University of Crete
Department of Computer Science

FO.R.T.H.
Institute of Computer Science

Speech Analysis/Synthesis Using an Adaptive Harmonic Model

(MSc. Thesis)

Gnostothea Veroniki Morfi

Thesis submitted in partial fulfillment of the requirements for the

Masters' of Science degree in Computer Science

February 2015

University of Crete
Computer Science Department

Speech Analysis/Synthesis Using an Adaptive Harmonic Model

Thesis submitted by
Gnostothea Veroniki Morfi
in partial fulfillment of the requirements for the
Masters' of Science degree in Computer Science

THESIS APPROVAL

Author: ______________________
Gnostothea Veroniki Morfi

Committee approvals: ______________________
Athanasios Mouchtaris
Assistant Professor, Thesis Supervisor

____
Panagiotis Tsakalidis
Professor, Committee Member

____
Georgios Tziritas
Professor, Committee Member

Departmental approval: ______________________
Antonis Argyros
Professor, Director of Graduate Studies

Heraklion, February 2015

Abstract

A speech production model that views speech as the result of passing a glottal excitation waveform through a time-varying linear filter (the latter modeling the resonant characteristics of the vocal tract) is widely used in digital speech signal processing. In many speech applications, two possible states of the glottal excitation can be assumed: voiced or unvoiced. Voice models often split the speech spectrum into these two (or even more) voiced/unvoiced frequency bands using respective cutoff frequencies. Voiced speech is usually modeled deterministically in the lower frequencies, while a stochastic approach is used for the upper frequency part. A so-called Maximum Voiced Frequency separates the deterministic and stochastic parts. However, it can be observed from the actual voice production mechanisms that the amplitude spectrum of the voice source decreases smoothly without any abrupt frequency changes that would justify such a classification of the spectrum in deterministic and stochastic components. Accordingly, it becomes a struggle for multiband models to estimate these cutoff frequencies. Consequently, artifacts produced by multiband methods can degrade the perceived quality. Moreover, the Fan Chirp Transformation (FChT), which uses a linear frequency basis adapted to the nonstationarities of the speech signal, has demonstrated that harmonicity is present at frequencies higher than those usually considered as voiced based on the Discrete Fourier Transform (DFT). This motivates alternative models which are based on a full-band modeling approach.

Sinusoidal and harmonic models aim to represent the speech signal with a set of parameters such as frequencies, amplitudes and phases. The accuracy and precision of the model parameters are key issues. All voice models have to be both precise and fast in order to represent the speech signal adequately and be able to process large amounts of data in a reasonable amount of time. So far, the Sinusoidal Model (SM), where the glottal excitation is represented as a sum of sine waves, has been widely used for many applications such as speech analysis, coding and modifications. However, as we show in our evaluations in this thesis, the estimated parameters are not as accurate as the ones computed by harmonic models. The adaptive Quasi-Harmonic Model (aQHM) has been proposed as an alternative and more adaptive method for speech analysis, that uses some of the attributes of the harmonicity of a signal. The aQHM offers even more flexibility than the FChT by using a set of adaptive non-linear basis functions. However, due to the assumption made by aQHM, that the initial frequency tracks can have a confined error, a frequency matching problem may occur. Hence, neither method is very suitable for full-band modeling of a speech signal.

Harmonic models were initially designed for representation of the deterministic part of the speech, but, as implied by the FChT, the need of a cutoff frequency limit is questionable. Thus, exploiting the properties of aQHM, the full-band adaptive

Harmonic Model (aHM) along with its corresponding algorithms for the estimation of harmonics up to the Nyquist frequency has been proposed. The aHM model uses the Least Squares (LS) solution in the Adaptive Iterative Refinement (AIR) algorithm in order to properly estimate a refinement of the f_0 curve without the problems caused by frequency errors. Even though aHM-AIR using LS allows for a robust estimation of the harmonic components, it lacks the computational efficiency that would make its use convenient for large databases, due to the use of the LS solution.

In this thesis, a Peak-Picking (PP) approach is suggested as a substitution to the LS solution used by the AIR algorithm. In order to integrate the adaptivity scheme of aHM in the PP approach, an adaptive Discrete Fourier Transform (aDFT), whose frequency basis can fully follow the variations of the f_0 curve, is also proposed. In order to evaluate the performance of the proposed method, the computational time has been calculated and an average time reduction of almost four times has been shown when comparing the proposed improvements to the original LS-based aHM-AIR algorithm. Additionally, the quality of the re-synthesis is preserved compared to the aHM-AIR using LS. With the use of Signal-To-Reconstruction-Error (SRER) and Perceptual Evaluation of Speech Quality (PESQ), we show that the speech reconstructed using aHM-AIR with PP and aDFT retains the quality of aHM-AIR using LS. Finally, formal listening tests show that the speech reconstructed by aHM-AIR with PP and aDFT is very similar to the one reconstructed by aHM-AIR using LS.

Περίληψη

Ένα μοντέλο παραγωγής ομιλίας το οποίο θεωρεί την ομιλία σαν το αποτέλεσμα του φιλτραρίσματος μιας κυματομορφής της γλωττιδικής διέγερσης από ένα χρονικά μεταβλητό γραμμικό φίλτρο το οποίο μοντελοποιεί τα κύρια χαρακτηριστικά της φωνητικής οδού χρησιμοποιείται ευρέως στην ψηφιακή επεξεργασία σημάτων ομιλίας. Σε πολλές εφαρμογές φωνής, δύο πιθανές καταστάσεις μπορούν να θεωρηθούν: η έμφωνη και η άφωνη. Τα μοντέλα φωνής συχνά διαχωρίζουν το φάσμα της ομιλίας σε αυτές τις δύο (ή ακόμη και περισσότερες) έμφωνες/άφωνες συχνοτικές ζώνες με τη χρήση ορίων στην συχνότητα. Ο έμφωνος λόγος μοντελοποιείται συνήθως ντετερμινιστικά στις χαμηλότερες συχνότητες, ενώ μια στοχαστική προσέγγιση χρησιμοποιείται για το ανώτερο μέρος των συχνοτήτων. Η Μέγιστη Έμφωνη Συχνότητα χωρίζει τα δύο αυτά μέρη. Ωστόσο, μπορεί να παρατηρηθεί από τους πραγματικούς μηχανισμούς παραγωγή φωνής ότι το φάσμα πλάτους της πηγής ελαττώνεται ομαλά χωρίς κάποια απότομη αλλαγή στην συχνότητα. Αναλόγως, χρειάζεται μεγάλη προσπάθεια από τη μεριά των μοντέλων πολλαπλών ζωνών για τον υπολογισμό αυτών τον ορίων. Συνεπώς, οι αλλοιώσεις που παράγονται από τις μεθόδους πολλαπλών ζωνών μπορούν να υποβαθμίσουν την ποιότητα μοντελοποίησης. Επιπλέον, ο μετασχηματισμός Fan Chirp (FChT), ο οποίος χρησιμοποιεί μια γραμμική βάση συχνοτήτων προσαρμοσμένη στις μη-στατικότητες του σήματος της φωνής, έχει επιδείξει αρμονικότητα σε υψηλότερες συχνότητες από αυτές που παρατηρούνται συνήθως από το μετασχηματισμό Fourier (DFT). Συνεπώς, μια προσέγγιση πλήρους ζώνης είναι επιθυμητή.

Τα ημιτονοειδή και τα αρμονικά μοντέλα στοχεύουν στην αναπαράσταση ενός σήματος φωνής με ένα σετ από παραμέτρους όπως συχνότητες, πλάτη και φάσεις. Η ακρίβεια αυτών των παραμέτρων του μοντέλου είναι ένα βασικό ζήτημα. Όλα τα μοντέλα φωνής πρέπει να είναι και ακριβή και γρήγορα έτσι ώστε να αναπαριστούν το σήμα φωνής επαρκώς και να είναι ικανά να επεξεργάζονται μεγάλη ποσότητα δεδομένων σε ένα λογικό χρονικό πλαίσιο. Ως τώρα, το ημιτονοειδές μοντέλο (SM), όπου η γλωττιδική διέγερση αναπαρίσταται σαν το άθροισμα ημιτονοειδών κυμάτων, χρησιμοποιείται ευρέως σε πολλές εφαρμογές όπως ανάλυση φωνής, κωδικοποίηση και τροποποίηση φωνής. Ωστόσο, όπως δείχνουμε στις αξιολογήσεις αυτής της εργασίας, οι παράμετροι που υπολογίζονται από το SM δεν είναι τόσο ακριβείς όσο αυτές που υπολογίζονται από τα αρμονικά μοντέλα. Ακόμη, το προσαρμοστικό Σχεδόν-Αρμονικό μοντέλο (aQHM) έχει προταθεί σαν μία εναλλακτική και πιο προσαρμοστική μέθοδος ανάλυσης φωνής, η οποία χρησιμοποιεί μερικές από τις ιδιότητες τις αρμονικότητας των σημάτων. Το aQHM παρέχει περισσότερη ευελιξία από το FChT χρησιμοποιώντας ένα σετ μη-γραμμικών συναρτήσεων βάσης. Παρόλα αυτά, λόγω της υπόθεσης της aQHM, ότι το αρχικό σφάλμα των συχνοτήτων είναι περιορισμένο, μπορεί να προκληθεί σφάλμα στην αντιστοίχηση των συχνοτήτων. Ως εκ τούτου, καμία από τις μεθόδους δεν είναι κατάλληλη για μοντελοποίηση πλήρους φάσματος ενός σήματος φωνής.

Τα αρμονικά μοντέλα είχαν σχεδιαστεί αρχικά για την αναπαράσταση του ντετερμινιστικού μέρους της ομιλίας, αλλά, όπως υποδηλώνεται από την FChT, η χρήση ενός ορίου συχνότητας είναι αμφισβητήσιμη. Ως εκ τούτου, αξιοποιώντας τις ιδιότητες της aQHM, το προσαρμοστικό Αρμονικό Μοντέλο (aHM) πλήρους ζώνης μαζί με τους αντίστοιχους αλγόριθμους για τον υπολογισμό των αρμονικών μέχρι την συχνότητα Nyquist έχει προταθεί. Το aHM μοντέλο χρησιμοποιεί την λύση των Ελάχιστων Τετραγώνων (LS) στον Προσαρμοστικό Επαναληπτικό αλγόριθμο Αναμόρφωσης (AIR) έτσι ώστε να γίνει μια σωστή εκτίμηση της αναμόρφωσης της καμπύλης f_0 χωρίς τα προβλήματα λόγω σφαλμάτων στην συχνότητα. Αν και η aHM-AIR που χρησιμοποιεί την μέθοδο LS επιτρέπει μια εύρωστη εκτίμηση των αρμονικών συνιστωσών, εξαιτίας της χρήσης της LS, της λείπει η υπολογιστική αποδοτικότητα η οποία θα έκανε την χρήση της ιδανική για μεγάλες βάσεις δεδομένων.

Στην εργασία αυτή, μια μέθοδος επιλογής κορυφών (PP) προτείνεται ως αντικατάσταση της LS στον AIR αλγόριθμο. Για να ενσωματωθεί η προσαρμοστικότητα του προσαρμοστικού Αρμονικού Μοντέλου στην PP προσέγγιση, προτείνεται επιπλέον ένας προσαρμοστικός Διακριτός Μετασχηματισμός Fourier (aDFT), του οποίου η συχνοτική βάση μπορεί να ακολουθήσει πλήρως τις εναλλαγές της f_0 καμπύλης. Για να γίνει η αξιολόγηση της απόδοσης της προτεινόμενης μεθόδου, μετρήσαμε τον υπολογιστικό χρόνο και δείξαμε ότι ο αλγόριθμος έχει γίνει τέσσερις φορές πιο γρήγορος. Ακόμη, η ποιότητα της επανασύνθεσης διατηρείται σε σύγκριση με αυτή της aHM-AIR που χρησιμοποιεί την LS. Με την χρήση του σφάλματος του σήματος προς την ανακατασκευή του (SRER) και την εκτίμηση της αντιληπτικής ποιότητας της ομιλίας (PESQ), δείχνουμε ότι η ομιλία που ανακατασκευάζεται με την χρήση της aHM-AIR με PP και aDFT διατηρεί την ποιότητα της aHM-AIR που χρησιμοποιεί την LS. Τελικά, επίσημα ακουστικά τεστ δείχνουν ότι η ομιλία που ανακατασκευάζεται από την aHM-AIR με PP και aDFT είναι παρόμοια με αυτήν που ανακατασκευάζεται από την aHM-AIR που χρησιμοποιεί την μέθοδο LS.

Acknowledgements

First of all I would like to thank my supervisor, Professor Athanasios Mouchtaris for giving me the opportunity to become a member of his team and showing belief in me, for all our constructive meetings, and for his great advice and support.

I am also especially grateful to have met and worked with Dr. Gilles Degottex. His continuous support, his ideas and his valuable help were a great contribution to this work.

Special thanks also go to the members of my dissertation committee, Professors Panagiotis Tsakalides and Giorgos Tziritas for their constructive comments and questions.

I would like to acknowledge the University of Crete and the Institute of Computer Science (FORTH-ICS) for providing financial support and all the necessary equipment during this work.

This work would not have been completed without the valuable help and patience of all the volunteers who participated in the listening tests. Guys, thank you all.

I would also like to thank all my colleagues at the lab. My warmest thanks to George, Maria, Olina, Sofia, Despoina, Tasos, Kostas for all the helpful discussions, the encouragement, nice atmosphere and for putting up with me, my singing and my painting.

A special thank you to my closest friends Kat, Dora, Gio, Antonis and Mina for always being there whenever I needed them, providing with their advice and support.

Last, but definitely not least, I would like to thank my family and of course my dog, Betty, life would not be the same without them.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Challenges of Speech Processing

Most human speech sounds can be classified as either voiced or unvoiced. Voiced sounds occur when air is forced from the lungs, through the vocal cords, and out of the mouth and/or nose. The vocal cords are two thin flaps of tissue stretched across the air flow, just behind the Adam's apple. In response to varying muscle tension, the vocal cords vibrate at frequencies between 50 and 1000 Hz, resulting in periodic puffs of air being injected into the throat. Vowels are an example of voiced sounds. In comparison, fricative sounds originate as random noise, not from vibration of the vocal cords. This occurs when the air flow is nearly blocked by the tongue, lips, and/or teeth, resulting in air turbulence near the constriction. Fricative sounds include: s, f, sh, z, v, and th.

After digitally analysing a speech signal, sinusoidal and harmonic models provide a set of sinusoidal parameters, such as amplitudes, phases and frequencies to represent the signal. These models have been widely used in speech coding and synthesis [1], [2], for hearing aids [3], speech enhancement [4], speech modeling [5] and voice transformation [6]. Additionally, the parameters can be later used to build higher-level representations [7] (eg. spectral envelopes) or to establish glottal source characteristics [8]. However, for this purpose, the accuracy and precision of the parameters are key issues.Furthermore, a representation that can produce sounds with sufficient perceived quality is of high importance for applications in synthesis, which need robust and precise estimates of f_0. There are plenty of real-time applications that need this high-quality synthesis, such as text-to-speech applications, analysis and synthesis techniques for quiet environments, etc. Additionally, speech signal analysis for voice production studies require a precision, that is higher than what can be perceived. Finally, even for offline computations, researchers need to test multiple ideas and parameters, various methods and large databases in a convenient time frame, hence, computationally efficient algorithms are preferred. Hence, models that are as computationally efficient as possible without any quality degradation are desired.

Sinusoidal and harmonic models are mainly designed for representing the periodic (or deterministic) part of speech, while they usually employ a random component in order to model the non-deterministic part. Alternatively, the voiced speech spectrum can be represented using multiple bands, with some bands representing the deterministic part and others the non-deterministic part of speech using noise components [9], [10]. Simpler models have also been suggested in which the speech spectrum is split into two bands. The separation occurs by using the so-called Maximum Voiced Frequency [9], [10]. The lower band represents the deterministic components, while the upper one represents the non-deterministic ones. For all multiband models, a reliable estimation of the voicing frequency limits is critical in order to avoid artifacts and provide a sufficient perceived quality of the synthesized sound.

Hence, when designing a digital signal processing model there are three questions that need to be asked: (1) how good does it need to sound?, (2)how precise should the parameter estimation be? and (3) how fast do we want it to be?

1.2 Motivation

It can be observed from the actual mechanism of voice production that the voice source is made of glottal pulses, as shown in Fig. 1.1, that are basically wideband signals whose amplitude spectrum is known to decrease smoothly without any abrupt frequency limit [11], [12]. Thus, it becomes a struggle for multiband models to estimate these frequency limits for separating the deterministic part from the non-deterministic one. Consequently, artifacts produced by an incorrect estimation of the frequency limits can degrade the perceived quality. Additionally, the following observation supports the presence of harmonic and deterministic content higher than usually observed with the DFT, hence it becomes apparent that the need for a frequency limit is questionable.

In voiced segments, the speech signal is usually assumed to be stationary in a small analysis window ($\approx$ 3 pitch periods). This hypothesis is fairly acceptable at low frequencies, because the variations of the fundamental frequency, f_0, of the glottal source are negligible compared to the stationary basis of most frequency analysis tools (e.g. DFT). However, the variations of f_0 are proportional to the harmonic number. The non-stationarity of the voiced signal is, therefore, highly increased as frequencies increase, making the validity of the stationarity hypothesis questionable for mid and high frequencies up to Nyquist. To alleviate this issue of modeling non-stationarities, the Fan Chirp Transform (FChT), which uses a chirp related frequency basis (i.e. linear frequency trajectories) adapted to the input signal, has been suggested in [13].

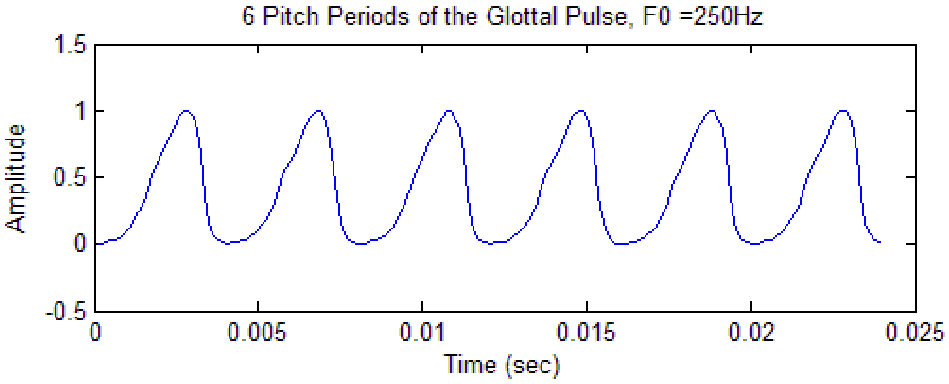

Figure 1.1: Glottal pulse.

