University of Crete
Department of Computer Science

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

Voice tremor detection using Adaptive Quasi-Harmonic Model

(MSc. Thesis)

Maria Koutsogiannaki

Heraklion

October 2010

Maria

Digitally signed by Maria
DN: cn=Maria, c=GR, email=mkoutsog@teemail.gr
Reason: I am the author of this document
Date: 2010.10.23 19:04:53 +03'00'

Department of Computer Science
University of Crete

Voice tremor detection using Adaptive Quasi-Harmonic Model

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

October 22, 2010

Author:

Maria Koutsogiannaki
Department of Computer Science

Committee

Supervisor

Yannis Stylianou
Associate Professor

Member

Athanasios Mouchtaris
Assistant Professor

Member

Panagiotis Tsakalides
Professor

Accepted by:

Chairman of the
Graduate Studies Committee

Panos Trahanias
Professor

Heraklion, October 2010

Abstract

Speech along with hearing is the most important human ability. Voice does not only audibly represents us to the world, but also reveals our energy level, personality, and artistry. Possible disorders may lead to social isolation or may create problems on certain profession groups. Most singers seek professional voice help for vocal fatigue, anxiety, throat tension, and pain. All these symptoms must be quickly addressed to restore the voice and provide physical and emotional relief.

Normophomic and dysphonic speakers have a mutual voice characteristic. Tremor, a rhythmic change in pitch and loudness, appears both in healthy subjects and in subjects with voice disorders. Physiological tremor or microtremor appears to be a derivative of natural processes. Pathological tremor, however, is distinguishable and characterized by strong periodical patterns of large amplitude that affect the quality of voice and influence the ability of patient's communication. However, researches examine not only pathological but physiological tremor as well, since they believe that it may be the first or only symptom of a neurological disease, or may indicate vocal fatigue. Therefore, the analysis of vocal microtremor in normophonic speakers is also important.

Traditional methods of vocal tremor detection involve visual inspection of oscillograms or spectrograms. A more accepted approach is the estimation of the fundamental frequency of the voice signal and then the extraction of the attributes of the signal that modulates the fundamental frequency, namely its amplitude and frequency. However, current methods for vocal tremor estimation are characterized by three limitations: a) the extraction only of the first harmonic for analysis, b) the short duration of the analyzed sustained vowel, and c) the use of a single value to represent a time-varying signal as tremor.

This thesis presents and validates a novel accurate method for the estimation of the vocal tremor characteristics on sustained vowels uttered by normophonic and dysphonic subjects and defines the attributes that define vocal tremor, that is, the leveled modulation amplitude of the harmonics of the signal and its deviation. The extraction of vocal tremor characteristics is performed in three steps. The first step consists of the estimation of the instantaneous amplitude and instantaneous frequency of every sinusoid-component of the speech signal using a recently proposed AM-FM decomposition algorithm, the so-called Adaptive Quasi-Harmonic Model. AQHM is an adaptive algorithm which is able to represent accurately multi-component AM-FM signals like speech. Moreover, AQHM estimates all the instantaneous components of speech, and thus, in contrast to previous studies, any of the instantaneous components can be used for the analysis of vocal tremor and not only the first harmonic. The second step concerns the subtraction from the instantaneous component of the very slow modulations that are derived from the pulsation of the heart. This is achieved by filtering the instantaneous component using a Savitzky-Golay

smoothing filter. Finally, at the third step the modulation frequency and the modulation level of the analyzed instantaneous component are estimated. The analyzed instantaneous component is assumed to contain time-varying features, since the modulations are primarily non-stationary. The estimation is performed by employing the AQHM algorithm and two distinct evaluation approaches namely, the Extended Kalman Smoother and the Hilbert transform. Finally, the efficiency of the algorithm is validated on four databases containing normophonic and dysphonic speakers.

Περίληψη

Μια από τις σημαντικότερες ανθρώπινες λειτουργίες είναι η φωνή, διότι όχι μόνο βοηθάει στην καθημερινή επικοινωνία του ατόμου, αλλά αποκαλύπτει το σύνολο των ιδιαίτερων ψυχικών, πνευματικών και καλλιτεχνικών του χαρακτηριστικών. Πιθανές διαταραχές στην φωνή μπορούν να οδηγήσουν στην κοινωνική απομόνωση του ατόμου ή να δημιουργήσουν πρόβλημα σε ορισμένες κατηγορίες επαγγελματιών. Πολλοί επαγγελματίες όπως π.χ τραγουδιστές απευθύνονται σε ειδικούς ώστε να ανακουφιστούν από τα συμπτώματα φωνητικής κούρασης, αισθήματος έντασης και πόνου στο λαιμό. Επομένως, η διάγνωση των συμπτωμάτων θα πρέπει να γίνεται άμεσα, ώστε να επέρχεται η ψυχική και η σωματική ανακούφιση του ασθενούς.

Ένα κοινό χαρακτηριστικό που συναντάται ανάμεσα σε φυσιολογικούς ομιλητές και ομιλητές που πάσχουν από δυσφωνία είναι το τρέμουλο. Το τρέμουλο της φωνής είναι μια ανεπαίσθητη ρυθμική αλλαγή του ηχοχρώματος και της έντασης της φωνής. Το φυσιολογικό τρέμουλο θεωρείται ότι προέρχεται από τις φυσιολογικές συσπάσεις των μυών που βρίσκονται στο ανώτερο αναπνευστικό σύστημα. Από την άλλη πλευρά, το παθολογικό τρέμουλο χαρακτηρίζεται από ευδιάκριτες και έντονες περιοδικές αλλαγές του σήματος της φωνής, με αποτέλεσμα να επηρεάζεται η ποιότητα και η ικανότητα ομιλίας του ασθενούς. Οι ερευνητές δεν ασχολούνται, όμως, μόνο με την μελέτη του παθολογικού τρέμουλου, αλλά και με την μελέτη του φυσιολογικού, διότι θεωρείται ως το μοναδικό σύμπτωμα ορισμένων νευρολογικών ασθενειών ή φωνητικής κούρασης. Επομένως, η μελέτη του τρέμουλου σε φυσιολογικούς ομιλητές είναι εξίσου σημαντική.

Για την ανίχνευση του τρέμουλου έχουν προταθεί κατά καιρούς διάφορες μέθοδοι. Οι πιο παραδοσιακές περιλαμβάνουν την οπτική ανίχνευσή του με την βοήθεια παλμογράφων ή φασματογράφων. Μια πιο διαδεδομένη και ευρέως αποδεκτή μέθοδος είναι η εκτίμηση του σήματος της θεμελιώδους συχνότητας που διαμορφώνει το σήμα της φωνής και η εξαγωγή των χαρακτηριστικών του σήματος που διαμορφώνει την θεμελιώδη συχνότητα, δηλαδή το πλάτους και της συχνότητας του σήματος αυτού. Οι μέθοδοι όμως που έχουν προταθεί μέχρι στιγμής εμφανίζουν τα εξής προβλήματα 1) μπορούν να εκτιμήσουν μόνο το σήμα της πρώτης αρμονικής 2) το σήμα της φωνής που αναλύεται πρέπει να είναι μικρό σε διάρκεια και 3) χρησιμοποιούν μια τιμή για να αντιπροσωπεύσουν ένα χρονικά μεταβαλλόμενο σήμα

Η διατριβή αυτή παρουσιάζει μια καινούρια και ακριβή μέθοδο για την εκτίμηση των χαρακτηριστικών του φωνητικού τρέμουλου σε διαρκή φωνήματα φυσιολογικών και παθολογικών ομιλητών και καθορίζει τα χαρακτηριστικά εκείνα που ορίζουν το τρέμουλο της φωνής, δηλαδή την συχνότητα διαμόρφωσης, το κανονικοποιημένο πλάτος διαμόρφωσης και τη διακύμανση του πλάτους διαμόρφωσης των αρμονικών του σήματος φωνής. Η εκτίμηση των χαρακτηριστικών του φωνητικού τρέμουλου υλοποιείται σε 3 βήματα. Το πρώτο βήμα περιλαμβάνει την εκτίμηση των σημάτων που διαμορφώνουν το σήμα φωνής, δηλαδή των χρονικά μεταβαλλόμενων πλατών και συχνοτήτων των σημάτων αυτών. Η ανάλυση αυτή γίνεται με ένα προσφάτως προτεινόμενο αλγόριθμο, το προσαρμοστικό σχεδόν-αρμονικό μοντέλο (Adaptive Quasi-Harmonic Model - AQHM). Ο αλγόριθμος αυτός έχει την ικανότητα να εκτιμά με ακρίβεια τις ημιτονοειδείς συνιστώσες που διαμορφώνουν ένα σήμα. Σε αντίθεση με άλλους

αλγορίθμους, μπορεί να εξάγει και να αναλύσει οποιοδήποτε σήμα διαμόρφωσης του σήματος φωνής και όχι μόνο την πρώτη αρμονική. Το δεύτερο βήμα περιλαμβάνει την απομάκρυνση από την αρμονική των σημάτων πολύ χαμηλών συχνοτήτων που προέρχονται από τους χτύπους της καρδιάς. Η μέθοδος που χρησιμοποιείται για την απομάκρυνση των συχνοτήτων αυτών είναι το φίλτρο εξομάλυνσης Savitzky-Golay. Τέλος, το τρίτο βήμα περιλαμβάνει την εκτίμηση των χρονικά μεταβαλλόμενων σημάτων συχνότητας και πλάτους που διαμορφώνουν το αρμονικό σήμα. Η εκτίμηση αυτή γίνεται χρησιμοποιώντας τον αλγόριθμο AQHM, ο οποίος και συγκρίνεται στο βήμα αυτό με δυο μεθόδους, τον μετασχηματισμό Hilbert και τον Extended Kalman Smoother αλγόριθμο. Η αποδοτικότητα του αλγορίθμου αξιολογείται πάνω σε 4 διαφορετικές βάσεις που περιέχουν σήματα φυσιολογικών ομιλητών και ομιλητών με σπασμωδική δυσφωνία.

Ευχαριστίες

Η διατριβή αυτή αποτελεί το επισφράγισμα της προσπάθειας δύο περίπου ετών για την απόκτηση του μεταπτυχιακού διπλώματος ειδίκευσης του τμήματος επιστήμης υπολογιστών του πανεπιστήμιου Κρήτης και ο απότοκος της ερευνητικής εμπειρίας που απέκτησα κατά την διάρκεια συνεργασίας μου με εργαστήριο τηλεπικοινωνιών και δικτύων του ινστιτούτου επιστήμης υπολογιστών στο Ιδρύματος Τεχνολογίας και Έρευνας.

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα της εργασίας μου καθηγητή Ιωάννη Στυλιανού, για την ευκαιρία που μου έδωσε να γίνω μέλος της ομάδας του και για την πολύτιμη καθοδήγηση του στην εκπόνηση της εργασίας μου. Επίσης τον συγχαίρω γιατί είναι πρωτίστως άνθρωπος και έδειξε κατανόηση σε προβλήματα που παρουσιάστηκαν στη διάρκεια των σπουδών μου.

Ευχαριστώ θερμά τον διδάκτορα Γιάννη Πανταζή, για την πολύτιμη συμβολή του στην ολοκλήρωση της εργασίας μου. Οι προτάσεις και συμβουλές του υπήρξαν καθοριστικές. Ένα μεγάλο κομμάτι της εργασίας αυτής στηρίζεται σε δική του δουλειά.

Θα ήθελα να ευχαριστήσω τον διδάκτορα Γιώργο Τζαγκαράκη για την καταλυτική βοήθειά του στην τελική μορφή της εργασίας.

Τέλος, ευχαριστώ τα παιδιά του εργαστηρίου για το ευχάριστο κλίμα που δημιούργησαν, την οικογενειά μου και τους φίλους μου για την συμπαράστασή τους και την ανοχή που έδειξαν στις δύσκολες στιγμές.

Οι ευχαριστίες αυτές γράφτηκαν στα ελληνικά για να μπορούν οι αξιολάτρευτοι γονείς μου να τις διαβάσουν. Μπορεί να μην έχουν την κατάλληλη μόρφωση, αλλά έχουν καλοσύνη, υπομονή και ανθρωπιά, αρετές πιο ισχυρές από την αρετή της γνώσης.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Speech along with hearing is the most important human ability. Voice does not only audibly represents us to the world, but also reveals our energy level, personality, and artistry. Possible disorders may lead to social isolation or may create problems on certain profession groups. Most singers seek professional voice help for vocal fatigue, anxiety, throat tension, and pain. All these symptoms must be quickly addressed to restore the voice and provide physical and emotional relief.

1.1 Importance of detecting tremor in voice

Tremor is an involuntary, rhythmic muscular contraction characterized by oscillations of one or more body parts such as hands, arms, head, face, vocal cords, trunk, and legs [1]. Tremor is either a symptom of a disease of the central nervous system or a neurological disorder itself called essential tremor (ET). As a symptom of a disease, tremor can be met in subjects with Parkinson disease [2], multiple sclerosis [3], stroke [4], or alcohol withdrawal [5]. However, tremor is not only associated with a pathological state. Physiological tremor is an inevitable side effect of any muscle activity. Factors that have been recognized as the major determinants of physiological tremor are rhythmic changes due to pulsatile blood flow, breathing and mechanically or neurally determined oscillations [6], [7]. In contrast to pathological tremor, physiological tremor is not distinguishable.

The previous paragraph refers to tremor of the limbs or other body parts. Tremor in voice is perceived as rhythmic changes in pitch and loudness. Voice production involves the participation of lungs, larynx and vocal tract. Since tremor is defined as an involuntary and rhythmic muscular contraction, oscillations in any of the muscles of the above speech organs may be the source of vocal tremor. Tremor can occur from oscillations of the muscles of the respiratory system [8], oscillations in tension or mass of the vocal folds [9], involuntary rhythmic movements of the intrinsic or extrinsic laryngeal oscillations in the upper vocal tract including lips, tongue, jaw and pharynx [9], [10], or oscillations in certain muscles such as thyroarytenoid and cricothyroid [11]. Natural oscillations of the muscles of the speech organs result to physiological vocal tremor. Phys-

iological vocal tremor, also referred as vocal microtremor [12] to distinguish it from pathological tremor, is not perceivable. In voice signals it appears as a low modulation of the fundamental frequency and a modulation of the amplitude [13], [14]. Pathological tremor, however, may be a disease (essential tremor of voice) or a symptom of a neurological disease and the corresponding voice signals are characterized by strong periodical patterns of large amplitude that affect the quality of voice and influence the ability of patient's communication.

Although vocal microtremor appears to be a derivative of natural processes, researches believe that it may be the first or unique symptom of a neurological disease [15] or may indicate vocal fatigue [16]. Therefore, the analysis of vocal microtremor in normophonic speakers is important.

1.2 Defining tremor attributes: Literature review

Traditional methods of vocal tremor detection involve visual inspection of oscillograms or spectrograms. Current methods involve the accurate estimation of the fundamental frequency of the voice signal and then the extraction of the attributes of the signal that modulates the fundamental frequency, namely its amplitude and frequency [12], [17]. The study of the amplitude modulation of a speech signal cannot provide safe conclusions relative to vocal tremor, since the amplitude of the signal is affected possibly by the vocal tract transfer function [12], [18]. Therefore, studies focus on the estimation of the fundamental frequency modulations.

Several methods have been proposed for extracting the time-varying fundamental frequency of a voice signal and its modulations. Ludlow et al. [19] used a low pass filter just below the first formant to extract the fundamental frequency for each speaker. Then, the modulation level was expressed in terms of the vocal jitter, while the modulation frequency was derived by determining the period between maximum positive slopes for each cycle. Winholtz and Ramig [17] proposed a device called vocal demodulator for tremor analysis. The fundamental frequency (F0) was detected by using a zero-crossing method. Then, a low pass filter isolated F0 from the other harmonics. The amplitude modulation was extracted by computing the peak variations of the amplitude in F0. For estimating the frequency modulation, the peak variations of F0's period were computed. Schoentgen [12] estimated the F0 signal by measuring the length of each vocal cycle. The auto-covariance function was used to detect the characteristic vocal cycle (mean F0). Then, the detection of the vocal cycles in the speech signal was performed by comparing this characteristic vocal cycle to the distance between adjacent prominent peaks. Dromey [20] employed the auto-correlation to express the F0 signal. The mean tremor rate was calculated from the time series of F0 as the inverse of the mean duration between peaks and the extent as the normalized mean of difference between the maximum and the minimum value among two successive peaks. Cnockaert et al. [21] used an enhanced method based on the continuous wavelet transform (CWT). CWT was applied in the speech signal to estimate F0. The wavelet central

frequency with the maximal CWT modulus gave an estimation of F0. Then, a CWT with a shorter mother wavelet was applied to estimate the instantaneous frequency. The modulation amplitude was computed by summing the squares of the CWT modulus over the frequency interval $2-15$Hz averaged by F0 and the modulation frequency was defined as the sum over the same frequency interval of the instantaneous frequencies of the CWT of the F0 trace, weighted by the wavelet transform energy.

1.3 Motivation

The algorithms presented briefly in the previous section estimate the fundamental component of speech along with its modulating attributes by measuring the length of each vocal cycle length ([12], [19]) or by estimating the instantaneous frequency of the fundamental frequency of speech ([17], [21]). For estimating vocal tremor, the second method appears to have many advantages compared to the first one. However, there are some issues not addressed in previous studies. Indeed, many studies exploit only the first harmonic, which is related with the fundamental frequency but not for the higher harmonics and the techniques they use to estimate the fundamental component impose the removal of the other harmonics. The first harmonic, however, may be modulated by the first formant especially in high vowels [22] which may lead to biased results. A more serious limitation of the previous studies is that the duration of the analyzed sustained vowel ranges from one to two seconds. The reason for using short duration is that the modulation frequencies, as well as the modulation levels, should be constant in order to apply classical frequency estimation analysis. This can be a main drawback, since the analysis of larger segments of speech may reveal interesting properties of vocal tremor [23], [24].

The objective of this work is to present and validate a novel method for the estimation of the vocal tremor characteristics on sustained vowels uttered by normophonic subjects. The proposed method models speech as a sum of sinusoids with time-varying amplitude and time-varying frequency or equivalently as a multi-component amplitude and frequency-modulated (AM-FM) signal. Moreover, each instantaneous component of speech can be considered to be modulated both in amplitude and frequency. Therefore, it can be modeled as an AM-FM signal as well. Low frequency modulations between $2-15$Hz of the instantaneous component are attributed to vocal tremor. Then, the demodulation of the instantaneous component provides an estimate of the acoustical vocal tremor characteristics, that is, the *modulation level* and the *modulation frequency*.

The extraction of vocal tremor characteristics is carried out in three steps. The first step consists of estimating the instantaneous amplitude and instantaneous frequency of every sinusoid component of the speech signal using a recently proposed AM-FM decomposition algorithm, the so-called Adaptive Quasi-Harmonic Model [22], [25]. AQHM is an adaptive algorithm which is able to represent accurately multi-component AM-FM signals like speech. Moreover, AQHM estimates all the instantaneous components of speech, and thus, in contrast to previous studies any

of the instantaneous components can be used for the analysis of vocal tremor and not only the first harmonic. In the second step, the very slow modulations (< 2Hz) derived from the pulsation of the heart are subtracted from the instantaneous component [26]. This is achieved by filtering the instantaneous component using a Savitzky-Golay smoothing filter [27]. During the final step, the modulation frequency and the modulation level of the analyzed instantaneous component, which is assumed to contain time-varying features and not constant since the modulations are primarily non-stationary, are estimated. The estimation is performed by employing the AQHM algorithm.

In the final step, we also examine the effect of different estimation methods. We implement the frequency tracker based on the Extended Kalman Smoother (EKS) [28] to extract the modulation frequency of the analyzed instantaneous component. EKS frequency tracking algorithm tries to estimate a state which minimizes the error between the observed values and the estimated ones in the previous state. The state to be estimated and the relationship between the state and the observed signal are described by the state-space model. In our case our observed signal is the instantaneous component and the state to be estimated is the modulation frequency. A third approach of extracting the modulation frequency and the modulation level of a monocomponent signal is to create its analytical version using the Hilbert transform and then to estimate the angle and subsequently the instantaneous frequency of the signal.

1.4 Contribution

The present thesis presents and validates a novel method for the estimation of the vocal tremor characteristics on sustained vowels uttered by normophonic and dysphonic subjects.

The contribution of this thesis can be summarized as follows:

1. The design of a three-step method, which estimates accurately the vocal tremor characteristics, that is, the tremor frequency and tremor level in various harmonics.

2. The introduction of a new attribute that defines tremor, namely, the deviation of the modulation level.

3. The accurate extraction of the phonatory frequency and other harmonics of speech without using filters, which may cut off other harmonics or affect the modulating frequencies of the components.

4. The removal of the limitations of the phoneme duration. The Savitzky-Golay smoothing filter used in our method may cut off frequencies below 2Hz, without affecting other modulating frequencies. Therefore, modulating frequencies, which result from the inability of the speaker to keep his voice steady in a specific phonatory frequency, are removed from the signal.

5. The modeling of the instantaneous component and the tremor signal as a monocomponent time-varying sinusoidal signal. Having a mathematical description of the signal is easier to handle and to reveal further properties of the signal. The modulation level and frequency are not single values but signals varying in time.

6. The disengagement of the user. More specifically, the user gives as input only the recorded speech signal of the speaker without defining any parameters.

