastic Decomposition. *Computer Music Journal*, 12:12–24, 1990.

[22] E. B. George and M. Smith. A new Speech Coding Model based on a Least-Squares Sinusoidal Representation. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, pages 1641–1644, Apr 1987.

[23] E. B. George and M. Smith. Analysis-by-Synthesis Overlap-Add Sinusoidal Modeling Applied to the Synthesis of Musical Tones. *Journal of the Audio Engineering Society*, 40:497–516, 1992.

[24] E. B. George and M. Smith. Speech Analysis/Synthesis and Modification using an Analysis-by-Synthesis/Overlap-Add Sinusoidal Model. *IEEE Trans. on Speech and Audio Processing*, 5:389–406, 1997.

[25] T. Kailath. *Linear Least-Squares Estimation*. Halsted Press, 1977.

[26] S. M. Kay. *Funtamentals of Statistical Signal Processing: Estimation Theory*. Prentice-Hall, Englewood Cliffs, NJ, 1993.

[27] V. F. Pisarenko. The retrieval of harmonics from a covariance function. *Geophysics Journal of Royal Astronomy Society*, 33:347–366, 1973.

[28] D. T. Tufts and R. Kumaresan. Estimation of Frequencies of Multiple Sinusoids: Making the Linear Predition Perform like Maximum Likelihood. *Proc. of IEEE*, 70:975–989, Sep 1982.

[29] P. Stoica, R. L. Moses, T. Soderstrom, and B. Friedlander. Maximum Likelihood Estimation of the Parameters of Multiple Sinusoids from Noisy Measurements. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 37, Mar 1989.

[30] P. Stoica, H. Li, and J. Li. Amplitude Estimation of Sinusoidal Signals: Survey, New Results, and an Application. *IEEE Trans. on Signal Processing*, 48:338–352, Feb 2000.

[31] P. Stoica and A. Nehorai. Statistical Analysis of two Nonlinear Least-Suares Estimators of Sine-wave Parameters in the Colored-Noise case. *Circuits Systems Signal Process*, 8, 1989.

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

[33] R. Badeau, R. Boyer, and B. David. Eds Parametric Modeling and Tracking of Audio Signals. In *Proc. of Int. Conf. on Digital Audio Effects*, pages 139–144, 2002.

[34] Kris Hermus, W. Verhelst, P. Lemmerling, P. Wambacq, and S. van Huffel. Perceptual audio modeling with exponentially damped sinusoids. *Signal Processing*, 85:163–176, 2005.

[35] M. G. Christensen, S. V. Andersen, and S. H. Jensen. Amplitude modulated sinusoidal models for audio modeling and coding. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 1334–1342. Springer Berlin, 2003.

[36] B. Friedlander and J.M. Francos. Estimation of Amplitude and Phase Parameters of Multicomponent Signals. *IEEE Trans. on Signal Processing*, 43, Apr 1995.

[37] M. Jabloun, N. Martin, F. Leonard, and M. Vieira. Estimation of the instantaneous amplitude and frequency of non-stationary short-time signals. *Signal Processing*, 88:1636–1655, 2008.

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

[39] J. S. Marques and L. B. Almeida. Frequency-varying Sinusoidal Modelong of Speech. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 39:763–765, 1989.

[40] M. Abe and J.O. Smith III. AM/FM Estimation for Time-varying Sinusoidal Modeling. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, pages III 201–204, Philadelphia, 2005.

[41] A. Robel. Parameter Estimation for Linear AM/FM Sinusoids using Frequency Domain Demodulation. In *Signal and Image Processing*, pages 162–166, 2007.

[42] S. Peleg and B. Friedlander. The Discrete Polynomial-Phase Transform. *IEEE Trans. on Signal Processing*, 43:1901–1914, Aug 1995.

[43] A. Francos and M. Porat. Analysis and Synthesis of Multicomponent Signals using Positive Time-Frequency Distributions. *IEEE Trans. on Signal Processing*, 47:493–504, 1999.

[44] M. Betser, P. Collen, G. Richard, and B. David. Estimation of Frequency for AM/FM Models using the Phase Vocoder Framework. *IEEE Trans. on Signal Processing*, 56:505–517, Feb 2008.

[45] F. Auger and P. Flandrin. Improving the Readability of Time-Frequency and Time-Scale Representations by the Reassignment Method. *IEEE Trans. on Signal Processing*, 43:1068–1089, 1995.

[46] S. A. Fulop and K. Fitz. Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. *J. Acoust. Soc. Am.*, 119:360–371, 2006.