Fig.1.2 shows the spectrograms of a short segment of voiced speech obtained by the DFT (left) and FChT (right). Although there seems to be a regular structure in the low frequencies in the DFT-based spectrogram, this is not the case for the frequencies around 3000 Hz where the frequency content is blurred. On the other hand, using the FChT, a regularity in the frequency content can be observed across almost all of the frequencies. This observation suggests that the current voice models often underestimate the voicing frequency and that a harmonic representation could be appropriate for both low and high frequencies.

Following the above arguments, we seek to follow a full-band harmonics-only representation of the speech spectrum. For sinusoidal models, the adaptive Quasi-Harmonic Model (aQHM), a quasi-harmonic representation of the speech spectrum that does not rely on a chirp frequency basis, has also been suggested in [14], [15]. Instead of limiting the frequency tracks to linear time evolution, as in FChT, aQHM relies on a more flexible frequency model. The frequency basis is adapted to the f_0 curve estimated from the speech signal. Thus, the adapted frequency basis can follow any non-linear variations of the frequency basis of the underlying signal. However, a proper estimation of the sinusoidal parameters can be obtained only if the input components of the frequency basis built from the f_0 curve are in a reasonable interval around the actual frequencies. Therefore, the tracking of the harmonic structure up to Nyquist can be easily compromised since any error on the f_0 curve is multiplied by the harmonic number.Furthermore, this generates a frequency matching problem, i.e. an ambiguity in terms of the connection between frequency components from neighbouring frames. A correct frequency matching is, however, vital for the preservation of the quality of the reconstructed signal, especially when this is applied during the analysis stage of aQHM. Consequently, from a point of view of either analysis or synthesis, an accurate f_0 estimate is critical in order to localize harmonic content in the high frequencies of the speech spectrum.

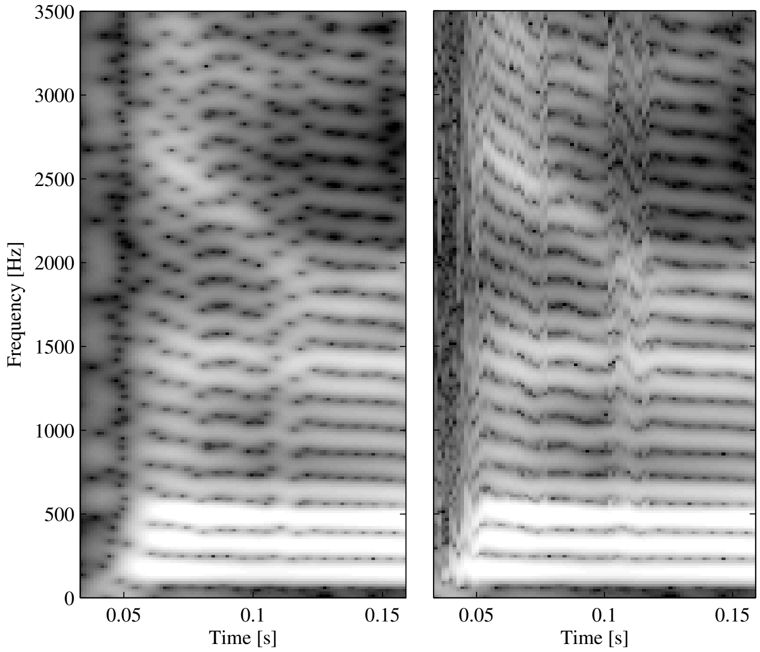

Figure 1.2: Time-frequency segments of spectrograms using DFT and FChT. The FChT clearly reveals a harmonic structure in higher frequencies than DFT.

If we want a full-band model of the speech signal in order to reveal the harmonics both at low and high frequencies, according to the phenomenon observed with the FChT, another method is necessary. In [16], an adaptive Harmonic Model (aHM) that uses adaptivity and a full-band representation was suggested. Also, an iterative algorithm, referred to as Adaptive Iterative Refinement (AIR) was proposed. The AIR algorithm starts with the lower frequency components, where the error is considerably small, and iteratively increases the number of harmonics. Additionally, it was shown that the quasi-harmonicity can be used for frequency correction and removed in the final representation of the signal. The whole method was called aHM-AIR, in order to distinguish it from aHM which could be used in many other ways. Managing the transients in speech model is always problematic since the detection of voiced/unvoiced transitions and the estimation of a maximum voiced frequency is a tricky task. A unified model covering both voiced and unvoiced segments is therefore an interesting solution. In aHM-AIR, the solution of the harmonic model is computed using the Least Squares (LS) solution. Since this model covers spectral content with regularly spaced components, the LS solution makes also sense in a random segment (i.e. fricative), especially used in the adaptive model thanks to flexibility of its non-stationary basis.

Compared to other approaches used for speech modeling (e.g. multiband models, HNM, mixed excitation models), aHM does not use a random component in voiced segments. Moreover, since aHM covers the whole spectrum and its frequency basis is not constraint to linear trajectories, it might also represent unvoiced segments properly. Thus, aHM can be used for the entire speech signal, whether or not the analyzed segments is voiced. Consequently, aHM-AIR's analysis/synthesis procedure does not need any detection of voiced/unvoiced transitions. However, even though aHM-AIR allows for a robust and full-band representation of the speech signal, the computational load due to the LS solution makes this method unsuitable for processing large databases in reasonable amount of time, which is a serious drawback.

1.3 Methodology

The issue of the computational efficiency is solved by replacing the LS solution with a Peak Picking (PP) approach in the AIR algorithm. The basic idea of the AIR algorithm is the following. It starts by first modeling the lowest harmonics, where errors in the f_0 measurements can easily be corrected by the correction mechanism of the QHM [17]. Next, the harmonic order of the model is iteratively increased by a continuous refinement of the f_0 trajectory. Consequently, the quasi-harmonicity is still used as a tool to estimate the adaptivity even though the quasi-harmonicity isn't kept at the final speech representation of aHM-AIR. Strict harmonicity is, hence, used as a constraint in aHM in order to avoid ambiguities during frequency matching.

Also, in order to integrate the adaptivity scheme of the aHM, the adaptive Discrete Fourier Transform (aDFT) is proposed. In contrast to the constant basis of the DFT, the frequency basis of the aDFT is fully adapted to the input f_0 curve of the signal as the aHM basis is adapted to the signal. We will be using this approach for both the refinement of the f_0 curve during the analysis process and the computation of the sinusoidal parameters used in the re-synthesis.

1.4 Thesis Contribution

In this thesis, we present and evaluate the aHM-AIR method that uses PP on aDFT and FChT in order to fully understand the results and their meaning. With the substitution of LS in the AIR algorithm by the Peak Picking approach a reduction on the computational load by a factor of 4 can be noticed. An example of this is, for instance, the analysis and synthesis of a 4 second sentence using the LS-based aHM-AIR takes about a whole minute. However, when using aHM-AIR with Peak Picking this process takes a bit over 10 seconds. Moreover, using synthetic signals, the accuracy and precision of the parameter estimation of all versions of aHM-AIR is evaluated, showing that the results of aHM-AIR using Peak Picking and aDFT are almost as robust as those of aHM-AIR using LS. Also, when using PP-aDFT the subjective and objective perceived quality of the reconstructed signal is preserved. Therefore, we provide a method that can indeed replace the original LS solution approach of aHM-AIR, while reducing the computational load by four times and keeping the high quality intact.

1.5 Thesis Organization

The remainder of this thesis is organized as follows: In Chapter 2 we discuss related sinusoid-type models. In Chapters 3 and 4 we describe the adaptive Harmonic Model(aHM) and the adaptive Discrete Fourier Transform(aDFT), respectively. In Chapter 5 we present our approach of the aHM-AIR method. An in depth validation and evaluation of our method is given in Chapter 6, where we compare it with the LS-based version of aHM-AIR and a few state-of-the-art methods of speech analysis and synthesis. Finally, Chapter 7 follows with the conclusions of this thesis.

Chapter 2

Related Work

In this chapter, the related work on the subject of speech analysis and synthesis is presented. Only the sinusoidal and harmonic parametric techniques will be described as they are close to the model-in-hand. In this work, the most important schemes will be presented. The description of all methods of speech analysis and synthesis will start from earlier approaches leading up to the latest ones, in order to show the evolution of the scientific area throughout the years in addition to highlighting major improvements over the methods.

Parametric techniques refer to methods that rely on a model of speech production, whose parameters are to be estimated. The type of parametric techniques discussed in this chapter is a model of time-series representation. Time-series based parametric representations include the decomposition of speech into components: a deterministic part, which is usually modelled as a sum of frequency and/or amplitude modulated components, and a non-deterministic (stochastic) part, which is modelled as frequency modulated Gaussian noise, usually weighted by a time-domain envelope. Typically, the deterministic part represents voiced speech, while the stochastic part represents unvoiced speech, friction noise, etc. Moreover, if the frequencies of the deterministic part are harmonically related, then the general model is called the Harmonic model. Various combinations have been made in literature: Deterministic plus Stochastic model [10], [18], Harmonic plus Noise model [18], Sinusoidal plus Noise model [19] and Quasi-Harmonic plus Noise model [14]. However, due to the inability of such models to represent highly non-stationary parts of speech, such as stop consonants or transient speech areas, extended models have been suggested, generally called Sinusoidal plus Noise plus Transients models [20], [21]. Typically, the parameters of these models include the harmonic (or not) frequencies, amplitudes and phases of the deterministic part, also, the number of sinusoids, whether an analysis frame is voiced or unvoiced, the time envelope of the noise, etc. Some of these parametric models, such as the Sinusoidal Model, the Harmonic plus Noise Model, and more, are described below.

2.1 The Sinusoidal Model (SM)

In 1986, McAulay and Quatieri suggested their famous Sinusoidal Model (SM). The speech waveform, $s(t)$, is assumed to be the output of a linear time-varying filter that has been excited by the glottal excitation waveform, $e(t)$. The filter has an impulse response denoted by $h(t,\tau)$ and is assumed to account for both the shape of the glottal pulse and the vocal tract impulse response. The speech waveform is given by

$$s(t) = \int_0^t h(t-\tau,t)e(\tau)d\tau \tag{2.1}$$

By representing the glottal excitation waveform as a sum of sine waves of arbitrary amplitudes, frequencies and phases, the model can be written as

$$e(t) = \sum_{k=1}^{N} a_k(t)cos(\Omega_k(t)) \tag{2.2}$$

where N is the number of sinusoids, $a_k(t)$ is the time-varying amplitude for the kth sinusoidal component and the excitation phase $\Omega_k(t)$ is the integral of the time-varying frequency $\omega_k(t)$.

$$\Omega_k(t) = \int_0^t \omega_k(\sigma)d\sigma + \phi_k \tag{2.3}$$

where ϕ_k is included to represent a fixed phase-offset because the sine waves will not necessarily be in phase.

The time-varying vocal-tract transfer function can be written as

$$H(\omega;t) = M(\omega;t)exp[j\psi(\omega;t)] \tag{2.4}$$

The system amplitude and phase along each frequency track $\omega_k(t)$ are given by

$$M_k(t) = M[\omega_k(t);t] \tag{2.5}$$

and

$$\psi_k(t) = \psi[\omega_k(t);t] \tag{2.6}$$

When the excitation signal $e(t)$ passes through the linear time-varying vocal-tract filter $h(t)$, the output is the sinusoidal representation of the speech signal

$$s(t) = \sum_{k=1}^{N} A_k(t)cos[\theta_k(t)] \tag{2.7}$$

where

$$A_k(t) = a_k(t)M_k(t) \tag{2.8}$$

and

$$\theta_k(t) = \Omega_k(t) + \psi_k(t) + \phi_k \tag{2.9}$$

represent the amplitude and phase of the kth sine wave along the frequency trajectory $\omega_k(t)$.

The above equations, (2.7), (2.8) and (2.9), are combined in order to provide a sinusoidal representation of a speech waveform. The validity of this representation is subject to the stationarity assumption of the excitation amplitudes and frequencies, compared to the vocal tract impulse response.

The analysis process of SM is performed in two steps. Firstly, the estimation of frequencies, composite amplitudes and phases. This first step is performed using a high-resolution Fourier Transform magnitude. This is done in a frame-by-frame scheme, after applying a window on the speech frame. For the first step, let $S(\omega, kR)$ be the short-time Fourier Transform of the speech signal, and R be the frame rate, so the estimated values are taken at kR sample indices. Thus, for the kth analysis frame, the lth frequency estimate is described by $\hat{\omega}_l^k$ and the corresponding amplitudes and phases are written as

$$\hat{A}_l^k = |S(\hat{\omega}_l^k, kR)| \tag{2.10}$$

and

$$\hat{\theta}_l^k = arg[S(\hat{\omega}_l^k, kR)] \tag{2.11}$$

where arg denotes the principal value.

The second step of the analysis accounts for the separation of the system and excitation components and it is done using homomorphic deconvolution, under the assumption of the vocal tract transfer function being minimum phase. Hence, the excitation components at each analysis frame boundary are given by

$$\hat{a}_l^k = \frac{\hat{A}_l^k}{\hat{M}_l^k} \tag{2.12}$$

and

$$\hat{\Omega}_l^k = \hat{\theta}_l^k - \hat{\psi}_l^k \tag{2.13}$$

Concerning the synthesis process, firstly, there is a matching procedure between the excitation frequencies measured on the kth frame with those of the $k+1$th one. Following is the matching of all other parameters which becomes easy, since they are measured at the excitation frequencies. In [22], an algorithm for matching the location of the spectral peaks by using a purely sinewave model was proposed.

Following parameter matching is the parameter interpolation. This is based on the assumption that the excitation and system functions are slowly varying across each frame along frequency tracks. System amplitudes, excitation amplitudes and system phases can be linearly interpolated, while a cubic polynomial is fitted on the excitation phases [22].

Finally, the synthetic waveform is given by

$$\hat{s}(n) = \sum_{l=1}^{L(n)} \hat{A}_l(n) cos(\hat{\theta}_l(n)) \tag{2.14}$$

where

$$\hat{A}_l(n) = \hat{a}_l(n)\hat{M}_l(n) \tag{2.15}$$

and

$$\hat{\theta}_l(n) = \hat{\Omega}_l(n) + \hat{\psi}_l(n) \tag{2.16}$$

where $L(N)$ is the number of sine waves estimated at time n.

2.2 The Harmonic Plus Noise Model (HNM)

A new model, called the Harmonic plus Noise Model, was proposed in the mid 90s by Stylianou [18]. In HNM, the speech signal is assumed to be composed of a harmonic part and a noise part. The harmonic part accounts for the quasiperiodic components of the speech signal and the noise part accounts for its nonperiodic components. A time-varying parameter, called the maximum voiced frequency, separates the two components in the frequency domain. In the lower band, the signal is assumed to be harmonic and is represented only by harmonics, while the upper band is represented by a modulated noise component and is modelled by an autoregressive (AR) model. In this model, both the analysis and synthesis is performed in a pitch-synchronous manner, inspired by PSOLA.

The lower band, or harmonic band, is modelled as a sum of harmonics

$$s_{hm}(t) = \sum_{h=-L(t)}^{L(t)} A_h(t) e^{jh\omega_0(t)t} = \sum_{h=1}^{L(t)} A_h(t) cos(h\theta(t) + \phi_h(t)) \tag{2.17}$$

where

$$\theta(t) = \int_{-\infty}^{t} \omega_0(u) du \tag{2.18}$$

and where $L(h)$ denotes the number of harmonics included in the harmonics part, $A_h(t)$ is the amplitude at time t of the hth harmonic and $\omega_0(t)$ is the fundamental frequency and $\phi_h(t)$ denotes the phase of the hth harmonic at time t.

The upper band is assumed to be dominated by modulated noise. In fact, in voiced speech, the noise part (high frequencies) exhibits a specific time-domain structure in terms of energy distribution (noise bursts), the energy of this high-pass information does not spread over the while speech period. Hence, the frequency components of the noise part are described by a time-varying AR model, and its time domain structured is formed by modulation using a parametric envelope. Thus, the noise part is given by

$$s_n(t) = e(t)[h(\tau, t) \star b(t)] \tag{2.19}$$

where $e(t)$ denotes the parametric envelope which modulates the noise components, $h(\tau, t)$ is the AR model used for describing the noise part, $\star$ denotes convolution and $b(t)$ is white Gaussian noise.

Before applying the model on speech, an estimation of the fundamental frequency and the maximum voiced frequency is required.A pitch estimation similar to the one in [9] is used. Then a voicing decision is made and a refined pitch is defined as the fundamental frequency. The position and duration of the analysis frames are set at a pitch-synchronous rate on the voiced parts of the speech and at a fixed rate on the unvoiced ones by using this stream of pitch values.

For the voiced part, the estimation of the parameters is performed by using weighted least squares

$$\epsilon = \sum_{t=-N}^{N} w(t)(x(t) - s_{hm}(t))^2 \tag{2.20}$$

where $2n+1$ represents the analysis window in samples and $x(t)$ is the original signal. The approach for the estimation of the HNM parameters is different from the one used in SM, which performs peak picking over the speech spectrum. In HNM, since the estimation is done only in the time domain, shorter windows can be used. While in SM a typical analysis window has a length of three to four pitch periods, in HNM two pitch periods are used. This is one very important aspect of HNM, because it registers the model convenient for modelling segments where speech exhibits high pitch or amplitude non-stationarity.

For the noise part, in each analysis frame, the power density function of the original signal is modelled by a pth-order all-pole filter, also the variance of the signal is calculated. The estimation of a parametric envelope in each frame follows. In [23], it was shown that an energy based time domain envelope outperforms the satisfactory results of the triangle type envelope.