The experimental results on synthetic signals reveal that the AQHM algorithm estimates accurately the instantaneous amplitude and instantaneous frequency of the signal, whereas the performance of the EKS tracker and the Hilbert transform is decreased. Results on sustained vowels uttered by normophonic speakers show that the proposed method estimates accurately the instantaneous components of speech signals and then extracts robustly the time-varying modulation frequency and modulation level attributed to vocal tremor. On the other hand, the other two approaches fail to estimate the modulation frequency from the instantaneous component.

1.5 Structure of the thesis

The thesis is organized as follows: In Chapter 2, we describe the Adaptive Quasi-Harmonic Model (AQHM) described in [22]. AQHM demodulates speech into components and estimates the tremor attributes of a speech component accurately.

In Chapter 3, we introduce the basic theory of Kalman Filters and their extension to non-linear models, the Extended Kalman Filters (EKF). Then, we use a statistical non-linear model that describes the Instantaneous Tremor Frequency and we apply the EKF to derive the equations for our model. The ITF is finally computed using a smoothed version of the EKF, the so-called Extended Kalman Smoother.

In Chapter 4, we evaluate the performance of the AQHM the EKS and the Hilbert transform on synthetic signals. We demonstrate the superiority of the AQHM algorithm against the other two approaches in the presence of noise and in the case of more than one modulating signals.

In Chapter 5, the proposed approach for detecting tremor in voiced speech signals is analyzed in detail.

In Chapter 6, we evaluate the performance of our method on three databases. Database 1 contains normophonic speakers and is used mostly for evaluation. Database 2 is used to analyze signals from dysphonic speakers and the tremor attribute that distinguishes normophonic to dysphonic speakers is estimated. Database 3, is employed to examine the relationship between the vocal tremor and voice fatigue.

Finally, in Chapter 7, we conclude this work and we propose future research directions.

Chapter 2

Adaptive Quasi-Harmonic Model (AQHM)

The Adaptive Quasi-Harmonic Model is an iterative method for the accurate estimation of the amplitude and frequency modulations (AM-FM) in time-varying, multi-component, quasi-periodic signals, such as the voiced speech. The suggested method is based on a time-varying, quasi-harmonic representation of speech referred to as Quasi-Harmonic Model (QHM). QHM assumes that an initial estimate for the frequencies of the components is provided, as well as that the number of components is known a priori. Then, the remaining parameters are estimated by minimizing the mean squared error between the speech signal and the model, which leads to a least squares (LS) solution. In practice, however, a frequency mismatch between the original and the initial estimates of the frequencies is inevitable, but QHM corrects the potential mismatches via an iterative estimation process.

Similarly to the sinusoidal modeling, QHM assumes speech to be locally stationary. However, QHM is limited in the sense that it can capture variations of frequencies and amplitudes but only up to a certain point. This limitation is overcome by an adaptive version of QHM (AQHM), where the speech signal is not assumed to be locally stationary by taking into account the trajectories of the instantaneous frequencies. Then, the speech signal is projected in a space generated by time-varying, non-parametric sinusoidal basis functions. Therefore, the basis functions are adapted to the local characteristics of the input signal. The basis functions are updated by minimizing the mean squared error between the input signal and the AQHM model at each adaptation step. This leads to a non-parametric AM-FM decomposition algorithm for speech signals.

2.1 AQHM and speech decomposition

A well-established approach in the theory of speech processing is to model voiced speech signals as a sum of time-varying sinusoidal components. More specifically, a voiced speech signal $s(t)$

can be expressed by a series expansion as follows:

$$s(t) = \sum_{k=1}^{K} A_k(t) cos(\phi_k(t)) \ , \tag{2.1}$$

where K is the number of components, while $a_k(t)$ and $\phi_k(t)$ are the instantaneous amplitude and instantaneous phase of the kth component, respectively. Moreover, the instantaneous frequency $f_k(t)$ is defined as the derivative with respect to time of the instantaneous phase scaled by $1/(2\pi)$:

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

AQHM estimates the instantaneous components of speech frame-by-frame, where the lth frame of the speech signal is given by

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

where t_l is the center of the frame and $w(t)$ denotes the analysis window function supported in the time interval $[-T, T]$. Typically, the window function vanishes at the limits of its support so as to alleviate the discontinuities at the boundaries of the frame.

AQHM has an initialization step for the estimation of the instantaneous components and an adaptation step for the refinement of the estimation. At the initialization step of AQHM, the speech frames are modeled by QHM which is able to correct small frequency mismatch errors and acts as a frequency tracker. During the adaptation step of AQHM, the speech frames are modeled by AQHM which is able to adjust its time-varying characteristics to the time-varying characteristics of the analyzed signal.

2.1.1 Initialization of AQHM

In the initialization step, a speech frame, $s_l(t)$, is modeled by the QHM as follows:

$$s_l(t) = \left(\sum_{k=-K}^{K} (a_k^l + t b_k^l) e^{j2\pi f_k^l t} \right) w(t) \ , \tag{2.4}$$

where K specifies the number of sinusoids, f_k^l and a_k^l denote the frequency and the complex amplitude, respectively, and b_k^l is the complex slope of the kth sinusoidal component. Note that K does not depend on the frame index, since we assume that the number of instantaneous components of speech is constant and known in advance.

The estimation of the model parameters $\{f_k^l, a_k^l, b_k^l\}_{k=-K}^{K}$ for the lth frame is performed in two steps. First, an initial frequency estimate $\hat{f}_k^l$ for each k is provided. Then $\{a_k^l, b_k^l\}_{k=-K}^{K}$ are estimated by minimizing the mean squared error between the speech frame and the associated

QHM which leads to a linear Least Squares (LS) solution. The initially estimated frequencies $\hat{f}_k^l$ of the lth frame are assigned as integer multiples of an estimated fundamental frequency $\hat{f}_0^{l-1}$ of the previous frame that is, $\hat{f}_k^l = k\hat{f}_0^{l-1}$. An estimate of the fundamental frequency for the lth frame is given by the following expression:

$$\hat{f}_0^l = \hat{f}_0^{l-1} + \frac{1}{K_f}\sum_{k=1}^{K_f}\frac{\rho_{2,k}^l}{k} \ , \tag{2.5}$$

where $\rho_{2,k}^l = \mathcal{R}e\left\{\frac{jb_k^l}{a_k^l}\right\}$. As it was shown in [25], $\rho_{2,k}^l$ can be viewed as an estimate of the frequency mismatch between the true frequency of the signal and its estimated value, while K_f is a small integer, typically 3 to 5. The fundamental frequency of the first frame is computed using the autocorrelation function of the first frame [29].

In the second step, the instantaneous components are computed at a given time-instant t from the parameters of the QHM as follows:

$$M_k(t) = |a_k + tb_k| = \sqrt{(a_k^R + tb_k^R)^2 + (a_k^I + tb_k^I)^2} \tag{2.6}$$

$$\Phi_k(t) = 2\pi f_k t + atan\frac{a_k^I + tb_k^I}{a_k^R + tb_k^R} \tag{2.7}$$

$$F_k(t) = \frac{1}{2\pi}\Phi_k'(t) = f_k + \frac{1}{2\pi}\frac{a_k^R b_k^I - a_k^I b_k^R}{M_k^2(t)} \tag{2.8}$$

where $M_k(t)$ is the instantaneous amplitude, $\Phi_k(t)$ is the instantaneous phase and $F_k(t)$ is the instantaneous frequency of the k^{th} component. If b_k is zero then we have the Sinusoidal Model where for each frame one value of amplitude, phase and frequency are calculated. However, in QHM the existence of the parameter b_k corrects for each frame the initial estimates of $\{M_k, \Phi_k, F_k\}_{k=-K}^K$ for every time instant t and not only at the center of the window t_l. However, in this step the instantaneous components are computed only at the center of the analysis window t_l.

2.1.2 Interpolation of the instantaneous components

If the above method proceeds by employing a time-step of one sample, then the estimation of the instantaneous components is completed. However, when the time-step takes values larger than one sample, the intermediate points of the instantaneous components should be computed by interpolation. For the instantaneous amplitude we suggest the use of linear interpolation, since it guarantees that the instantaneous amplitude will be always positive, which is a necessary condition for the well-positiveness of the instantaneous amplitude. On the other hand, cubic or spline interpolation schemes do not guarantee the positiveness of the instantaneous amplitude. For the instantaneous frequency, the use of spline interpolation is preferred, since it provides smooth estimates of the frequency trajectories, which is considered to be representative of the typical voiced speech. However, such simple solutions are not possible for the interpolation of

the instantaneous phase.

According to the non-parametric approach, the instantaneous phase of the kth component between two consecutive analysis time instants t_{l-1} and t_l can be estimated as the integral of the estimated instantaneous frequency:

$$\check{\phi}_k(t) = \hat{\phi}_k(t_{l-1}) + \int_{t_{l-1}}^{t} 2\pi \hat{f}_k(u) du \tag{2.9}$$

However, this solution does not take into account the frame boundary conditions at the time instant t_l, which means that there is no guarantee that $\check{\phi}_k(t_l) = \hat{\phi}_k(t_l) + 2\pi M$, where M is the closest integer to $|\hat{\phi}_k(t_l) - \check{\phi}_k(t_l)|/(2\pi)$. The phase continuity at the frame boundaries is guaranteed by modifying (2.9) as follows:

$$\hat{\phi}_k(t) = \hat{\phi}_k(t_{l-1}) + \int_{t_{l-1}}^{t} \left(2\pi \hat{f}_k(u) + r_k^l \sin\left(\frac{\pi(u - t_{l-1})}{t_l - t_{l-1}} \right) \right) du \ . \tag{2.10}$$

Note that the derivative of the instantaneous phase over time in both formulas gives the instantaneous frequency estimate in the interval $[t_{l-1},\ t_l]$. Moreover, as it can be seen in (2.10) the instantaneous phase at the upper bound t_l will be equal to $\hat{\phi}_k(t_l) + 2\pi M$, by setting r_k^l to be equal to:

$$r_k^l = \frac{\pi(\hat{\phi}_k(t_l) + 2\pi M - \check{\phi}_k(t_l))}{2(t_l - t_{l-1})} \tag{2.11}$$

where M is computed as before.

2.1.3 Adaptation of the AQHM

During the adaptation step of the AQHM algorithm, the lth frame of a speech signal centered at a time-instant t_l is modeled as follows:

$$s_l(t) = \sum_{k=-K_l}^{K_l} (a_k^l + t b_k^l) e^{j(\hat{\phi}_k(t_l+t) - \hat{\phi}_k(t_l))} w(t) \ , \tag{2.12}$$

where b_k^l plays the same role as in QHM, providing a way to update the frequency of the underlying sine wave at the center of the analysis window, t_l. Note also that the old phase value at t_l, $\hat{\phi}_k(t_l)$, is subtracted from the instantaneous phase, so as the argument of the basis function vanishes at the center of the analysis window. Thus, a new phase estimate at time-instant t_l is obtained again from the argument of a_k^l.

Comparing QHM with AQHM, we first note that the argument of the QHM basis functions is parametric and stationary, while the argument of the AQHM basis functions is neither parametric nor necessarily stationary. Moreover, since the AQHM basis functions use the instantaneous phases, which have been estimated from the analyzed signal, these are also adaptive to the cur-

rent characteristics of the signal. Thus, it is expected to provide more accurate estimates.

The adaptation step can be iterated until changes in the Signal-to-Reconstruction Error Ratio (SRER) are not significant. The SRER is defined by

$$SRER = 20log_{10}\frac{\sigma_{s(t)}}{\sigma_{s(t)-\hat{s}(t)}} \ , \tag{2.13}$$

where $\sigma_{v(t)}$ denotes the standard deviation of a signal $v(t)$, while $\hat{s}(t)$ is the reconstructed speech signal, which is given by

$$\hat{s}(t) = \sum_{k=1}^{K} \hat{A}_k(t)cos(\hat{\phi}_k(t)) \ . \tag{2.14}$$

Chapter 3

Extended Kalman Filter

In this chapter we will present the method which uses the Extended Kalman Filter and the Extended Kalman Smoother to track the instantaneous frequency of a signal. We thoroughly explain the proposed method, we prove from scratch the Extended Kalman Filter equations (see Appendix A) and we derive the equations for our model.

3.1 Tracking frequency using the Extended Kalman Smoother

3.1.1 Model definition and derivation of the Extended Kalman Filter equations

A speech component is modeled as a signal modulated by a time-varying frequency. Therefore, a speech component can be modeled as follows:

$$\psi(n) = \overline{\psi} + \psi_d sin(\varphi(n)) \ , \tag{3.1}$$

where ψ is the mean value of the harmonic, ψ_d is the maximum deviation of the mean value and $\varphi(n)$ is the instantaneous phase of the signal. The objective is to estimate the instantaneous frequency of this signal. Given that our observed signal is

$$y(n) = \psi(n) - \overline{\psi} \ , \tag{3.2}$$

we need a model to describe this signal. The statistical model for $y(n)$ is described by the following equation:

$$y(n) = asin(2\pi T_s \overline{f} n + \theta(n)) + v(n) \tag{3.3}$$

where $T_s = \frac{1}{f_s}$ is the sampling period, $v(n)$ is a white noise process with zero-mean and variance r, a models the ψ_d of the signal, while the phase $\varphi(n)$ of the signal is approximated by the term $2\pi T_s \overline{f} n + \theta(n)$ where $\overline{f}$ denotes the mean value around which the instantaneous frequency fluctuates and is provided by the user. In case of tremor frequencies we set $\overline{f} = 6$Hz.

Finally, we need a model for the instantaneous phase $\theta(n)$. The instantaneous phase of the signal is defined as the integral of the instantaneous frequency. If we estimate this integral with a Riemann sum, we have:

$$\theta(n+1) = \theta(n) + 2\pi T_s f_{in}(n) \tag{3.4}$$

Therefore, the instantaneous frequency $f_i(n)$ is estimated as follows:

$$f_i(n) = \overline{f} + f_{in}(n) \ . \tag{3.5}$$

In our case the tremor frequencies, $f_i(n)$, should not exceed specific limits. More specifically, to prevent $f_i(n)$ from exceeding the interval $[f_{min},\ f_{max}]$ a clipping function s[f] is used, which is defined by

$$s[f] = \begin{cases} f_{max} - \overline{f} & \text{if } f_{max} - \overline{f} \leq f \\ f & \text{if } f_{min} - \overline{f} \leq f < f_{max} - \overline{f} \\ f_{min} - \overline{f} & \text{if } f < f_{min} - \overline{f} \end{cases} \tag{3.6}$$

Taking into account the above clipping function and by combining the instantaneous frequency $f_i(n)$ and angular frequency $u(n)$, (3.4), and (3.5) take the following form:

$$\theta(n+1) = \theta(n) + 2\pi T_s s[\frac{u(n)}{2\pi}] \bmod 2\pi \tag{3.7}$$

$$f_i(n) = \overline{f} + s[\frac{u(n)}{2\pi}] \ . \tag{3.8}$$

The modulus operator has no effect on the model from a mathematical perspective, but keeps $\theta(n)$ bounded and reduces the round off error.

We can now use a model for $u(n)$. In the following, $u(n)$ is modeled as a first-order autoregressive process. Given a time-series X_t, a first-order AR model is given by

$$X_t = c + \phi_1 X_{t-1} + \epsilon_t \ , \tag{3.9}$$

where ϵ_t is a white noise process with zero mean and variance σ_ϵ^2, ϕ_1 is a parameter of the model and c is a constant. If $\phi_1 = 0$ then $var(X_t) = \sigma_\epsilon^2$ and the process reduces to a white noise model. If $\phi_1 = 1$, then $var(X_t) \to \infty$ and the process results to a random walk model. Therefore, we can model $u(n)$ as follows:

$$u(n+1) = \gamma u(n) + w(n) \tag{3.10}$$

where $\gamma = \phi_1$ and $w(n) = \epsilon_t$ with zero mean and variance $\sigma_w^2 = T_s q$. Depending on the signal to be estimated we set $\gamma = 0.001$ or $\gamma = 0.9987$. If the ITF of the signal has a high variance the random walk model ($\gamma = 0.9987$) provides a good estimate of the ITF, whereas if the ITF has low variance then the white noise model ($\gamma = 0.001$) is preferred.

If we define the state vector $x(n) = [\theta(n)\ u(n)]^T = [x_1(n)\ x_2(n)]^T$, then the state-space model can be expressed as:

$$\begin{aligned} x(n+1) &= \begin{bmatrix} x_1(n) + 2\pi T_s s[\frac{x_2(n)}{2\pi}]\ mod\ 2\pi \\ \gamma x_2(n) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix} w(n) = F(n, x(n)) + Gw(n)\ , \\ y(n) &= asin(2\pi T_s \overline{f} n + x_1(n)) + v(n) = C(n, x(n)) + v(n)\ , \end{aligned} \tag{3.11}$$

which is a non-linear state-space model described by (A.29), (A.30). Therefore, the appropriate equations for our state-space model (3.11) can be derive from the equations of Table on Section A.2.2.

Input vector process:

$$Observations = y(1), y(2), ..., y(n)$$

Initial conditions:

$$\hat{x}(1|0) = E[x(1)] = 0$$
$$K(1,0) = \Pi_1 = 0.1I$$

Known parameters:

$$Transition\ matrix = F(n, x(n))$$
$$Measurement\ matrix = C(n, x(n))$$
$$Correlation\ matrix\ of\ noise\ process = Q_1(n)$$
$$Correlation\ matrix\ of\ measurement\ process = Q_2(n)$$

Computed parameters:

$$C(n) = [a\ cos(2\pi T_s \overline{f} n + \hat{x}(n|n-1))\ 0]$$
$$F(n+1, n) = \begin{bmatrix} 1 & T_s s'[\frac{\hat{x}_2(n|n)}{2\pi}] \\ 0 & \gamma \end{bmatrix}$$

Computation for n=1,2,3... :

$$G_f(n) = K(n, n-1)C^H(n)[C(n)K(n, n-1)C^H(n) + Q_2(n)]^{-1}$$
$$a(n) = y(n) - C(n, \hat{x}(n|n-1)) = y(n) - a\ sin(2\pi T_s \overline{f} n + \hat{x}_1(n|n-1))$$
$$\hat{x}(n|n) = \hat{x}(n|n-1) + G_f(n)a(n)$$
$$\hat{x}(n+1|n) = F(n, \hat{x}(n|n)) = \begin{bmatrix} \hat{x}_1(n|n) + 2\pi T_s s[\frac{\hat{x}_2(n|n)}{2\pi}]\ mod\ 2\pi \\ \gamma \hat{x}_2(n|n) \end{bmatrix}$$
$$K(n) = [I - G_f(n)C(n)]K(n, n-1)$$
$$K(n+1, n) = F(n+1, n)K(n)F^H(n+1, n) + Q_1(n)$$
$$\hat{f}_i(n|n) = \overline{f} + s\left[\frac{\hat{x}_2(n|n)}{2\pi}\right]$$

The linearized matrices $F(n+1,n)$, $C(n)$ are updated from the matrices $F(n,x(n))$, $C(n,x(n))$ (A.25) as follows:

$$C(n) = \frac{\partial C(n,x(n))}{\partial x}\Big|_{x=\hat{x}(n|n-1)} = \frac{\partial[a\ sin(2\pi T_s\overline{f}n + x_1(n))\ 0]}{\partial x}\Big|_{x=\hat{x}(n|n-1)}$$
$$= [a\ cos(2\pi T_s\overline{f}n + \hat{x}(n|n-1))\ 0]$$

$$F(n+1,n) = \frac{\partial F(n,x(n))}{\partial x}\Big|_{x=\hat{x}(n|n)} = \frac{\partial[x_1(n) + 2\pi T_s s[\frac{1}{2\pi}x_2(n)]\ mod\ 2\pi\ \ \gamma x_2(n)]^T]}{\partial x}\Big|_{x=\hat{x}(n|n)}$$
$$= \begin{bmatrix} 1 & 2\pi T_s \frac{\partial s[\frac{x_2(n)}{2\pi}]}{\partial x}\Big|_{x=\hat{x}(n|n-1)} \\ 0 & \gamma \end{bmatrix} = \begin{bmatrix} 1 & T_s s'[\frac{\hat{x}_2(n|n)}{2\pi}] \\ 0 & \gamma \end{bmatrix}$$

3.1.2 Applying the Extended Kalman Smoother

After applying the Extended Kalman Filter we apply the Extended Kalman Smoother (EKS) in order to refine our estimation. The EKS takes the estimated matrices from the EKF and estimates the state vector using a backward iteration. The associated equations for our model are given by [28]:

$$\psi(N+1|N) = 0\ , \tag{3.12}$$
$$G(n) = (F(n+1,n)K(n,n-1)C(n)^T)(r + C(n)K(n,n-1)C(n)^T)^{-1}\ , \tag{3.13}$$
$$\psi(n|N) = (F(n+1,n) - G(n)C(n))^T\psi(n+1|N) + C(n)^T(r + C(n)K(n,n-1)C(n)^T)^{-1}a(n)\ , \tag{3.14}$$
$$\hat{x}(n|N) = \hat{x}(n|n-1) + K(n,n-1)\psi(n|N)\ , \tag{3.15}$$
$$\hat{f}(n|N) = \overline{f} + s[\frac{\hat{x}_2(n|N)}{2\pi}]\ , \tag{3.16}$$

where ψ is a variable called adjoined variable, N is the entire record and n=N, N-1, N-2, ..., 1. The final estimate of the instantaneous tremor frequency is given by (3.16).

