University of Crete
Department of Computer Science

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

VOICE QUALITY ASSESSMENT USING PHASE INFORMATION: APPLICATION ON VOICE PATHOLOGY

(MSc. Thesis)

Olympia Simantiraki

Heraklion

September 2014

Department of Computer Science
University of Crete

VOICE QUALITY ASSESSMENT USING PHASE INFORMATION: APPLICATION ON VOICE PATHOLOGY

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

September 29, 2014

Author:

Olympia Simantiraki
Department of Computer Science

Committee

Supervisor

Yannis Stylianou
Professor

Member

Antonis Argyros
Professor

Member

Athanasios Mouchtaris
Assistant Professor

Accepted by:

Chairman of the
Graduate Studies Committee

Antonis Argyros
Professor

Heraklion, September 2014

Abstract

One of the most important human abilities is speech along with hearing. Speech is the primary way in which we attune to the society. Our voice can uncover several information about us to other people. It reveals our energy level, our emotions, our personality and our artistry. Voice abnormalities may cause social isolation or may create problems in the professional field. Due to this significance of the voice, the early detection of a voice pathology is essential.

A well-known voice abnormality is called Spasmodic Dysphonia (SD). SD is a neurological disease primarily affecting the regular contraction of the muscles around vocal cords, causing their undesirable vibration. This abnormal vibration of muscles of the glottis has an impact on speech. One that suffers from SD speaks more tremulous and makes disruptions during speech. Similar indications appear also to normophonic speakers usually related to stress, voice fatigue, etc. Even for the normophonic cases, these indications may be a first symptom of a neurological disease, so an early diagnosis is necessary. Therefore, algorithms that measure the intensity of the symptoms are very useful.

Traditional methods that detect and quantify voice pathologies use the amplitude information of the speech signal. More refined approaches make essential the isolation of the glottal source signal as the glottis is related to voice abnormalities. However, in both cases the amplitude based methods are not very reliable because the amplitude spectrum cannot capture characteristics of the glottis. A better indicator of voice irregularities is the phase information. Nevertheless, very few studies use the phase information because of its difficulty in the manipulation. Moreover, studies which work with the phase information, use inverse filtering techniques for extracting the glottal source signal and then they extract features from the phase spectrogram of the glottal source. In this thesis, an innovated phase-based method for voice quality assessment is presented. The proposed method is less complex than the state-of-the-art methods which use the inverse filtering for extracting the glottal source. Firstly, the instantaneous amplitudes, phases and frequencies are estimated from the speech signal by an adaptive harmonic model. From the instantaneous phases of the speech signal through mathematical formulas, a new phase spectrum, the Phase Distortion (PD) spectrum, is extracted, highly correlated with the shape of the glottal source. From the time variance of the PD spectrum (PDD), a new metric called Regularity Ratio (RR) is proposed to capture the irregularities of the glottal source.

Finally, the efficiency of our method is validated on a database containing speakers with SD before and after the botulinum toxin injection. The results show that the obtained ranking is

highly correlated with the subjective evaluations provided by medical doctors not only on the overall severity of SD but also on other features like tremor and jitter, revealing that our proposed feature, the RR, can be applied on other voice pathologies.

Περίληψη

Μια από τις σημαντικότερες λειτουργίες του ανθρώπου είναι η φωνή. Η ομιλία είναι ο πρωταρχικός τρόπος εναρμόνισής μας με την κοινωνία. Επίσης, η φωνή μας μπορεί να αποκαλύψει αρκετές πληροφορίες για μάς. Κάποιες από αυτές είναι η ενέργειά μας, τα συναισθήματά μας, η προσωπικότητά μας, καθώς επίσης και τα καλλιτεχνικά χαρακτηριστικά μας. Πιθανές διαταραχές στη φωνή μπορεί να προκαλέσουν κοινωνική απομόνωση του ατόμου ή να δημιουργήσουν προβλήματα σε ορισμένες κατηγορίες επαγγελμάτων. Λόγω λοιπόν της σημαντικότητας της φωνής, η έγκαιρη ανίχνευση παθολογιών σχετιζόμενων με τη φωνή είναι απαραίτητη.

Μια πολύ γνωστή πάθηση της φωνή είναι η σπασμωδική δυσφωνία. Η πάθηση αυτή είναι νευρολογική και κατά κύριο λόγο επηρεάζει την ομαλή λειτουργία των μυών που βρίσκονται κοντά στη γλωττίδα, προκαλώντας ανεπιθύμητες συσπάσεις. Αυτές οι μη ομαλές συσπάσεις των μυών της γλωττίδας έχουν αντίκτυπο στην παραγόμενη ομιλία. Η φωνή του ατόμου που πάσχει από σπασμωδική δυσφωνία εμφανίζει τρέμουλο και διακόπτεται απότομα κατά την διάρκεια της ομιλίας του. Παρόμοιες ενδείξεις μπορεί να παρουσιάσουν και νορμοφωνικοί ομιλητές, οι οποίες σχετίζονται με το άγχος, την φωνητική κούραση κλπ. Ακόμα όμως και για τις περιπτώσεις αυτών των νορμοφωνικών ομιλητών, αυτές οι ενδείξεις μπορεί να είναι ένα πρώτο σύμπτωμα για την εμφάνιση κάποιας νευρολογικής ασθένειας. Επομένως, γίνεται αντιληπτό ότι οι αλγόριθμοι επεξεργασίας φωνής, που ποσοτικοποιούν τη σοβαρότητα των συμπτωμάτων, είναι πολύ χρήσιμοι για την έγκαιρη διάγνωση των ασθενειών.

Διάφορες μέθοδοι, που έχουν προταθεί κατα καιρούς για την ανίχνευση και ποσοτικοποίηση παθολογιών φωνής, χρησιμοποιούν την πληροφορία από το φασματικό πλάτος του σήματος ομιλίας. Άλλες εγκυρότερες μέθοδοι, απομονώνουν το σήμα της γλωττίδας η οποία και σχετίζεται με την φυσιολογική ή μη λειτουργία της φωνής. Όμως οι μέθοδοι, που βασίζονται στο φασματικό πλάτος του σήματος, δεν είναι αξιόπιστες, γιατί το φάσμα πλάτους δεν απεικονίζει τα χαρακτηριστικά της γλωττίδας. Ένας καλύτερος δείκτης για την ανίχνευση ανωμαλιών φωνής είναι η φάση του σήματος της γλωττίδας. Όμως πολύ λίγες μελέτες χρησιμοποιούν την πληροφορία της φάσης, λόγω της δυσκολίας εξαγωγής της από το σήμα φωνής. Οι μελέτες που χρησιμοποιούν την πληροφορία φάσης, χρησιμοποιούν τεχνικές αντίστροφου φιλτραρίσματος για την εξαγωγή του σήματος της γλωττίδας και έπειτα εξάγουν χαρακτηριστικά από το φασματογράφημα της φάσης του σήματος της γλωττίδας. Στην εργασία αυτή παρουσιάζεται μια καινούρια μέθοδος για την εκτίμηση της ποιότητας της φωνής που βασίζεται στη φάση. Η μέθοδος αυτή είναι λιγότερο πολύπλοκη από άλλες μεθόδους που για να εξάγουν το σήμα της γλωττίδας χρησιμοποιούν τεχνικές αντίστροφου φιλτραρίσματος. Αρχικά,

εφαρμόζοντας στο σήμα ένα αρμονικό προσαρμοστικό μοντέλο εκτιμώνται τα στιγμιαία χαρακτηριστικά του σήματος φωνής (πλάτος, φάση, συχνότητα). Από τις στιγμιαίες φάσεις του σήματος φωνής μέσω μαθηματικών τύπων, ένα καινούριο φασματογράφημα φάσης, το φασματογράφημα παραμόρφωση φάσης (PD-Phase Distortion) εξάγεται, το οποίο είναι συσχετισμένο με το σήμα της γλωττίδας. Από την διακύμανση του PD φασματογραφήματος, μια καινούρια μετρική, ο Δείκτης Κανονικότητας, προτείνεται για να συλλάβει τις ανωμαλίες του σήματος της γλωττίδας.