The synthesis is performed in a pitch-synchronous way. The analysis time instants coincide with the synthesis time instants. For the harmonic part, the estimated amplitudes and phases are linearly interpolated between successive frames, with the phases being previously unwrapped. The unwrapping of the phases happens by predicting the phase of the current frame, using the phase of the previous one and the average instantaneous frequency. For the noise part, the synthesis is done by using an Overlap-Add (OLA) procedure, in order to avoid discontinuities at the frame boundaries. At a synthesis time instant, two pitch periods are synthesized by filtering a unit variance, white Gaussian noise through a normalized lattice filter, and multiplying the output by the variance estimated at the corresponding analysis time instant. If the frame is voiced, then the lower part, up to the maximum voiced frequency, is synthesized using harmonics while the noise part is filtered by a high-pass filter with a cut-off frequency equal to the maximum voiced frequency of that analysis time instant. Then, the synthetic noise part is obtained by applying OLA on two noise parts, one synthesized at the current synthesis time instant and the other at the previous one. Finally, the triangular time domain envelope is applied on the synthetic noise part. The synthetic signal can be written as

$$\hat{s}(t) = s_{hm}(t) + s_n(t) \tag{2.21}$$

2.3 Quasi-Harmonic Model (QHM)

Similar to sinusoidal models, the Quasi-Harmonic Model (QHM) assumes a local stationarity for speech. Even thought QHM is not an adaptive model, it provides the mechanism of adaptation, with the frequency correction mechanism, which yields an estimate of the mismatch between the actual and estimated frequencies. However, due to the assumption of local stationarity, QHM can only capture variations of frequencies and amplitudes up to a certain degree.

In an analysis window, QHM is written as

$$s(t) = \Big(\sum_{h=-H}^{H} (a_h + tb_h) e^{(j2\pi t)} \Big) w(t) \tag{2.22}$$

where H specifies the order of the model, i.e., the number of harmonics, a_h, are the complex amplitudes and. b_h are the complex slopes, f_h refers to the initial estimates of the frequency that are considered to be known, and $w(t)$ is the analysis window, which is typically a Hamming window and zero outside a symmetric interval $[-T, T]$. Hence, $t = 0$ denotes the center of the analysis window.

In the frequency domain, the hth component is written as

$$S_h(f) = a_h W(f - f_h) + \frac{jb_h}{2\pi} W'(f - f_h) \tag{2.23}$$

where $W(f)$ is the Fourier transform of the analysis window, $w(t)$, and $W'(f)$ is the derivative of $W(f)$ over f.

In order to implement a correction of frequency mismatches, it was shown in [24], that QHM can do so by projecting b_h onto a_h. Accordingly,

$$b_h = \rho_{1,h} a_h + \rho_{2,h} j a_h \tag{2.24}$$

where ja_h denotes the perpendicular (vector) to a_h, and the parameters $\rho_{1,h}$ and $\rho_{2,h}$ are computed as

$$\rho_{1,h} = \frac{\Re a_h \Re b_h + \Im a_h \Im b_h}{|a_h|^2} \tag{2.25}$$

and

$$\rho_{2,h} = \frac{\Re a_h \Im b_h - \Im a_h \Re b_h}{|a_h|^2} \tag{2.26}$$

where $\Re a_h, \Re b_h$ and $\Im a_h, \Im b_h$ are the real and imaginary parts of a_h and b_h, respectively.

If we consider the Taylor series expansion of $W(f - f_h - \rho_{2,h}/2\pi)$ and the value of the term $W''(f)$ at f_h as small, then for small values of $\rho_{2,h}$, it can be shown that the hth component of $S_h(f)$ can be written as

$$S_h(f) \approx a_h \Big[W(f - f_h - \frac{\rho_{2,h}}{2\pi}) + j\frac{\rho 1, h}{2\pi} W'(f - f_h) \Big] \tag{2.27}$$

which in the time domain can be expressed as

$$s_h(t) \approx a_h \Big[e^{j(2\pi f_h + \rho_{2,h})t} + \rho_{1,h} t e^{j2\pi f_h t} \Big] w(t) \tag{2.28}$$

Thus, from (2.28), it is clear that $\frac{\rho_{2,h}}{2\pi}$ accounts for the frequency mismatch between the hth component and the initial estimate of frequency, f_h, hence, $\frac{\rho_{2,h}}{2\pi}$ is an estimator of the frequency error

$$\frac{\rho_{2,h}}{2\pi} = \frac{1}{2\pi} \frac{\Re a_h \Im b_h - \Im a_h \Re b_h}{|a_h|^2} \tag{2.29}$$

while $\rho_{1,h}$ accounts for the normalized slope of the amplitude for the kth component, considering the instantaneous amplitude at the center of the analysis window.

In [24], it has been shown that this correction depends on the magnitude of $\rho_{2,h}$ and the value of the term $W''(f)$ at f_h. Finally, the estimation of a_h and a_h is performed via Least Squares (LS) in the following way:

Let's define the parameter vector $x = \begin{bmatrix} a \\ b \end{bmatrix}$. The error is defined in discrete time as

$$\epsilon(a,b) = \sum_{n=-N}^{N} |s[n] - s_q[n]|^2 \tag{2.30}$$

$$= \sum_{n=-N}^{N} (s[n] - s_q[n]) * (s[n] - s_q[n]) \tag{2.31}$$

where $s[n]$ is the original windowed signal, $s_q[n]$ is the Quasi-Harmonic representation of it, and the window size is $2N+1$.

In matrix notation, by separating the window values from the samples, the above equation, (2.31), can be written as

$$\epsilon(a,b) = (Ws - Ws_q)^H (Ws - Ws_q) \tag{2.32}$$

$$= (W(s - s_q))^H W(s - s_q) \tag{2.33}$$

$$= (s - s_q))^H W^H W(s - s_q) \tag{2.34}$$

where W is a square matrix having the analysis window values in its diagonal, s is the original signal samples in a vector, and ^{H} denotes the Hermitian operator. While s_q follows derives from the following:

The QHM representation can be written as

$$s_q[n] = \sum_{n=-N}^{N} (a_h + nb_h) e^{j2\pi f_h n/f_s} \tag{2.35}$$

$$= \sum_{n=-N}^{N} a_h e^{j2\pi f_h n/f_s} + \sum_{n=-N}^{N} nb_h e^{j2\pi f_h n/f_s} \tag{2.36}$$

In matrix notation the above, Eq. (2.36), can be written as

$$s_q = E_0 a + E_1 b = [E_0, E_1] \begin{bmatrix} a \\ b \end{bmatrix} = Ex \tag{2.37}$$

where

$$E_0 = (E_0)_{n,h} = e^{j2\pi f_h n/f_s} \tag{2.38}$$

$$E_1 = (E_1)_{n,h} = n(E_0)_{n,h} = ne^{j2\pi f_h n/f_s} \tag{2.39}$$

and

$$E = [E_0, E_1] \tag{2.40}$$

Thus, the minimization happens when

$$\frac{\partial \epsilon(x)}{\partial x} = 0 \tag{2.41}$$

$$\frac{\partial}{\partial x}(s - Ex)^H W^H W(s - Ex) = 0 \tag{2.42}$$

The solution for the above is given by

$$x = \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = (E^H W^H W E)^{-1} E^H W^H W s \tag{2.43}$$

Finally, a local approximation of the signal is expressed as

$$s(t) = \sum_{h=-H}^{H} |\hat{a}_h| e^{j(2\pi(\hat{f}_h + \frac{\hat{\rho}_{2,h}}{2\pi}}) t + \hat{\phi}_h) w(t) \tag{2.44}$$

where

$$\hat{\phi}_h = \angle \hat{a}_h \tag{2.45}$$

Although the QHM has been proved to perform better than the standard Sinusoidal or Harmonic Models [25], it still assumes signal stationarity inside the analysis window.

2.4 Adaptive Quasi-Harmonic Model (aQHM)

In [15], an adaptive Quasi-Harmonic Model (aQHM) has been proposed in order to alleviate the issue of non-stationarity. In aQHM the speech signal is represented as

$$s(t) = \Big(\sum_{h=-H}^{H} (a_h + t b_h) e^{j(\hat{\phi}_h(t+t_h) - \hat{\phi}_h(t_i))} \Big) w(t) \tag{2.46}$$

where $t \in [-T, T]$, $\phi_h(t)$ denotes the instantaneous phase function of the hth component and t_i is the center of the analysis window, while everything else plays the same role as in QHM.

Moreover, the model parameters are found via LS, as in QHM, given the samples of the input signal in vector s:

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = (E^H W^H W E)^{-1} E^H W^H W s \tag{2.47}$$

where W is the matrix containing the window values in the diagonal, s is the input signal vector and $E = [E_0, E_1]$, E_0 and E_1 have elements given by

$$E_0 = (E_0)_{n,h} = e^{j(\phi(t_n + t_i) - \phi_h(t_i))} \tag{2.48}$$

$$E_1 = (E_1)_{n,h} = t_n (E_0)_{n,h} = t_n e^{j(\phi(t_n + t_i) - \phi_h(t_i))} \tag{2.49}$$

and the instantaneous phase of the kth component can be computed as

$$\hat{\phi}_h(t) = \hat{\phi}_h(t_i) + \int_{t_i}^{t_i + t} 2\pi f_h(u) du \tag{2.50}$$

where $t \in [-T, T]$ and $f_h(t)$ is the frequency trajectory of the hth component.

The instantaneous phase of a single component, $\phi(t)$, can be computed as the integral of the instantaneous frequency, $f(t)$, based on the definition of phase. Furthermore, the instantaneous phase is obtained by an initial parameter estimation, such as in QHM. In order to interpolate phase values between two successive time instants, t_i and t_{i+1}, the following equation has been proposed

$$\phi(t) = \hat{\phi}(t_i) + \int_{t_i}^{t_i + t} 2\pi \hat{f}(u) du \tag{2.51}$$

where $\hat{\phi}(t_i)$ is the instantaneous phase estimate at the instant t_i.

However, this solution does not taken into account the frame boundary conditions at time instant t_{i+1}. Thus, it cannot be guaranteed that the phase value at t_{i+1} is

$$\phi(t)|_{t=t_{i+1}} = \hat{\phi}(t_{i+1}) + 2\pi M \tag{2.52}$$

where M is an integer appropriately selected to be as close as possible to

$$M = \mathbf{round}\Big(\frac{\phi(t_{i+1}) - \hat{\phi}(t_i)}{2\pi}\Big) \tag{2.53}$$

where $\mathbf{round}(\cdot)$ is the rounding to closest integer function.

In order to ensure phase continuation over frame boundaries, it has been suggested in [15] to modify Eq. (2.51) as

$$\phi(t) = \hat{\phi}(t_i) + \int_{t_i}^{t_i+t} 2\pi \hat{f}(u) + c(u) du \tag{2.54}$$

where $c(u)$ is expressed as

$$c(u) = r(t_{i+1}) \text{sin}\Big(\frac{\pi(u - t_i)}{t_{i+1} - t_i}\Big) \tag{2.55}$$

Hence, if $r(t_{i+1})$ is chosen as

$$r(t_{i+1}) = \frac{\pi(\phi(t_{i+1}) + 2\pi M - \hat{\phi}(t_{i+1})}{2(t_{i+1} - t_i))} \tag{2.56}$$

where M follows Eq. (2.53), then Eq. (2.52) is met.

While, in QHM the argument of the basis functions is parametric and stationary, in aQHM it is neither parametric nor necessarily stationary. From the aforementioned it can be observed that the basis functions of aQHM are adaptive to the estimates of the current phase characteristics of the signal.

Chapter 3

Adaptive Harmonic Model(aHM)

In [16], another adaptive model, this time based on the Harmonic Model, called the adaptive Harmonic Model (aHM) has been proposed. The main difference between the Harmonic Model (HM) and the adaptive Harmonic Model (aHM) is that the first uses random noise components (i.e. HNM [18]) or multiple bands in order to represent the non-deterministic part of speech while aHM is a full-band model that uses the adaptive scheme of aQHM. For aHM, an a priori knowledge of the fundamental frequency curve $f_0(t)$ is assumed, though a potential error is considered. Given the speech waveform $s(t)$, the following aHM model of $s(t)$ is used in a single window of 3 pitch periods:

$$x(t) = \sum_{h=1}^{H} a_h(t) \cdot e^{jh\phi_0(t)} \tag{3.1}$$

where $a_h(t)$ is a complex function of time representing both the amplitude and the instantaneous phase of the hth harmonic and $\phi_0(t)$ is a real function defined by the integral of $f_0(t)$:

$$\phi_0(t) = \frac{2\pi}{f_s} \int_0^t f_0(\tau)\, d\tau \tag{3.2}$$

where the time reference $t = 0$ is the center of the window, and f_s is the sampling frequency. According to the adaptive scheme proposed in [15], $a_h(t)$ and $f_0(t)$ are obtained by linear and spline interpolation of anchor values a_h^i and estimated f_0^i at specific instants t_i, respectively. Therefore, aHM will provide estimates of these parameters, which are sufficient for the complete representation of the speech signal. However, the number of anchors has to be properly chosen, since too many anchors may overfit the signal and represent variations that are not related to a deterministic component in voiced segments. A behavior like that has no true meaning for statistical modeling and may even cause the voice characteristics to be difficult to control in voice transformation. On the other hand, underfitting the signal with too few anchors should also be avoided. For speech, it can safely be assumed that the frequency modulation is related to a change of pulse duration and not to any modulation inside a single pulse. Hence, one anchor per period should suffice. A pitch synchronous analysis in which the distance between anchors respects an input f_0 curve, is assumed.

For the aHM parameter estimation with the presence of potential errors in the f_0 curve, the frequency correction mechanism of aQHM is used [15]. Within a single window, this model is represented as:

$$x(t) = \sum_{h=1}^{H} (a_h + tb_h) e^{jh\phi_0(t)} \tag{3.3}$$

where $\phi_0(t)$ is still defined by Eq. 3.2 and a_h, b_h are complex values that are constant in the window, in contrast to $a_h(t)$ in Eq. 3.1. To estimate a_h and b_h the following squared error is minimized by discrete sampling between the windowed speech segment $s[n]$ and its model $x[n]$

$$\epsilon = \sum_{n=0}^{N-1} (s[n] - x[n])^2 \tag{3.4}$$

where N is the number of samples in the analysis window. The solution of this minimization is found as in QHM via the LS solution, given the samples of the input signal in a vector s:

$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = (E^H W^H W E)^{-1} E^H W^H W s \tag{3.5}$$

where W is the diagonal matrix containing the window values in the diagonal, s is the input signal vector and $E = [E_0, E_1]$ is the adapted frequency basis, which have elements given by

$$E_0 = (E_0)_{n,h} = e^{jh\phi_0(t_n)} \tag{3.6}$$

$$E_1 = (E_1)_{n,h} = t_n (E_0)_{n,h} = t_n e^{jh\phi_0(t_n)} \tag{3.7}$$

It becomes clear that in order to compute the LS solution and estimate the aHM parameters there is a great computational load. In this thesis, another way of making the above computations was used, namely a Peak Picking approach, in order to decrease the computational load. The following chapter introduces a new, fully adaptive Fourier Transform where the Peak Picking is applied on. Following, Chapter 5 gives a detailed description of the Adaptive Iterative Refinement (AIR) algorithm that uses aHM for speech analysis and synthesis. Finally, Chapter 6 includes a full evaluation of the different approaches proposed for aHM.

Chapter 4

Adaptive Discrete Fourier Transform

The core and novelty of the suggested approach, used in adaptive harmonic speech analysis and synthesis, for the estimation of the sinusoidal parameters lies in the adaptive Discrete Fourier Transform (aDFT), as proposed in [26]. In order to properly describe the aDFT and emphasize the importance of adaptivity for the Adaptive Iterative Refinement (AIR) algorithm, a comparison between the DFT, FChT and aDFT is first presented in this chapter. In Fig. 4.1, the frequency basis for the three transformations mentioned above, is made visible, for a single analysis window. The results obtained by these three methods for a longer time period are depicted in Fig. 4.2.

4.1 Discrete Fourier Transform (DFT)

In order to compare all three transforms, we need to start with the frequency basis of the DFT. For a windowed signal $x[n]$ of length N, the DFT is defined as

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2\pi \frac{k}{N} \phi[n]} \tag{4.1}$$

where N represents the DFT length, $k = 0, 1, ..., N-1$. An example of the frequency basis and the results produced by DFT in a single analysis window is displayed in the first row of Fig. 4.1. In the DFT, there is the assumption of stationarity in the analyzed signal, since the frequency basis $\phi[n]$ used to compute the DFT is constant inside the analysis window:

$$\phi[n] = n \tag{4.2}$$

with time derivative:

$$\phi'[n] = 1 \tag{4.3}$$

However, in speech signals, this assumption of stationarity is valid only when the variations of the fundamental frequency, f_0, are negligible compared to the stationary basis of the DFT. Moreover, the variations of the harmonics are proportional to those of f_0 multiplied by the harmonic number. Hence, as frequencies increase so does the non-stationarity of the voiced signal, making the validity of the stationarity hypothesis questionable for mid and high frequencies. Fig. 4.2 presents the spectrograms obtained by using all three transforms and in its first row, one can see that in the results of DFT, the frequency content is highly blurred around 2.5kHz.

4.2 Fan Chirp Transform (FChT)

To alleviate the issues caused by DFT, the Fan Chirp Transform (FChT) has been proposed in [13]. In this method, a chirp related frequency basis (i.e. linear frequency trajectories) is used, with its slope adjusted to the average slope of the f_0 curve in the analysis window. For a windowed signal $x(n)$ of length N, the FChT is defined as

$$X_a[k] = \sum_{n=0}^{N-1} x[n]\xi^*(n,k,a) \tag{4.4}$$

where N also stands for the FChT length, $k = 0, 1, ...N-1$, * denotes the complex conjugate and $\xi(n,k,a)$ is the frequency basis of the FChT defined as

$$\xi(n,k,a) = \sqrt{|\phi_a'[n]|}e^{-j2\pi\frac{k}{N}\phi_a[n]} \tag{4.5}$$

where $\phi_a[n]$ rules the time dependence of the frequency basis exponent

$$\phi_a[n] = \left(n + \frac{1}{2}an^2\right) \tag{4.6}$$

whose time derivative is:

$$\phi_a'[n] = (1 + an) \tag{4.7}$$

where a is the chirp rate of the f_0 slope. The frequency basis and the respective spectrogram produced by FChT for a single analysis window is shown in the second row of Fig. 4.1. It can be observed that with the linearly adapted frequency basis of FChT, the harmonics become more definite compared to the ones produced by DFT. While in the second row of Fig. 4.2, one can notice that even around 2.5kHz the harmonics can be easily traced, especially compared to the results of the DFT spectrogram, in the first row. Hence, there is a regularity in the frequency content even in mid/high frequencies when the FChT is used, which was not visible before. Still, even though the FChT basis adapts to the frequency modulations better than the DFT, the frequency basis is constrained to linear trajectories only.