Chapter 4

Performance evaluation on synthetic signals

Before applying the two algorithms on real voice signals to extract the tremor characteristics, we use synthetic signals to evaluate the accuracy of the algorithms. The synthetic signals simulate an instantaneous component modulated by a stochastic instantaneous tremor frequency.

4.1 Construction of synthetic signals with ITF

To create a synthetic signal, we generate a signal $x(t)$ with an instantaneous frequency $f_i(n)$, by lowpass ($f_c = 0.5$Hz) filtering of white Gaussian noise $v(n)$ with variance σ_v^2. The noise variance determines the variance of the tremor frequency. Then, a mean value $\overline{f} = 6$Hz is added to the resultant signal $v_i(n)$ to create a stochastic tremor frequency $f_i(n)$ around 6Hz, which is an accepted tremor value. Besides, the instantaneous phase is calculated as the integral of the instantaneous frequency using the Riemann sum as follows:

$$\phi(n) = \frac{2\pi}{f_s} \sum_{k=1}^{n} f_k t \ , \tag{4.1}$$

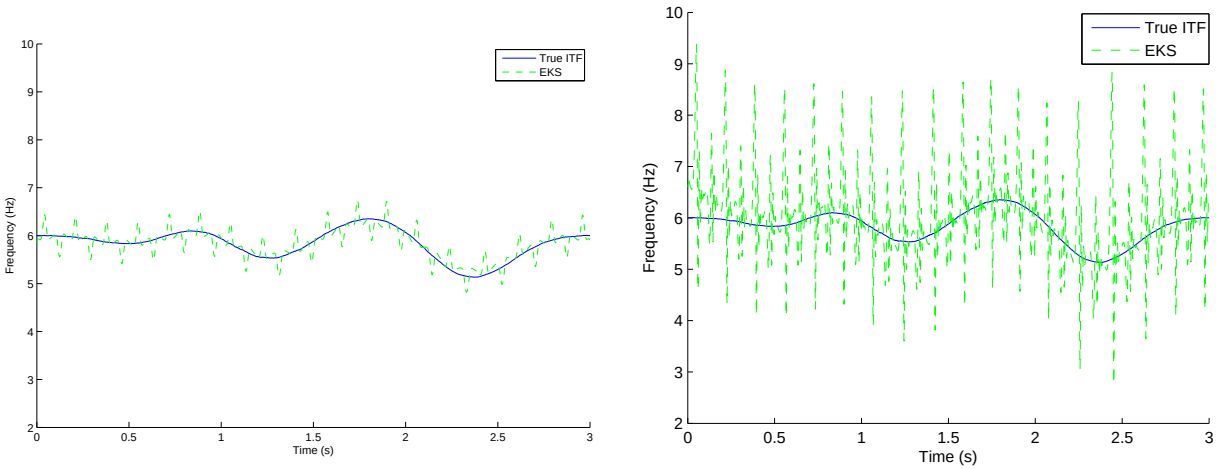
where a sample rate $f_s = 1000$Hz is selected. Then, the instantaneous component,which is modulated by the instantaneous frequency $f_i(n)$, is given by

$$r(n) = \overline{r} + r_d cos(\phi(n)) \ , \tag{4.2}$$

where $\overline{r}$ is the mean value of the instantaneous component and r_d is the maximum deviation of the mean value. In the subsequent derivations we set $r_d = 80$Hz and $\overline{r} = 100$Hz.

4.2 Accuracy enhancement of EKS algorithm on synthetic signals

The efficiency of the EKS tracker, presented in Chapter 3, depends mainly on three parameters. The first parameter is the mean frequency $\overline{f}$ of the modulating signal, which must be estimated accurately. For our synthetic signals $\overline{f}$ is provided by the user ($\overline{f} = 6$Hz). The second model parameter denoted by γ, is defined in the interval $(0, 1)$. In particular, smaller values of γ make it easier for the EKS to track frequencies close to $\overline{f}$, but make it more difficult to track frequencies far from $\overline{f}$. This is demonstrated in the graphs below, where we have created a synthetic signal with low variance (Figure 4.1) and a synthetic signal with high variance (Figure 4.2), while we estimated the ITFs using the two different models, namely, the white noise model ($\gamma = 0.001$) and the random walk model ($\gamma = 0.9987$). The efficiency of the EKS algorithm is evaluated using the signal-to-reconstructed error ratio (SRER) defined in (2.13). Figure 4.1 shows that the white noise model is appropriate for signals with low frequency variance, since it achieves a high SRER value. On the other hand, Figure 4.2 reveals the ability of the random walk model to capture signals with high frequency variance.

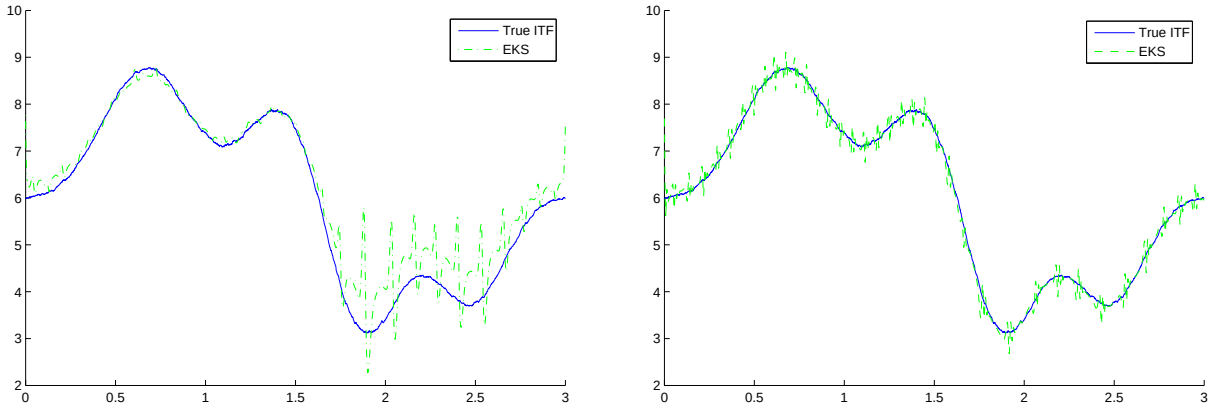


(a) $\gamma = 0.001$, $SRER = 17.0882$

(b) $\gamma = 0.9987$, $SRER = -4.1711$

Figure 4.1: EKS estimation of the tremor frequency for a synthetic signal with low variance $\sigma_v^2 = 10$ a) white noise model b) random walk model.

The third parameter that affects the efficiency of EKS is denoted by $\lambda = \frac{r}{q}$. This parameter determines how quickly the tracker adapts the state variables to changes in the observed signal. More, specifically, if λ is large the tracker fails to adapt to rapid changes in the instantaneous tremor frequency (ITF), while for a small λ the tracker becomes more sensitive to changes in the ITF but also to artifact and noise. In the subsequent analysis we set $\lambda = 100$. The value's selection is based on experimental results during the development of the tracker.



(a) $\gamma = 0.001$, $SRER = -1.3926$

(b) $\gamma = 0.9987$, $SRER = 16.6140$

Figure 4.2: EKS estimation of the tremor frequency for a synthetic signal with high variance $\sigma_v^2 = 60$ a) white noise model b) random walk model.

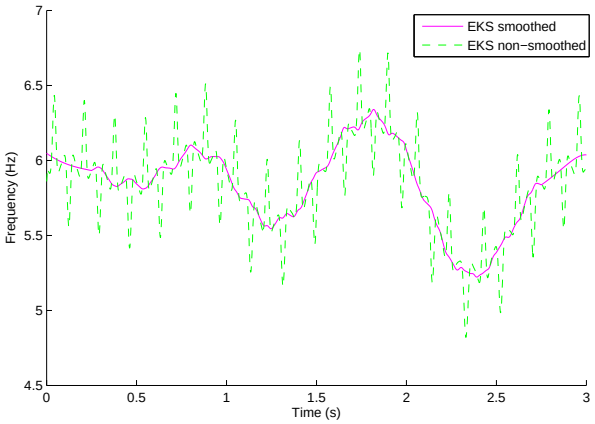
Since tremor frequency may range from 2 to 15 Hz, in our implementation we use both models, that is, the white noise model and the random walk model for the estimation of the instantaneous frequency. First, we compute the instantaneous frequency for $\gamma = 0.001$ and for $\gamma = 0.9987$.

As a second step, we apply a Savitzky-Golay (S-G) filter to enhance the EKS estimate of the ITF, since the estimated ITF may include noise. The noise is introduced by the EKS tracker in an attempt to track the correct frequency of the signal. The input noise depends on the parameter λ. For the Savitzky-Golay filter the order of the local polynomial is set to 4, while the frame size is set to $0.5s$ (501 samples). The S-G filter estimates the trend of the ITF and ignores the fast noisy variations (Figure 4.3(a)). The smoothed ITF, as Figure 4.3(b) shows, is very close to the true ITF.

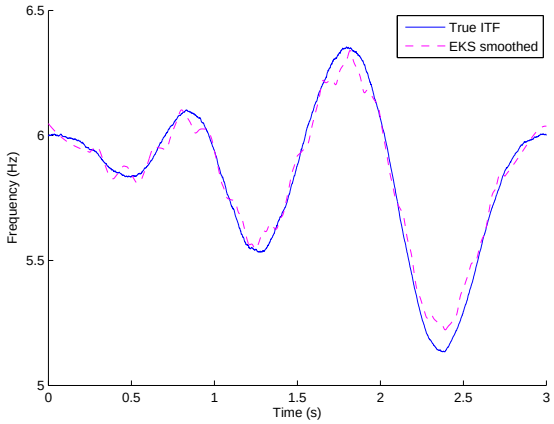
In order to examine if there is an actual improvement with the smoothing technique we generated synthetic signals of different variance (one hundred signals for each variance value σ_v^2 as Table 4.1 shows), for which we estimated the ITF using the EKS approach and then we reconstructed two signals: a) the signal from the ITF estimated using the EKS and b) the signal from the smoothed ITF estimated by the EKS. We compare each reconstructed signal $\hat{x}$ with the true signal x using the Mean Absolute Error (MAE) defined by

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{x}_i - x_i| \ , \tag{4.3}$$

where n is the number of the signal's samples. Moreover, in order to examine the efficiency of



(a) EKS estimate and smoothed EKS estimate of the ITF



(b) Smoothed EKS estimate of the and true ITF

Figure 4.3: EKS estimate and smoothed EKS estimate of the tremor frequency for a synthetic signal with high variance $\sigma_v^2 = 60$.

the smoothing technique, we compute the percentage change of MAE defined by

$$\text{Percentage change} = \frac{MAE_2 - MAE_1}{MAE_1} \ , \tag{4.4}$$

where MAE_2 and MAE_1 are the MAE values with and without applying the smoothing technique, respectively. Negative percentage change indicates a decrease in the MAE and therefore, an improvement in the estimated signal when applying the smoothing technique. Table 4.1 shows that the improvement is apparent in signals with different variance and with noise (25dB) since the percentage change is negative. Moreover, as the σ_v^2 value increases the percentage change increases, which means that the smoothing method is more efficient for low values of the σ_v^2. Furthermore, the percentage change has lower values in signals with noise than in signals without noise, but this change is not significant. In any case, the smoothing method with the S-G filter makes EKS tracker more efficient.

Finally, after computing the SRER value of the reconstructed component from the smoothed instantaneous frequency for $\gamma = 0.001$, and the SRER value of the reconstructed component from the smoothed instantaneous frequency for $\gamma = 0.9987$, we compare the SRER values and select the smoothed instantaneous frequency which results to the greatest value of SRER.

4.3 Accuracy enhancement of Hilbert transform algorithm on synthetic signals

The Hilbert transform is useful in calculating instantaneous attributes of a time series, especially the amplitude and frequency. More specifically, if we want to estimate the instantaneous

		MAE		
		without smoothing	with smoothing	Percentage change %
variance $\sigma_v^2 = 10$	without noise	0.079	0.018	-77
	with noise	0.238	0.038	-84
variance $\sigma_v^2 = 30$	without noise	0.118	0.049	-58
	with noise	0.327	0.135	-59
variance $\sigma_v^2 = 60$	without noise	0.445	0.379	-15
	with noise	0.806	0.637	-21

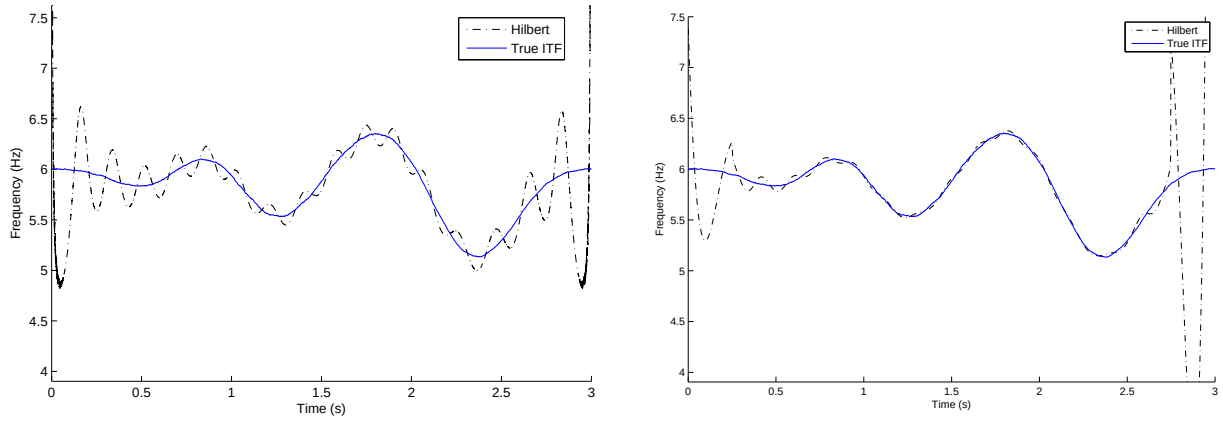
Table 4.1: EKS smoothing efficiency for synthetic signals

frequency of a mono-component signal ψ we create its analytical signal as follows:

$$y = \psi + iH(\psi) \ , \tag{4.5}$$

where $H(\psi)$ is the Hilbert transform of the signal ψ. If $\psi(n) = Acos(2\pi fn)$, then $H(\psi(n)) = Asin(2\pi fn)$, which yields $y(n) = Acos(2\pi fn) + iAsin(2\pi fn) = Ae^{i2\pi fn} = Ae^{\theta(n)}$. The instantaneous amplitude is the amplitude of the complex Hilbert transform and the instantaneous frequency is the time rate of change of the instantaneous phase angle: $f(n) = \frac{\theta(n)-\theta(n-1)}{2\pi f_s}$, where $\theta(n) = 2\pi fn$.

Figure 4.4(a) illustrates the Hilbert estimate of the ITF for a low variance synthetic signal. The performance of the Hilbert algorithm decreases at the beginning and at the end of the signal. To remove the fast noisy variations from the signal we apply the S-G filter. On the other hand, Figure 4.4(b) presents the smoothed Hilbert estimate of ITF. As it can be seen, this approach is quite efficient mainly at the "body" of the signal.



(a) Hilbert estimate of the ITF

(b) Smoothed Hilbert estimate of the ITF

Figure 4.4: Hilbert estimation of the tremor frequency for a synthetic signal with low variance $\sigma_v^2 = 10$. a) without the S-G filter b) after applying the S-G filter.

We construct a similar table as in Table 4.1 in order to examine if there is an actual improvement with the smoothing technique. Table 4.2 shows that the improvement is apparent in signals with different variance and with noise (25dB) since the percentage change (defined in (4.4)) is negative. Moreover, as the σ_v^2 value increases the percentage change remains at the same levels which means that the efficiency of the Hilbert algorithm does not depend on the variance σ_v^2 of the signal. On the other hand, the percentage change has significant lower values in signals with noise than in signals without noise. This reveals that the smoothing method is more efficient, when applied in signals with noise. In any case, the smoothing method with the S-G filter makes the Hilbert algorithm more efficient.

		MAE		
		without smoothing	with smoothing	Percentage change %
variance $\sigma_v^2 = 10$	without noise	0.151	0.111	-26.5
	with noise	6.622	0.128	-98.1
variance $\sigma_v^2 = 30$	without noise	0.168	0.131	-22.0
	with noise	6.636	0.143	-97.8
variance $\sigma_v^2 = 60$	without noise	0.178	0.126	-29.2
	with noise	6.667	0.151	-97.7

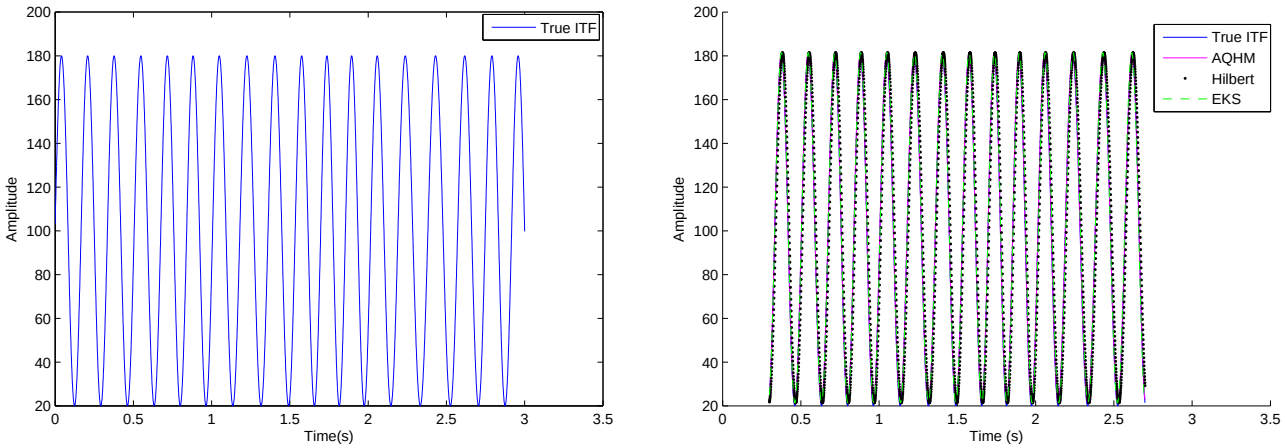
Table 4.2: Hilbert smoothing efficiency for synthetic signals

4.4 Performance evaluation of EKS, Hilbert transform and AQHM on synthetic signals

In the previous section, we enhanced the performance of the EKS and the Hilbert algorithm aiming at a more accurate estimate of the instantaneous tremor frequency. In this section, we will evaluate the performance of these enhanced algorithms, as well as the AQHM algorithm. For the rest of the thesis, when referring to the EKS and Hilbert algorithms, we imply the corresponding enhanced versions.

4.4.1 Synthetic signals of low variance

To evaluate the performance of the three algorithms we generate signals of low variance. Figure 4.5 depicts a synthetic signal of low variance and the reconstructed signals computed by each one of the three algorithms. In Figure 4.6(a) we can see a specific area in more detail. The signal reconstructed using the AQHM and the true signal are almost identical. Hilbert algorithm performs a little bit worse than the EKS. Figure 4.6(b) depicts the ITF estimated by the three algorithms. The first and the last analysis window is omitted, since both the AQHM and the Hilbert methods fail to estimate them accurately.



(a) Synthetic signal

(b) Reconstructed signal

Figure 4.5: a) The synthetic signal with low frequency variance $\sigma_v^2 = 10$ and b) the reconstructed signal using AQHM, Hilbert and EKS.

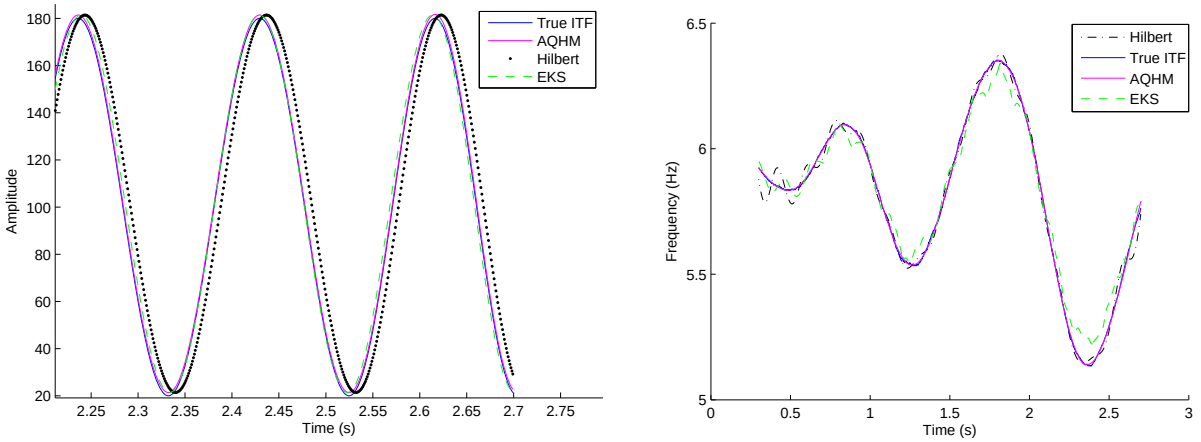
4.4.2 Changing the variance of ITF and adding Gaussian noise

Each synthetic signal simulates an instantaneous component modulated by an instantaneous frequency. The synthetic signals differ, as mentioned in section 4.1, in the variance of the tremor frequency σ_v^2. A greater variance causes a greater fluctuation of the tremor frequency. We generated synthetic signals of different variance σ_v^2 (one hundred signals for each variance value σ_v^2) as shown in Table 4.3. We computed the instantaneous frequency using the AQHM, the Hilbert and the EKS algorithms and we compared their efficiency using the SRER and the MAE, as performance metrics. Each synthetic signal is corrupted with additive white Gaussian noise $w(n)$ of 25dB in order to evaluate the efficiency of the algorithms in the presence of noise as well. The SRER and the MAE are calculated again for these signals. Table 4.3 summarizes the results. Each cell contains the average value from one hundred randomly generated signals. In the last two columns of Table 4.3 we can see the percentage of the SRER loss as we add noise to the signal for the three algorithms. The SRER loss is defined by

$$\text{SRER loss} = \frac{SRER_2 - SRER_1}{SRER_2} \ , \tag{4.6}$$

where $SRER_2$ and $SRER_1$ are the SRER values without and with noise, respectively. Positive SRER loss shows the negative influence of noise's presence on the original signal's estimation. Table 4.3 reveals:

1. The dependency of the AQHM algorithm's performance on the presence of noise and higher variance values. The SRER decreases in the presence of noise and as the variance increases. However, AQHM gives satisfactory SRER values and performs better than the EKS and the Hilbert algorithm in any case.