[47] S. Mann and S. Haykin. The Chirplet Transform: A Generalization of Gabor's Logon Transform. *Proc. Vision Interface*, pages 205–212, 1991.

[48] D. Mihovilovic and R. N. Bracewel. Adaptive Chirplet Representation of Signals in the Time-Frequency Plane. *IEEE Electronic Letters*, 27:1159–1161, 1991.

[49] L. B. Almeida. The fractional Fourier Transform and Time-Frequency Representations. *IEEE Trans. on Signal Processing*, 42:3084–3091, 1994.

[50] H. M. Ozaktas, Z. Zalevsky, and M. A. Kutay. *The Fractional Fourier Transform with Applications in Optics and Signal Processing*. John Wiley & Sons, 2001.

[51] L. Weruaga and M. Kepesi. The Fan-chirp Transform for Non-stationary Harmonic Signals. *Signal Processing*, 87:1504–1522, 2007.

[52] R. Dunn and T. F. Quatieri. Sinewave Analysis/Synthesis based on the Fan-Chirp Transform. In *Workshop on Appl. of Signal Proc. to Audio and Acoustics (WASPAA)*, pages 16–19, Oct 2007.

[53] L. Cohen. *Time-Frequency Analysis*. Prentice Hall, New York, 1995.

[54] A. V. Oppenheim and R. W. Schafer with J. R. Buck. *Discrete-Time Signal Processing*. Signal Processing Series. Prentice-Hall, 2nd edition, 1989.

[55] S. L. Hahn. *Hilbert Transforms in Signal Processing*. Artech House Publishers, 1996.

[56] P. Maragos, J. Kaiser, and T. Quatieri. On Separating Amplitude from Frequency Modulations using Energy Operators. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, pages 1–4, San Francisco, USA, Mar 1992.

[57] P. Maragos, J. F. Kaiser, and T. F. Quatieri. On Amplitude and Frequency Demodulation using Energy Operators. *IEEE Trans. on Signal Processing*, 41:1532–1550, 1993.

[58] A. Potamianos and P. Maragos. A comparison of the energy operator and Hilbert transform approaches for signal and speech demodulation. *Signal Processing*, 35:95–120, 1994.

[59] J. K. Gupta, S.C. Sekhar, and T. V. Sreenivas. Performance Analysis of AM–FM Estimators. In *TENCON 2003*, pages 954–958, Oct 2003.

[60] W. C. Lindsey and C. M. Chie. A survey of digital phase-locked loops. *IEEE Proc.*, pages 410–431, 1981.

[61] W. C. Pai and P. C. Doerschuk. Statistical AM-FM Models, Extended Kalman Filter Demodulation, Cramer-Rao Bounds and Speech Analysis. *IEEE Trans. on Signal Processing*, 48:2300–2313, Aug 2000.

[62] L. Cohen. What is a multicomponent signal. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, pages 113–116, 1992.

[63] L. Cohen and C. Lee. Instantaneous frequency, its standard deviation and multicomponent signals. *Proc. SPIE*, 975:186–208, 1988.

[64] B. Picinbono. On Instantaneous Amplitude and Phase of Signals. *IEEE Trans. on Signal Processing*, 45:552–560, 1997.

[65] P. Maragos, J. F. Kaiser, and T. F. Quatieri. Energy Separations in Signal Modulations with Application to Speech Analysis. *IEEE Trans. on Signal Processing*, 41:3024–3051, 1993.

[66] T. F. Quatieri T. E. Hanna and G. C. O'Leary. AM-FM Separation using Auditory-Motivated Filters. *IEEE Trans. on Speech and Audio Processing*, 5:465–480, 1997.

[67] J. H. L. Hansen and D. T. Chappel. An Auditory-based Distortion Measure with Application to Concatenative Speech Synthesis. *IEEE Trans. on Speech and Audio Processing*, 6:489–495, Sep 1998.

[68] J. L. Flanagan and R. M. Golden. Phase vocoder. *Bell Syst. Tech Journal*, 45:1493–1509, Nov 1966.

[69] B. Santhanam and P. Maragos. Multicomponent AM-FM demodulation via periodicity-based algebraic separation and energy-based demodulation. *IEEE Trans. on Communications*, 48:473–490, 2000.

[70] F. Gianfelici, G. Biagetti, P. Crippa, and C. Turchetti. Multicomponent AM-FM Representations: An Asymptotically Exact Approach. *IEEE Trans. on Audio, Speech and Language Processing*, 15:823–837, 2007.