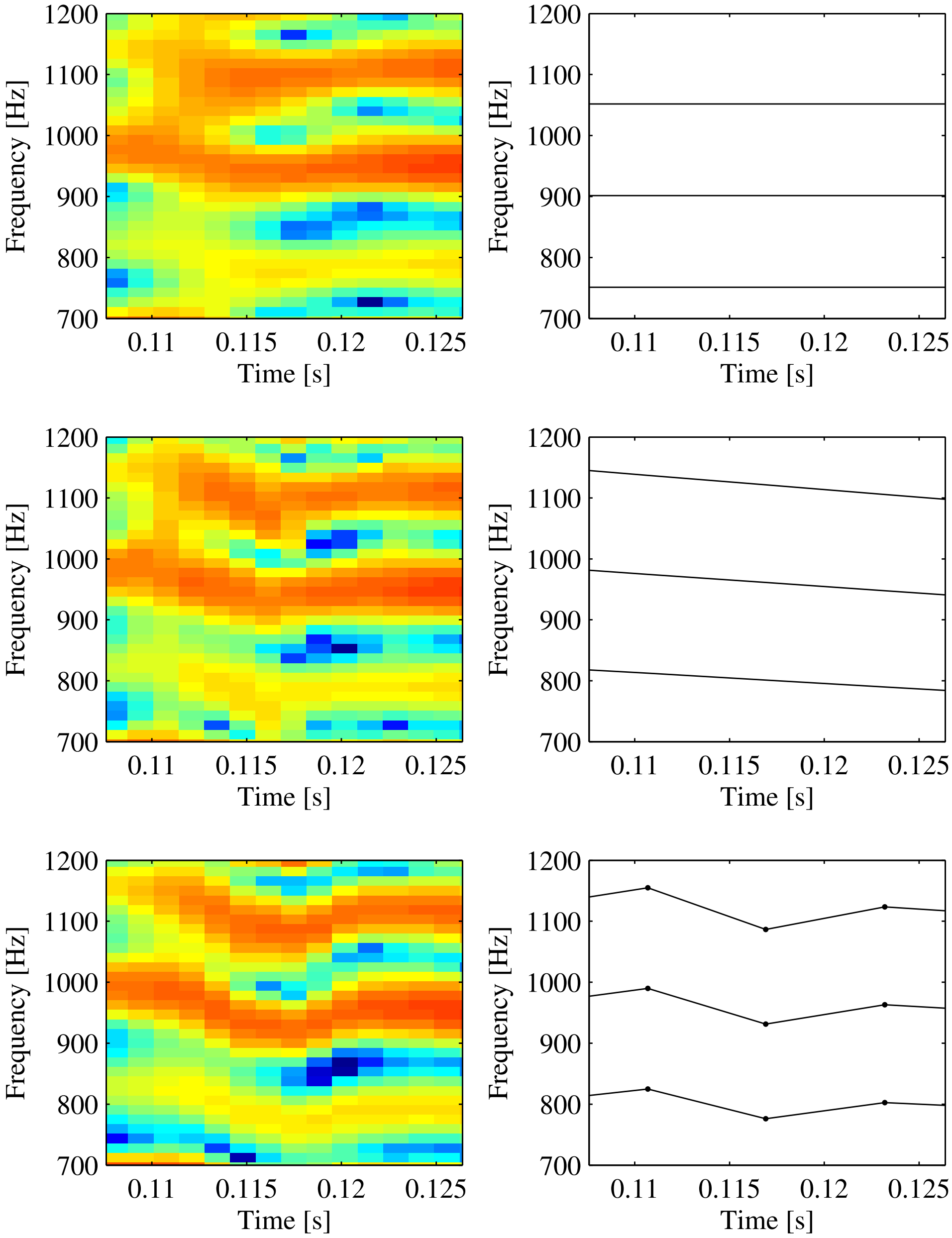

Figure 4.1: Three different transforms and their respective frequency bases for a single analysis window. First row depicts the spectrogram and frequency basis of DFT, second row of FChT and third row of aDFT.

4.3 Adaptive Discrete Fourier Trandform (aDFT)

In [26], in order to better follow the non-linear variations of f_0, the adaptive Discrete Fourier Transform (aDFT), based on the adaptivity scheme of aQHM [14] and aHM [16], was proposed. The frequency basis used for the aDFT follows completely the f_0 curve variations. Since the tracking of real sinusoids needs only the positive frequencies, the following representation is limited to the positive part of the aDFT. For a windowed signal $x[n]$ of length N, the aDFT of the positive frequencies, is defined as

$$X[k] = \sum_{n=0}^{N/2} x[n] e^{-j2\pi \frac{k}{N} \phi_0[n]} \tag{4.8}$$

where N, also, refers to the aDFT length, $k = 0, 1, ..., N/2$ and $\phi_0[n]$ is the "fundamental phase" of the frequency basis, whose values are obtained from the discrete sampling of the continuous real function, $\phi_0(t)$, defined by the normalized integral of the fundamental frequency $f_0(t)$:

$$\phi_0(t) = \int_0^t \frac{f_0(\tau)}{f_0(0)} \, d\tau \tag{4.9}$$

where the time reference $t = 0$ is the center of the window, f_s denotes the sampling frequency. In (4.9), $f_0(0)$ normalizes the frequency basis so that in the center of the window, where $t = 0$, it corresponds to that of the DFT as shown through the time derivative:

$$\phi_0'(t) = \frac{f_0(t)}{f_0(0)} \tag{4.10}$$

According to the adaptivity scheme, $f_0(t)$ is obtained by linear interpolation of the consecutive f_0^i values around specific instants t_i.

The third row of Fig. 4.1, shows an example of the results of aDFT applied on an analysis window and its respective frequency basis. It can be noticed that the frequency basis of aDFT compared to the other two methods is fully adapted on the variations of the f_0 curve, hence, it can produce more distinct results.

As mentioned above, in the second row of Fig. 4.2 the harmonics around 2.5kHz can be more easily traced compared to the ones in the first row. This creates a regularity in the frequency content even in mid/high frequencies when FChT is used. In the third row of Fig. 4.2, where the aDFT is used, this regularity can be noticed even more in mid/high frequencies.

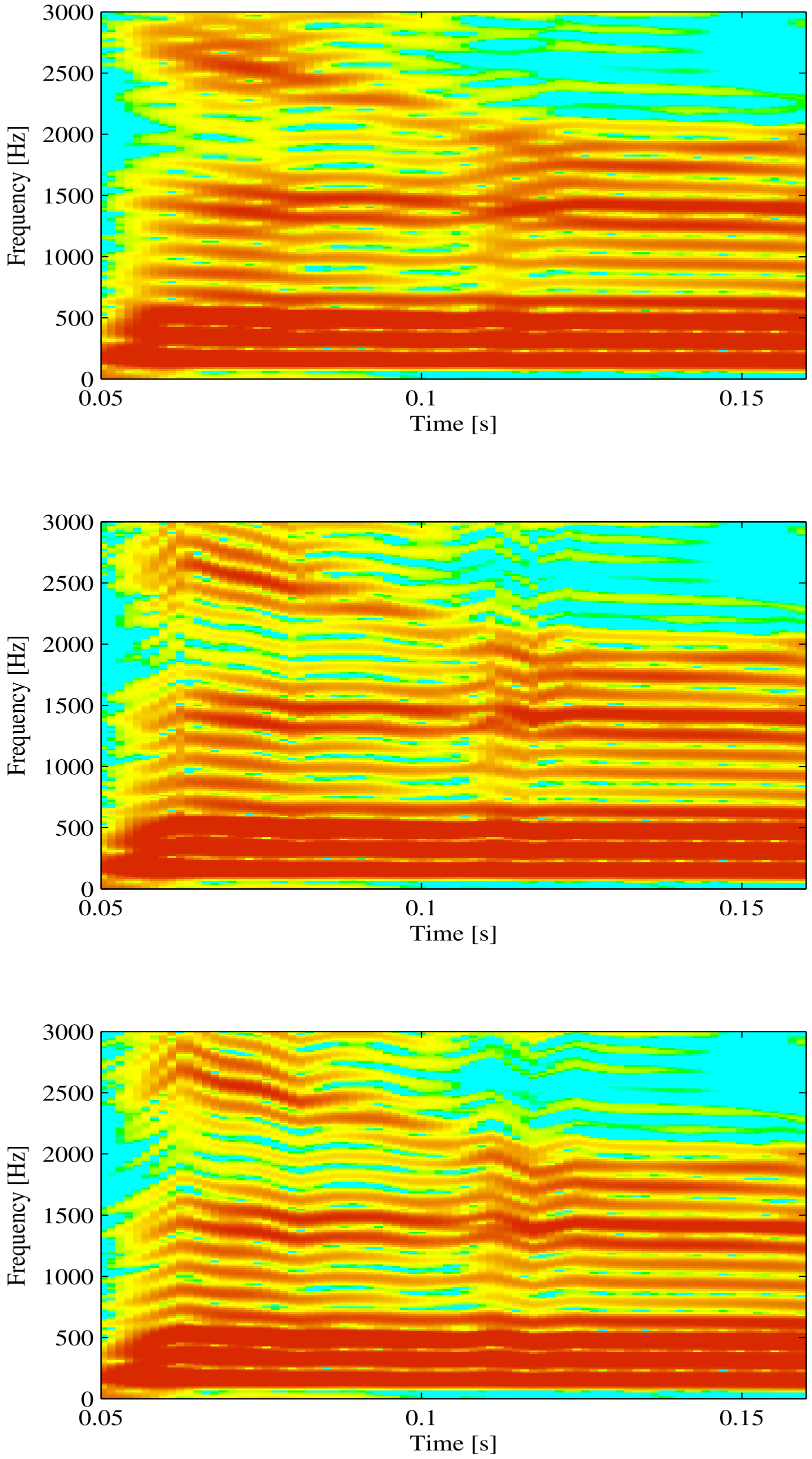

Figure 4.2: Spectrograms produced by DFT, FChT and aDFT depicted in the first, second and third row, respectively.

Chapter 5

Adaptive Iterative Refinement (AIR)

The Adaptive Iterative Refinement (AIR) algorithm is used to refine the incorrect localization of sinusoidal components due to the potential error in f_0, in order to allow a robust estimation of harmonic components up to the Nyquist frequency. In this thesis, three different methods (LS, FChT, aDFT) were used for the AIR algorithm and the refinement of the f_0 curve. These three methods were, then, used for the estimation of the sinusoidal parameters in the last analysis step. In the rest of this thesis, the analysis process will be separated in the two aforementioned steps and, for clarity purposes, they will be referred to as the *Refinement of* f_0 step and the *Sinusoidal Parameters Estimation* step. Combining these methods (LS, FChT, aDFT) for the two different steps of the analysis process results in the five methods of Table 5.1 that will be later on discussed in Chapter 5.

	Analysis Process Steps	
Method Name	**Refinement of f_0**	**Sinusoidal Parameters Estimation**
LS-LS	LS	LS
aDFT-aDFT aDFT-LS	aDFT	aDFT LS
FChT-FChT FChT-LS	FChT	FChT LS

Table 5.1: Methods Used for Both Steps of the Analysis Process

The AIR algorithm is used for the *Refinement of* f_0 step of the analysis process. The basic idea of the algorithm is that it begins with modeling the lower harmonics, where the error in the f_0 measurements can easily be corrected. Then, the harmonic order of the model is iteratively increased, and there is a refinement of the f_0 trajectory based on the estimations of the f_0 values for each frame.

During analysis, a parametrization of the speech signal at time instants t_i takes place. Using a rough estimate of the input f_0 curve, a sequence of instants t_i is first created, with distance of one pitch period between each of them. A Blackman window of 3 local pitch periods is then applied to the speech signal centered around each t_i, with the aDFT length (N) being defined as twice this window's length in order to make the main lobes appear in the aDFT. Consequently, voices with high pitch (e.g. female voices) will need a smaller aDFT length than voices with low pitch (e.g. male voices).

Before presenting the new, computationally efficient approach of AIR that uses Peak Picking on a frequency transform, a brief description of the previous version using the LS solution follows.

5.1 AIR Using the Least Squares Solution

The original AIR algorithm [16] will be presented in this chapter. This version of AIR uses the LS solution in order to compute the aHM parameters, a_h^i, b_h^i of the ith frame, as well as the frequency correction df_h and the fundamental correction f_{corr}.

The frequency correction is evaluated by:

$$df_h = \frac{f_s}{2\pi} \cdot \frac{\Re a_h \Im b_h - \Im a_h \Re b_h}{|a_h|^2} \tag{5.1}$$

where $\Re(\cdot)$ and $Im(\cdot)$ denote the real and imaginary parts, respectively. Using this correction, each anchor frequency f_0^i can be iteratively refined.

The basic idea of the proposed iterative algorithm is the following. For a single analysis window, we first assume that the initial predicted frequencies $f_h = h \cdot f_0$ for a small number of harmonics, H, are close enough to the actual frequencies of the signal. This means that the initial pitch is assumed to be free of octave errors. Then, estimating the aHM parameters, the correction term related to the fundamental frequency f_{corr} can be estimated as the mean of correction terms df_h relative to f_0:

$$f_{corr} = \frac{1}{H} \sum_{h=1}^{H} df_h / h \tag{5.2}$$

The number of harmonics H can then be updated, taking into account this fundamental correction f_{corr}. Indeed, if $|f_{corr}|$ is low, the current set of H harmonics converges to the actual values. Hence, it can be assumed that a few harmonics above H are now in a reasonable interval around their actual frequencies and H can thus be increased. To control the number of new harmonics added at each iteration, we propose linking H to f_{corr} in the following way. We first assume that the f_0 error remaining to be corrected is smaller or equal to f_{corr}. Therefore, the highest predicted harmonic frequency inside an interval of size $2N_w$ around the actual frequency is:

$$H = \lfloor N_w / |f_{corr}| \rfloor \tag{5.3}$$

According to [15], equation (5.1) holds only if the frequency to be corrected lies in a reasonable interval around the actual frequency. According to experiments, the size of this interval is about $B_w/3$ where B_w is the bandwidth of the squared window's main lobe [15]. Additionally, the highest frequency of the new set of harmonics has to be closer to its actual frequency than one of its neighboring frequencies (which are located $0.5 \cdot f_0$ around the actual frequency). Consequently, we chose N_w as the minimum between $B_w/3$ and $0.5 \cdot f_0$. Using (5.3), the initial harmonics number H can be chosen based on an assumed initial fundamental error (e.g. 20Hz). Using the mechanism of frequency correction of aQHM, $|f_{corr}|$ will be reduced progressively along the iterations and H will thus be increased up to the Nyquist frequency.

Algorithm 1 summarizes the analysis procedure:

Algorithm 1 AIR for aHM using the LS solution

```
Create a sequence of time instants $t_i$ according to $f_0(t)$
Initiate each $f_0^i = f_0(t_i)$
Initiate each $H_i$ using $f_{corr} = 20Hz$ and (5.3)
while $\exists i$ such as $f_0^i \cdot H_i < f_s/2$
  Compute $\phi_0(t)$
  for each anchor $c$ do
    Create a segment of 3 periods around $t_c$ using $f_0^c$
    Compute LS solution $(a_h^c, b_h^c)$
    Compute $df_h$ and $f_{corr} = \text{median}(df_h/h)$
    Compute $\hat{f}_0^c = f_0^c + f_{corr}$
    if $\hat{f}_0^c \cdot H_c < f_s/2$
      Update $H_c = \lfloor 0.5 N_w / |f_{corr}| \rfloor$
    end if
  end for
  Set $f_0^i = \hat{f}_0^i \; \forall i$
end while
```

In the Algorithm 1, a few more points have to be considered. Firstly, concerning the consistency of the correction terms, a df_h term whose harmonics lies in a frequency band made of noise cannot be interpreted as frequency correction. Therefore, it is necessary to ignore df_h values which may degrade the f_0 curve instead of refine it. Any df_h which does not satisfy the following three tests is discarded from the computation of f_{corr}: One, $|df_h|$ has to be smaller than $f_0/2$, otherwise two components may be close to each other turning the LS solution unstable. Two, $hf_0 + df_h$ has to be higher than 50kHz, this limit is assumed to be a minimum for f_0.

Also, even though Algorithm 1 stops when the model is full-band, extra iterations may still improve the representation of the signal. The iterations stop when the following two convergence criteria are met: i) the correction at the highest harmonic level $H \cdot |f_{corr}|$ has to be smaller than 10% of f_0 to ensure that the modifications of the frequency grid are negligible and ii) the maximum improvement of Signal to Reconstruction Error Ratio (SRER) for all of the frames is smaller than 0.1dB.

Finally, Algorithm 1 provides parameters of aQHM and not aHM, the former having bigger flexibility than latter because of the quasi-harmonicity in aQHM. Consequently, in order to ensure the consistency between the analysis and synthesis models, the aHM model is used in the last iteration.

For the synthesis procedure, each harmonic is generated successively for the whole signal, without the use of any synthesis window [27]. Below, the way to generate each harmonic from its estimated parameters, namely its amplitudes $|a_h^i|$, its phases $\angle a_h^i$ and the fundamental frequency f_0^i. First, we obtain the instantaneous amplitude $|a_h(t)|$ by means of linear interpolation across time of the anchor amplitudes $|a_h^i|$ using a logarithmic scale. For the computation of the instantaneous phase there is a linear phase term related to the time advance between each anchor instant. Hence, in order to compute them, first this term needs to be removed using the integral of $f_0(t)$:

$$\angle \hat{a}_h^i = \angle a_h^i - k\phi_0(t_i) \tag{5.4}$$

With this preprocessing, the phase values change smoothly from one anchor to the next. In order to obtain $\angle \hat{a}_h(t)$, $\angle \hat{a}_h^i$ can be interpolated. To avoid phase jumps in the interpolation, real and imaginary parts of $e^{j \cdot \angle \hat{a}_h^i}$ are interpolated independently and the interpolated values are recovered through the arctangent function. Additionally, a spline or cubic interpolation is necessary so that the time-derivative of $\angle \hat{a}_h^i$ is still continuous. Finally, $\phi_0(t)$ is obtained using equation (3.2), with $t = 0$ being the start of the signal, and $a_h(t)$ is $|a_h(t)| \cdot e^{j \cdot \angle \hat{a}_h(t)}$. All harmonics are finally summed as in (3.1) while discarding time segments of harmonics whose frequency is above the Nyquist.