(a) Synthetic signal

(b) Reconstructed signal

Figure 4.6: a) Instance of the reconstructed signal and b) ITF estimate using AQHM, Hilbert and EKS.

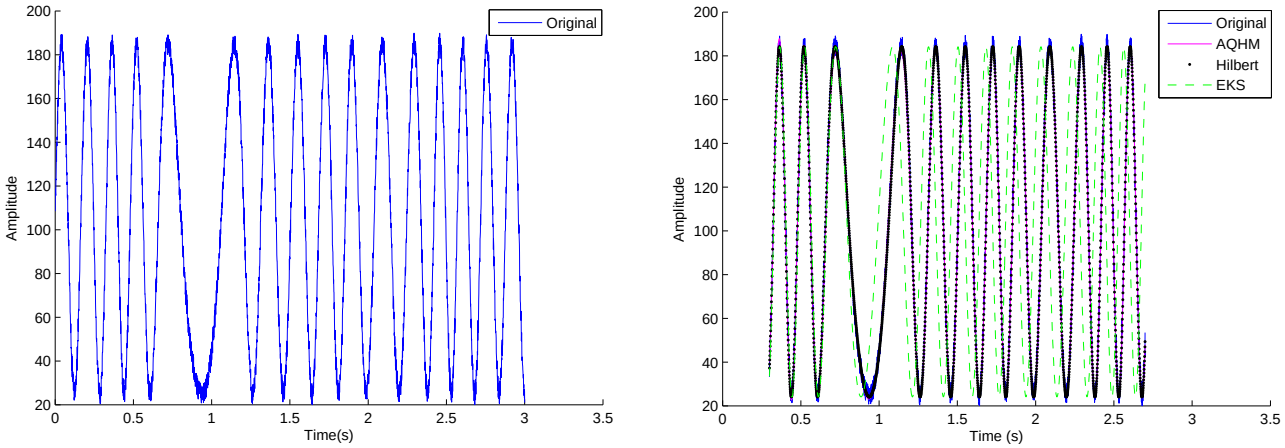
2. The dependency of the EKS algorithm's performance on the presence of noise and higher variance values. The SRER decreases in the presence of noise and as the variance increases. The EKS algorithm gives very low SRER values almost in all cases.

3. The dependency of the Hilbert algorithm's performance on the presence of noise. The Hilbert algorithm is not affected by the variance σ_v^2. On the other hand, the Hilbert algorithm is negatively influenced by the presence of noise but relative to the EKS and AQHM this influence is weaker.

Figure 4.7(a) shows a synthetic signal with high variance $\sigma_v^2 = 60$ in the presence of white Gaussian noise, while Figure 4.7(b) depicts the reconstructed signal for each one of the three algorithms. The presence of noise causes the EKS algorithm to loose track of the ITF (Figure 4.7(b) and Figure 4.8(a)). On the other hand, the AQHM (Figure 4.9(a)) and the Hilbert algorithm(Figure 4.8(b)) detect the ITF accurately. Figure 4.9(b), shows the true ITF along with the three estimates.

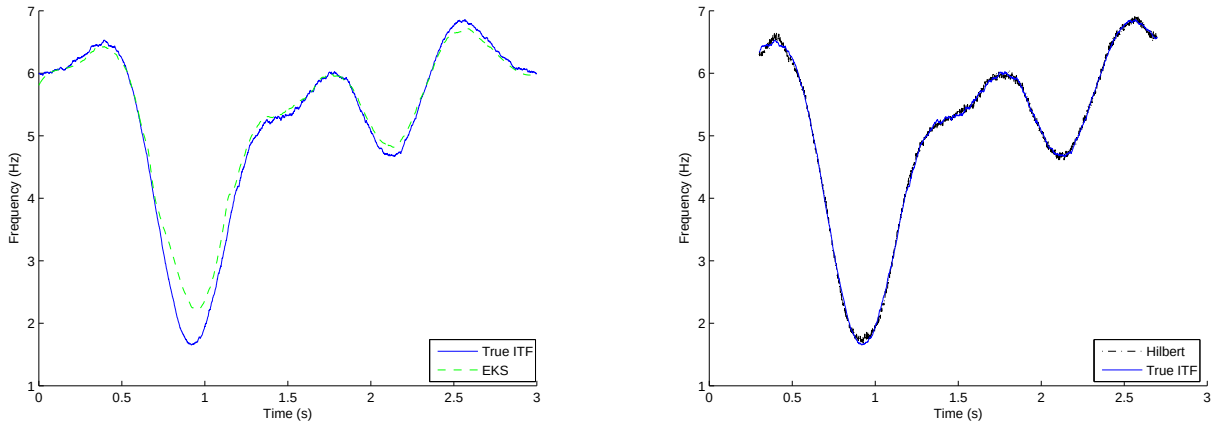
		SRER						SRER loss %		
		without noise			with noise					
σ_v^2	[min,max] ITF	AQHM	EKS	Hilbert	AQHM	EKS	Hilbert	AQHM	EKS	Hilbert
10	$[5.7, 6.6]$	59.17	25.83	32.86	25.37	14.30	23.33	57	45	29
30	$[4.8, 7.2]$	41.86	15.51	31.97	24.95	4.54	23.21	40	71	27
60	$[2.7, 9.6]$	34.23	8.52	30.10	23.66	1.45	23.01	31	83	24

Table 4.3: AQHM, EKS and Hilbert efficiency for synthetic signals



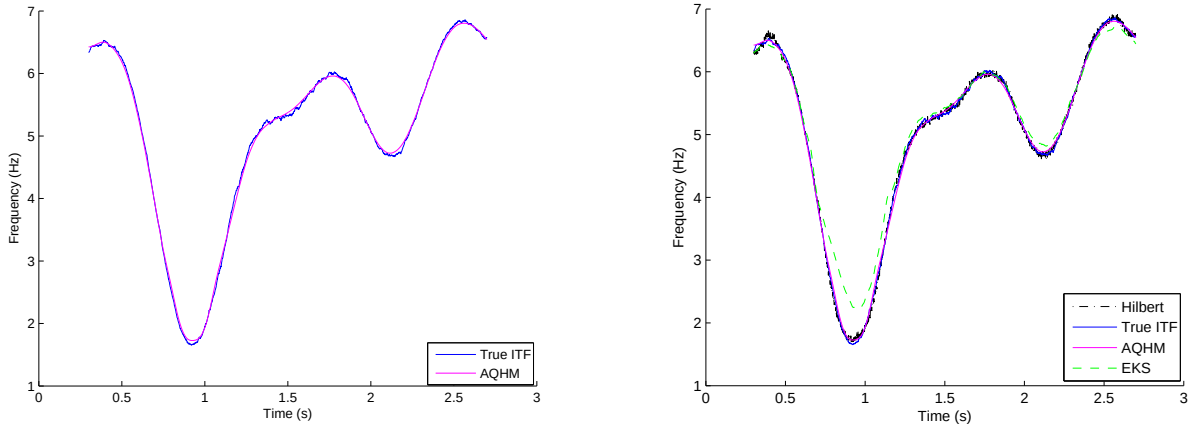
(a) Synthetic signal (b) Reconstructed signal

Figure 4.7: a) Synthetic signal of high variance with white Gaussian noise 25dB. b) Reconstructed signal using AQHM, EKS and Hilbert.



(a) EKS estimate of the ITF (b) Hilbert estimate of the ITF

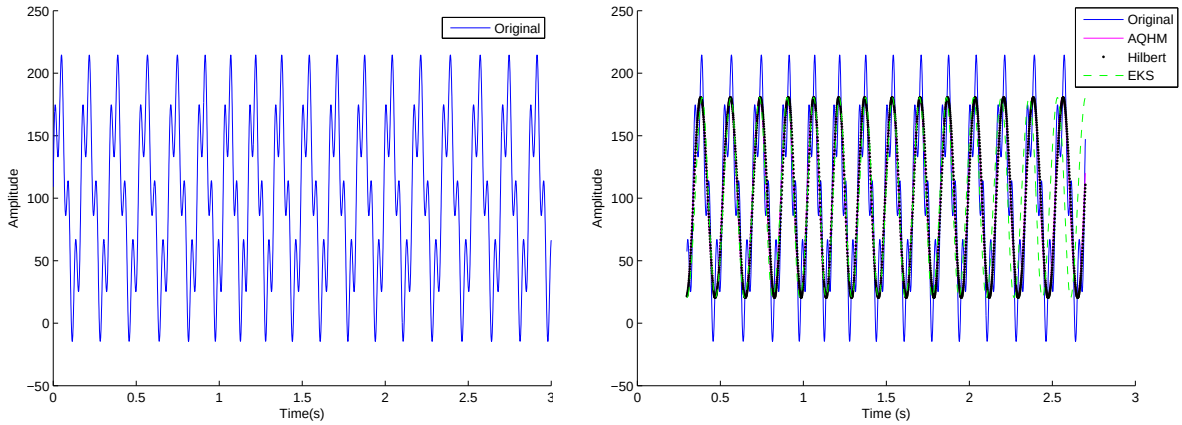
Figure 4.8: ITF estimation using a) EKS, b) Hilbert.



(a) AQHM estimate of the ITF (b) ITF estimation by the three algorithms

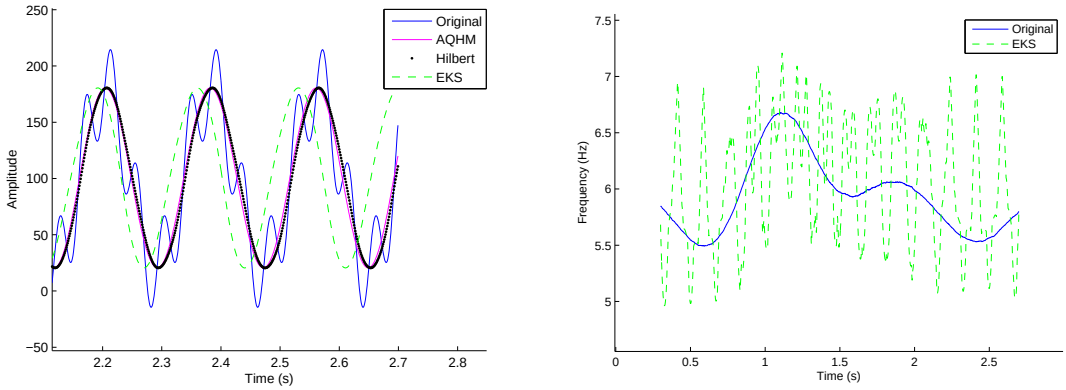
4.4.3 Detecting ITF on multicomponent synthetic signals

In order to evaluate the performance of the algorithms in multi-component signals we generate the two synthetic signals: $r_1(n) = \overline{r} + r_d sin(\varphi(n)) + \frac{r_d}{2} sin(4\varphi(n))$ and $r_2(n) = \overline{r} + r_d sin(\varphi(n)) + \frac{r_d}{2} sin(4\varphi(n)) + \frac{r_d}{2} sin(10\varphi(n))$, modulated by two and three modulating signals, respectively. Figures 4.10 - 4.13 demonstrate the performance of the three algorithms in the first multi-component signal r_1. The Hilbert approach and the AQHM perform well, whereas the EKS fails to estimate the ITF accurately. Despite the smoothing procedure, fast noisy variations are present in the ITF signals estimated by the EKS (Figure 4.11(b)) and the Hilbert (Figure 4.12(a)). On the other hand, the AQHM method captures accurately the modulation frequency (Figure 4.12(b)). Figures 4.14, 4.15 are referred to the second signal. The presence of the third modulation component causes the Hilbert algorithm to give a wrong estimate of the ITF.



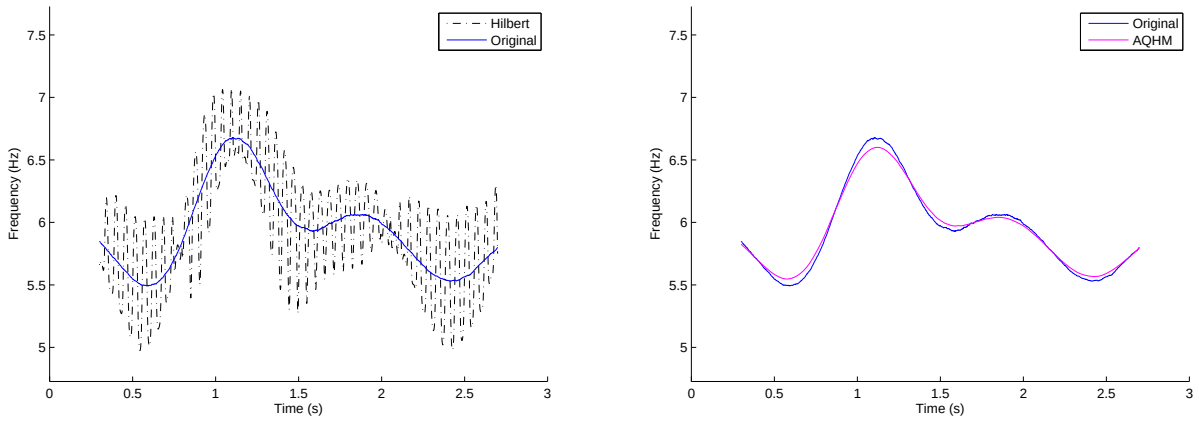
(a) Synthetic signal (b) Reconstructed signal

Figure 4.10: a) The synthetic signal r_1. b) The reconstructed signal using AQHM, EKS and Hilbert.



(a) Synthetic signal (b) ITF

Figure 4.11: a) Instance of the reconstructed signal, b) EKS estimate of the ITF for the signal r_1.



(a) ITF (b) ITF

Figure 4.12: ITF of the signal r_1 estimated using a) the Hilbert, b) the AQHM.

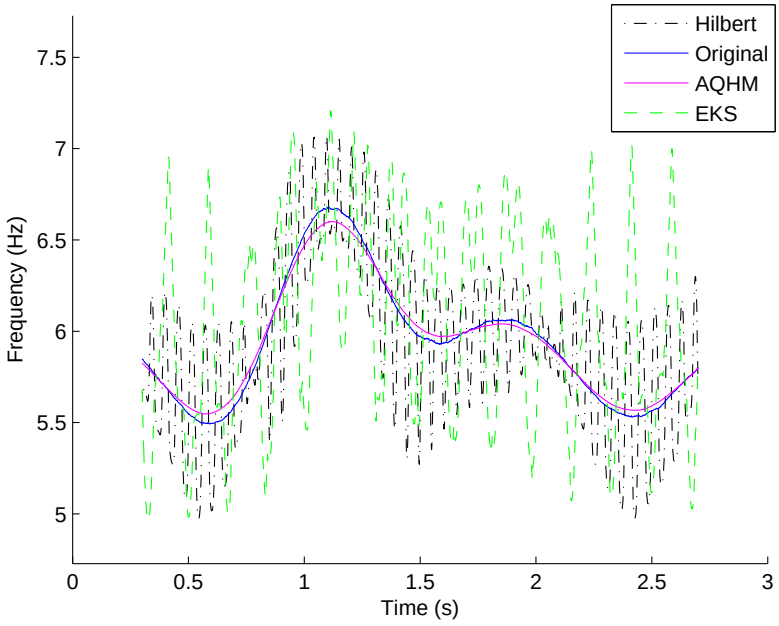
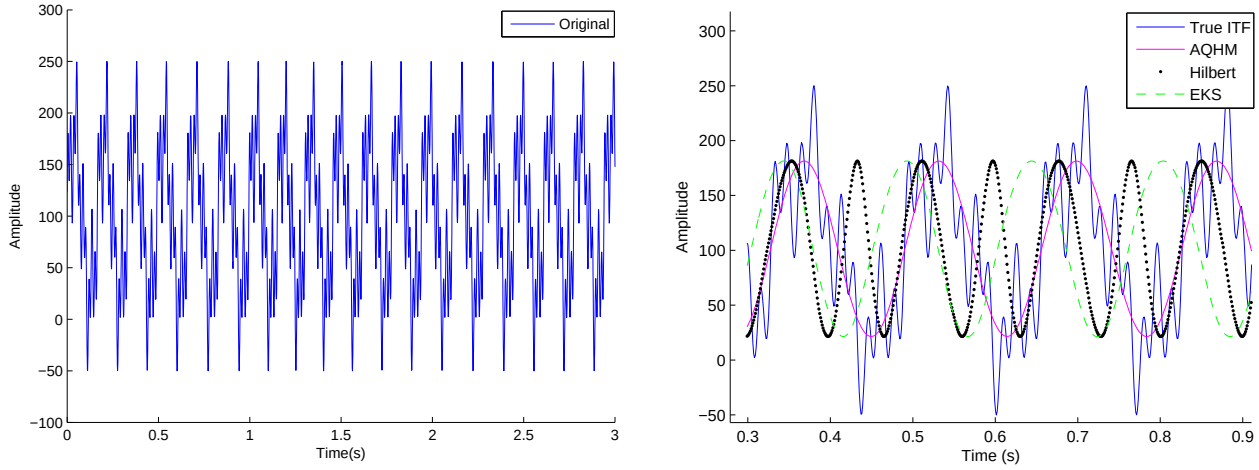


Figure 4.13: ITF of the signal r_1 estimated using AQHM, EKS and Hilbert.



(a) Synthetic signal

(b) Reconstructed signal

Figure 4.14: a) The synthetic signal r_2. b) The reconstructed signal using AQHM, EKS and Hilbert.

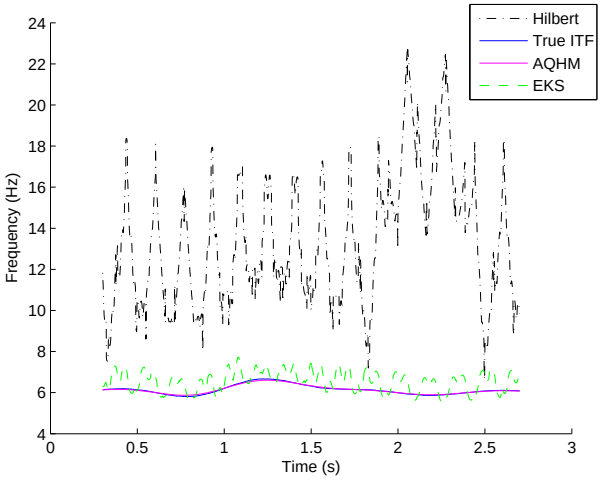


Figure 4.15: ITF of the signal r_2 estimated using AQHM, EKS and Hilbert

4.4.4 Examining the possibility of enhancing ITF detection

From the above analysis we can see that the AQHM outperforms the EKS and the Hilbert. The EKS algorithm is affected by the variance of the ITF and by the presence of noise and the Hilbert transform cannot be applied to multi-component signals. Therefore, we will introduce a method which employs the AQHM in order to extract vocal tremor characteristics.

For the first and last analysis window the AQHM applies an interpolation method to estimate the modulation signal. As a result, it fails to estimate the correct ITF for these frames. The initial concept for correcting this error was to apply the Hilbert transform in the first and last frame and combine the estimation with the estimation of the AQHM for the rest of the

signal. However, we showed that Hilbert transform also fails to estimate correctly the ITF for the first and last part of the signal, while the presence of more than one modulating frequencies in the analyzed speech harmonic may cause the Hilbert transform to estimate the wrong ITF.

On the other hand, the EKS algorithm seems to adapt more quickly and captures the ITF of the first and last frame of the signal. Therefore, in the next chapter we will apply the EKS algorithm in real speech signals in order to evaluate its efficiency.

Chapter 5

Application on speech - Voice tremor detection

This section describes in detail a three-step method for the estimation of the acoustical vocal tremor characteristics. The first step involves the estimation of the instantaneous components performed by the AQHM, while the second step implements the removal of the very slow modulations of the analyzed instantaneous component employed by the Savitzky-Golay smoothing filter. The third and final step concerns the estimation of the modulation frequency and the modulation level from the processed instantaneous component. The estimation of the tremor attributes, namely, the modulation frequency and level, is employed again by the AQHM and is compared with two distinct evaluation approaches, the Extended Kalman Smoother and the Hilbert transform.