Τέλος, η αποδοτικότητα της μεθόδου μας εκτιμάται πάνω σε μια βάση που περιέχει ομιλητές με σπασμωδική δυσφωνία πριν και μετά την έγχυση βουτουλινικής τοξίνης στους μύες της γλωττίδας. Τα αποτελέσματα από την κατάταξη που προέκυψαν έδειξαν ότι η μέθοδος που προτείνει η εργασία αυτή, είναι συσχετισμένη σε μεγάλο βαθμό όχι μόνο με τη συνολική σοβαρότητα της σπασμωδικής δυσφωνίας αλλά και με άλλα υποκειμενικά χαρακτηριστικά παθολογίας όπως το τρέμουλο σε χαμηλές και υψηλές συχνότητες (θόρυβος), που σημαίνει ότι η προτεινόμενη μετρική, Δείκτης Κανονικότητας, μπορεί να εφαρμοστεί και σε άλλες παθολογίες φωνής.

Acknowledgements

This thesis is the outcome of a creative effort for the acquisition of the Master of Science degree in Computer Science Department of the University of Crete. It is also the result of my experience in the field of research which I obtained during my collaboration with the Institute of Computer Science (ICS-FORTH).

First of all, I would like to thank my supervisor, Professor Yannis Stylianou, for giving me the opportunity of becoming a member of his team and for his great advice and support.

I would especially like to thank the PhD student Maria Koutsogiannaki who showed me how to face research-related issues and for her priceless help and patience during my MSc studies.

Many thanks to my parents who have been standing by my side with whatever I am trying to fulfil in my life.

I am also very grateful to my sisters (Penny, Natalia, Ismini) for helping me with the formation of my thesis and for being a very important part of my life.

I would like to thank all of the laboratory members for the pleasant atmosphere that we created.

Definitely, I could not miss out to thank my closest friends, Eleni, Kallia, Paulos, Tasos, Maria, Giorgos, who are always there for me.

And of course .... my lovely dogs Flor and Snoopy!!

Last but not least, I would like to thank the girls of Gregory's cafeteria, who made my coffee with a smile on a daily basis and enhanced the beginning of my day.

To my beloved grandmother, Olympia, after whom I was named

Στην αγαπημένη και συνονόματη γιαγιά μου, Ολυμπία

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Early investigations on the perceptual relevance of the phase information concluded that human ear is phase-deaf. In 19th century, Ohm(1843) demonstrated that the human ear is able to capture sinusoidal oscillations and performing a Fourier analysis showed that only the magnitude spectrum is perceived [3]. A few years later in 1875, the Ohm's acoustic law was verified by Helmholtz [4]. After a long period of time, the investigations about this issue were continued [5],[6] and through human perception experiments [7] it was showed that the short-time phase spectrum contributes to speech intelligibility the same as the corresponding power spectrum. It has also been showed in speech coding and psychoacoustic research [8] that humans can distinguish very well the different phase spectra and that the human auditory system is sensitive to the difference between zero and non-zero phase signals [9].

The majority of studies, focusing on voice quality assessment, use features from the amplitude parameters because of the difficulty on phase manipulation. The difficulty is based on the wrapping of the phase. In this work we propose a phase based feature, named Phase Distortion Deviation (PDD). This feature is computed through mathematical formulas which solve the phase wrapping problem. As we know speech is obtained by exciting a minimum phase (vocal tract filter) and a maximum phase component (glottal source). Therefore, the amplitude spectrum cannot capture maximum phase characteristics. Voice quality is connected to the glottal source so the extracted features should be linked with the maximum phase component of speech.

So, due to the significant importance of the phase spectrum of the speech signal and the direct relation of the phase with the glottal source, a new metric is proposed. This metric is based on the phase spectrum for characterizing the maximum-phase component of the glottal source. The new phase representation proposed in this work is used as a voice quality assessment metric for pathological voices. The results show that it can automatically detect voice irregularities revealing the importance of the phase.

Speech is important for a variety of reasons, such as that it is one of the main ways we use to convey messages, sentiments and certain feelings. For some people such as singers and teachers, the voice is a necessary tool for their work. So, having a voice abnormality may cause trouble in communication and/or even in work. The extended use of the voice can cause problems like vocal fatigue, pain etc.. All these symptoms must be quickly diagnosed in order to restore the voice and provide the patients with physical and emotional relief. Medical doctors have to do a clinical examination so that to detect the disease and evaluate the progress of the patient. The successful diagnosis is obviously based on the experience and the skills of the clinical doctor. This is a difficult task considering the fact that every disease may consist of many different symptoms and signs, some of which may become overlapped by others or not be significant during the clinical examination. However, due to the development of the voice technologies, different algorithms can be applied which could objectively detect and quantify the existing pathology. The development of such techniques would be a useful tool for medical doctors, helping them end up with the right diagnosis. In addition, the objectivity of these methods makes their results undoubtable.

1.1 Literature review

Up to now, various objective measures have been proposed for speech quality evaluation. In speech modeling the distortion metrics that are used are based on time-domain features, like Signal to Error Reconstruction Ratio and Signal to Noise Ratio, or frequency domain features extracted from the amplitude spectrum, like Spectral Distance and Cepstrum distance measures. To quantify the quality of natural speech, a state of the art method is Harmonic to Noise Ratio which is commonly used. HNR is computed on the speech signal and uses amplitude - spectrum based features. However, as it will be shown in Chapter 2, HNR is not ideal for quantifying the quality of voice pathology. Moreover, many techniques focus on the glottal features and amplitude - spectrum based features computed on the glottal signal (HRF , $H_1 - H_2$) as they are shown to be linked with speech quality. Features extraction is more complicated when the signal comes from disordered voices, while the glottal source estimation is a rather complex and delicate problem.

The phase information has not received much attention as a quality indicator in literature. In this study an objective measure for voice quality assessment based on the phase spectrum is proposed. Several studies have shown the connection of the maximum phase component of speech

with the glottal source and it's importance on maintaining the perceived quality of speech. Some methods used to estimate the glottal source are: the Iterative Adaptive Inverse Filtering method (IAIF) [10], the minimum/maximum - phase decomposition methods (Complex Cepstrum (CC) and Zeros of the Z-Transform (ZZT) based methods [11]).

Phase is often neglected due to its wrapping which results from the linear phase shift. This linear phase mismatch problem is solved when the notion of the center of gravity is introduced [12]. In [12] the phase structure of speech is also revealed, being used ever since in various applications with great success. Moreover, in [13] the phase difference between two frequency components has been shown to have perceived characteristics. In [14] the Phase Distortion (PD) is used to characterize the shape of periodic pulses of the glottal source independently of the source - filter characteristics, such as the duration of the glottal pulse, the position of analysis window and the influence of minimum phase component of the speech (e.g the vocal-tract filter). Due to the assumption that the glottal shape is connected with the voice quality , the phase distortion can be used as a quality assessment metric.

1.2 Voice pathology detection using phase information

In this thesis, information based on the instantaneous phase of a voice signal is used innovatively in order to detect voice pathologies. Despite the fact that the phase of a speech signal includes a lot of information, until the last decade it was not widely used by the researchers due to its complexity (normally it comes out in a wrapped form from the voice output signal). Lately, ways of simplifying the management of the phase have been developed ([15], [16]), which is the reason why more and more scientists work on it. Also, analysing the phase we could export useful features which can keep the whole information about the signal.

This study proposes a new phase representation which automatically detects voice irregularities. Other researches that use phase-based features for automatic voice pathology detection [17] have been proposed. However, these researches use inverse filtering methods for estimating the phase components that correspond to the glottis. These glottal estimation is a complex process and much more for irregular voice signals, in which there is almost no harmonicity. Our proposed method, estimates the glottal source signal from the speech signal using a Harmonic model without explicit glottal estimation.