5.2 AIR Using the Peak Picking Approach

In this chapter, the method for estimating the aHM parameters up to Nyquist is described, namely the Adaptive Iterative Refinement (AIR) algorithm which uses the proposed Peak Picking (PP) approach on the adaptive Discrete Fourier Transform (aDFT). The global structure of the original AIR algorithm is kept the same. In the original version of the AIR algorithm [16], the refinement of the f_0 trajectory was computed by using the Least Squares (LS) solution, while in this approach, instead of the LS solution a Peak Picking approach is used. Every other aspect of the AIR algorithm was kept the same. A full description of the AIR algorithm and a more detailed explanation of the methods used in it follows.

The AIR algorithm works first for each time instant t_i separately, estimating the value of the f_0 at that time instant, namely the f_0^i, where the original estimate of the f_0 curve is provided by the SWIPEP [28] algorithm. At the end of each iteration, the f_0 curve is redefined by all these values. The algorithm begins at a low harmonic level, $H_i = 8$, for each time instant, meaning that only harmonics up to the 8th one will be taken into account for the refinement of the f_0 curve during the first iteration. For each iteration, the corrected $\hat{f}_0^i$ is estimated for each time instant t_i from the Peak Picking on the aDFT computed from the segment created around that time instant. For the computation of $\hat{f}_0^i$, the harmonic peaks, f_h^i, computed by PP, where h corresponds to the harmonic number, are taken into account. More specifically, the value of $\hat{f}_0^i$ derives from the median of those harmonic peaks, divided by each peak's harmonic number. It was assumed that some peaks are representing noisy components. Thus, some peaks might be unreliable and the median value is an efficient way to discard outliers in the computation of a mean.

$$\hat{f}_0^i = median(f_h^i/h) \tag{5.5}$$

At the end of each iteration, all f_0^i values are replaced by the new $\hat{f}_0^i$. Before the next iteration begins, H_i is updated for each time instant t_i, as in the original AIR algorithm [16]. Eventually, this process is repeated for all frames until the Nyquist frequency is reached for all of them. Algorithm 2 describes this analysis procedure:

Algorithm 2 AIR for aHM using Peak Picking

```
Create a sequence of time instants $t_i$ according to $f_0(t)$
Initiate each $f_0^i = f_0(t_i)$
Initiate each $H_i = 8$
while $\exists i$ such as $f_0^i \cdot H_i < f_s/2$
  for each $i$ for which $f_0^i \cdot H_i < f_s/2$ is true
    Create a segment of 3 periods around $t_i$
    Compute the aDFT of the segment
    Pick the harmonic peaks $f_h^i$ up to $H_i$ from aDFT
    Compute $\hat{f}_0^i$ = median($f_h^i/h$)
    if $\hat{f}_0^i \cdot H_i < f_s/2$
      Compute $f_{corr}^i = \hat{f}_0^i$ - $f_0^i$
      Update $H_i = \lfloor 0.5N_i/|f_{corr}^i| \rfloor$
    end if
  end for
  Set $f_0^i = \hat{f}_0^i \; \forall i$
end while
```

In Algorithm 2, f_{corr}^i is the correction of f_0^i estimated in each iteration (i.e. $f_{corr}^i = |\hat{f}_0^i - f_0^i|$) and N_i is the aDFT length of frame i. The updated value of H_i has as upper limit the Nyquist frequency.

A brief comparison with the previous version of aHM-AIR [16] can clarify the ways in which this new version (i.e. Algorithm 2) is more computationally efficient. Originally, in the algorithm proposed in [16], in every iteration for each time anchor t_i, which hasn't yet converged, the LS solution was used for the minimization of Eq. 3.4 and in order to compute the a_k and b_k parameters of aHM (i.e. Eq. 3.5). On the other hand, in this approach, with the substitution of the LS solution with a Peak Picking method, this computationally heavy estimation becomes unnecessary. In Algorithm 2, instead of computing the aHM parameters in each iteration, the f_0 refinement for each time instant t_i, namely $\hat{f}_0^i$, is computed via Peak Picking in an aDFT transform and Eq. 5.5. This substitution reduces the computational load, making the new version of the AIR algorithm more efficient timewise.

Taking into account that the main reason behind the replacement of the LS solution with Peak Picking and aDFT approach is to improve the computational load of the aHM-AIR method while preserving the quality of the re-synthesis, a few more modifications were made. The following chapter, 5.2.1, presents all the refinements used during the aHM-AIR to reduce the computational load, describes a faster version of aDFT, called limited-aDFT and explains how the use of this function affects the Peak Picking approach. Furthermore, a more detailed description of the techniques used for the unvoiced segments follows in 5.2.2.

5.2.1 Reduction of Computational Load (Limited-aDFT)

When the aHM-AIR method begins, the harmonic level is set for each time instant t_i, at a low count. For the next iterations, this level is always limited until the Nyquist frequency is reached. Hence, only the part of the aDFT containing the necessary harmonics needs to be calculated, avoiding the computation of bins above the current harmonic level. This optimization cannot be done using the LS solution, because the corrections made by the aHM are not meaningful for LS when not applied up to the Nyquist frequency.

Another improvement regarding the method's complexity is based on the fact that the f_0^i values refined in each iteration, eventually converge. It can be noted that the frequency basis remains almost the same for the low frequencies, as the harmonic level, H_i, increases. Hence, the aDFT in low frequencies is very similar between iterations and the correction of the frequency basis for lower frequencies becomes more and more negligible. Thus, it can be assumed that below a specific extent of correction for each f_0^i, the peaks estimated during the previous iteration would remain almost the same in the lower frequencies, so they can be maintained for all following iterations. In order to implement this idea in the proposed method, a threshold, B_i, in the frequency bins of the aDFT, needs to be decided upon. The use of following relation is suggested:

$$B_i = \left\lfloor tol \cdot \frac{f_0^i \cdot N_i}{f_{corr}^i \cdot f_s} \right\rfloor \tag{5.6}$$

where f_0^i is the frequency at the time instant t_i, N_i is the aDFT length for frame i, f_{corr}^i is the correction of f_0^i computed from the previous iteration (i.e. $f_{corr}^i = |\hat{f}_0^i - f_0^i|$). A tolerance factor of 10% of the f_0^i value (i.e. $tol = 0.1 f_0^i$) was chosen, which provided an important reduction of the computational time without altering drastically the results. This tolerance factor of 10% roughly means that 10% of the previously computed lower peaks, depending on the correction f_{corr}^i made during this step, can be kept the same in the next aDFT. Hence, it is assumed that computing the new values of these lower peaks in the next iteration will have a negligible influence in the computation of $\hat{f}_0^i$.

Utilizing the above threshold, the bins of the aDFT below (5.6) would be kept the same for the following iterations, thus, the aDFT is only computed for the rest of the bins. This is the core of the limited-aDFT idea. As shown in Fig. 5.1 the lower bins of the aDFT were obtained by previous iterations. With each recursion, more of the lower bins of the aDFT are maintained for the next iterations.

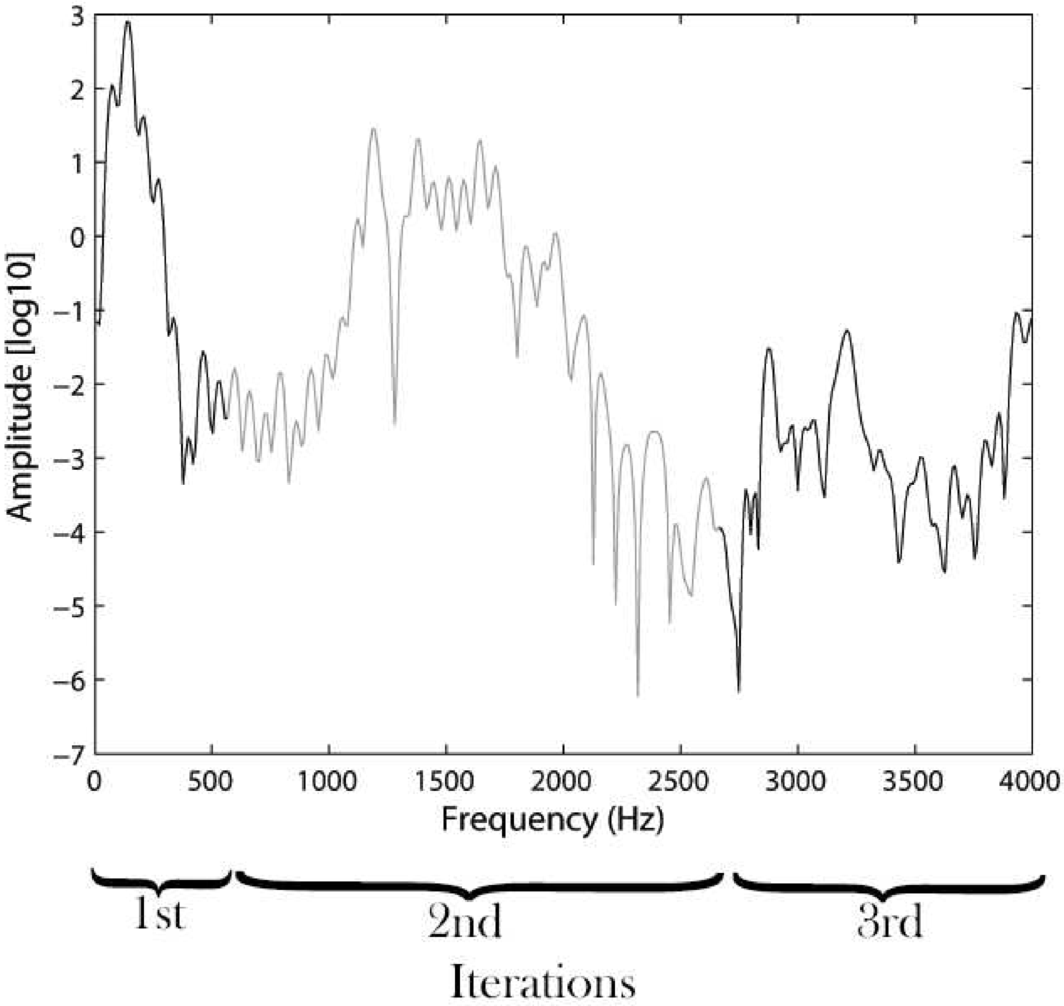

Figure 5.1: Illustration of how limited-aDFT works through iterations. Each part of the aDFT is computed in a different iteration, marked at the bottom of the figure.

It becomes apparent, that by using the limited-aDFT, thus, keeping part of the aDFT intact from iteration to iteration, the harmonic peaks inside that part, also, remain the same. This has an interesting affect on the PP approach. The previously used Peak Picking method can adapt to keep the harmonic peaks obtained from the frequency bins below the threshold and only compute the peaks in the rest of the frequencies. Later on, both the old and new peaks are used for the computation of $\hat{f}_0^i$, that will replace f_0^i at the end of each iteration. The conditions used by PP to determine whether a peak in the aDFT is harmonic or not, and which harmonic peak it corresponds to are explained below.

First, considering the f_0^i of the segment as the first harmonic, the method tries to find the rest of the harmonic peaks. All the peaks in the aDFT are, then, obtained. The harmonic peaks are defined as multiples of f_0^i with the harmonic order h $(h = 1, 2, ..., H_i)$. In order to determine which of the aDFT peaks are harmonics, the minimum distance between each harmonic peak and the peaks of the aDFT is computed, thus, finding the closest peak to that harmonic peak similarly to [29]. However, a peak is identified as a harmonic peak under some restrictions. More conditions are taken into account in order to determine if this peak can be used to refine the f_0 curve. If neither of the following conditions is met, then the peak under consideration is used in the estimation of $\hat{f}_0^i$, otherwise it is discarded. The first condition is whether the peak has already been identified as a harmonic peak, hence used in the refinement of f_0. The second condition examines whether the distance of the peak to the harmonic peak surpasses $f_0^i/2$. Also, every time the first condition is met, instead of discarding the peak immediately, the second closest peak to the harmonic under consideration is obtained. If this peak does not meet either one of the above conditions, it is identified as the current harmonic peak and used in the refinement of the f_0 curve. After the maximum harmonic level, H_i, of this iteration is reached, the refined $\hat{f}_0^i$ is computed following equation (5.5).

The results of this method have a satisfying perceived quality compared to the ones given by the LS solution for the Refinement of f_0 step, but its robustness is based on the assumption that the input f_0 curve is fairly correct. That is not the case when there is a substantial amount of noise in the curve. In order to solve this problem and make the method more robust, instead of computing the $\hat{f}_0^i$ at the end of each iteration of the PP method, $\hat{f}_0^i$ is evaluated whenever a new harmonic peak is identified, following (5.5). With every new peak, the value used for the first harmonic base changes (i.e. $f_0^i = \hat{f}_0^i$), resulting to a more precise estimation of the rest of the harmonics. More precisely, in the first iteration the harmonic base derives from the input f_0 curve which, as mentioned above, could have some noise. In the PP method only the first harmonic $(h = 1)$ is obtained, namely f_1^i, and based on the input frequency basis, PP will look for this harmonic around the value $1 \cdot f_0^i$. Then, in the next iteration of PP the method will search for the second harmonic f_2^i around the value $2 \cdot f_0^i$ and so forth. Consequently, at the end of PP all the harmonics collected will be almost multiples of the frequency basis, f_0^i, hence its error will be carried in all the following estimations, too, which may lead to skipping harmonics and wrongly identifying others. However, if after computing the first couple of harmonics, we consider recomputing the harmonic base, f_0^i, as in equation (5.5), after each iteration, then the original error will be significantly reduced. This is based on the fact that not all harmonics are an exact multiple of the harmonic base, hence with each recomputation of the harmonic base its value will converge to the real one. As a drawback, the algorithm becomes a little slower but the results become more robust.

5.2.2 Unvoiced Segments

In unvoiced segments, no harmonic structure exists, hence using a harmonic model in those parts becomes questionable. However, as it has been shown in [16], it is possible to use aHM for both voiced and unvoiced segments, thus providing a uniform representation across time which does not need any voicing decision. However, often, while using the suggested PP approach in unvoiced segments, either substantial deviations from the input f_0 curve occurred or the $\hat{f}_0^i$ value computed for an unvoiced segment ended up not converging. This is caused by the lack of harmonic structure in addition to the low harmonic level used during the first steps (e.g. $H_i = 8$ for the first iterating step), which prevent convergence of the $\hat{f}_0^i$ values. However, it was observed that using a higher harmonic level this was not the case, even for unvoiced segments.

Ideally, while dealing with unvoiced frames, an estimator should favour low frequencies, so that there is enough frequency resolution for representing the noise. In this thesis, the estimator considers a higher harmonic count in the unvoiced frames, thus, it doesn't favour the lower frequencies, but it tries to fit the most harmonic structure it can find closer to the initial f_0 curve values. We suggest to discard $\hat{f}_0^i$ values with any substantial deviations from the previous f_0^i values of each time instant t_i. Additionally, when a value is discarded, a forced increase of the harmonic level, before the next iteration, is used. In the current implementation, a deviation threshold of 8% from f_0^i is used to decide whether or not each $f_0'^i$ will be discarded. It was observed, after experimentation, that any $f_0'^i$ value that surpassed the 8% threshold either ended up converging in a extremely erroneous value or did not converge at all. In the case of a discard, the forced increase of the harmonic level takes place according to the following equation:

$$H_i' = |\hat{f}_0^i - f_0^i| \cdot H_i \tag{5.7}$$

This allows to force the harmonic level for the next iteration high enough that even the unvoiced frames will have enough harmonic peaks to compute a fairly correct estimation of f_0^i to eventually converge.

Chapter 6

Evaluation

Alongside the evaluations for the suggested Peak Picking approach in AIR, some of the results produced by the comparison between AIR using the LS solution and other well-known methods (i.e. SM, HM, aQHM) [16] will be presented, in order to better understand the importance and results of the AIR algorithm.

For the following evaluations, three different implementations of aHM-AIR were taken into account, namely the AIR algorithm can use either the LS solution [16], a Peak Picking approach using FChT or a Peak Picking approach using aDFT, for the *Refinement of* f_0 step of the analysis process. From this refined f_0 curve, the sinusoidal parameters of the harmonic model are, then, evaluated in the last step of the analysis process, namely the *Sinusoidal Parameters Estimation* step of analysis. For this step, all three methods mentioned above were, again, applied. This led to the comparison of the 5 methods from Table 5.1, depicted by the line styles of Fig. 6.1 in the following evaluations.

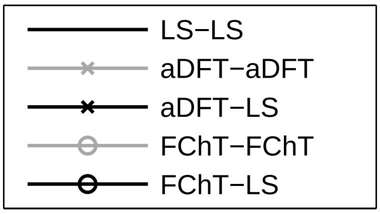

Figure 6.1: Line styles for all methods show in Fig. 6.4, Fig. 6.7 and Fig. 6.9. The first method in all line styles denotes the method used for the refinement of f_0 and the second one denotes the method used for the estimation of the sinusoidal parameters.

The evaluations were applied on a small database of 32 utterances (16 male and 16 female, originating from 16 different languages, between 2s and 4s length, with sampling frequency varying between 16kHz and 44kHz). The different phonemes and origins of these languages are assumed sufficient to provide a voice variability for supporting the validity of the results. For FChT, the chirp-factor a for each time instant t_i, was estimated based on the slope factor of the linear interpolation of the two neighboring f_0 values, f_0^{i-1} and f_0^{i+1}, around t_i and f_0^i.

6.1 Computational Time

For each method, the running time has been measured for each recording and the time reduction ratios, with respect to the LS-based method (Table 6.1 and Table 6.2), were averaged among all sentences. Table 6.1 presents the ratios for the *Refinement of* f_0 step of the analysis process. While, Table 6.2 displays the ratios for the *Sinusoidal Parameters Estimation* step of the analysis, where the parameters are estimated by all three methods.