5.1 Step 1: Estimation of the instantaneous components of speech

The estimation of the instantaneous components involves the decomposition of the speech signal into its harmonics. The decomposition includes the initialization of the AQHM, the interpolation of the instantaneous components and the adaptation of the AQHM, as described in Chapter 2. In the AQHM, the time resolution is determined by the time step of the algorithm and frequency resolution by the window type and the window length. In this step, we choose a time step of 5ms and a Hamming window as an appropriate window function. The window's duration is adaptive and is chosen to be two times the estimated pitch period. The number of components K is set to 30.

Figure 5.1 shows the first five harmonics of a sustained vowel /a/ of a male speaker extracted using the AQHM. The signal which is reconstructed from the instantaneous components, has a SRER value of about 31dB, which indicates that the analysis is very accurate. In addition, Figure 5.1 reveals that the modulations of higher harmonics are more evident, which motivates the use of the modulation level, which is relative to the mean value of the instantaneous component.

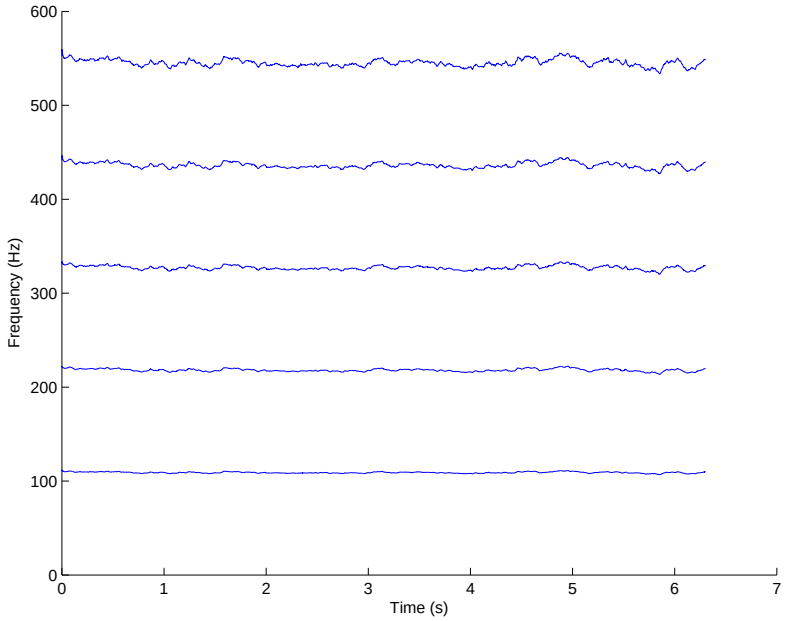


Figure 5.1: Analysis of a speech signal into 5 harmonics

5.2 Step 2: Removal of the very slow modulations

After decomposing the speech signal into K components we choose the instantaneous component to be analyzed. The second step of the analysis consists of eliminating the modulations which are less than 2Hz from the analyzed instantaneous component. The removal of the trend is necessary in order to reveal the quasi-periodical modulations attributed to vocal tremor. However, as a pre-processing step, before the removal of the trend, we down-sample the instantaneous component at sampling frequency of 1000Hz. Indeed, since we are interested in modulations which are less than 20Hz, the down-sampled instantaneous component does not omit any important information.

The smoothing of the instantaneous component is performed using the Savitzky-Golay (S-G) filter [27], [30]. The S-G smoothing filter performs essentially a local polynomial regression on a distribution of equally spaced points to determine the smoothed value for each point. The main advantage of this approach is that it tends to preserve features of the distribution, such as relative maxima, minima and width, which are usually "flattened" by other adjacent averaging techniques, such as the moving averages. Using different parameters for the S-G filter, the smoothed signal will capture more or less of the signal's frequencies. The order of the local polynomial used in this study is equal to 4 while the frame size is set to $1s$ (1001 samples). The smoothed instantaneous component is subtracted from the non-smoothed one in order to obtain the remaining modulations of the component.

Figure 5.2(a) shows the instantaneous frequency of the first harmonic after removing its mean

value along with its filtered version using the S-G smoothing filter. The S-G smoothed signal contains information about the frequencies which are less than 2Hz. The smoothed instantaneous component is then removed from the first component in order to reveal the modulations which are attributed to vocal tremor. Figure 5.2(b) shows the remaining component without the low modulations. Figure 5.3 shows the single-sided spectrum of the initially analyzed component and the remaining component. The frequencies below 2Hz have been removed from the remaining component. Notice that in 2.8Hz the S-G filter estimates the peak wrongly. However, the mismatch is below 0.01 and therefore the amplitude of the signal is not affected. The remaining component can be further analyzed to reveal tremor properties. The analysis is performed on the next step by the AQHM, the EKS and the Hilbert algorithm.

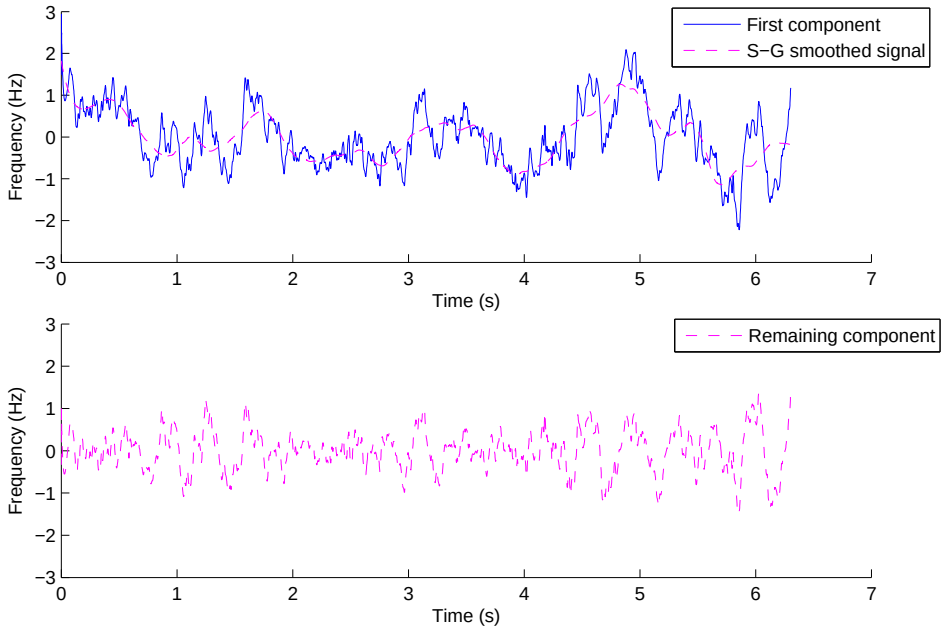


Figure 5.2: (a) The first component and the S-G smoothing filter capturing frequencies less than 2Hz. (b) The final component without the low modulation frequencies

5.3 Step 3: Vocal tremor characteristics extraction

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

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

where $x(t)$ are the remaining modulations of the analyzed instantaneous component, $m(t)$ is the instantaneous amplitude, which with the appropriate scaling corresponds to the modulation

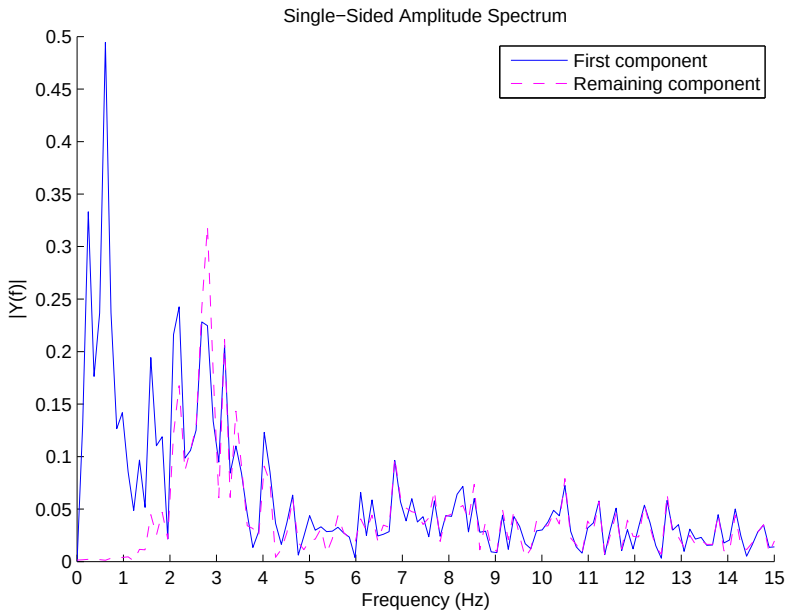


Figure 5.3: (a) The Fourier transform of the first component and the remaining component without the low modulation frequencies. The low modulating frequencies are eliminated by employing the S-G filter

level and $\psi(t)$ corresponds to the instantaneous phase. Once again, instantaneous frequency is given by $\zeta(t) = \frac{1}{2\pi}\frac{d\psi(t)}{dt}$ and corresponds to the modulation frequency. We suggest three different approaches to extract the tremor characteristics, that is, the AQHM, the EKS and the Hilbert transform.

The AQHM algorithm can be applied for the estimation of the instantaneous attributes, $m(t)$ and $\zeta(t)$. Besides, the AQHM needs an initial frequency estimate for the first frame. The largest peak of the Fourier transform of the first frame provides the initial estimate. In this step, the time-step is set to 1ms. Hamming window and its duration is set as above in Step 1.

For the estimation of the instantaneous attribute $\zeta(t)$ we also use the EKS method described in Chapter 3. The third method we use for the estimation of $\zeta(t)$ is the smoothed Hilbert transform described in Section 4.3.

Figure 5.4(a) indicates that the AQHM decomposition algorithm adapts to the non-stationary modulations of the signal. Extended tests on four databases confirmed the ability of AQHM to adapt to the signal characteristics. EKS captures some modulations, but generally it does not perform well (Figure 5.4(b)) which is also verified by the negative SRER value. The Hilbert algorithm performs well, when only one modulation frequency is dominant (first 1.5 sec of Figure 5.4(c)). The extracted time-varying modulation frequency and modulation level estimated by the AQHM and the EKS methods are shown in Figure 5.5. The modulation frequency ranges from 2Hz to 4Hz. Modulation level is not estimated by EKS. However, in order to compare

the two algorithms we use the modulation amplitude computed by the AQHM and the modulation frequency of the EKS to reconstruct the signal estimated by the EKS (Figure 5.4(b)). Figure 5.6 depicts the time evolution of the modulation frequency as estimated by the Hilbert transform. The existence of more than one modulation frequencies, as Figure 5.3 demonstrates, makes Hilbert inappropriate for tremor estimation.

The EKS algorithm results in a decreased performance when applied on real signals. Therefore, the initial concept of applying the EKS algorithm on the first and last frame of a harmonic is abandoned. Our method suggests to ignore these frames in order to compute the correct tremor attributes. This imposes one limitation, that is, the signals to be processed must have a duration of more that 1.5 seconds, since the duration of the frame varies from 0.3 to 0.5 seconds.

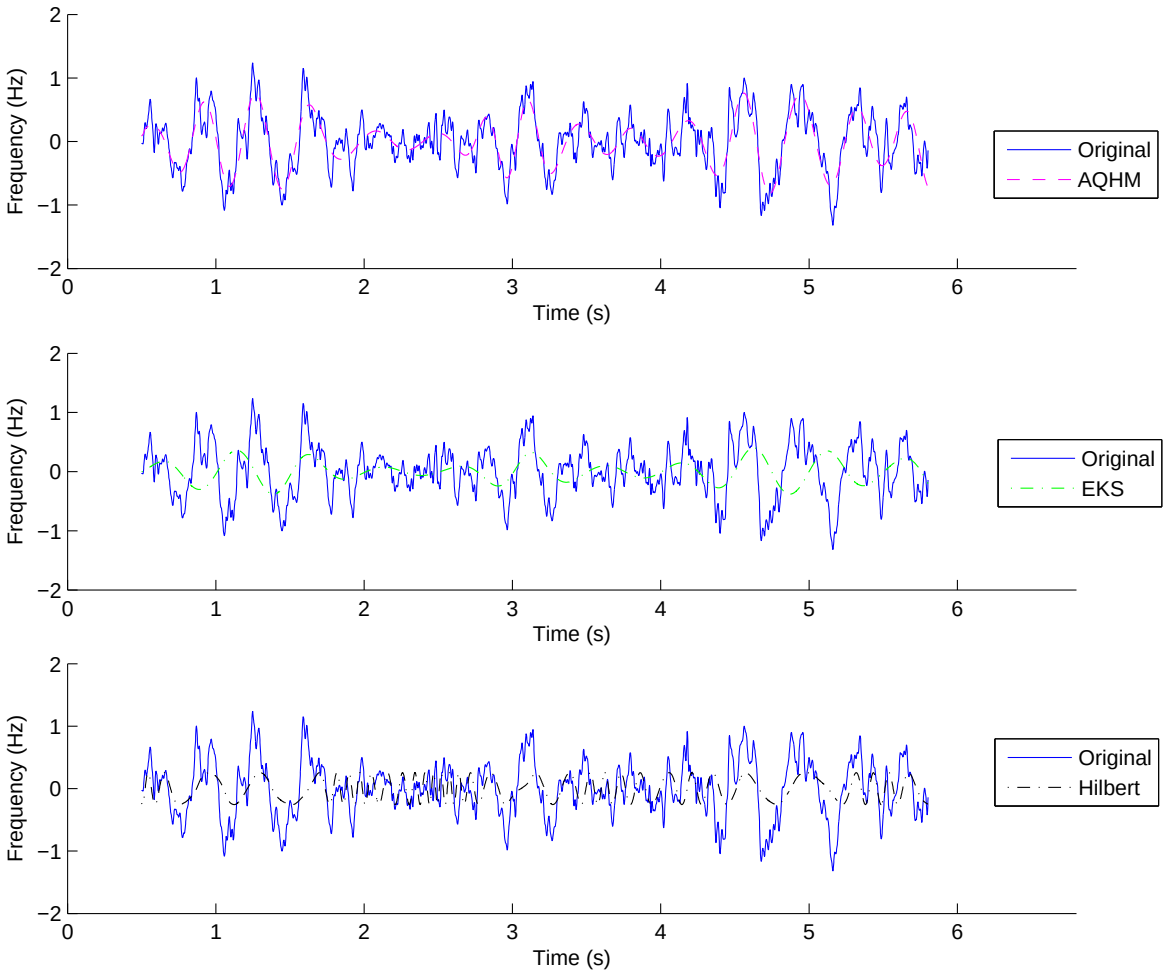


Figure 5.4: Estimation of the modulation tremor signal using (a) the AQHM with SRER=5.78, (b) the EKS with SRER=-0.9942 and (c) the Hilbert with SRER=0.96.

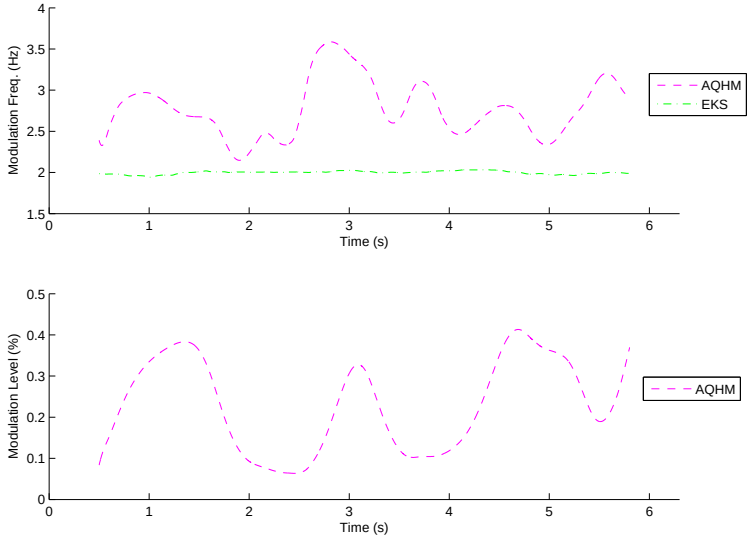


Figure 5.5: (a) Tremor frequency in time estimated by AQHM and EKS. (b) Tremor level in time estimated by AQHM.

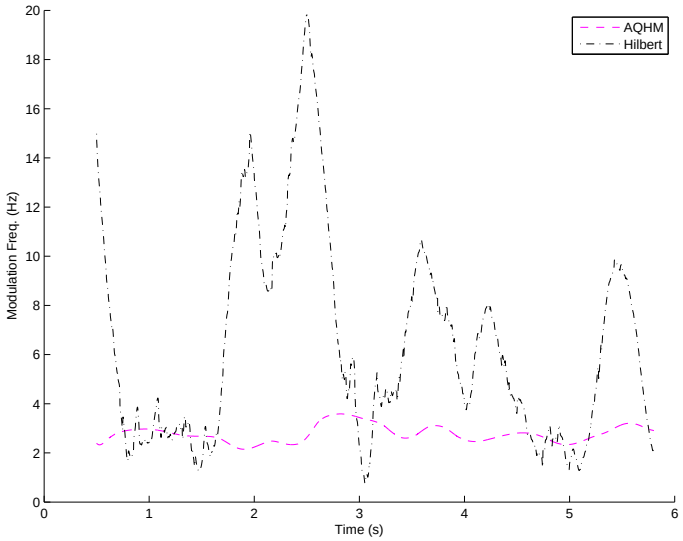


Figure 5.6: Tremor frequency in time estimated by Hilbert and AQHM.

Chapter 6

Databases

In the previous chapter we have introduced a method for the extraction of vocal tremor characteristics from a speech signal. In this chapter we evaluate this method on a database consisting of normophonic speakers. The high signal-to-reconstruction error ratios reveal the accuracy of our method. Moreover, the proposed method is applied to a database, which consists of subjects who suffer from spasmodic dysphonia. Our analysis suggests that the significant attributes that distinguish the normophonic from the dysphonic subjects and subsequently determine the voice tremor are the *modulation level* and the *deviation of the modulation level*. Finally, we process two more databases in order to reveal the relationship between vocal tremor and vocal loading.

6.1 Database 1: Evaluation on normophonic speakers

Our proposed method is validated on a database of normal voices developed in our recording lab. Eleven male and five female healthy subjects, whose age varies between 23 and 45 years old were participated. Sustained vowels /a/, /e/, /i/, /o/ and /u/ have been recorded at 48kHz and then downsampled at 16kHz. The duration of sustained vowels varies from 2sec to 8sec depending primarily on the speaker.

Figure 6.1 shows the SRER values that our proposed method achieved on decomposing every speech signal. For the total database, which contained 279 phonemes, the average SRER was 27.41dB with a standard deviation of 5.07.

Figure 6.2 presents the mean tremor frequency and mean tremor level as well as the corresponding deviations for every speaker in our database as computed by the AQHM technique. Specifically, for each speaker and for every phoneme /a/, /e/, /i/, /o/, /ou/ we calculated a median value of the time-varying modulation level and a median value of the time-varying modulation frequency along with the corresponding 25 and 75 percentiles. Then, we calculated the deviations by subtracting the 25th from the 75th percentile. We estimated a mean value of these median values for all the phonemes and a mean value of the deviations for all phonemes, produc-

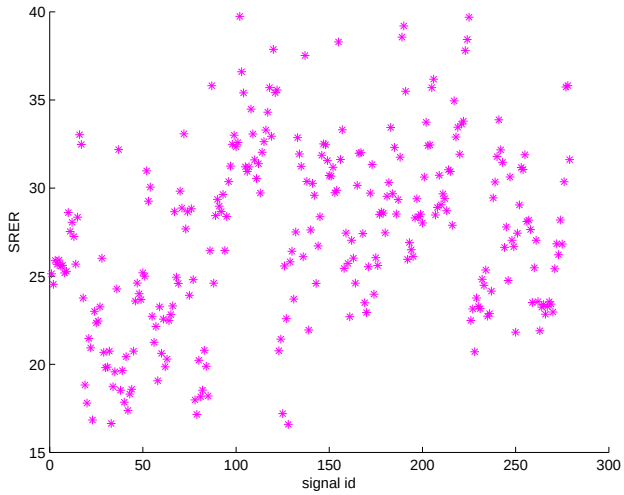


Figure 6.1: Signal-to-Reconstruction Error Ratio for every signal in the database estimated by AQHM.

ing a representative value for each speaker. For instance, as shown in Figure 6.2, the speaker 1 produced 5 phonemes which appear to have a mean value of tremor frequency at around 3.2Hz. The mean deviation from this value during the utterance of the phonemes ranges from 2.4 to 4.1Hz. Note that, apart from the speaker with speaker id eleven, the mean modulation level is below 1. In the next section, we show that dysphonic speakers have mean modulation level much greater than value 1.