The methodology is based on Phase Distortion (PD) proposed in [13] but the estimation of the phase features is done by a harmonic model, thus giving to PD similar to the group delay

characteristics ([18], [19], [20], [17]). The estimation of the instantaneous phases is performed by the adaptive Harmonic Model (aHM) [21]. Then, for revealing the phase structure of speech, the linear phase shift and the Phase Distortion (PD) are estimated on the signal after removing its minimum phase component. Our feature which better describes voice disorders than PD, is the standard deviation of the PD computed over time for each harmonic. PDD describes the phase variability of the voice source [22] which is more evident in pathological voices. The advantage of our technique over other phase - based techniques is that it eliminates the necessity of reliable estimation of the glottal closure instants (GCI).

PDD is evaluated in two databases consisting of normophonic and dysphonic speakers with spasmodic dysphonia [1]. The objective ranking performed by PDD on the database of dysphonic speakers is compared with the subjective ranking performed by medical doctors who ranked the patients according to three features: jitter, tremor and overall severity of SD [1]. Also, our metric is compared with another objective metric, based on the appearance of tremor on SD, called WMTV [23] which quantifies the severity of SD. Finally, HNR, H1-H2 and HRF metrics are compared to the overall severity of SD in order to prove that these metrics are not appropriate for characterizing the pathology of SD.

1.3 Contributions

This thesis presents a novel method for quantifying objectively the voice quality of normophonic and dysphonic subjects [24].

The contribution of this thesis can be accurately summarized by the following highlights:

1. The importance of the phase information for the voice quality assessment is revealed.

2. The Phase Distortion [14] combined with an Adaptive Harmonic Model [21] is used to remove the linear phase influence due to the misalignments of the glottal closure instants with the analysis window. This makes the accurate estimation of the glottal source signal unnecessary as it describes it using the instantaneous phases extracted from the harmonic model.

3. The proposed quality descriptor is based on the time deviation of Phase Distortion and can

objectively quantify the severity of spasmodic dysphonia, revealing the connection between the phase and voice quality.

Our results from the experiments show that the proposed method (PDD) can capture the dysphonia -if it exists- and can also quantify the noisy and harmonic parts of speech. This arises from the high correlations of PDD-based method with the subjective and objective evaluations.

1.4 Structure of the thesis

The thesis is organized as following:

In **Chapter 2**, the mechanism of human speech production is analysed. Also, in this Chapter the phase structure for harmonic speech models is shown as well as why the phase spectrum is preferred over the magnitude spectrum for characterizing pathological signals.

In **Chapter 3**, the PD is thoroughly analysed and a new feature based on the PD is proposed, named Phase Distortion Deviation (PDD).

In **Chapter 4**, the two databases which contain normophonic speakers (the first database) and dysphonic speakers suffering from Spasmodic Dysphonia (the second database) are described. A new metric for quantifying the quality of the voice is proposed, named Regularity Ratio(RR). The performance of the proposed metric is compared to other subjective and objective metrics that also quantify Spasmodic Dysphonia and then the results are discussed, showing that the proposed metric is highly correlated with medical doctor's results.

Finally, in **Chapter 5**, the conclusions are summed up and ideas for future work are suggested.

Chapter 2

Voice pathology detection techniques

In this Chapter we propose the extraction of the phase spectrum from the speech signal and not from the glottal signal. Specifically, the mechanism of human speech production is analysed and then the speech production system (source-filter model) is shown as a model of minimum and maximum phase components. A series of methods for detecting voice pathologies is analysed, using signal processing techniques. In the literature, most of them are based on the harmonicity detection of the amplitude spectrum of speech signal because of the ease of its manipulation. More accurate methods separate the glottal source signal from the vocal tract(filter). This separation is essential for isolating the glottal source which is related to the voice quality and as a consequence, to the voice pathologies, in which we are interested. Most studies in order to estimate the glottal source from the output/speech signal use inverse filtering techniques. Inverse filtering is a complex method even more for pathological voices. Due to the difficulty of the phase manipulation, very few studies focused on extracting information from the phase spectrum. However, if we want to deal with detection and quantification of voice pathologies, we should seriously consider the information provided by the phase of the glottal signal. This arises because the pathologies where we practice, due to abnormal muscle movement in the larynx, affect the signal produced in the glottis (glottal source) and not in the vocal tract. The superiority of using phase information for voice pathology is explained. Lastly, the possibility of extracting phase information from the speech signal directly and not from the glottal signal is presented, leading to our proposed method for voice pathology detection.

2.1 The mechanism of human speech production

The mechanism of human speech production is normally complex. It can be classified into three main groups: lungs, larynx (vocal source) and vocal tract. The lungs is an air pump that provides

airflow to the larynx and to the vocal tract (Fig. 2.1).

For the speech source there are three general categories: periodic, noisy, impulsive. The discrete-

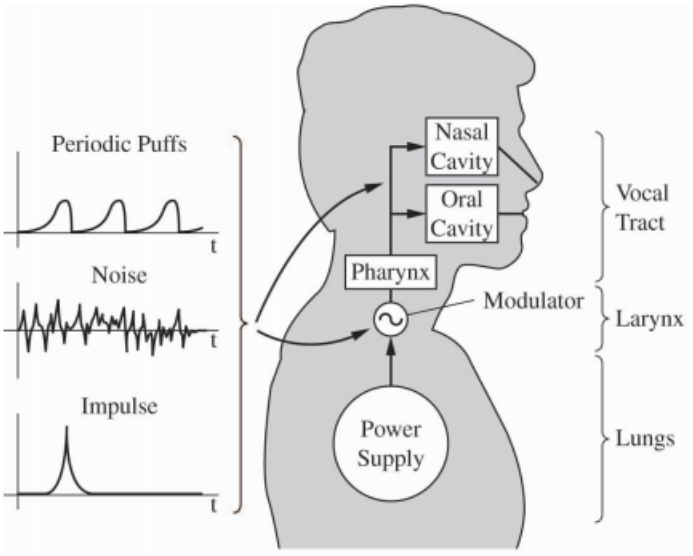


Figure 2.1: Speech Production

time speech production model, is illustrated in Fig. 2.2 for periodic, noise and impulsive inputs. In a discrete-time point of view, the speech pressure output - speech signal $s[n]$ is composed by the convolution $(*)$ of the impulse response of : the vocal tract $v[n]$, the excitation signal $e[n]$ and the lip radiation $r[n]$:

$$s[n] = e[n] * v[n] * r[n] \tag{2.1}$$

In Z-transform :

$$S(z) = E(z)V(z)R(z) \tag{2.2}$$

where $R(z)$ is the discrete-time radiation impedance, $V(z)$ is the discrete-time all-pole vocal transfer function from the volume velocity at the glottis to volume velocity at the lips and $E(z)$ is the discrete-time input-source signal.
$R(z) \approx 1 - z^{-1}$ and is derived as a single zero on the unit circle, $V(z)$ has poles inside the unit circle, but may have zeros inside and outside the unit circle so the vocal transfer function is a mixed phase model and $E(z)$ is maximum-phase having two poles outside the unit circle. The above speech model is the typical, non-parametric system representation, proposed by Quatieri [25].