Methods	**Male Voices**	**Female Voices**	**All**
$\frac{FChT}{LS}$	0.11	0.15	0.13
$\frac{aDFT}{LS}$	0.23	0.28	0.25

Table 6.1: Average Time Reduction Ratios for the Refinement of f_0 Step

On Table 6.1, it can be noticed that, on average, when using FChT, aHM-AIR becomes 7.69 (i.e. $\frac{LS}{FChT} = \frac{1}{0.13} \approx 7.69$) times faster, while, with aDFT, it becomes 4 (i.e. $\frac{LS}{aDFT} = \frac{1}{0.25} = 4$) times faster compared to when using the LS solution approach. Among the used sentences, the maximum ratio of time improvement observed was 21.67 for FChT and 7.67 for aDFT compared to the LS solution. The reason why there is such a difference between the improvement caused between FChT and aDFT is due to the fact that the frequency basis for FChT is less flexible than for aDFT and the slope parameters of FChT converge quicker than the actual f_0 values. On one hand, the aDFT keeps on changing as long as the f_0 values change. Thus, if the f_0 values change from one iteration to the next, the frequency basis of the aDFT will also be different, hence, the peak picking will find different peaks and the next f_0 correction will be proportional to these changes. On the other hand, for FChT, even though the f_0 values can change between two refinement iterations, the slope can be extremely similar, since many different sets of f_0 values have the exact same linear regression. Thus, the FChT may not change, and as a consequence the peaks remain the same and the f_0 correction can be almost zero for the next step. Thus, one can, indeed, expect a faster convergence with FChT than with aDFT.

For the *Sinusoidal Parameters Estimation* step of the analysis, all three methods were, also, used. By studying Table 6.2 it can be observed that using the LS solution is faster than using either FChT or aDFT in the *Sinusoidal Parameters Estimation* step. This is mainly due to the fact that in this step of the analysis process, both FChT and aDFT are computed for each frame up to the maximum harmonic level (i.e. Nyquist), while during the *Refinement of* f_0 step of analysis only parts of them are computed in each iteration, as discussed in chapter 5.2.1. Thus, the approaches using transforms are, according to our experiments, not faster than the LS solution for the *Sinusoidal Parameters Estimation* step. Table 6.2 shows that in this step, on average, LS is 2.10 times faster than FChT and 3.27 times faster than aDFT.

Methods	**Male Voices**	**Female Voices**	**All**
$\frac{FChT}{LS}$	1.98	2.23	2.10
$\frac{aDFT}{LS}$	3.16	3.38	3.27

Table 6.2: Average Time Reduction Ratios for the Sinusoidal Parameters Estimation Step

6.2 Parameters Estimation Error

The purpose of studying the parameter estimation error is to evaluate the precision of the estimated parameters in terms of a sinusoidal representation, compared to aHM-AIR using the LS solution. In the following tests, the estimated frequency, amplitude and phase values are compared to ground truth values of synthetic signals. A synthetic signal, which is as close as possible to a natural speech signal, is obtained by using a Liljencrants-Fant glottal model [11] to synthesize the glottal source. To obtain a realistic vocal tract filter, a digital simulator is used [30] that allows production of 13 different voiced phonemes, including nasalized sounds.

The synthetic signal is obtained as:

$$s(t) = 2\Re\Big(\sum_{k \in \mathbb{R}^+} G^{f_0(t)}(kf_0(t)) \cdot C(kf_0(t)) \cdot e^{jk\phi_0(t)} \Big) \tag{6.1}$$

where $G^{f_0(t)}(f)$ is the spectrum of the Liljencrants-Fant model, $C(f)$ is the vocal tract filter representing a random phoneme among 13 covering the vocalic triangle, and $\phi_0(t)$ follows (3.2), except that, now, $t = 0$ corresponds to the beginning of the signal. The pulse shape of the glottal model is controlled by a random parameter $Rd \in [0.3; 2.7]$ as in [11] and its period is defined by $f_0(t)$.

The following test evaluates the robustness of the different methods when the initial f_0 curve has errors which should be alleviated by the AIR algorithm. In the following tests, the original $f_0(t)$, in (6.1), is synthesized by using 5 anchors per second with random values in $[80; 400]$ Hz. A zero-mean Gaussian noise with various STandard-Deviation (STD) is, then, added to this curve which results to the input curve to the methods. In Fig. 6.4 and Fig. 6.7, the estimation error of the sinusoidal parameters is plotted as a function of the STD of this additive f_0 error. Using a sampling frequency of 44.1kHz, 320 test samples of 500ms duration each are generated. The samples are analyzed at regular intervals of 5ms and the differences between the estimated parameters computed by each method and the reference parameters, are determined. Fig. 6.4 and Fig. 6.7 show the mean and the STD (using a base-10 logarithmic scale) of the estimation error, in the first three and the last three rows, respectively. The phase error was computed by the wrapped difference between the unwrapped real and estimated values of the phase for these synthetic signals.

First, a comparison of aHM-AIR using the LS solution and other state-of-the-art methods (i.e. SM, HM, aQHM) is presented. For Fig. 6.3, we follow the same line style convention which is shown in Fig. 6.2.

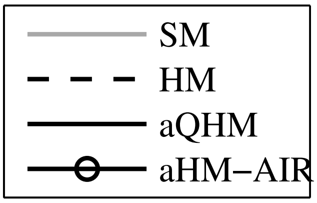

Figure 6.2: Line styles for all methods show in Fig. 6.3. aHM-AIR denotes the LS solution approach for AIR.

Fig. 6.3 shows the mean of the estimation error on the first three rows and the STandard Deviation (STD) using a base-10 logarithmic scale on the last three rows. In the last three rows, where the STD is shown, aHM-AIR always shows a smaller STD than the other methods except for the amplitude estimation under 4kHz. The estimation of the frequency grid is thus more precise when using aHM-AIR (fourth line). The estimation of the phase is also more precise especially above 4kHz (last line, right column). Globally, the improvement provided by aHM-AIR compared to the other methods is most apparent when considering the upper band of the signal. The aHM-AIR method thus provides better parameter precision in the high frequencies. For the Harmonic Model (HM), the error increases quickly as the f_0 error increases because no correction method is used to reduce the influence of the f_0 errors. On the other hand, the SM method selects the observed peaks in the amplitude spectrum even though the input f_0 values can be erroneous. Also, both aQHM and aHM-AIR use an iterative method for the refinement of the input f_0. Concerning the precision of SM in the estimation of the amplitudes below 4kHz (fifth row, left column), an explanation could be the following. The Sinusoidal Model always modifies the integer multiples of f_0 by means of quadratic interpolation in order to fit the maximum amplitude of a peak. Even though the frequency can be modified towards an erroneous value, this behavior ensures that the amplitude is always maximized. However, for aHM and aQHM, if the harmonic frequency, $h \cdot f_0$, is not properly aligned with the peak before the LS solution is computed and it slides down the main lobe of the window, the estimated amplitude can be substantially erroneous and consequently have higher variability than the maximized amplitude provided by SM.

For Figs. 6.4, 6.7 and 6.9, the same line style convention is followed, which is shown in Fig. 6.1. In the line style names, the first method mentioned denotes the method used the *Refinement of f_0* step of the analysis and the second one is the method used for the *Sinusoidal Parameters Estimation* step, as shown in Table 5.1. The mean and the STD values were computed through the median and the interquartile range, respectively, to avoid the influence of outliers.

It can be overall observed that the results produced by the five different methods used in Fig. 6.4 and Fig. 6.7 are, in some cases, very similar. Thus, arises the question of whether or not the difference between the different systems is significant. In order to better understand their difference, the 95% confidence intervals were computed for each method for both mean and STD, prior to the parameter estimation error. The intervals were computed by using 464870 and 2073504 samples for frequencies below 4kHz and above 4kHz, respectively. The width of these intervals was approximately 0.1, 0.01 and 0.003 for the mean error of frequencies, amplitudes and phases, respectively, and 0.0015 (base-10 logarithmic scale) for the STD error. Additionally, in most cases, there was no overlap between the different methods and even when there was, it occurred for intervals of a very small width. From all the above we can conclude that there is in fact a significant difference between the different methods. And even when there is not a big difference in the parameter estimation error one can consider the results of the computational time (6.1), the SRER (6.3) and the perceived quality tests (6.4).

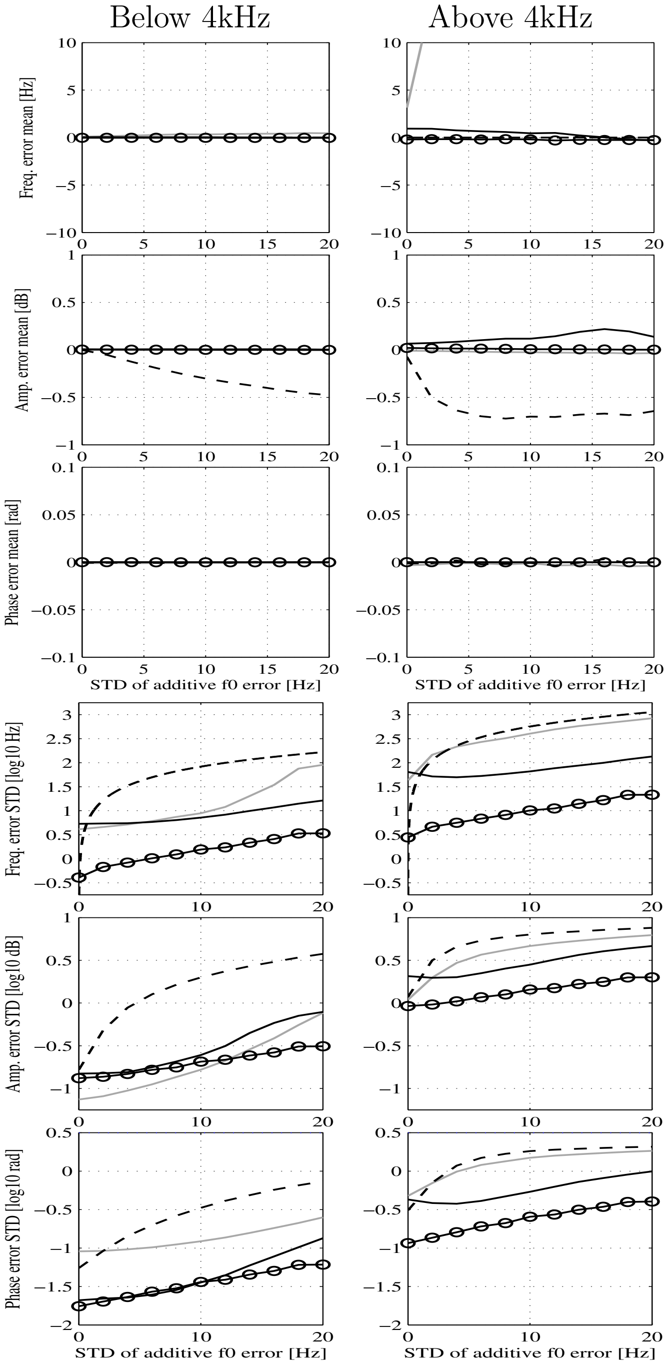

Figure 6.3: Error of sinusoidal parameters with respect to a potential error on the f_0 curve provided to the analysis method.

6.2.1 Refinement of f_0: Full Adaptivity vs. Linear Adaptivity (LS-LS vs. aDFT-LS vs. FChT-LS)

In Fig. 6.4, the results of the parameter estimation error for aHM-AIR when the LS solution is replaced by a Peak Picking method in the *Refinement of* f_0 step of the analysis process, are shown. These values are obtained by the estimation of the sinusoidal parameters, with the Sinusoidal Parameters Estimation step of analysis being performed by the LS solution. In the last three rows, the differences between the three methods can be observed more clearly. In the frequency error, row four, it can be observed that for the lower additive noise LS-LS has a smaller STD than the other two methods and as the noise increases LS-LS becomes indistinguishable from FChT-LS until 18Hz of additive noise STD are reached. FChT-LS begins with the same STD as aDFT-LS but, as the additive noise increases, its results correspond to the ones produced by LS-LS, while aDFT-LS has a slightly bigger STD than the other two methods, below 18Hz STD of noise. However, still for the same row, for the higher values of additive noise (above 18Hz STD), LS-LS and aDFT-LS have a smaller STD than FChT-LS. Finally, in the amplitude and phase errors, rows five and six respectively, it can be observed that FChT-LS has better results for the lower values of noise, while the results of aDFT-LS and LS-LS become better than those of FChT-LS as the noise increases and these two methods, aDFT-LS and LS-LS, have very similar results to each other. The behavior of FChT-LS in the higher values of the additive error can be contributed to its linear frequency basis. The more additive noise there is in the input f_0 curve, the harder it becomes for FChT-LS to find linear trajectories that can follow the adaptations of the f_0 values. On the other hand, this is not the case for LS-LS and aDFT-LS that are fully adaptive.

Fig. 6.5 gives a visual representation of what the *Refinement of* f_0 looks like, for an actual speech signal of almost 4s length. The first subfigure shows the input f_0 curve to the AIR algorithm, this curve has been estimated by the SWIPEP [28] algorithm and there is no additive noise. It can be observed that all three methods compute a very similar refinement of the f_0 curve by the end of AIR. However, in the results produced by the FChT, it can be noticed that a couple of f_0 values between 1.25s and 1.5s haven't converged properly, in fact this values reach up to 1400Hz, which can't be displayed in this subfigure since it would register the rest of the f_0 curve estimated by FChT unreadable.

In Fig. 6.6, the results that the three methods (LS, aDFT, FChT) provide for the *Refinement of* f_0 step when there is the maximum zero-mean Gaussian additive noise with 20Hz STD, are depicted. As expected from the observations made for Fig. 6.4, the results produced by the LS solution are the ones closer to the refinement when there is no noise (Fig. 6.5), with aDFT a close second. While the f_0 refinement estimated by FChT appears to be the most noisy one.

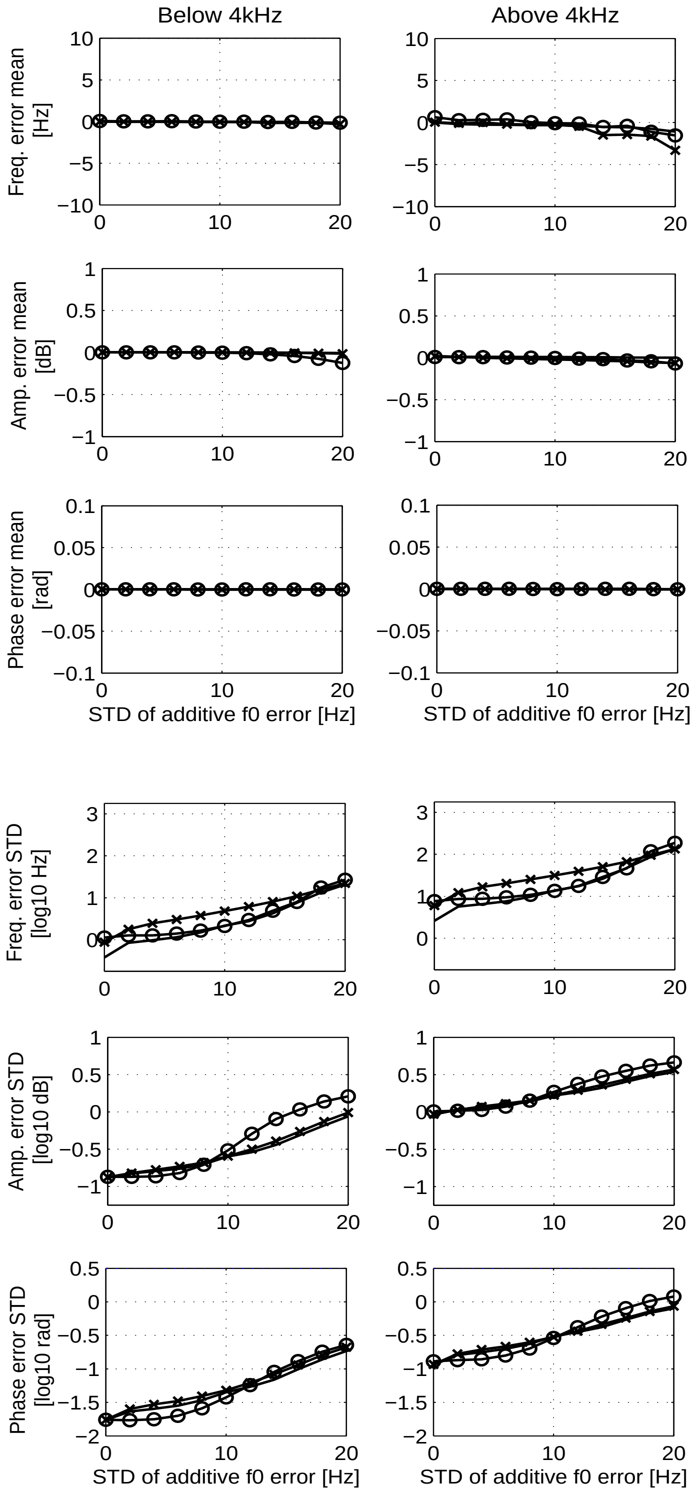

Figure 6.4: Error of sinusoidal parameters with respect to a potential error on the f_0 curve provided to the analysis methods, comparing full adaptivity with linear adaptivity during the f_0 refinement steps.