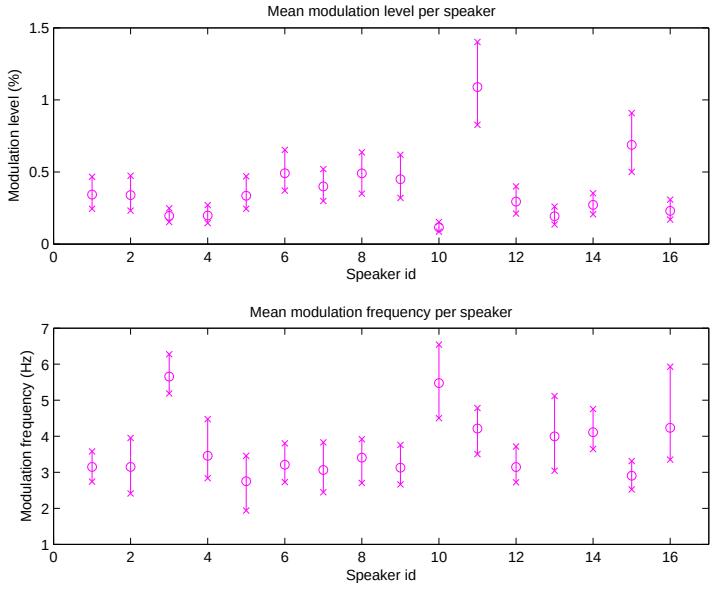


Figure 6.2: Mean tremor frequency and mean tremor amplitude for each speaker in our database as computed by the AQHM.

6.2 Database 2: Relationship between vocal tremor attributes and spasmodic dysphonia

In this section, we examine tremor in speakers who suffer from spasmodic dysphonia and we try to find if there is a relation between tremor attributes and the severity of spasmodic dysphonia. For this purpose, we analyze a database which consists of speech signals of twenty speakers, sixteen male and four female. For every speaker the sustained vowels of /a/ are extracted to create the signals for our analysis. Five speakers cannot be analyzed since, due to the severity of the problem, the duration of the phonemes is less than one second. Some phonemes of other speakers cannot be analyzed for the same reason.

Figure 6.3 illustrates the decomposition of a phoneme /a/ of a dysphonic speaker. The harmonics appear to have large amplitude variations compared to the harmonics of normophonic speakers (Figure 5.1).

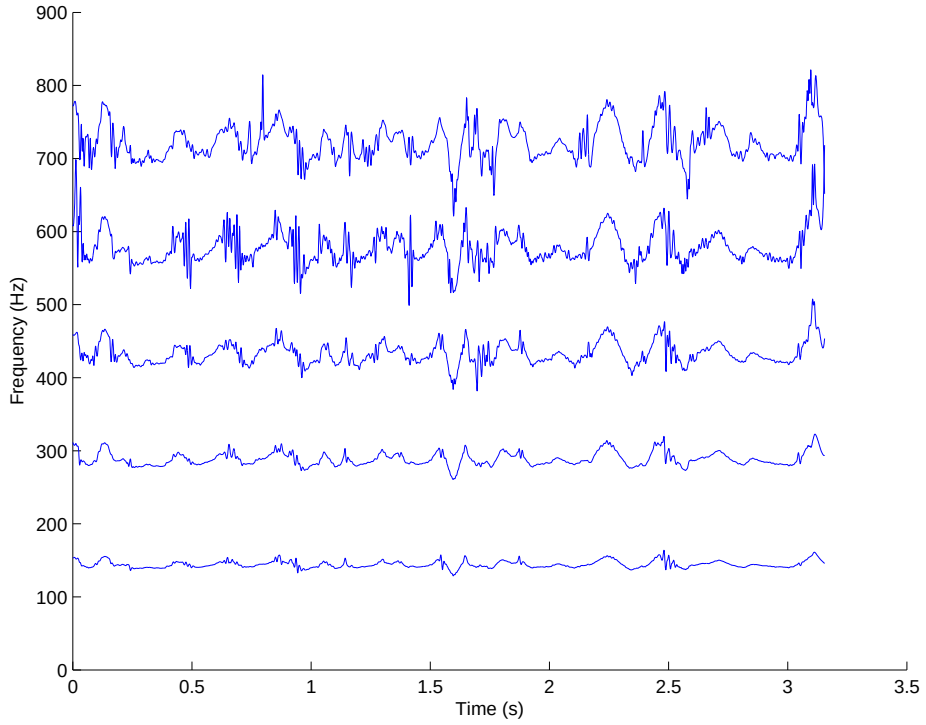


Figure 6.3: First five harmonics of a speech signal of the dysphonic speaker Burpre estimated by the AQHM.

We perform the same analysis as in normophonic speakers and we extract the desired modulating signal (Figure 6.5) from the first instantaneous component (Figure 6.4) . The time-varying tremor attributes are depicted in Figure 6.6. Notice the very high modulation level. In normophonic speakers the modulation level ranges from 0 to 1, whereas for this dysphonic speaker the modulation level varies from 3 to 8. Table 6.1 summarizes this observation, where we have calculated the median values of the modulation frequency and the modulation amplitude and the

corresponding 25th and 75th percentiles.

Table 6.1: Tremor characteristics for normophonic and dysphonic speakers

Vowel 'a'		
	normophonic speaker	dysphonic speaker
median modulation frequency	6.6726	4.1788
25th percentile of modulation frequency	4.3721	3.7168
75th percentile of modulation frequency	8.2103	4.7716
min modulation frequency	3.1402	2.5680
max modulation frequency	8.9469	5.9674
median modulation level	0.0963	1.9918
25th percentile of modulation level	0.0641	1.3240
75th percentile of modulation level	0.1644	3.5561
min modulation level	0.0431	0.9021
max modulation level	0.1937	6.0419

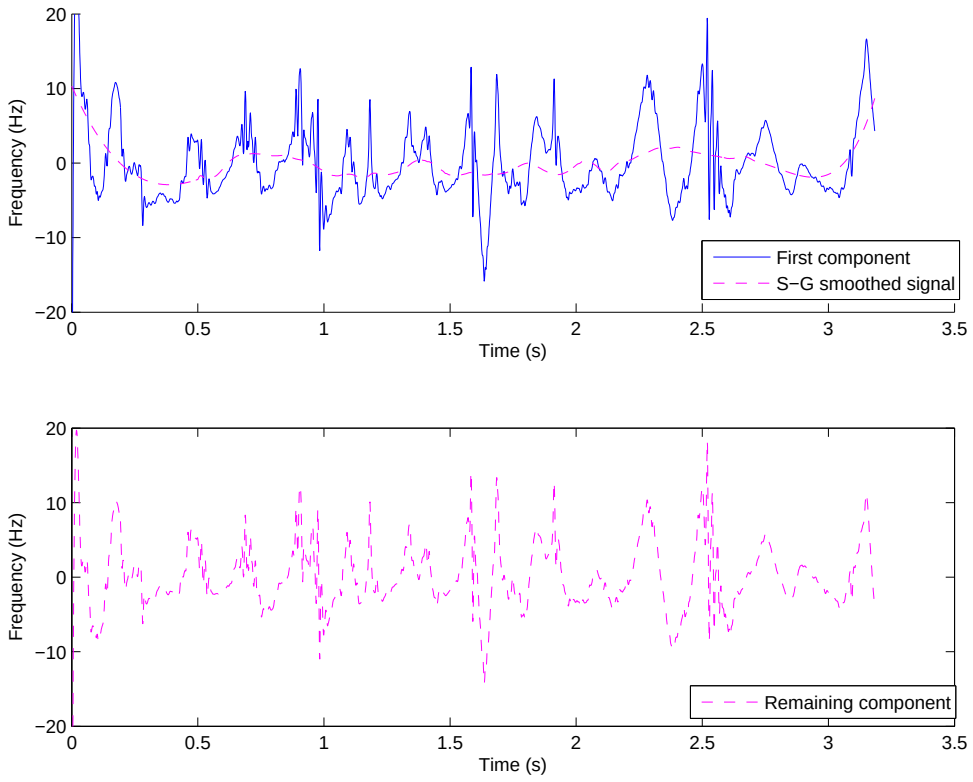


Figure 6.4: (a) The leveled first instantaneous component and the low modulating signal computed by the Savitzky-Golay filter. (b) The remaining instantaneous component without the very low modulations.

Figure 6.7 shows the median values of the modulation level for every dysphonic speaker of the database. Figure 6.8 shows the median values of the modulation level for every normophonic speaker of our database. The tremor level in dysphonic speakers is more prominent than in nor-

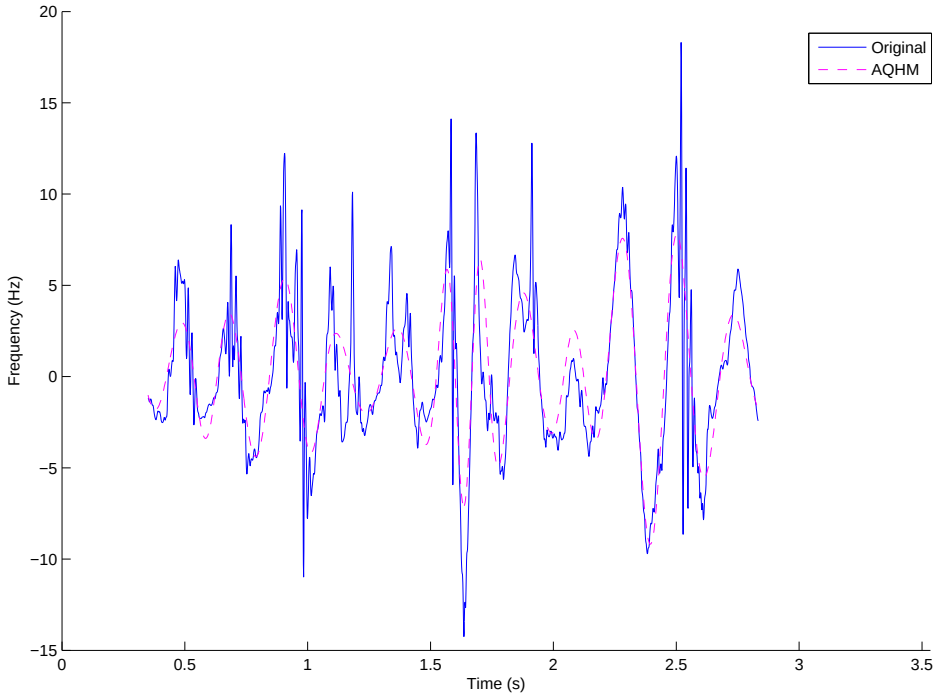


Figure 6.5: The modulating signal of the instantaneous component of the Figure 6.4(b) as computed by AQHM.

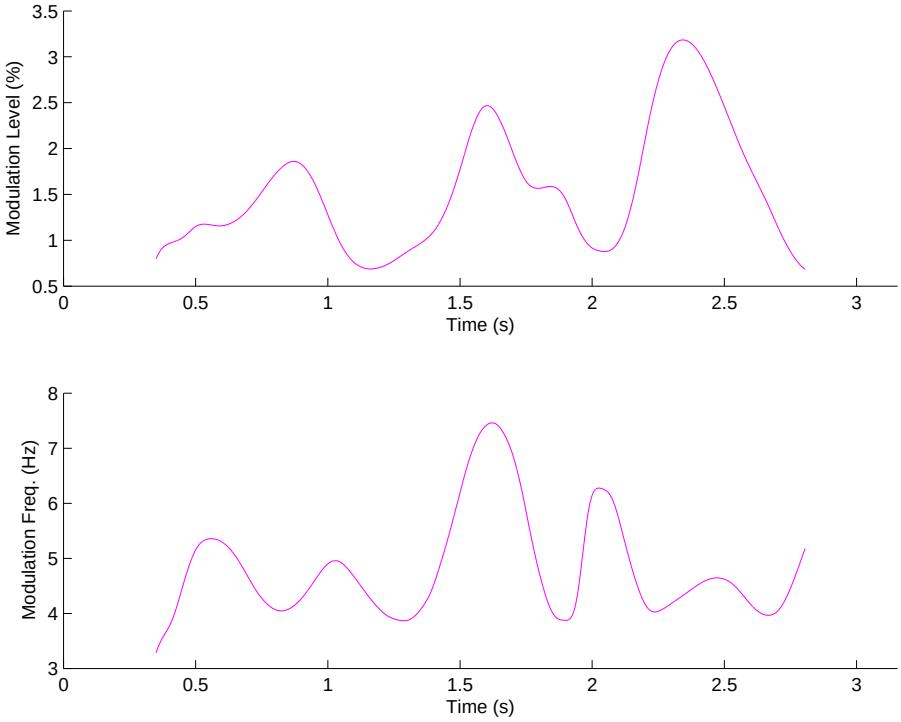


Figure 6.6: Modulation level and modulation frequency of the first instantaneous component.

mophonic speakers and seems to distinguish normophonic from dysphonic speakers. Figure 6.9(a) presents the median modulation level as a function of the median modulation frequency, where each point refers to a speaker. As we can see, normophonic speakers appear to have lower modula-

tion levels than the dysphonic speakers and the same modulation frequency values. Figure 6.9(b) shows the deviation of the modulation level as a function of the deviation of the modulation frequency for normophonic and dysphonic speakers. The deviation is computed by subtracting the 25^{th} percentile from the 75^{th} percentile. We can see that the level's deviation for normophonic speakers is less than one for dysphonic speakers. This is demonstrated in Figure 6.10, which depicts the modulation level and the modulation frequency as a function of their deviations for both groups of speakers. The normophonic speakers(Figure 6.10(a)), occupy the lower left area of the graph where the modulation level and its deviation take low values.

The speakers in our database were dysphonic speakers who were subjected to treatment. Recordings have been made before and after the treatment. Figure 6.11 shows the estimated tremor level and the deviation of the tremor level for every speaker before and after the treatment. For some speakers there seems to be an improvement in the modulation level after the surgery, while in others an improvement in the deviation of the modulation level occurs. Figure 6.12 shows the tremor level coordinates, namely, the modulation level as a function of its deviation. We draw arrows to show the change of tremor level coordinates for each speaker before and after the treatment. Three of the speakers appear to have obvious improvement as they approach the normophonic area. This also applies for the speakers whose signals could not be analyzed due to the severity of the problem. The corresponding green cycles approach also the normophonic area.

In Figure 6.7, we can see that the speakers appear to have different values of modulation amplitude and different deviations from the median value. The aim is to classify the speakers according to the median value of their tremor level and the deviation of the median value, in order to verify if there is an actual correlation between these attributes and the severity of spasmodic dysphonia, as evaluated subjectively by the doctors. To perform the classification, we calculate for every speaker a weighted mean tremor value, defined as the 80% of the median value and the 20% of the deviation. The classification is performed according to this weighted mean tremor value (WMTV). The suggested classification with descending weighted mean tremor values is shown in Table 6.2. To compare the subjective classification with our classification we normalized the tremor rating value and the WMTV.

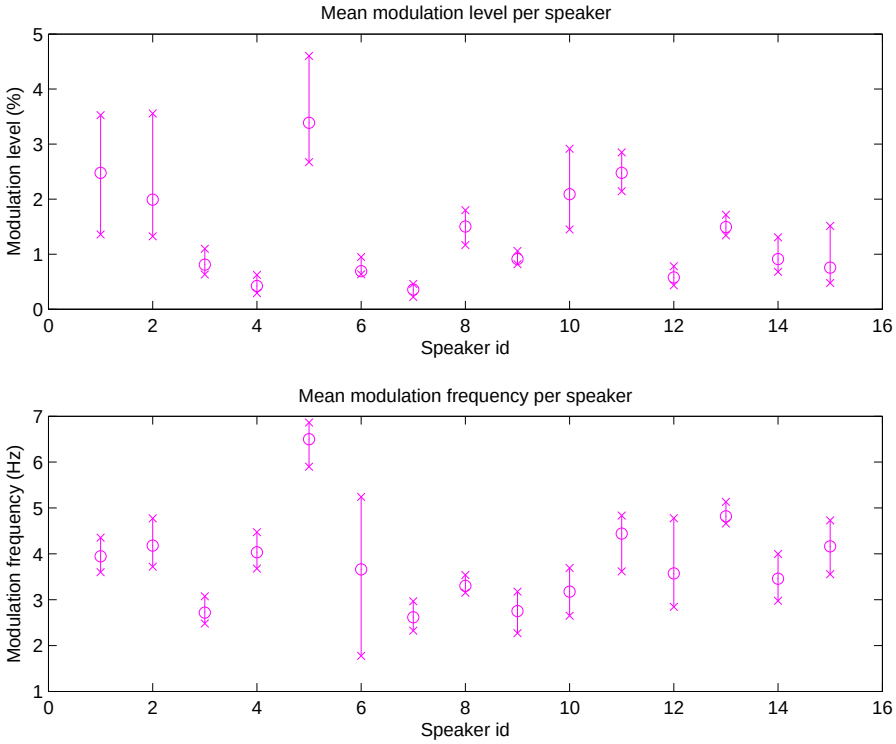


Figure 6.7: Modulation level and modulation frequency of dysphonic speakers

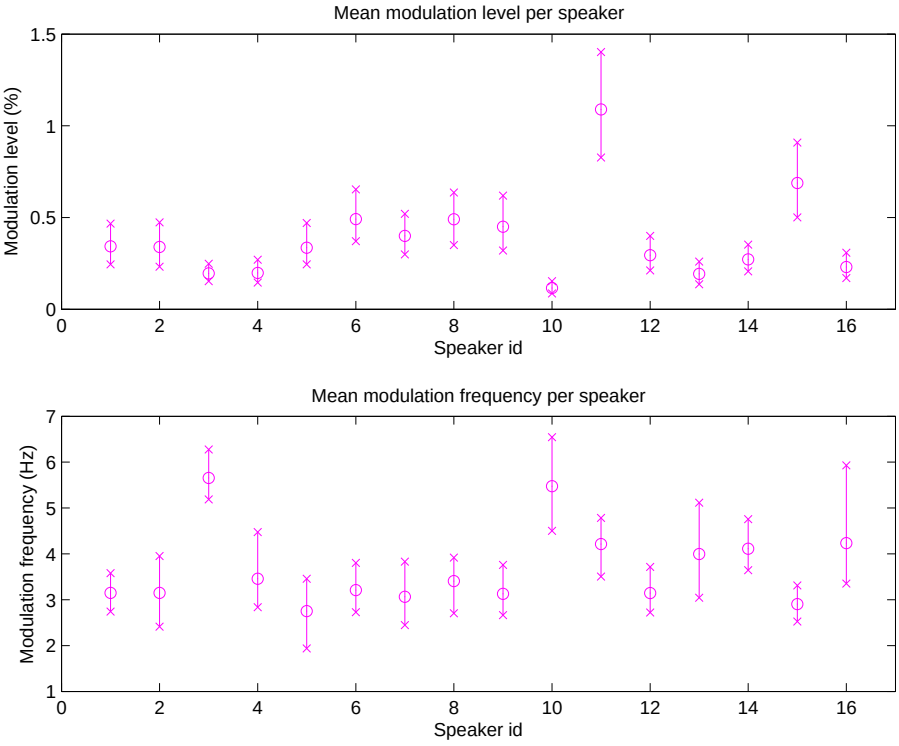


Figure 6.8: Modulation level and modulation frequency of normophonic speakers

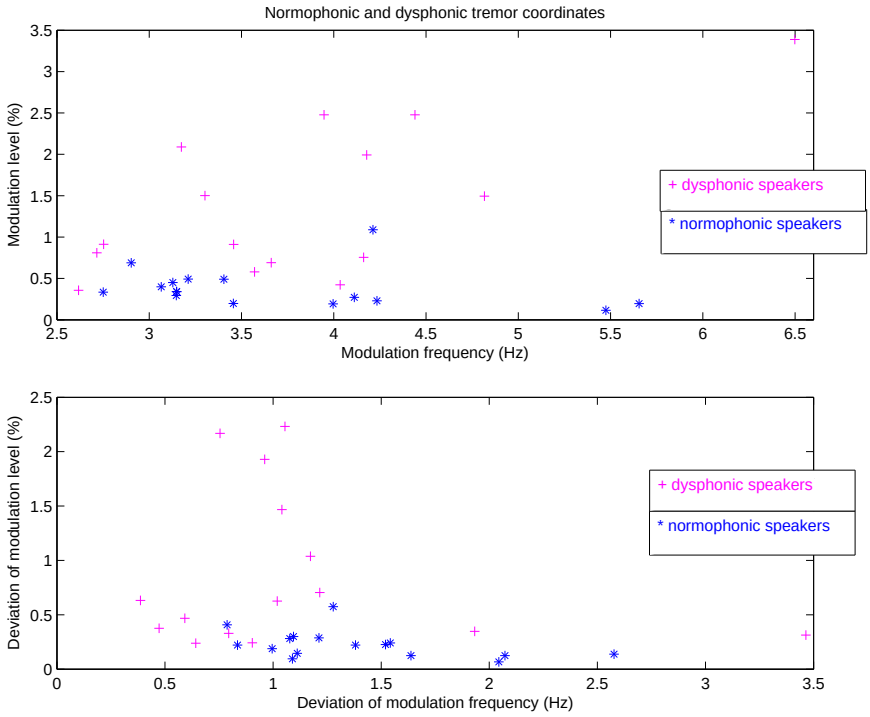


Figure 6.9: (a) Modulation level as a function of modulation frequency for normophonic and dysphonic speakers (b) Deviation of the modulation level as a function of the deviation of the modulation frequency for normophonic and dysphonic speakers

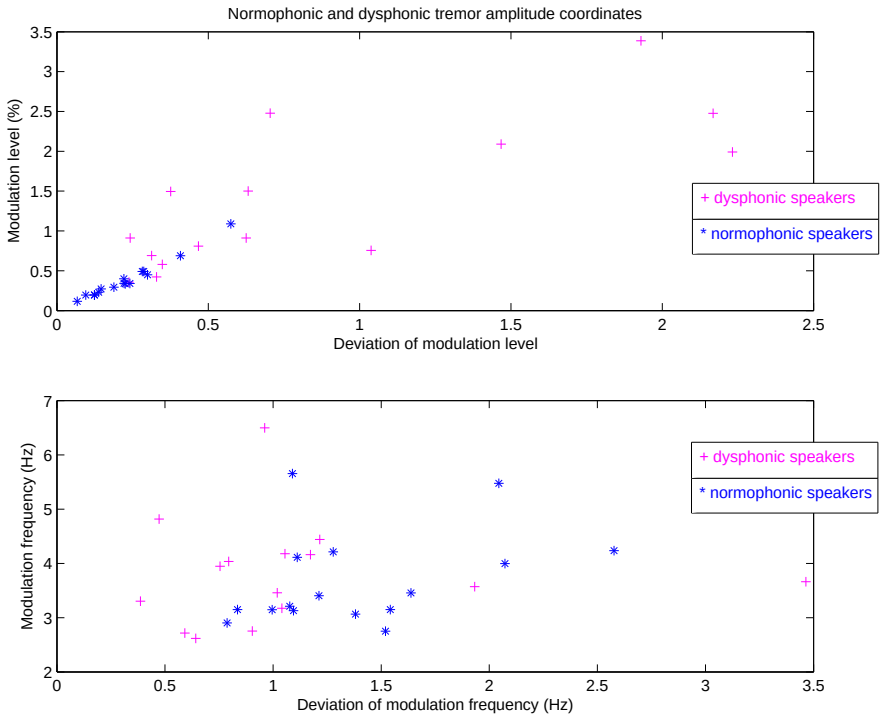


Figure 6.10: (a) Modulation level as a function of its deviation for normophonic and dysphonic speakers. (b) Modulation frequency as a function of its deviation for normophonic and dysphonic speakers.