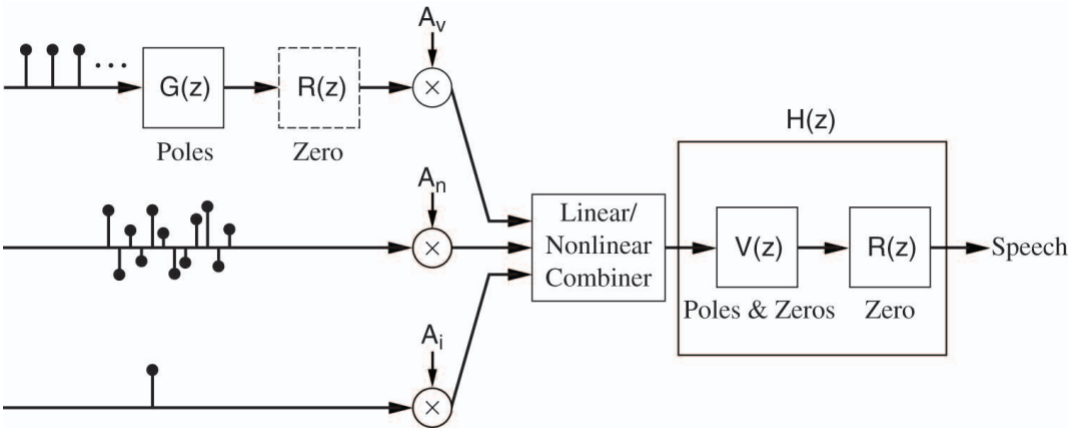


Figure 2.2: Complete discrete-time speech production model

In this basic source - filter model, speech production is a "forward only" system without interaction between source and vocal tract filter. Glottal excitation signal is a maximum phase model and vocal tract is a minimum phase model. Speech signal is a combination of these two systems and so it provides an all-pole system (mixed-phase). For the reason that voice quality is connected with the glottal source, it seems logical to find a way to process with the signal produced in the glottis. The main problem is to extract the vocal tract part of the all pole system that is required for the inverse filtering in order to get the glottal flow signal. If we estimate the glottal flow signal, then we have to calculate glottal flow parameters e.g voice quality variations. One way to estimate glottal parameters is by using the original time domain signal and detecting landmarks of the signal (minima, maxima, zero crossings), but this method is not very robust in noisy data because time domain waveform and landmarks vary a lot with the noise. So, another more preferable method is to transform the signal to frequency domain as glottal flow characteristics contribute to the speech signal spectrum. The source spectrum is typically computed using the fast Fourier transform (FFT). In most of the cases the spectrum is computed pitch-asynchronously over several fundamental periods and the spectral decay is measured from the levels of the harmonics. In addition, it is possible to quantify the spectral decay of the glottal source by using the FFT that is computed pitch-synchronous over a single glottal cycle.

In Fig. 2.3 it is shown that the glottal shape is directly connected with the quality of the speech. Both the glottal pulses (green line) and the original speech signal (blue line) are illustrated. Fig. 2.3(up) shows the glottal pulses of a normophonic speaker with small changes in

period, in contrast to Fig. 2.3(down) which shows a dysphonic speaker whose glottal pulses are changing both in period and in shape. Glottal pulses are strongly related to voice dysphonia.

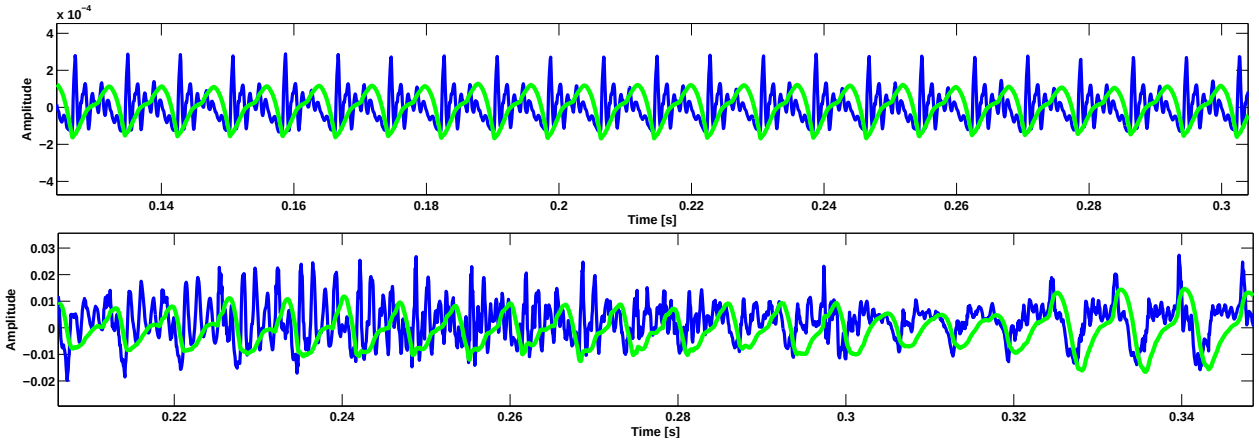


Figure 2.3: Glottal pulses (green line) of a normophonic male speaker (Fig. 2.3(up)) and a dysphonic male speaker who suffers from SD (Fig. 2.3(down)). Both speakers uttered the sustained vowel /a/. The original speech signal is depicted with the blue line.

2.2 Amplitude spectrum based techniques

2.2.1 Amplitude spectrum based techniques directly on the speech signal

Most studies, in order to assess the quality of the voice, use amplitude spectrum based features computed on the speech signal or on the glottal source signal.

HNR [26]: Amplitude spectrum based techniques use the magnitude spectrum of the Fourier Transform (FT) to detect voice pathology. An objective and very common way to quantify the amount of noise in a signal is **HNR (Harmonic to Noise Ratio)**. This state of the art method is also used widely in speech signals, for example to calculate the degree of hoarseness [27]. Generally HNR computes the energy of the harmonics of the output speech signal and compares it to the energy of the noise. Many HNR methods have been proposed for the quantification of the voice pathology. In this thesis an HNR method is implemented based on the algorithm description that follows. This HNR method as well as other methods proposed have two drawbacks: firstly, the voice abnormalities are related to the glottal source and the HNR metric is applied on the magnitude spectrum of the speech signal which is also affected by the vocal tract, secondly another problem lies on the fact that for dysphonic speakers, HNR doesn't select only the peaks that correspond to harmonics but selects also peaks from noise, so the HNR returns unreliable

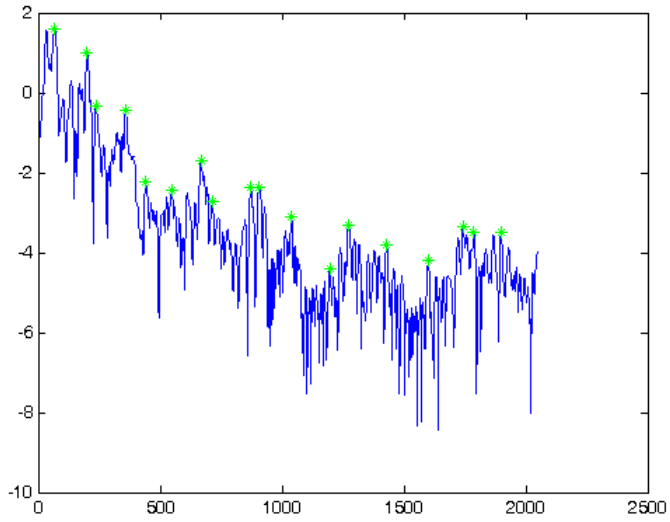
results. This is shown below in Fig. 2.4(a) where incorrect peak picking of a dysphonic speaker is performed by HNR. Fig. 2.4(b) depicts the peak picking of a normophonic speaker's signal. We can easily observe that selecting the peaks that correspond to harmonics is complicated in the case of the dysphonic speaker since the peaks are not integer multiplies of the fundamental frequency(the frequency that corresponds to the first peak of this frame) and we do not know which peaks should be considered harmonics and which noise. The disadvantages of HNR method are presented in Chapter 4, where an HNR implementation is compared to subjective evaluation metrics.