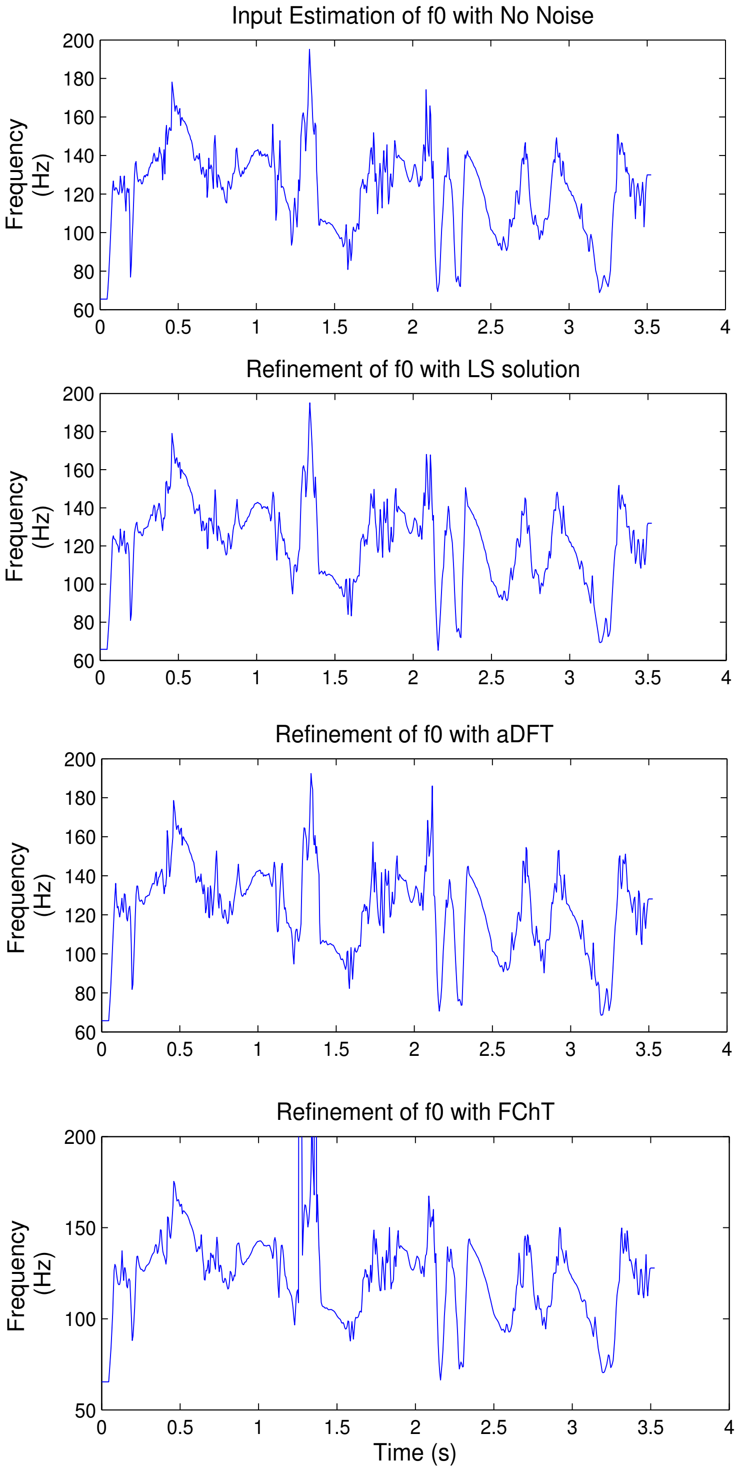

Figure 6.5: The results of the three different methods, namely the LS solution, aDFT and FChT, used in the Refinement of f_0 when there is no noise in the input curve.

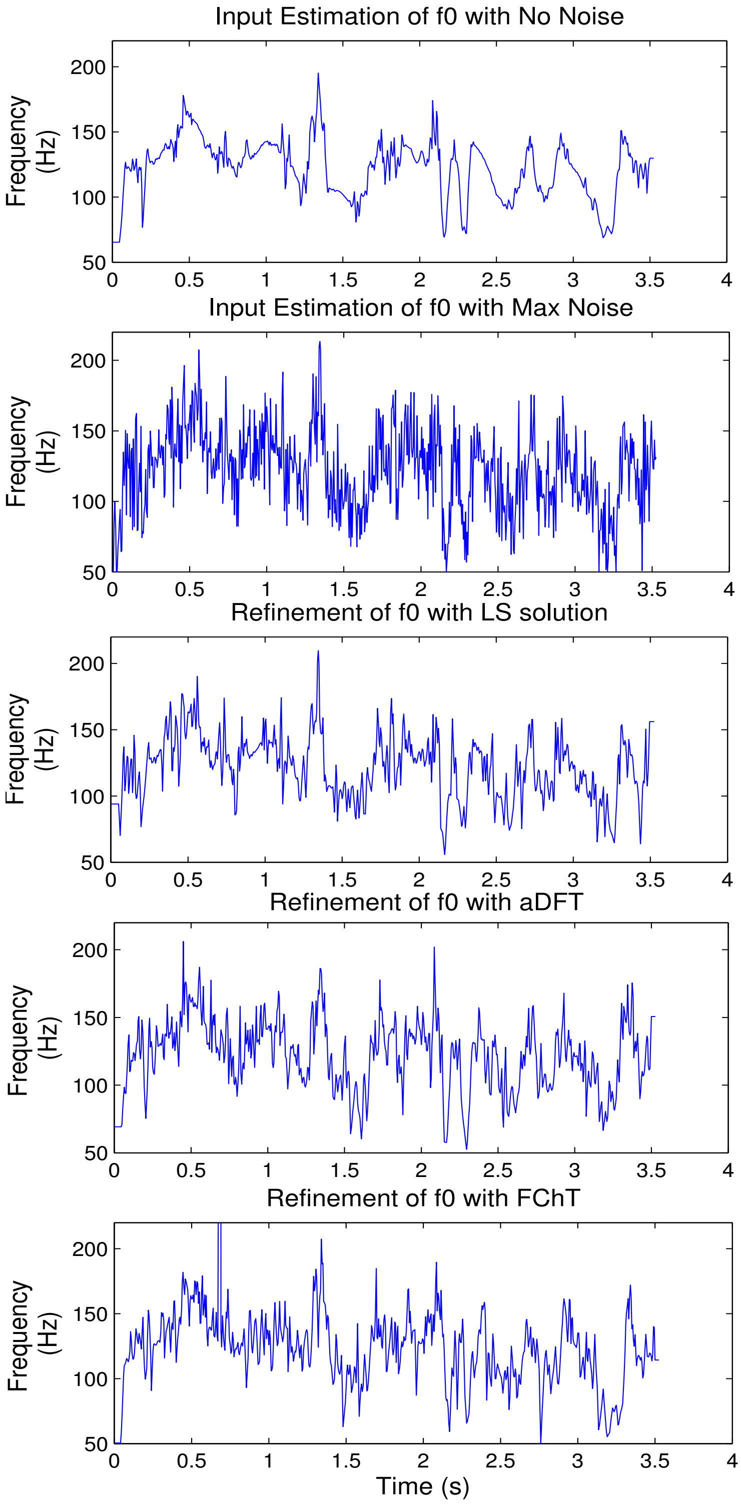

Figure 6.6: The results of the three different methods, namely the LS solution, aDFT and FChT, used in the Refinement of f_0 when there is the maximum additive noise in the input curve.

6.2.2 Sinusoidal Parameters Estimation: LS Solution vs. Peak Picking (aDFT-aDFT vs. aDFT-LS vs. FChT-FChT vs. FChT-LS)

The results shown in Fig. 6.7 can be studied in order to better understand the influence of the method used in the *Sinusoidal Parameters Estimation* step of the analysis process. For this test, either aDFT or FChT was used for the *Refinement of* f_0 step, while all three methods (LS, aDFT, FChT) were combined with them, as shown in Table 5.1, for the *Sinusoidal Parameters Estimation* step. It can be observed that when using a Peak Picking method in the *Sinusoidal Parameters Estimation* step instead of the LS solution the results of the parameter estimation are not always the best. In the first row, displaying the frequency mean error, it can be noticed that, in high frequencies, both aDFT-aDFT and FChT-FChT present an erroneous behavior, especially the latter with a mean error over 20Hz in most of the cases. Another great deviation for FChT-FChT from the results of the rest of the methods can be observed in the phase error estimation in third row. There, both in low and high frequencies, FChT-FChT demonstrates a highly erroneous behavior, having the biggest error estimated in all four methods. In concern to the STD of the parameters estimation error, aDFT-aDFT has either almost the same or better results than aDFT-LS, while FChT-FChT experiences some further difficulties. Namely, in the fourth row, the STD of the frequency error is almost the same for aDFT-aDFT and aDFT-LS while FChT-FChT has the worst results out of all four of them. In the fifth row, the amplitude error of aDFT-aDFT is the smallest one. Finally, in the last row, the phase error of aDFT-aDFT is the smallest out of all four methods in low frequencies and almost the same as aDFT-LS in higher frequencies. The good results produced by aDFT-aDFT are due to the PP which always catches the summit of the peaks, whereas LS can miss the peaks leading to higher amplitude and phase errors.

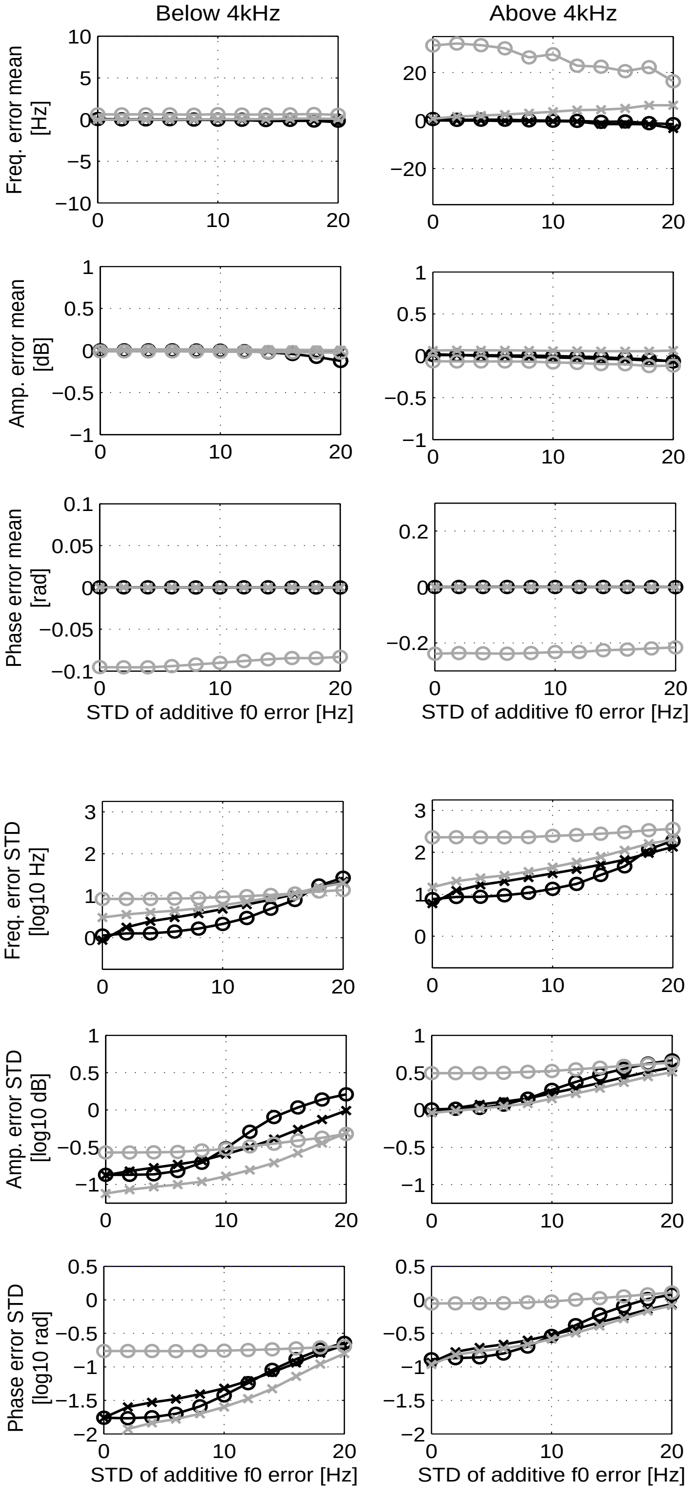

Figure 6.7: Error of sinusoidal parameters with respect to a potential error on the f_0 curve provided to the analysis methods, comparing LS solution with Peak Picking in the last analysis step.

6.3 Signal-to-Reconstruction Error Ratio (SRER)

The segmental Signal-to-Reconstruction Error Ratio (SRER) between the recorded utterances and their models was computed using equation 6.2 in order to evaluate the global reconstruction accuracy of the suggested methods. The SRER between an original signal s and its reconstruction $\hat{s}$ can be written as

$$SRER = 20\log_{10}\left(\frac{\sigma_s}{\sigma_{s-\hat{s}}}\right) \tag{6.2}$$

where σ_s denotes the standard deviation of a signal s. It can also be observed that the results of SRER are converted into decibel (dB). The higher the result of the above equation the better similarity $\hat{s}$ has to the original signal s.

A sliding window of 10ms with 50% overlap was used. In order to evaluate both the impact of the AIR algorithm, which refines the fundamental frequency, and the best method to compute the final sinusoidal parameters used for the synthesis, all five previously mentioned methods are compared. The SRER was computed using the full-band of the recordings and its distribution of the voiced and unvoiced segments is shown on the top and bottom plots of Fig. 6.8 and Fig.6.9. The sole 32 sentences were sufficient to obtain more than 10000 values for each distribution.

In Fig. 6.8, it can be observed that the three models, HM, aQHM and aHM, have very similar distributions compared to the SM model. For the voiced frames, the mean of these distributions is clearly higher than that of SM. The mean corresponding to aQHM is more than 10dB above the one of SM. One the other hand, the three models, HM, aQHM and aHM, use the LS solution, which explicitly minimizes the reconstruction error during the parameters estimation. While in the SM method, it is only assumed that estimating sinusoidal parameters by peak picking provides a set of sinusoids which properly represent the signal. Hence, the observed difference between the Harmonic Models and SM. Finally, the aQHM model has a slightly better SRER compared to aHM. This results is also expected since aQHM is more flexible than aHM, thanks to quasi-harmonicity. Concerning the unvoiced frames, the average SRER is obviously lower for all methods since the limited number of sinusoids of the models cannot properly cover the noise that fills the whole spectrum. The aHM model provides a better fitting of the noise than HM because of its adaptivity. However, as for voiced segments, one could expect that aQHM would provide a better SRER than aHM, which is not the case in Fig. 6.8. This is due to the fact that the correction terms df_h are meaningless for noise and lead to misplaced quasi-harmonics. One the other hand, the strict harmonicity of aHM ensures, at least, that the full-band is regularly sampled.

It can be observed that the distributions of all three methods using the LS solution in the Sinusoidal Parameters Estimation step are very similar to each other for voiced segments. This means that the reconstruction quality is preserved and, as was shown

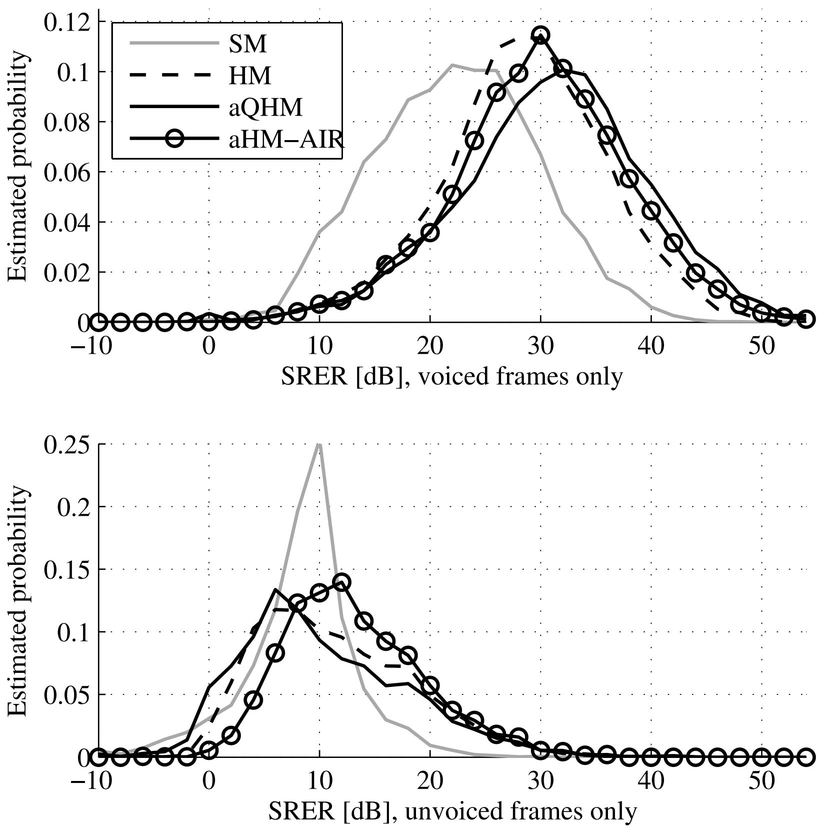

Figure 6.8: Estimation of the full-band SRER distributions for voiced and unvoiced frames on top and bottom plots respectively.

in chapter 6.1, the computation load has a considerable decrease. On the other hand, both FChT-FChT and aDFT-aDFT present some issues with both voiced and unvoiced frames which can be explained by the higher frequency errors when not using the LS solution. It is very interesting to notice the behavior of FChT-LS in the unvoiced segments, where a smaller SRER is observed compared to the other two methods using LS in the Sinusoidal Parameters Estimation step. This is due to the fact that the frequency basis in FChT is constrained to linear trajectories and cannot fully adapt to the input f_0 curve, in contrast to the fully adaptive methods.