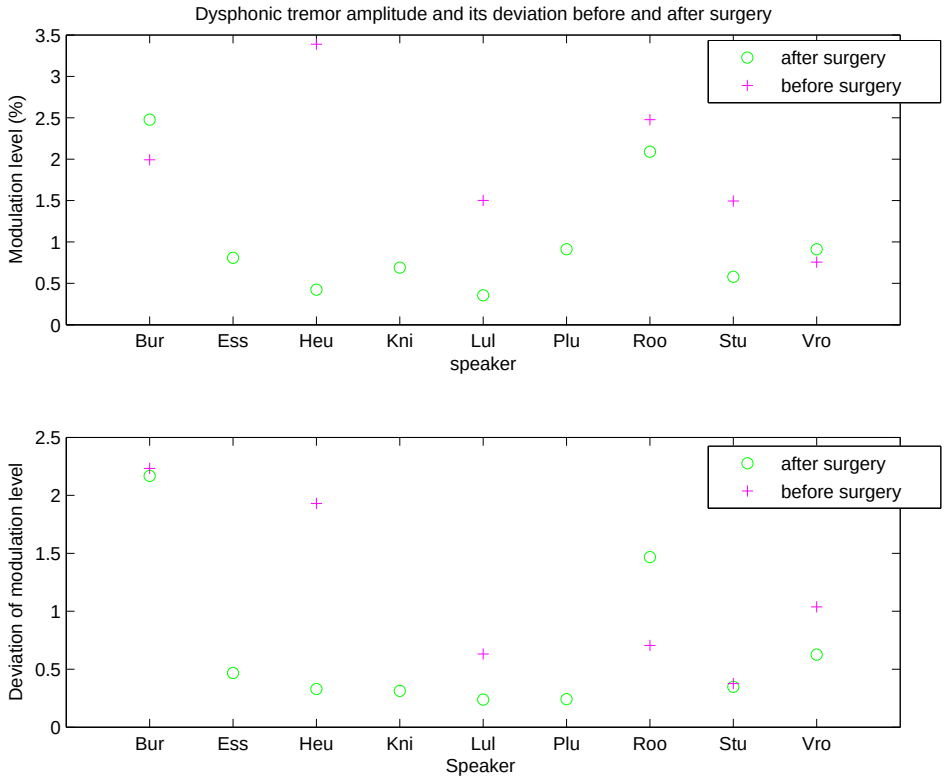


Figure 6.11: (a) Modulation level for dysphonic speakers before and after surgery. (b) Deviation of modulation level for dysphonic speakers before and after surgery.

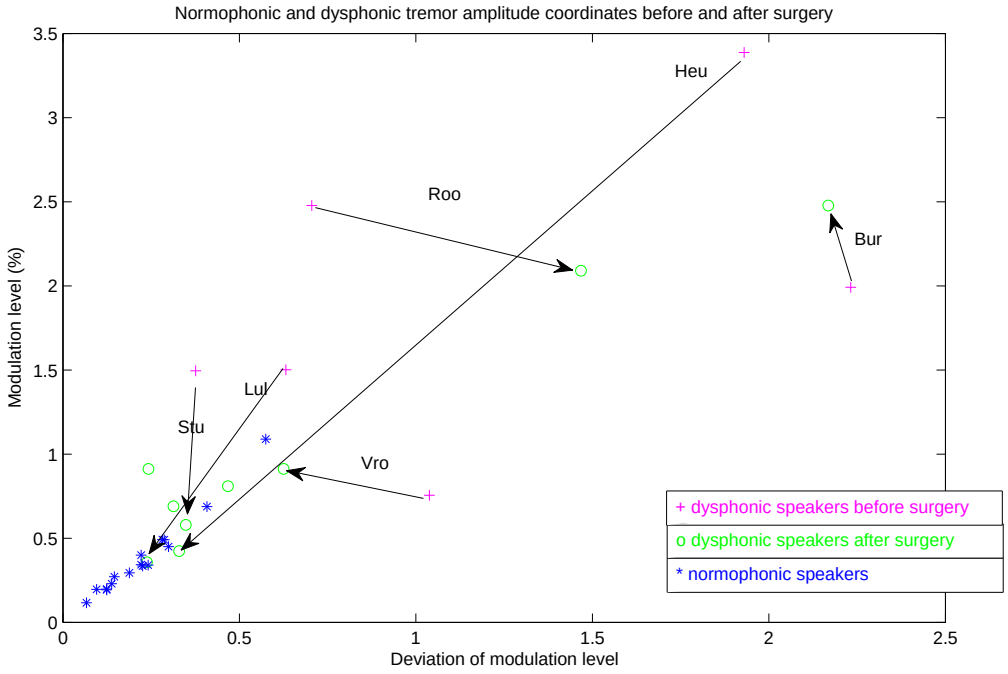


Figure 6.12: Normophonic and dysphonic speakers' tremor level coordinates before and after surgery.

Subjective classification			Our classification		
	Tremor rating (TR)	Normalized TR		WMTV	Normalized WMTV
Burpre	8.5	1.00	Heupre	3.16	1.00
Burpos	8	0.94	Burpos	2.42	0.78
Roopre	7	0.82	Roopre	2.12	0.69
Stupre	6	0.71	Burpre	2.04	0.66
Roopos	5	0.59	Roopos	1.97	0.63
Vropre	4.5	0.53	Lulpre	1.33	0.43
Vropos	4	0.47	Stupre	1.27	0.41
Heupre	3.5	0.41	Vropos	0.85	0.28
Knipos	3.5	0.41	Vropre	0.81	0.26
Lulpre	2	0.24	Plupos	0.78	0.25
Plupos	1	0.12	Esspos	0.74	0.24
Esspos	0.5	0.06	Knipos	0.61	0.20
Heupos	0.5	0.06	Stupos	0.53	0.17
Lulpos	0.5	0.06	Heupos	0.40	0.13
Stupos	0	0.0	Lulpos	0.33	0.11

Table 6.2: Dysphonic speakers classification based on: a) subjective evaluation, b) descending weighted mean tremor value (weighting factor = 80%)

If we normalize the tremor evaluations for the two classifications, then we can calculate, by varying the weights the MSE between the normalized tremor values of the same speaker for the two classifications. Figure 6.13 shows that the weight which minimizes the MSE, and therefore achieves the best matching between the two classifications, is defined as the 40% of the mean and the 60% of the standard deviation. However, we would need a different database to evaluate this weight before generalizing it and using it in the classification of other databases.

In order to examine how correlated the two classifications are, namely the subjective classification and the WMTV classification, we compute the correlation between the classifications and the p-value. The p-value is the probability of getting a correlation as large as the observed value by random chance, when the true correlation is zero. If the p-value is small, say less than 0.05, then the correlation is significant. If we calculate the correlation between the two classifications for $w = 40\%$, the correlation coefficient is 0.72 and the p-value is equal to 0.0024. This means that the probability of getting by random chance a correlation as large as 0.72 when there is no correlation between our classifications is 0.0024, that is very very small. The correlation between the two classifications for $w = 80\%$ is equal to 0.68 and the p-value 0.0051. This means that in both cases the correlation of our classifications is significant, since the p-value is smaller than 0.05. The classification based on the new descending weighted mean tremor values is shown in Table 6.3.

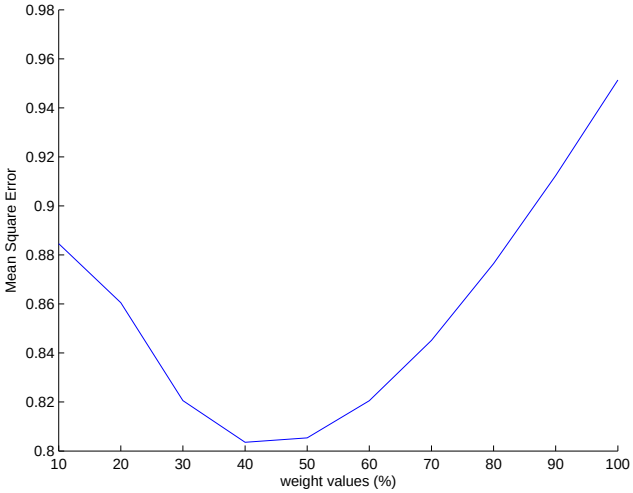


Figure 6.13: (a) MSE between the two classifications for different weight values. Weight 40% gives the minimum MSE and the best matching between the two classifications.

Subjective classification		Our classification	
	Normalized TR		Normalized WMTV
Burpre	1.00	Heupre	1.00
Burpos	0.94	Burpos	0.91
Roopre	0.82	Burpre	0.85
Stupre	0.71	Roopos	0.68
Roopos	0.59	Roopre	0.56
Vropre	0.53	Lulpre	0.39
Vropos	0.47	Vropre	0.37
Heupre	0.41	Stupre	0.33
Knipos	0.41	Vropos	0.30
Lulpre	0.24	Esspos	0.24
Plupos	0.12	Plupos	0.20
Esspos	0.06	Knipos	0.18
Heupos	0.06	Stupos	0.18
Lulpos	0.06	Heupos	0.15
Stupos	0.0	Lulpos	0.11

Table 6.3: Dysphonic speakers classification based on: a) subjective evaluation, b) descending weighted mean tremor value (weighting factor = 40%)

6.3 Database 3: Relationship between vocal tremor attributes and vocal loading

In the following, we analyze physiological tremor in normophonic speakers and we examine if there is a relation between tremor and voice fatigue. For this purpose, we analyze two databases of normal voices.

In the first database the subjects that participated in the recordings were university students between 23 and 26 years old. They attended a vocally loading test, which consisted of shouting numbers for five minutes. The recordings have been made in a well-damped studio, however,

there is some low frequency noise, especially in channel 0, probably related to the ventilation system. For this reason, we analyzed the second channel, which contained less noise. Sustained vowels of /a/, /i/, /ou/ had been recorded for each student twice. The first recording took place before the beginning of the loading test and the second recording right after the end. Their voice was evaluated by experts.

In the second database we examine physiological tremor in 8 normophonic male teachers. Sustained vowels of /a/, /i/, /ou/ had been recorded for each speaker twice. The first recording took place before the beginning of the class and the second recording right after the end. Four of them complained about feeling voice-fatigued.

6.3.1 Database 3a: Tremor evaluation of students' voice via vocal loading tests

The subjects which participated in the recordings were asked to determine of how tired their throat felt after the loading test on a scale from 0 (no tired) to -3 (very tired). Two speech trainers evaluated their samples perceptually on a scale from -3 (being very poor voice) to +3 (being excellent). Table 6.4 shows the subjective evaluations as reported from the speakers and the speech trainers. In the following, our tremor analysis for every speaker is presented and the subjective evaluations with our tremor rating are compared.

Speaker id	Gender	Speaker's evaluation	Trainer's evaluation
HA	Female	-1/worse	from 0 to +1/better
HK	Female	-1/worse	from -1 to -1.5/worse
PH	Female	-3/worse	from -1.5 to -1.5/same
KU	Male	-3/worse	from -2 to -2.5/worse
MI	Male	0/same	from -1 to -0.5/better
PP	Male	0/same	from -1 to -2/worse

Table 6.4: Subjective evaluation of speakers' voice before and after vocal loading.

The analysis is performed on the second harmonic, since the first harmonic was corrupted by noise. For every speaker we plot the modulation level and its deviation before and after the loading test. Figures 6.14, 6.15 reveal a deterioration for speakers PP, HK and an improvement for speakers PH, KU, HA. These results do not agree with the subjective evaluations of Table 6.4

Our results show no evident correlation between the voice fatigue and the voice tremor. The voices, which have been evaluated worse after the vocal loading may indicate a vocal fatigue but the tremor attributes remained almost the same. Notice in Figure 6.15 that the maximum level difference before and after the loading test is less than 0.15, while the maximum level deviation difference is less than 0.05. In subjects with spasmodic dysphonia one of these values was at least 0.5 whereas the maximum level difference marked was 2.5 and the maximum difference of

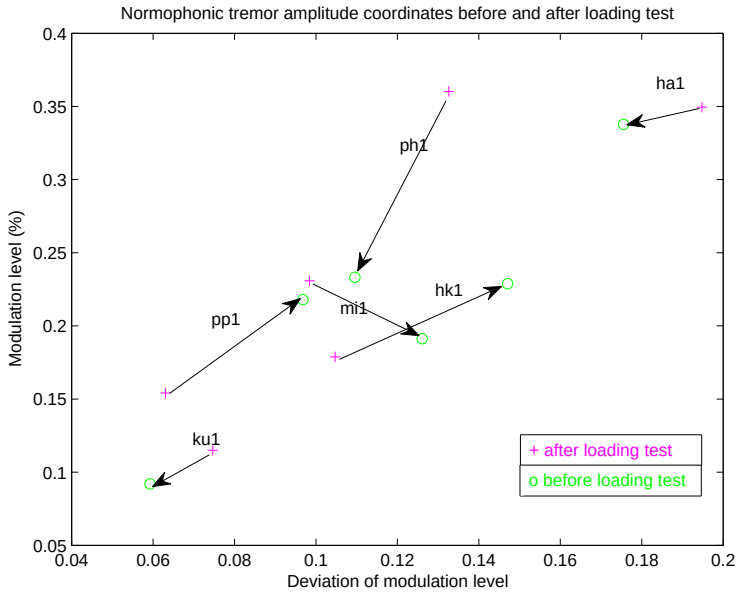


Figure 6.14: Students' tremor level coordinates before and after the loading test.

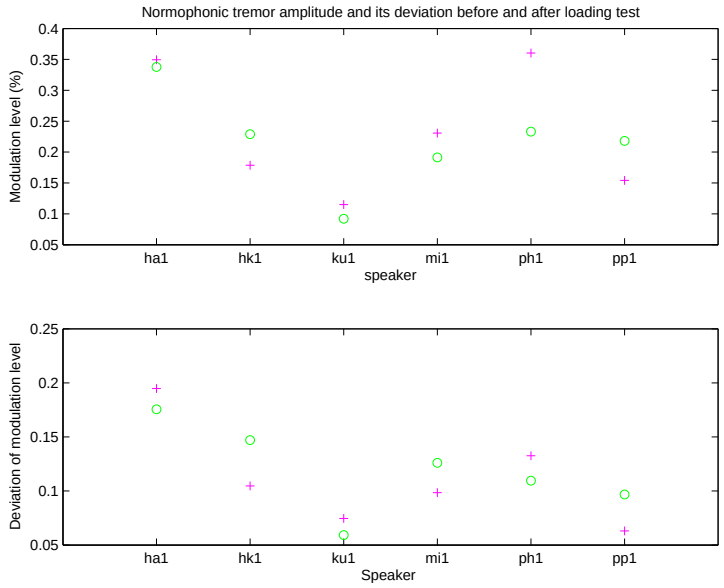


Figure 6.15: Modulation level and its deviation before (magenta cross) and after (green cycle) loading test for every speaker

the deviation level was equal to 1.5. Our analysis in this database indicates that there is no correlation between the vocal fatigue and the voice tremor.

6.3.2 Database 3b: Tremor evaluation of teachers' voice

Finally we examine the physiological tremor in 8 normophonic male teachers and we verify whether there is a relation between the tremor and the voice fatigue. For this purpose, we analyze a database of normal voices consisting of eight male teachers. Sustained vowels of /a/, /i/, /ou/ had been recorded for each speaker twice. The first recording took place before the

beginning of the class, while the second recording right after the end. Four of them complained about feeling voice-fatigued. Although each voice recording contains two channels, however we analyze only the first channel. Note that one speaker could not be analyzed, since the duration of the phonemes was too short.

Figure 6.16 shows that six out of seven subjects appear to have an improvement in their voice tremor. The tremor level and the deviation of the tremor level increased by 0.1 for speaker me1 only.

Based on the information that four of them complained about feeling voice fatigued, the inference of our analysis is that there seems to be no relation between the tremor and the voice fatigue.

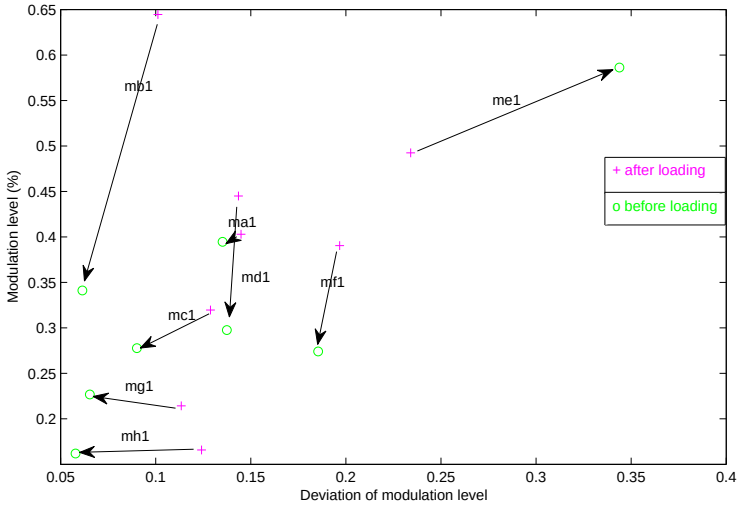


Figure 6.16: Teachers' tremor level coordinates before and after loading test

Chapter 7

Conclusions and future work

In the present thesis we worked on the accurate estimation of the vocal tremor in normophonic and dysphonic speakers and we proposed an accurate method for extracting vocal tremor characteristics.

We employed two existent methods for this purpose, namely a) the Adaptive Quasi Harmonic Model to extract accurately the speech components and the modulation frequency and level of the analyzed component and b) the Savitzky-Golay smoothing filter to cut off frequencies below 2Hz without affecting the desired modulation signals.

The advantages of our work can be summarized as follows:

1. The proposed method estimates voice tremor attributes accurately as reveled by the SRER values of the reconstructed signals.

2. All components can be analyzed to extract the tremor attributes. This is important since in case of existence of noise in the first harmonic other harmonics can be analyzed.

3. The tremor signal is modeled as a monocomponent time-varying sinusoidal signal. The modulation level and frequency do not correspond to single values but they are time-varying signals. Having a mathematical description of the signal is easier to handle and to reveal further properties of the signal, such as the significance of the deviation of the modulation level.

4. We achieved to remove the limitation of the duration of the phoneme. The Savitzky-Golay smoothing filter removes the low modulations which result from the inability of the speaker to keep his voice steady in a specific phonatory frequency.

5. The disengagement of the user. More specifically, the user has to give as input only the recorded speech signal of the speaker without defining any parameters.

6. We introduced a new significant attribute that defines tremor, namely, the deviation of the

modulation level. Normophonic from dysphonic speakers differed both in the modulation level and its deviation.

7. We introduced the weighted mean tremor value, defined by a weighted mean between the tremor level and the deviation of the tremor level to represent the voice tremor of a speaker.

The limitation of our work concerns the duration of the phoneme. The AQHM algorithm uses an autocorrelation method to estimate the first and last frame of the analyzed signal. As a result it fails to capture the signal for these frames. Therefore, we omit these frames from our analysis. However, this imposes a limitation on the duration of the phoneme. Phonemes cannot be shorter than 1.5 sec. Subjects who suffered from spasmodic dysphonia found it hard to speak and thus their vowels had a duration less than 1.5 sec. We proposed other methods for frequency tracking in order to apply them to the first and last frame of the AQHM. However, those methods we have implemented failed to extract the modulation signal, either due to the presence of other modulating components or due to the presence of noise in the signal.

Future work on voice tremor detection may involve the enhancement of the AQHM algorithm to capture the first and the last frame of the analyzed signal. This is significant for another reason despite the significance in analysis of dysphonic speakers' vowels of short duration. Tremor values may be more prominent at the beginning of the speech, where the voice cords start to vibrate. Another expansion of this work may be the analysis of other modulating frequencies of the analyzed harmonic. During our analysis we detected more than one modulating signals on some analyzed components. These modulating signals had frequencies in the tremor range. Therefore, there may not exist only one tremor modulation signal but others as well of lower amplitude. The examination of these frequencies may be important since the voice tremor, as mentioned in Chapter 1, is produced by different muscles of the respiratory system.