HNR Algorithm

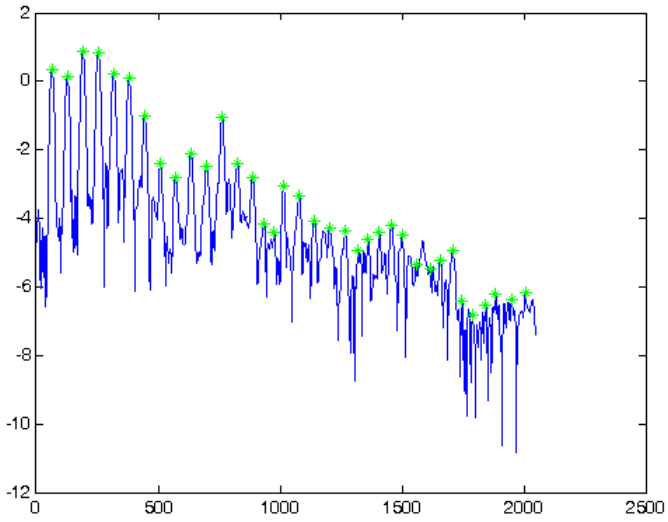
- make the STFT of the signal
- take the absolute value
- take the logarithmic magnitude spectrum of the single-sided magnitude spectrum
- apply peak picking in the magnitude spectrum (**quasi-periodic peaks**)
- compute the cepstrum coefficients by applying fft on the log magnitude
- define the liftering window
- take the low-time liftered cepstrum (LT)
- and the high-time liftered cepstrum (HT)
- from the HT liftered cepstrum find the peak location of the peak gives pitch period in quefrency samples
- take non-cepstral coefficients from the LT
- make the fft for non-cepstral and calculate the noise peaks (**noise peaks**)
- the HNR is given by :

$$HNR = 20log(\frac{energy\ of\ the\ \textbf{periodic components}\ of\ speech}{energy\ of\ the\ \textbf{noise components of speech}})$$

Cepstral Peak Prominence CPP [28], [29]: Another method used for quantifying the dysphonia is the Cepstral Peak Prominence (CPP). This method is an amplitude-based spectrum



(a) dysphonic speaker

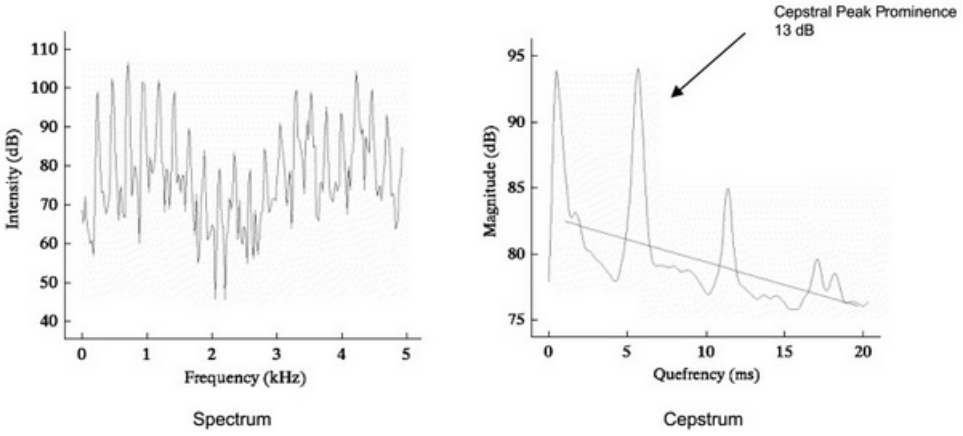


(b) normophonic speaker

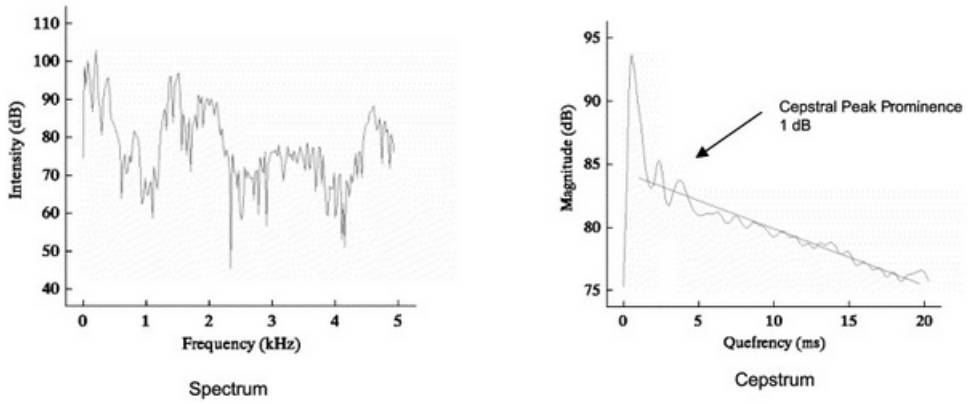
Figure 2.4: peak peaking using HNR

technique computed on the speech signal. Firstly, a Fourier transform is applied on the acoustic signal to create the amplitude spectrum. Thus, the intensity of each frequency within the signal is represented in the spectrum. Then, a Fourier transformation is performed on the amplitude spectrum to produce the cepstrum. In the quefrency domain, a better visual picture of the degree of harmonic organization is produced. A linear regression line is fitted relating quefrency to

cepstral magnitude. The CPP is the difference in amplitude between the cepstral peak and the corresponding value on the regression line that is directly below the peak (Fig. 2.5). So CPP is a measure of the degree of harmonic organization, which shows how far the cepstral peak emanates from the cepstral "background-noise". A normal voice which has a well-defined harmonic structure, will have a strong cepstral peak. On the other hand, signals that don't have harmonic structure have small CPPs. Also, the individual cepstra are averaged over a given number of frames before extracting the cepstral peak and calculating the peak prominence. This CPP smoothing (CPPS) is prove to be a better predictor.



(a) Spectrum and cepstrum of a normal voice.



(b) Spectrum and cepstrum of a severely dysphonic voice.

Figure 2.5: CPP method.

Global Energy [30]: The Global Energy of the speech signal is used also as a discrimination

tool for voice pathologies. The global energy is computed by the spectrogram in which the variations of the energy in time and in frequency appear.

2.2.2 Amplitude spectrum based techniques on the glottal signal

Inverse filtering method is used by the majority of the voice quality assessment techniques in order to extract features from the glottal source signal. In [31] the inverse filtered waveforms (glottal pulses) were used as the excitation for several voice types and the following features were measured: (1) instant of maximum closing slope of the glottal pulse, (2) glottal pulse width and (3) the ratio of the duration of the glottal opening phase factors to assess voice to duration of the glottal closing phase. In that study, it was observed that the glottal spectra of different voice types have distinctive amplitude relationships between the fundamental and higher harmonics. So they defined a parameter **HRF (Harmonic Richness Factor)** to measure this relationship, which is expressed in Eq. 2.3.

$$HRF = \sum_{i>2} H_i/H_1 \tag{2.3}$$

where H_1 is the amplitude of fundamental frequency and H_i is the amplitude of the ith harmonic. HRF is the ratio between the sum of the amplitude of harmonics above the fundamental and the amplitude of the fundamental.

Other studies analyse the spectral decay of the voice source by computing the difference between the amplitude of the fundamental and the second harmonic. The voice source is isolated by estimating the glottal excitation signal using inverse filtering. The open part of the glottal vibratory (the open quotient (OQ)) is connected to the voice quality [32]. To estimate changes in OQ, its amplitude spectrum is computed. The changes in the relative amplitudes of the first two harmonics of the voice source, **H1-H2**, denote changes in OQ (Fig. 2.6). As OQ increases, energy in the first harmonic (and thus H1-H2) is assumed to increase and this increase is the presumptive cause of the change in vocal quality. In Fig. 2.6, the glottal area waveform was generated from high-speed images. H1-H2 can be measured in two ways. The first way is directly from the spectrum of the voice source (usually obtained by inverse filtering or from a computational model) and the second way is from the audio signal at the mouth after cancelling the influence of vocal tract resonances (H1*-H2*).