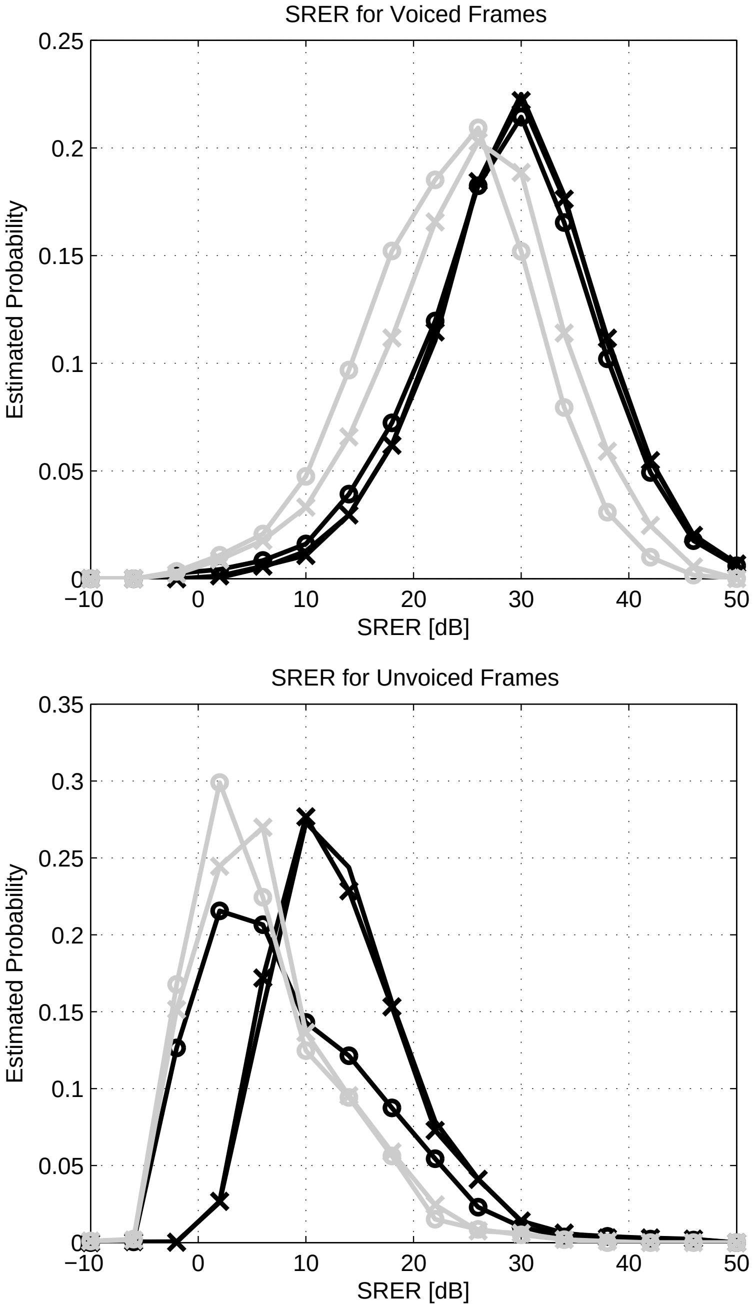

Figure 6.9: Estimation of the full-band SRER distributions for voiced and unvoiced frames. Line styles from Fig. 6.1

6.4 Perceived Quality: Listening Test, PESQ

In this part of the evaluations, the perceived quality of the reconstructed signals using the five methods was evaluated subjectively and objectively, using listening tests and the PESQ method [31] respectively.

6.4.1 Subjective Perceptual Evaluation

For the first listening test, the listeners were asked to listen to one original recording among 24 utterances (12 languages, one male and one female speaker for each). Then, they had to rate the impairment of five sounds: four of which were the synthesized ones made with SM, HM, aQHM, aHM, while the fifth sound was the original recording, which was added to the comparison set in order to check the consistency of the answers. In the test, each listener was asked to grade only 2 languages randomly selected from the set of the 12 languages. Since each language was represented by one male and one female voice, each listener evaluated the resynthesis of 4 recordings. The following rating scale of impairment was used: (5)Imperceptible, (4)Perceptible but not annoying, (3)Slightly annoying, (2)Annoying and (1)Very Annoying. Only the answers given by listeners who used earphones or headphones were kept. Additionally, answers from listeners who did not rate the original recordings between 4 and 5 were discarded. In total, 48 people answered the test and the answers given by 44 of them were kept. Since the sounds to evaluate were selected randomly, the number of occurrences of each sound was not uniform (even though it tends to be when the number of listeners increases). In order to remove any possible bias, the mean and confidence intervals of the results were normalized according to the number of occurrence of each sound.

Fig. 6.11 shows the results of this listening test. Firstly, the SM method has been clearly graded lower than the other methods. Significant artifacts seem to appear in the high frequencies of the resynthesis using SM. Globally, the three remaining methods use a harmonic or quasi-harmonic grid which ensures minimal continuity of the sinusoidal components. Conversely, in SM, a component can disappear from one frame to the next which generates a persistent artifact mainly in high frequencies. The slight downward trend of the aQHM method compared to aHM and HM can be explained by some musical sounds which can be sparsely perceived along the resynthesis. Having the frequency components completely independent, as in aQHM, may provide better flexibility, though it also adds a risk that components leave the frequency band in which they are supposed to be. On the other hand, the strict harmonicity may oversimplify the representation, even though it offers a global constraint stabilizing the resynthesis. Even though the SRER of aQHM is higher than that of aHM in voiced segments (Fig. 6.8), the SRER difference around 10dB in unvoiced segments is easier to perceive than that around 30dB in voiced segments.

The global difference can therefore explain the slight downward trend seen in the listening test. Finally, the results specific to gender show that the resynthseis of the male voices made by the HM methods is clearly indistinguishable from the original recordings.

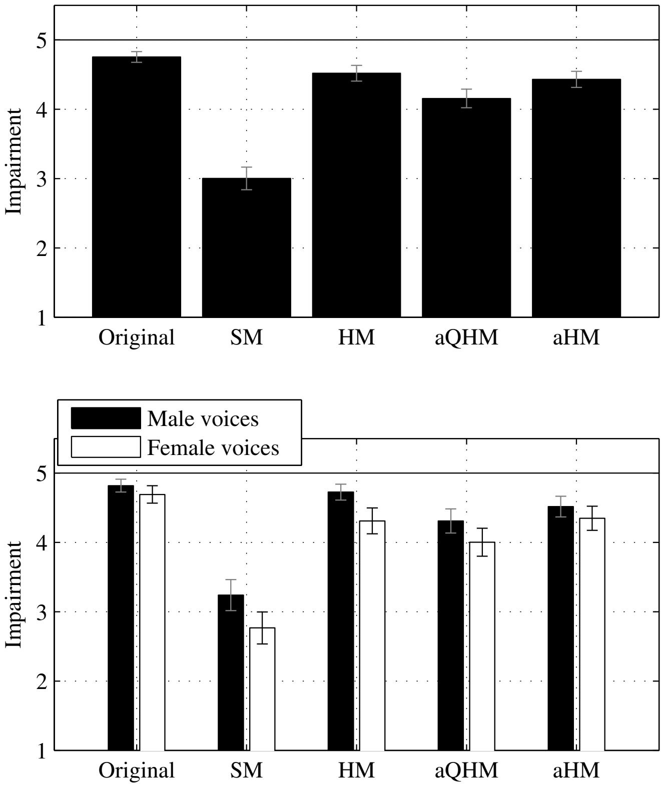

Figure 6.10: Impairment evaluation of the resynthesis quality by 44 listeners using 24 utterances of 12 different languages, with the 95% confidence intervals. The used f_0 values are those provided by the aHM-AIR method.

The purpose of the following listening test is to evaluate the methods which are used for both steps of the analysis process (i.e. *Refinement of* f_0 and *Sinusoidal Parameters Estimation*). In order to do so, the same 32 utterances of 16 different languages as in 6.2 were used. Listeners were asked to evaluate the quality of sound files compared to an original recording using a web interface. Among the six files they had to rate, five of them were synthesized with LS-LS, aDFT-aDFT, aDFT-LS, FChT-FChT and FChT-LS, while the sixth file was the original recording, which was added to the comparison set in order to check the consistency of the answers. In this test, each listener was asked to grade only 3 languages randomly selected from the 16 languages. Each language was represented by one male and one female voice, hence, each listener evaluated the resynthesis of (6 recordings) × (the 6 different methods). The following grading scale of quality was used: (5)Excellent, (4)Good, (3)Fair, (2)Poor and (1)Bad. In order to optimize the results of the listening test, the listeners were asked what device they used to listen to the signals, and only the answers from listeners who used headphones or earphones were kept. Moreover, answers by listeners who did not rate the original recordings systematically between 4 and 5 were discarded, considering that they did not understand the instructions or they were not focused enough. After all the above answers were discarded, the quality evaluation of the resynthesis was computed by the answers of 24 listeners. Since the sounds to evaluate were selected randomly, the number of occurrences of each sound was not uniform. In order to remove any possible bias, the mean and confidence intervals of the results were normalized according to the number of occurrence of each sound. Fig. 6.10 shows the results of this listening test.

According to Fig. 6.10, it can be noticed that only three methods have a global score under 4, aDFT-aDFT, FChT-FChT and FChT-LS. This is caused by the fact that all three methods cannot adapt really well to the unvoiced parts of a signal, as shown in Fig. 6.9, hence creating artifacts in the resynthesis. On the other hand, aDFT-LS and LS-LS have very similar scores overall, very close to the results of the Original signal.

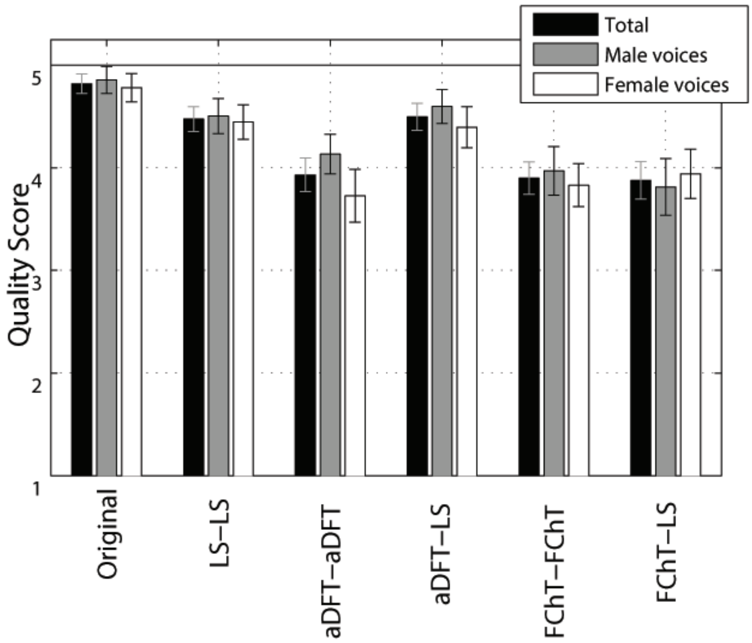

Figure 6.11: Quality evaluation of the resynthesis quality by 24 listeners using 32 utterances of 16 different languages, with 95% confidence intervals.

6.4.2 Objective Perceptual Evaluation of Speech Quality using PESQ

It is expected that, since the results of SRER for the LS-LS and aDFT-LS methods are very similar, an objective measure of perception would give the same results. In order to verify this, the PESQ method [31] is used to assess the perceived quality of the reconstructed signals compared to the originals. Table 6.3 presents the PESQ scores for the five methods of Table 5.1, using the same database as in the previous tests. Due to the fact that the sampling frequency for the signals in the database varied from 16kHz to 44kHz, a re-sampling of all signals to 16kHz was performed in order for the PESQ measurement to be used. The results show that the LS solution has the best PESQ score with aDFT-LS being a close second. On the other hand, FChT-LS and FChT-FChT have the worst results of them all, with aDFT-aDFT being a little better than them, as is expected from the SRER and listening test results.

PESQ Ratings (up to 4.5)	
LS -LS	4.18
aDFT - aDFT	3.92
aDFT - LS	4.15
FChT - FChT	3.73
FChT - LS	3.82

Table 6.3: PESQ scores assessing the overall quality of the re-synthesized signals of the methods compared to the original signal.

Chapter 7

Conclusions

Taking advantage of the good perceived quality provided by aHM-AIR, a Peak Picking approach was suggested in order to replace the LS solution for the f_0 refinement. The main reason behind this substitution is the reduction of the heavy computational load of the AIR algorithm, mainly caused by the LS solution. Two different transforms were used for Peak Picking, namely the Fan Chirp Transform (FChT) and the adaptive Discrete Fourier Transform (aDFT). Evaluations have shown that by performing this substitution, the computational load of the AIR algorithm decreases, in average, by a factor of 7.69 and 4, for the FChT and aDFT, respectively. Moreover, using synthetic signals, the accuracy and precision of the parameter estimation of all versions of aHM-AIR was evaluated showing that the results of aDFT-LS are almost as robust as those of LS-LS. Concerning the methods using the very fast FChT approach, a slightly erroneous parameter estimation was observed, registering them inadequate for applications where high quality parameter estimation is required. Also, a listening test was carried out in order to assess the subjective perceived quality provided by the suggested analysis/synthesis procedure. According to this listening test, the resynthesis of aHM-AIR using Peak Picking and aDFT for the f_0 refinement and LS for the final sinusoidal parameter estimation (aDFT-LS), has globally the same high quality as aHM-AIR using the LS solution, which is also confirmed by an objective measurement (i.e. PESQ). Therefore, an approach using Peak Picking applied on aDFT can indeed replace the original LS solution approach of aHM-AIR, while reducing the computational load by four times and keeping the high quality intact.

Future work, based on the methods presented in this thesis, could include an extended version of aHM based on the principles presented in the extended adaptive Quasi-Harmonic Model (eaQHM) [32]. Additionally, a more in depth exploration of the adaptive Discrete Fourier Transform (aDFT) is in order (i.e. definition for both positive and negative frequencies). Furthermore, it would be interesting to find other applications where applying aDFT instead of DFT could provide better results.

Bibliography

[1] L. Almeida and J. Tribolet, "Harmonic coding: A low bit-rate, good-quality speech coding technique," *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '82*, 1982.

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

[3] Y. Hu and P. C. Loizou, "On the importance of preserving the harmonics and neighboring partials prior to vocoder processing: Implications for cochlear implants," *Journal of Acoustic Society of America*, vol. 127, no. 1, p. 427–434, 2010.

[4] J. Jensen and J. Hansen, "Speech enhancement using a constrained iterative sinusoidal model," *Speech and Audio Processing, IEEE Transactions on*, vol. 9, no. 7, pp. 731–740, 2001.

[5] Y. Stylianou, *Modeling Speech based on Harmonic Plus Noise Models.* Springer Berlin / Heidelberg, 2005, pp. 244–260.

[6] G. Kafentzis, G. Degottex, O. Rosec, and Y. Stylianou, "Time-scale modifications based on a full-band adaptive harmonic model," in *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, Canada, August 2013.

[7] M. Campedel-Oudot, O. Cappe, and E. Moulines, "Estimation of the spectral encelope of voiced sounds using a penalized likelihood approach," *IEEE Trans. on Speech and Audio Processing*, vol. 9, pp. 469–481, 2001.

[8] G. Degottex, A. Roebel, and X. Rodet, "Phase minimization for glottal model estimation," *IEEE Trans. on Audio, Speech and Lang. Processing*, vol. 19, pp. 1080–1090, 2011.

[9] D. W. Griffin and J. S. Lim, "Multiband Excitation Vocoder," *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 36, no. 8, pp. 1223–1235, 1988.

[10] X. Serra, “A system for sound analysis, transformation, synthesis based on a determistic plus stochastic decomposition,” Ph.D. dissertation, Stanford University, 1989.

[11] G. Fant, “The lf-model revisited. transformations and frequency domain analysis,” *STL-QPSR*, vol. 36, pp. 119–156, 1995.

[12] B. Doval, C. D’Alessandro, and N. Henrich, “The Spectrum of Glottal Flow Models,” *Acta Acustica*, vol. 92, pp. 1009–1025, 2006.

[13] M. Kepesi and L. Weruaga, “Adaptive chirp-based time-frequency analysis of speech,” *Speech Communication*, vol. 48, pp. 474–492, 2006.

[14] Y. Pantazis, G. Tzedakis, O. Rosec, and Y. Stylianou, “Analysis/Synthesis of Speech based on an Adaptive Quasi-Harmonic plus Noise Model,” in *Proc. IEEE ICASSP*, Dallas, Texas, USA, Mar 2010.

[15] Y. Pantazis, O. Rosec, and Y. Stylianou, “Adaptive AM-FM signal decomposition with application to speech analysis,” *IEEE Trans. on Audio, Speech, and Lang. Proc.*, vol. 19, pp. 290–300, 2011.

[16] G. Degottex and Y. Stylianou, “Analysis and synthesis of speech using an adaptive full-band harmonic model,” *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 10, pp. 2085–2095, 2013.

[17] Y. Pantazis, O. Rosec, and Y. Stylianou, “Iterative estimation of sinusoidal signal parameters,” *IEEE Signal Processing Letters*, no. 5, pp. 461–464, May 2010.

[18] Y. Stylianou, “Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification,” Ph.D. dissertation, E.N.S.T - Paris, 1996.

[19] W. Oomen and A. C. den Brinker, “Sinusoids plus noise modelling for audio signals,” in *International Conference of Audio Engineering Society Conference*, 1999.

[20] S. Levine, “Audio representations for data compression and compressed domain processing,” Ph.D. dissertation, Stanford University, 1999.

[21] H. Thornburg, “Detection and modeling of transient audio signals with prior information,” Ph.D. dissertation, Stanford University, 2005.

[22] R. J. McAulay and T. F. Quatieri, “Speech Analysis/Synthesis based on a Sinusoidal Representation,” *IEEE Trans. on Acoust., Speech and Signal Processing*, vol. 34, no. 4, pp. 744–754, 1986.

[23] Y. Pantazis and Y. Stylianou, “Improving the Modeling of the Noise Part in the Harmonic plus Noise Model of Speech,” in *Proc. IEEE ICASSP*, Las Vegas, Nevada, USA, Apr 2008.

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

[25] ——, “AM-FM estimation for speech based on a time-varying sinusoidal model,” in *Interspeech*, Brighton, Sep 2009.

[26] V. Morfi, G. Degottex, and A. Mouchtaris, “A computationally efficient refinement of the fundamental frequency estimate for the adaptive harmonic model,” in *Proc. IEEE ICASSP*, Florence, Italy, May 2014.

[27] G. Degottex and Y. Stylianou, “A full-band adaptive harmonic representation of speech,” in *Interspeech*, Portland, Oregon, U.S.A, 2012.

[28] A. Camacho, “Swipe: A sawtooth waveform inspired pitch estimator for speech and music,” Ph.D. dissertation, University of Florida, USA, 2007.

[29] D. B. Paul, “The spectral envelope estimation vocoder,” *IEEE Trans. on Acoust., Speech and Signal Processing*, vol. 29, pp. 786–794, 1981.

[30] S. Maeda, “A digital simulation method of the vocal-tract system,” *Speech Communication*, vol. 1, pp. 199–229, 1982.

[31] A. Rix, J. Beerends, M. Hollier, and A. Hekstra, “Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs,” *Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP ’01). 2001 IEEE International Conference on*, vol. 2, 2001.

[32] G. P. Kafentzis, Y. Pantazis, O. Rosec, and Y. Stylianou, “An Extension of the Adaptive Quasi-Harmonic Model,” in *Proc. IEEE ICASSP*, Kyoto, March 2012.