Part I

Appendix

Appendix A

From Kalman Filter to Extended Kalman Filter equations

A.1 Kalman filter overview and standard linear models

The Kalman filter is a set of mathematical equations that provides an efficient computational (recursive) means to estimate the state of a process, in a way that minimizes the mean of the squared error. The Kalman filter estimates a process by using a form of feedback control: the filter estimates the process state at some time and then obtains feedback in the form of (noisy) measurements. As such, the equations for the Kalman filter fall into two groups: time update equations and measurement update equations. The time update equations are responsible for projecting forward (in time) the current state and error covariance estimates to obtain the a priori estimates for the next time step. The measurement update equations are responsible for the feedback - i.e. for incorporating a new measurement into the a priori estimate to obtain an improved a posteriori estimate.

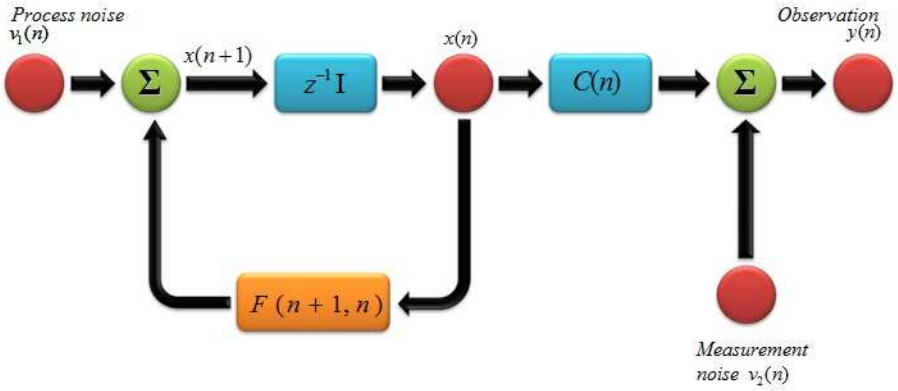


Figure A.1: Signal-flow graph representation of a linear, discrete-time dynamical system

The graph A.1 represents a linear, discrete-time dynamical system. In order to describe the systems behavior and therefore predict its future behavior we need to know the state vector $x(n)$. However, the state $x(n)$ is unknown and to estimate it we use the observed data $y(n)$. From the graph we can derive the following equations:

$$x(n+1) = F(n+1, n)x(n) + v_1(n) \tag{A.1}$$

$$y(n) = C(n)x(n) + v_2(n) \tag{A.2}$$

-where $v_1(n)$, $v_2(n)$ are uncorrelated zero-mean white processes with correlation matrices Q_1, Q_2 respectively. Equation A.3 is the process equation and equation A.4 is the measurement equation. The process equation models an unknown physical stochastic process $x(n)$ as the output of a linear dynamic system excited by the white noise $v_1(n)$. The measurement equation relates the observable output of the system $y(n)$ to the state $x(n)$. The Kalman filtering problem, namely, the problem of jointly solving the process and measurements equations for the unknown state in an optimal manner may formally be stated as follows:

Use the entire observed data, consisting of the observations $y(1), y(2), ..., y(n)$ to find for each $n \geq 1$, the minimum mean-square estimate of the state $x(i)$. The problem is called filtering if $i = n$, prediction if $i > n$ and smoothing if $1 \leq i < n$.

The transition matrix $F(n+1, n)$ and the measurement matrix $C(n)$ depend on the system and are known. Also known, are considered the correlation matrices Q_1, Q_2. Therefore, if we initialize the state $x(n)$ at time $n = 0$ given no previous observations, we will have an estimate $\hat{y}(n|y_{n-1})$ of the observation $y(n)$ at $n = 0$ and a forward prediction error $a(n) = y(n) - \hat{y}(n|y_{n-1})$. The forward prediction error (or innovation error) occurred from the wrong prediction/ estimation of the $x(n)$. Therefore, if we estimate the next state $x(n+1)$ from the process equation, this estimation will be wrong since the initialization of the state $x(n)$ was incorrect. Therefore, we can we estimate the next state $x(n+1)$ from the process equation plus a correction term $G(n)a(n)$ where G(n) is called Kalman gain.
To summarize, for the linear model described by

$$x(n+1) = F(n+1, n)x(n) + v_1(n) \tag{A.3}$$

$$y(n) = C(n)x(n) + v_2(n) \tag{A.4}$$

the Kalman filter equations are [31]:

Input vector process:

$$Observations = y(1), y(2), ..., y(n)$$

Initial conditions:

$$\hat{x}(1|y_0) = E[x(1)]$$
$$K(1,0) = E[(x(1) - E[x(1)])(x(1) - E[x(1)])^H] = \Pi_0$$

Known parameters:

$$Transition\ matrix = F(n+1,n)$$
$$Measurement\ matrix = C(n)$$
$$Correlation\ matrix\ of\ noise\ process = Q_1(n)$$
$$Correlation\ matrix\ of\ measurement\ process = Q_2(n)$$

Computation for n=1,2,3... :

$$G(n) = F(n+1,n)K(n,n-1)C^H(n)[C(n)K(n,n-1)C^H(n) + Q_2(n)]^{-1} \quad \text{(A.5)}$$
$$a(n) = y(n) - C(n)\hat{x}(n|y_{n-1}) \quad \text{(A.6)}$$
$$\hat{x}(n+1|y_n) = F(n+1,n)\hat{x}(n|y_{n-1}) + G(n)a(n) \quad \text{(A.7)}$$
$$K(n) = K(n,n-1) - F(n+1,n)G(n)C(n)K(n,n-1) \quad \text{(A.8)}$$
$$K(n+1,n) = F(n+1,n)K(n)F^H(n+1,n) + Q_1(n) \quad \text{(A.9)}$$

A.2 Extended Kalman filter and non-linear models

Kalman filtering problem addresses the estimation of a state vector in a linear model of a dynamic system. Since we are interested in non-linear models we may extend the use of Kalman filtering. The result filter is referred to as Extended Kalman Filter (EKF). EKF is the nonlinear version of the Kalman filter which linearizes about the current mean and covariance.

A.2.1 Kalman filter: standard linear state-space model and a two-step update process

We modify slightly the equations of the linear state-space model described in the previous section. Specifically, we modify the equation (A.6) which updates the state estimate $\hat{x}(n+1|y_n)$ in one step from the $\hat{x}(n|y_{n-1})$. The update of the state estimate is now performed in two steps: a) the first step uses $\hat{x}(n|y_{n-1})$ to update $\hat{x}(n|y_n)$ and b) the second step uses $\hat{x}(n|y_n)$ to update $\hat{x}(n+1|y_n)$ as Figure A.2 shows.

The corresponding equations are:

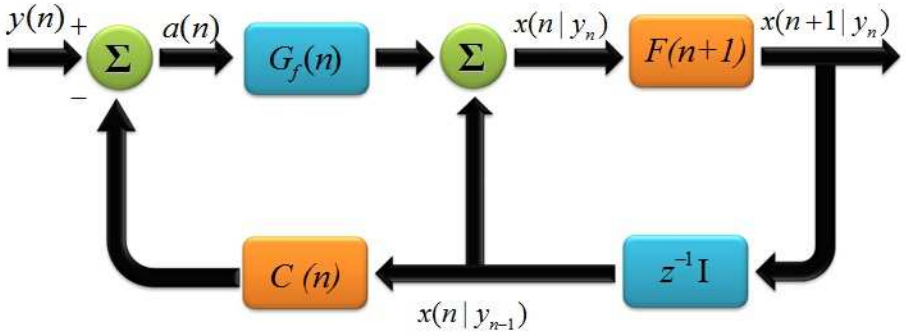


Figure A.2: Signal-flow graph representation of a linear, discrete-time dynamical system updated in two steps

$$\begin{aligned}
\hat{x}(n+1|y_n) &= F(n+1,n)\hat{x}(n|y_{n-1}) + G(n)a(n) = \\
&= F(n+1,n)[\hat{x}(n|y_{n-1}) + F(n+1,n)^{-1}G(n)a(n)] \\
&= F(n+1,n)\hat{x}(n|y_n) \\
\hat{x}(n|y_n) &= \hat{x}(n|y_{n-1}) + F(n+1,n)^{-1}G(n)a(n) = \\
&= \hat{x}(n|y_{n-1}) + G_f(n)a(n)
\end{aligned}$$

where we defined $G_f(n) = G_f(n)a(n)$ as a new gain matrix. Therefore, the equations change to:

Input vector process:

$$Observations = y(1), y(2), ..., y(n)$$

Initial conditions:

$$\hat{x}(1|y_0) = E[x(1)]$$
$$K(1,0) = E[(x(1) - E[x(1)])(x(1) - E[x(1)])^H] = \Pi_0$$

Known parameters:

$$Transition\ matrix = F(n+1,n)$$
$$Measurement\ matrix = C(n)$$
$$Correlation\ matrix\ of\ noise\ process = Q_1(n)$$
$$Correlation\ matrix\ of\ measurement\ process = Q_2(n)$$

Computation for n=1,2,3... :

$$G_f(n) = K(n, n-1)C^H(n)[C(n)K(n, n-1)C^H(n) + Q_2(n)]^{-1} \tag{A.10}$$

$$a(n) = y(n) - C(n)\hat{x}(n|y_{n-1}) \tag{A.11}$$

$$\hat{x}(n|y_n) = \hat{x}(n|y_{n-1}) + G_f(n)a(n) \tag{A.12}$$

$$\hat{x}(n+1|y_n) = F(n+1, n)\hat{x}(n|y_n) \tag{A.13}$$

$$K(n) = K(n, n-1) - G_f(n)C(n)K(n, n-1) = [I - G_f(n)C(n)]K(n, n-1)] \tag{A.14}$$

$$K(n+1, n) = F(n+1, n)K(n)F^H(n+1, n) + Q_1(n) \tag{A.15}$$

Instead of having the state space-model described above, we consider the linear state-space model:

$$x(n+1) = F(n+1, n)x(n) + v_1(n) + d(n) \tag{A.16}$$

$$y(n) = C(n)x(n) + v_2(n) \tag{A.17}$$

where $d(n)$ is a known vector. Then the above equations remain the same except for the equation A.36 that changes to:

$$\hat{x}(n+1) = F(n+1, n)\hat{x}(n|y(n)) + d(n) \tag{A.18}$$

A.2.2 Non-linear state-space model

A non-linear state-space model is described by the equations:

$$x(n+1) = F(n, x(n)) + v_1(n) \tag{A.19}$$

$$y(n) = C(n, x(n)) + v_2(n) \tag{A.20}$$

The functional $F(n, x(n))$ denotes a non-linear transition matrix function that is possibly time varying. The functional $C(n, x(n))$ denotes a non-linear measurement matrix function that may be time-varying too. In the linear case we simply have:

$$F(n, x(n)) = F(n+1, n)x(n) \tag{A.21}$$

$$C(n, x(n)) = C(n)x(n) \tag{A.22}$$

The basic idea of the EKF is to LINEARIZE the state-space model at each time instant around the most recent state estimate which is taken to be either $\hat{x}(n|y_n)$ or $\hat{x}(n|y_{n-1})$ depending on which particular functional is being considered. Once a linear model is obtained the standard Kalman filter equations are applied.

From the Taylor series we have:

$$F(n, x(n)) \approx F(n, \hat{x}(n|y_n)) + \left.\frac{\partial F(n,x)}{\partial x}\right|_{x=\hat{x}(n|y_n)} [x(n) - \hat{x}(n|y_n)] \tag{A.23}$$

$$C(n, x(n)) \approx C(n, \hat{x}(n|y_{n-1})) + \left.\frac{\partial C(n,x)}{\partial x}\right|_{x=\hat{x}(n|y_{n-1})} [x(n) - \hat{x}(n|y_{n-1})] \tag{A.24}$$

If we construct the matrices $F(n+1, n)$ and $C(n)$ to be respectively:

$$F(n+1, n) = \left.\frac{\partial F(n,x)}{\partial x}\right|_{x=\hat{x}(n|y_n)} \tag{A.25}$$

$$C(n) = \left.\frac{\partial C(n,x)}{\partial x}\right|_{x=\hat{x}(n|y_{n-1})} \tag{A.26}$$

then A.23, A.24 can be written as:

$$F(n, x(n)) \approx F(n+1, n)x(n) + F(n, \hat{x}(n|y_n)) - F(n+1, n)\hat{x}(n|y_n) = F(n+1, n)x(n) + d_1(n) \tag{A.27}$$

$$C(n, x(n)) \approx C(n)x(n) + C(n, \hat{x}(n|y_{n-1})) - C(n)\hat{x}(n|y_{n-1}) = C(n)x(n) + d_2(n) \tag{A.28}$$

If we substitute these equations to A.29, A.30 respectively we have:

$$x(n+1) \approx F(n+1, n)x(n) + d_1(n) + v_1(n) \tag{A.29}$$

$$y(n) \approx C(n)x(n) + d_2(n) + v_2(n) \tag{A.30}$$

where $d_1(n) = F(n, \hat{x}(n|y_n)) - F(n+1, n)\hat{x}(n|y_n)$ and $d_2(n) = C(n, \hat{x}(n|y_{n-1})) - C(n)\hat{x}(n|y_{n-1})$ Apparently in the linear case $d_1(n) = 0$ and $d_2(n) = 0$ if we have the linear state-space model (A.3), (A.4), or $d_1(n) = d(n)$ and $d_2(n) = 0$ if we have the linear state-space model (A.16), (A.17). Since $d_2(n) \neq 0$ we cannot apply the known Kalman filter equations. However, we can apply these equations to the following model:

$$x(n+1) \approx F(n+1, n)x(n) + d_1(n) + v_1(n) \tag{A.31}$$

$$\overline{y(n)} \approx C(n)x(n) + v_2(n) \tag{A.32}$$

and then compute $y(n)$ from the equation $y(n) = \overline{y(n)} + d_2(n)$. The entries in the term $\overline{y(n)}$ are all known at time t, therefore, $\overline{y(n)}$ can be regarded as the observation vector at time n. Likewise, in the term $d_1(n)$ are all known at time n. Therefore, the approximate state-space model (A.31), (A.31) is a linear state-space model of the same mathematical form as described in Eqs. (A.33)-(A.35), (A.9)-(A.38), (A.18). If we apply this approximation model to above the equations we have:

$$\begin{aligned}
a(n) &= \overline{y(n)} - C(n)\hat{x}(n|y_{n-1}) = \\
&= y(n) - d_2(n) - C(n)\hat{x}(n|y_{n-1}) = \\
&= y(n) - C(n, \hat{x}(n|y_{n-1})) + C(n)\hat{x}(n|y_{n-1}) - C(n)\hat{x}(n|y_{n-1}) = \\
&= y(n) - C(n, \hat{x}(n|y_{n-1})) \\
x(n+1) &= F(n+1)\hat{x}(n|y_n) + d_1(n) = \\
&= F(n+1)\hat{x}(n|y_n) + F(n, \hat{x}(n|y_n)) - F(n+1, n)\hat{x}(n|y_n) = \\
&= F(n, \hat{x}(n|y_n))
\end{aligned}$$

We notice that the only difference between the extended kalman filter equations and the kalman filter equations are in the computation of the vector $a(n)$ and the updated estimate $\hat{x}(n+1|y_n)$.

Input vector process:

$$Observations = y(1), y(2), ..., y(n)$$

Initial conditions:

$$\hat{x}(1|y_0) = E[x(1)]$$
$$K(1,0) = E[(x(1) - E[x(1)])(x(1) - E[x(1)])^H] = \Pi_0$$

Known parameters:

$$Transition\ matrix = F(n, x(n))$$
$$Measurement\ matrix = C(n, x(n))$$
$$Correlation\ matrix\ of\ noise\ process = Q_1(n)$$
$$Correlation\ matrix\ of\ measurement\ process = Q_2(n)$$

Computation for n=1,2,3... :

$$G_f(n) = K(n, n-1)C^H(n)[C(n)K(n, n-1)C^H(n) + Q_2(n)]^{-1} \quad \text{(A.33)}$$
$$a(n) = y(n) - C(n, \hat{x}(n|y_{n-1})) \quad \text{(A.34)}$$
$$\hat{x}(n|y_n) = \hat{x}(n|y_{n-1}) + G_f(n)a(n) \quad \text{(A.35)}$$
$$\hat{x}(n+1|y_n) = F(n, \hat{x}(n|y_n)) \quad \text{(A.36)}$$
$$K(n) = [I - G_f(n)C(n)]K(n, n-1)] \quad \text{(A.37)}$$
$$K(n+1, n) = F(n+1, n)K(n)F^H(n+1, n) + Q_1(n) \quad \text{(A.38)}$$

the linearized matrices $F(n+1, n)$, $C(n)$ are computed from $F(n, x(n))$, $C(n, x(n))$.

Bibliography

[1] D. Fucci and L. Petronsino. The practical applications of neuroanatomy for the speech-language pathologist. *Speech and Language, Advances in Basic Research and Practice*, 11:249–317, 1984.

[2] W.J. Wiener and A.E. Lang. *Movement disorders: A coprehensive survey.* Mount Kisco, NY: Futura Publishing Co., 1989.

[3] J.D. Speelman and J. Van Manen. Stereotactic thalamonotomy for the relief of intension tremor of multiple sclerosis. *Journal of Neurology, Neurosurgery and Psychiatry*, 47:596–599, 1984.

[4] C. Ohye, T. Shibazaki, T. Hirai, Y. Kawashima, M. Hirato, and M. Matsumura. Plastic change of thalamic organization in patients with tremor after stroke. *Applied Neurophysiology*, 48:288–292., 1985.

[5] W.L. Thompson. Management of alcohol withdrawal syndromes. *Archives of Internal Medicine*, 138:278–283, 1978.

[6] M. Mackey L. Rensing, U. van der Heiden, editor. *Central rhythmicities in motor control and perturbances.* Springer, Berlin, 1987.

[7] J. Schoentgen. Stohastic models of jitter. *Journal of Acoustic Society of America*, 109:1631–1650, 2001.

[8] H. Tomoda, H. Shibasaki, Y. Kuroda, and T. Shin. Voice tremor: Dysregulation of voluntary expiratory muscles. *Neurology*, 37(117–122), 1987.

[9] M. Critchley. Observations on essential (heredofamilial) tremor. *Brain*, 72:113–139, 1949.

[10] J. Jiang, E. Lin, J. Wu, C. Gener, and D.G. Hanson. Effects of simulated source of tremor on acoustic and airflow voice measures. *Journal of Voice*, 14:47–57, 2000.

[11] J. Koda and C.L. Ludlow. An evaluation of laryngeal muscle activation in patients with voice tremor. *Otolaryngology - Head and Neck Surgery*, 107:684–696, 1992.

[12] J. Schoentgen. Modulation frequency and modulation level owing to vocal microtremor. *Journal of Acoustic Society of America*, 2002.

[13] O. Fujimura and M. Hirano, editors. *Definitions and nomenclature related to voice quality*, 1995.

[14] R. T. Satallof. *Proffessional Voice: The Science and Art of Clinical Care.* Singular, San Diego, 1997.

[15] J. R. Brown and J. Simonson. Organic voice tremor. *Neurology*, 13:520–523, 1963.

[16] V.J. Boucher and T. Ayad. Physiological attributes of vocal fatigue and their acoustic effects: A synthesis of findings for a criterion-based prevention of acquired voice disorders. *Journal of Voice*, pages 1–13, 2008.

[17] W. S. Winholtz and L. O. Ramig. Vocal tremor analysis with the vocal demodulator. *Journal of Speech Hearing Research*, 35:562–573, 1992.

[18] Ph. Panter. *Modulation, noise and spectral analysis*. McGraw-Hill, New York, 1965.

[19] C. Ludlow, C.J. Bassich, N.P. Connor, and C. Coulter. Phonatory characteristics of vocal fold tremor. *Journal of Phonetics*, 14:509–515, 1986.

[20] C. Dromey, P. Warrick, and J. Irish. The influence of pitch and loudness changes on the acoustics of vocal tremor. *Journal of Speech, Language, and Hearing Research*, 45:879–890, 2002.

[21] L. Cnockaert, J. Schoentgen, P. Auzou, C. Ozsancak, Defebvre, and F. Grenez a. Low frequency vocal modulations in vowels produced by. *Speech Communication*, 50:288–300, 2008.

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

[23] J. Kreiman, B. Gabelman, and B.R. Gerratt. Perception of vocal tremor. *Journal of Speech, Language and Hearing Research*, 46:203–214, 2003.

[24] H. Ackermann and W. Zeigler. Acoustic analysis of vocal instability in cerebellar dysfunctions. *Annals of Otology, Rhinology and Laryngology*, 103:98–104, 1994.

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

[26] R. Orlikoff and R. Baken. Fundamental frequency modulation of the human voice by the heartbeat: preliminary results and possible mechanisms. *Journal of Acoustic Society of America*, 85:888–893, 1989.

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

[28] S.Kim and J. McNames. Tracking tremor frequency in spike trains using the extended kalman smoother. *IEEE Transactions on Biomedical Engineering*, 53(8), 2006.

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

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

[31] Simon Haykin. *Adaptive Filter Theory*. Prentice Hall, fourth edition.