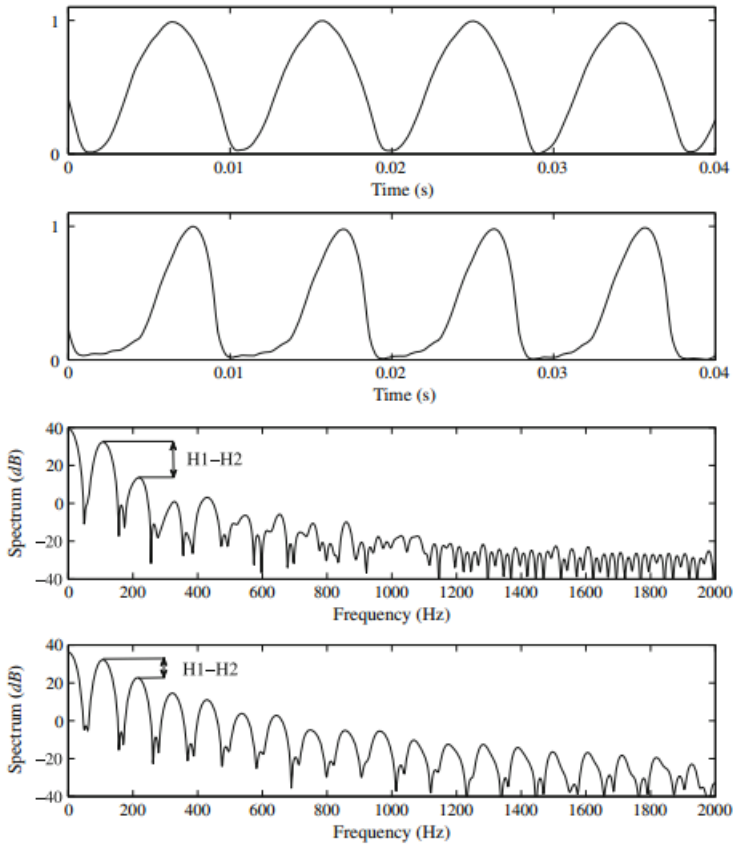


Figure 2.6: The acoustical speech pressure waveform pronounced by a male speaker. Time-domain waveforms of the glottal flow are shown in breathy (top panel) and pressed (second panel from top) phonation. Below these, voice source spectrum is shown for breathy (third panel from top) and pressed (bottom panel) phonation. Different spectral decay of the two phonation is quantified by H1-H2: the value of H1-H2 is 18.4 dB and 9.6 dB in breathy and pressed phonation, respectively [2].

2.3 Phase spectrum based techniques on the glottal signal

As mentioned on the previous Section, the glottal excitation signal is modelled as a maximum phase component and vocal tract as a minimum phase component. Speech signal is a combination of these two systems and so it provides an all-pole system (mixed-phase). The amplitude spectrum of the Fourier transform assumes that speech is a minimum-phase component. Therefore, the amplitude spectrum cannot reveal useful information of a maximum-phase component, the glottal signal. This denotes that amplitude spectrum based techniques not only are difficult to manipulate in case of dysphonic voices (due to the non-harmonicity of the amplitude spectrum) but also do not describe sufficiently the glottal source. Furthermore, the glottal source estimation

is a rather complex and delicate problem. Therefore, other techniques should be used to detect voice irregularity. These techniques should focus on the phase spectrum, as the maximum-phase characteristics of the glottal signal can be observed on the mixed-phase component of FT of the speech signal and not on the amplitude spectrum.

The function of the group delay has been shown to reveal voice source characteristics. Specifically, the **Group Delay Function** is the derivative of the unwrapped phase spectrum and can be used for extracting phase-based features for detecting voice abnormalities. Mixed - phase characteristics can only be observed on the phase component of the FT spectrum but not on the amplitude spectrum. So, group delay processing (first derivative of the unwrapped phase) is important for studying characteristics of the glottal flow. By nature, phase component of the Fourier Transform spectrum is in wrapped form and the first derivative of the unwrapped phase spectrum is much easier to be studied. When minimum phase assumption for speech signal is used, the glottal flow characteristics are mixed with that of vocal tract characteristics in the phase/group spectrum domain.
This mixing is more obvious on (Fig. 2.7). Due to the anti-causality of the glottal flow, the corresponding group delay function includes a negative peak, which also appears in the speech group delay function. This peak shows that the glottal formant is of maximum phase. In the above figures (Fig. 2.7), we can clearly see that mixed-phase characteristics (spectral components from maximum and minimum phase parts) can only be observed on the group delay but not on the magnitude spectrum, by comparing magnitude spectrum and group delay, which are functions of speech. This shows that studying phase/group delay information is especially important if we want to capture characteristics of glottal flow and of the vocal tract components separately. In our case, where we are interested only for the glottal excitation signal, the vocal tract influence should be removed.

In [17] a more refined than Group delay phase-based technique has been used for detecting the voice disorders automatically. They estimate the glottal source characteristics by using an inverse filtering technique called Complex Cepstrum-based Decomposition (CCD) [11]. CCD deconvolves the mixed-phase signal into a minimum-phase and a maximum-phase component, providing an estimation of the glottal source. This method requires the accurate estimation of the Glottal Closure Instants, GCIs. Then, the group delay of the glottal signal is taken to detect voice pathology.

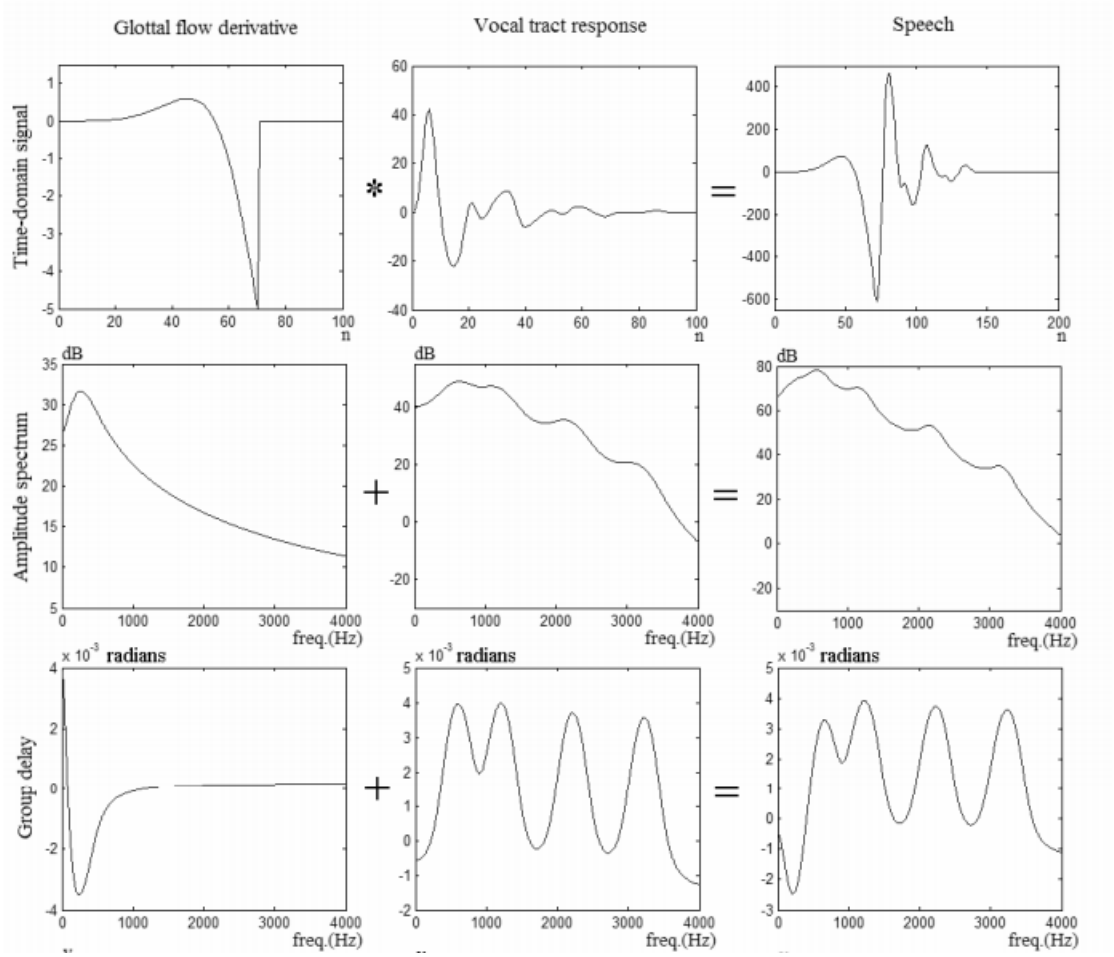


Figure 2.7: The mixed - phase speech model

2.4 Revealing the phase structure of speech: the notion of center of gravity or Relative phase shift

As mentioned above, glottal excitation signal is a maximum phase component and vocal tract is a minimum phase component, while their addition, the speech signal is a mixed-phase component. The aim is to remove the influence of the phase imposed by the vocal tract. Moreover, another issue should be taken into account, the wrapping of the phase. A phase model [33] that describes the above relation is:

$$\phi_k^i = \theta_k^i + k \int_{t_c^i}^{t^i} \omega_0(\tau) d\tau + \angle V^i(k\omega_0(t_i)) \tag{2.4}$$

where k denotes the number of kth harmonic components, therefore suggesting the use of a harmonic analysis model. The extracted phases ϕ_k from the speech waveform, are modelled as a

summation of i) the phase of the glottal pulses θ_k, ii) the linear phase imposed by the delay of the center of the analysis window t_c and the position of the glottal pulse and iii) the minimum phase component which models the vocal tract influence $\angle V^i(k\omega_0(t_i))$. Many studies have shown the link of the maximum-phase component of speech with the glottal source ([34], [35], [36], [11]) and its importance on maintaining the perceived quality of speech ([12], [37], [38]). Therefore, aiming on describing only the glottal source shape, all factors should be eliminated apart from θ_k. The linear phase shift [12] creates an undesirable phase wrapping which prevents the disclosure of the phase structure. However, in [12] the notion of the center of gravity or relative phase shift (RPS, [37]) has been introduced as an attempt to remove the linear phase mismatch which is attributed to the excitation phase and reveal the phase structure in speech, leading to its successful incorporation in various applications ([39], [17]). Using the definition of RPS and applying it on ϕ_k the linear phase is discarded. The RPS is defined as:

$$RPS_k^i = \phi_k^i - k\phi_0^i \tag{2.5}$$

Fig. 2.8 shows the reveal of this phase structure after removing the linear phase shift. In Fig. 2.8(b) we can see that the instantaneous phase seems to have random values. However, for the voiced segments it can be observed in Fig. 2.8(a) that relative phase shift has a structure. In voiced segments the phase movement is stable through the frequencies (harmonics). The linear phase component in the instantaneous phase (Eq. 2.4) is responsible for the randomness that appeared in the instantaneous phase spectrogram in Fig. 2.8(b). After the subtraction of this term using the RPS the structure of the phase is revealed (Eq. 2.5). Based on this observation and on the fact that the phase variations of the unvoiced segments are high, we can quantify the voice quality. This phase information can be useful for voice pathology detection where the sustained phonation is problematic, in order to capture the voice abnormalities which are connected to the glottal source. Therefore RPS seems a useful tool for voice pathology detection.

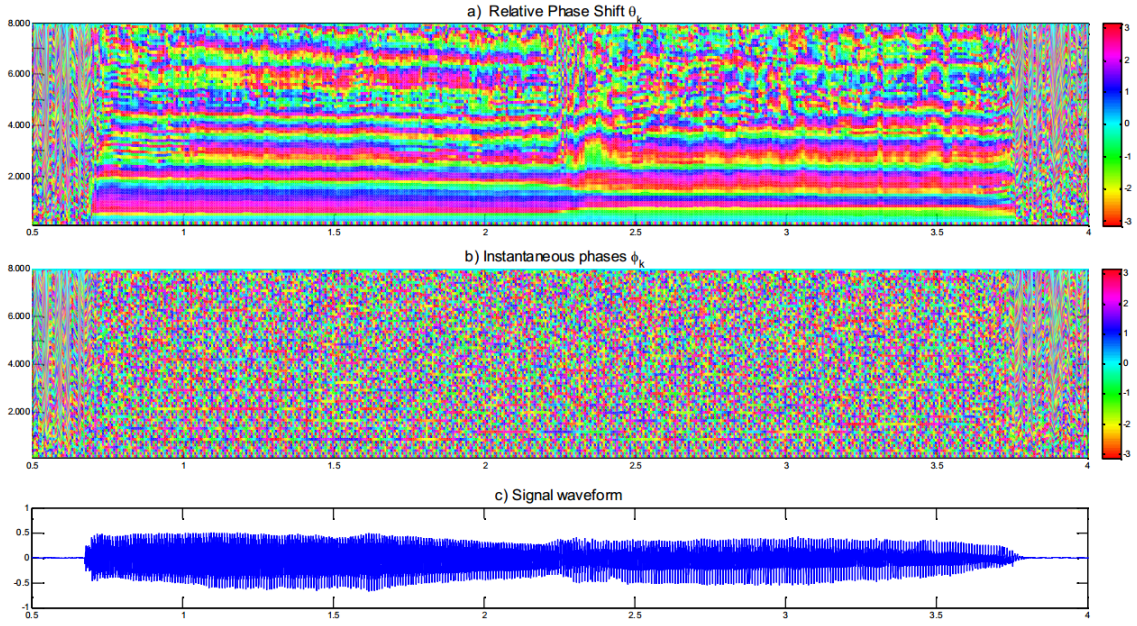


Figure 2.8: Phasegrams of voiced speech segment /ea/ sampled at 16kHz (a) Relative phase shift (θ_k) (b) Instantaneous phases (ϕ_k) (c) Signal waveform

Chapter 3

Extracting the phase structure of the glottal source from speech using adaptive Harmonic analysis

Our proposed method for detection and quantification of voice pathologies is based on extraction of the phase spectrum of the glottal signal using the instantaneous phases computed on the speech signal by an adaptive Harmonic Model. This method can capture the regularity of the glottal source signal without its accurate estimation and without the utilization of complex techniques like inverse filtering.

At the beginning of this Chapter the two models are presented. The adaptive Harmonic model for modelling speech (mix-phase component) and the mixed-phase model for modelling the maximum and minimum phase components of speech. aHM computes the instantaneous phases of the mixed-phase component. Then, the idea of the notion of Center of Gravity or the Relative Phase Shift is applied on the mixed-phase model to estimate the instantaneous phases of the maximum-phase component, the glottal signal. Then, to remove harmonic dependences the idea of Phase Distortion [14] is incorporated. Applying PD on pathplogical samples, a new feature the Phase Distortion Deviation (PDD) is introduced. PDD is a way to quantify in 2-Dimensions the variations that are obtained by the PD. This feature makes easy the separation between dysphonic and normophonic speakers as it will be shown in the following Section's figures.

3.1 The Phase Distortion

The voice quality assessment method which is proposed in this thesis is based on phase and more specifically on the variations of the phase in time and in frequency. Due to this, in this Section the structure of the phase will be fully explained as well as how the phase of the glottal source signal arises from the speech signal's phase.
Voice Models are used for matching the perceptual properties of the speech signal to meaningful parameters on which we are able to process. Some well known, voice model-examples are the STRAIGHT [40] and the Sinusoidal Model (SM) [41]. In this thesis the voice model that is used for estimating the useful parameters is the aHM [21]. The abbreviation aHM denotes the full-band **a**daptive **H**armonic **M**odel. This model uses properties of a previous model named adaptive Quasi Harmonic Model (aQHM) [42]. It also demonstrates voiced and unvoiced segments uniformly, without the necessity of using methods for speech separation, whereas in aQHM such methods were needed. In aHM in order to estimate the components in higher frequencies (up to Nyquist frequency) the **A**daptive **I**terative **R**efinement is used (the algorithm is shown in the Appendix). For estimating the unvoiced segments, the condition is that the distance between the two anchors is short enough. In this case $20ms$ is used and the anchor instants are generated by the f_0 of the unvoiced segments. This means that there is no need to use a VUF (Voiced Unvoiced Frequency) point to separate in two bands the frequencies, the harmonic part and the noisy part (and in the reconstruction step there is no need to insert random noise). The used model is a more accurate and simple way for extracting the components. The most significant difference between aHM-AIR and aQHM is that aHM-AIR is a homogeneous model across frequency and time. Also, using aHM, the reconstructed signal is nearly indistinguishable from the original. Another advantage of aHM is that it performs better than previous models as far as the accuracy of sinusoidal parameters is concerned.