[71] Y. Pantazis, O. Rosec, and Y. Stylianou. On the Properties of a Time-Varying Quasi-Harmonic Model of Speech. In *Interspeech*, pages 1044–1047, Brisbane, Sep 2008.

[72] Y. Pantazis, O. Rosec, and Y. Stylianou. Iterative Estimation of Sinusoidal Signal Parameters. *IEEE Signal Processing Letters*, 17(5):461.

[73] Y. Pantazis, O. Rosec, and Y. Stylianou. Chirp Rate Estimation of Speech based on a Time-varying Quasi-Harmonic Model. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, 2009.

[74] Y. Pantazis, O. Rosec, and Y. Stylianou. Adaptive AM-FM Signal Decomposition with Application to Speech Analysis. *IEEE Trans. on Audio, Speech and Language Processing*, submitted.

[75] Y. Pantazis and Y. Stylianou. Improving the Modeling of the Noise Part in the Harmonic plus Noise Model of Speech. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, 2008.

[76] Y. Pantazis, M. Koutsogiannaki, and Y. Stylianou. A Novel Method for the Extraction of Vocal Tremor. In *MAVEBA*, Florence, 2009.

[77] J. Laroche. A new Analysis/Synthesis System of Musical Signals using Prony's Method. Application to Heavily Damped Percussive Sounds. In *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, pages 2053–2056, Glasgow, UK, May 1989.

[78] J.-M. Valin, D. V. Smith, C. Montgomery, and T. B. Terriberry. An iterative linearised solution to the sinusoidal parameter estimation problem. *Computers & Electrical Engineering*, In Press, Corrected Proof, 2008.

[79] D. C. Rife and R. R. Boorstyn. Single-tone Parameter Estimation from Discrete-time Observations. *IEEE Trans. on Information Theory*, 20:591–598, 1974.

[80] D. Griffin and J. Lim. Multiband-Excitation Vocoder. *IEEE Trans. on Acoustics, Speech and Signal Processing*, 36:236–243, 1988.

[81] Y. Stylianou, J. Laroche, and E. Moulines. High-Quality Speech Modification based on a Harmonic + Noise Model. *Proc. EUROSPEECH*, 1995.

[82] C. d'Alessandro, V. Darsinos, and B. Yegnanarayana. Effectiveness of a periodic and aperiodic decomposition method for analysis of voice sources. *IEEE Trans. on Speech and Audio Processing*, 6:12–23, Jan 1998.

[83] G. Bailly. Accurate estimation of sinusoidal parameters in an Harmonic+Noise Model for speech synthesis, 1999.

[84] A. McCree. A 14kb/s Wideband Speech Coder with a Parametric Highband Model. pages II1153–II1156, Instambul, 2000.

[85] H. Kawahara. Speech Representation and Transformation using adaptive Interpolation of Weighted Spectrum: Vocoder revisted. *Proc. of IEEE Int. Conf. on Audio, Speech and Signal Processing (ICASSP)*, 2:1303–1306, Apr 1997.

[86] I. R. Titze. Motor and Sensory Components of a Feedback Control Model of Fundamental Frequency. *Producing Speech: Contemporaty Issues*, pages 309–320, 1995.

[87] H.J. Freund. Central Rhythmicities in Moter Control and its Pertubances. In L. Resning, U. Heiden, and M.C. Mackey, editors, *Temporal Disorder in Human Oscillatory Systems*, pages 79–82. Springer, Berlin, 1987.

[88] J. Schoentgen. Modulation Frequency and Modulation Level owing to Vocal Microtremor. *J. Acoust. Soc. Am.*, pages 690–700, Aug 2002.

[89] C.A. Meeuwis and E.A. Baarsma. Essential (Voice) Tremor. *Clinical Otolaryngology*, 5, 1985.

[90] L.J. Findley and M.A. Gresty. Head, Face and Voice Tremor. In J. Jankovic and E. Tolosa, editors, *Facial dyskinesias: Advanses in neurology*, pages 239–253. New York: Raven, 1988.

[91] W. Winholtz and L. Ramig. Vocal Tremor Analysis with the Vocal Demodulator. *Journal of Hearing Research*, 35:562–573, 1992.

[92] A. Savitzky and M.J.E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36:1627–1639, 1964.

[93] J. Steinier, Y. Termonia, and J. Deltour. Comments on smoothing and differentiation of data by simplified least square procedure. *Analytical Chemistry*, 44:1906–1909, 1972.