More practically, the aHM assumes that the fundamental frequency's functions of the signal are integer multiples of f_0 (harmonics, $f_0, 2f_0$ etc.). Given this situation, the sinusoidal parameters (phase, amplitude) are calculated at the given frequencies. If the model of speech was completely harmonic the calculated phases in each harmonic would be an integer multiply of ϕ_0 (the phase of the first harmonic). However, the model of speech is not perfect harmonic, so there is a phase difference from the ideal case $\phi_k - k\phi_0$. The analytic speech signal $s(t)$ is represented by the aHM in an analysis window of 4 pitch periods (Eq. 3.1). The parameters of the model were computed by the Least Squares (LS) method which is appropriate for the unvoiced segments

as well.

$$s^i(t) = \sum_{k=1}^{k^i} a_k^i e^{j(k\phi_0(t)+\phi_k^i)} \tag{3.1}$$

From this cumulative representation of the signal, the instantaneous amplitudes a_k and the instantaneous phases ϕ_k of each harmonic k, were extracted.
The variable i referred to the frame index, $K^i = \lfloor \frac{f_s}{2f_0(t^i)} \rfloor$, the f_s is the sampling frequency, $\phi_0(t)$ is the integral of $f_0(t)$:

$$\phi_0(t) = \int_{t_c^i}^{t^i} \omega_0(\tau) d\tau \qquad\qquad \omega_0(t) = f_0(t) 2\pi / f_s \tag{3.2}$$

and ϕ_k^i is the instantaneous phase.
The phase is directly connected with the perceptual characteristics of the speech signal. Below there is a step by step analysis on how the glottal phase was separated by the instantaneous phases of the output speech signal. Firstly, the sinusoidal parameters have been estimated. The fundamental frequency curve $f_0(t)$ is computed by STRAIGHT method (Speech Transformation and Representation by Adaptive Interpolation of weiGHTed spectrogram), which is a state of the art method for modifying the fundamental frequency. It is based on a simple channel VOCODER. It decomposes input speech signals into source parameters and spectral parameters. The Eq. 3.3 is a representation of the instantaneous phase of the output speech signal. This phase model from [33] is used in order to distinguish the maximum phase component of speech, which corresponds to the glottal source signal.

$$\phi_k^i = \theta_k^i + k \int_{t_c^i}^{t^i} \omega_0(\tau) d\tau + \angle V^i(k\omega_0(t_i)) \tag{3.3}$$

In Eq. 3.3 the instantaneous phase ϕ_k^i is shown as a summation of the source shape term θ_k^i (phase of glottal pulses), the linear phase is imposed by the delay of the center of analysis window t_c and the position of the glottal pulse and the last term of the sum is the vocal tract filter response $\angle V(k\omega)$ which is assumed to be of minimum phase. The $\angle V(k\omega)$ term has to be eliminated because the purpose is to describe the maximum phase component that is connected to the glottal source. To remove the influence of the vocal tract, the amplitude spectral envelope is firstly estimated through linear interpolation across frequency of the amplitude parameters a_k of the harmonic model in Eq. 3.1. Then, the minimum phase response corresponding to this amplitude spectral envelope is estimated through cepstrum liftering [43]. In Fig. 3.1 the complex cepstrum determination of the vocal tract (VT) $\hat{h}[n]$ is depicted. For calculating the phase of the amplitude of the VT, the Fourier transform of $\hat{h}[n]$ is made, $\hat{H}(e^{j\omega})$ and the imaginary part

is extracted (Eq. 3.4). The produced phase is an unwrapped phase function.

$$\angle H(e^{j\omega}) = \Im\{\hat{H}(e^{j\omega})\} \tag{3.4}$$

And finally, the term of the vocal tract phase is subtracted by the instantaneous phase ϕ_k using

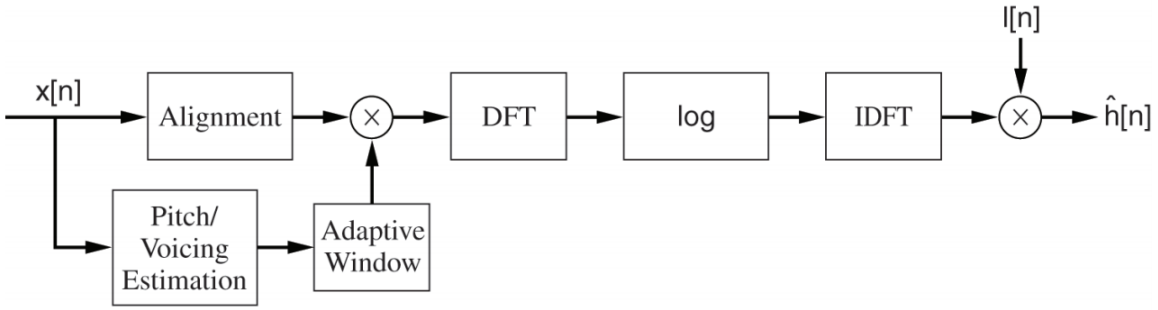


Figure 3.1: Determination of the complex cepstrum for minimum-phase signal, where $x[n]$ is the original waveform, $l[n]$ is the cepstral lifter and $\hat{h}[h]$ is the estimated cepstral representation of the vocal tract impulse response

subtraction in log-spectral domain (i.e deconvolution in the time domain).
So the Eq. 3.3 after the subtraction will become:

$$\widetilde{\phi_k^i} = \theta_k^i + k \int_{t_c^i}^{t^i} \omega_0(\tau) d\tau \tag{3.5}$$

The aim of processing the instantaneous phase is to raise the structure of the glottal that lies in it. Thus, the linear phase term has to be discarded as well.
To arise the phase structure, a method for transforming the instantaneous phases into initial phase shift differences (Relative Phase Shifts) is used, after the subtraction of the VT, $\widetilde{RPS_k^i}$. The instantaneous phase $\phi_k(t)$ extracted from aHM, for the different analysis instants and for being time synchronous with the signal, the center of gravity technique [12] was used. As a center of gravity in a voiced speech frame it is assumed to be the point of the GCI due to the high energy around this point. The frame synchronization applied, is that the center of gravity of each frame is moved to the center of each frame window.
The phase wrapping attributed to the linear phase term $k \int_{t_c^i}^{t^i} \omega_0(\tau) d\tau$ of the instantaneous phase ϕ_k^i from one t^i to the next t^{i+1}. The phase wrapping arises from the movement of the analysis window. The signals contained in the analysis' windows didn't have the same period T. They are time/pitch asynchronous. The calculated instantaneous phases do not have the same values as those they would have if the same period was included in every window. This phase difference

$\Delta\phi$ is illustrated in Fig. 3.2. The $t1, t2, t3$ correspond to the centres of the analysis window, with T_0 the period of each sinusoidal and with $\phi1, \phi2, \phi3$ the instantaneous phases computed by aHM. Thus the linear phase term $k\int_{t_c^i}^{t^i}\omega_0(\tau)d\tau$ is created in the Eq. 3.3. This is happening normally for both dysphonic and normophonic speakers.

The RPS metric has been suggested in [12],[37] for discarding the linear phase component. Due to the reason that we want to capture only the glottal source characteristics, first the influence of the vocal tract is subtracted and then the RPS is used $\widetilde{RPS}_k^i$. As it can be shown in Eq. 3.6 applying RPS on the instantaneous phases $\widetilde{\phi_k^i}$ the linear phase component is removed.

$$\widetilde{RPS}_k^i = \widetilde{\phi_k^i} - k\widetilde{\phi_0^i} = \theta_k^i + k\int_{t_c^i}^{t_i}\omega_0(\tau)d\tau - k(\theta_1^i + \int_{t_c^i}^{t_i}\omega_0(\tau)d\tau) = \theta_k^i - k\theta_1^i \tag{3.6}$$

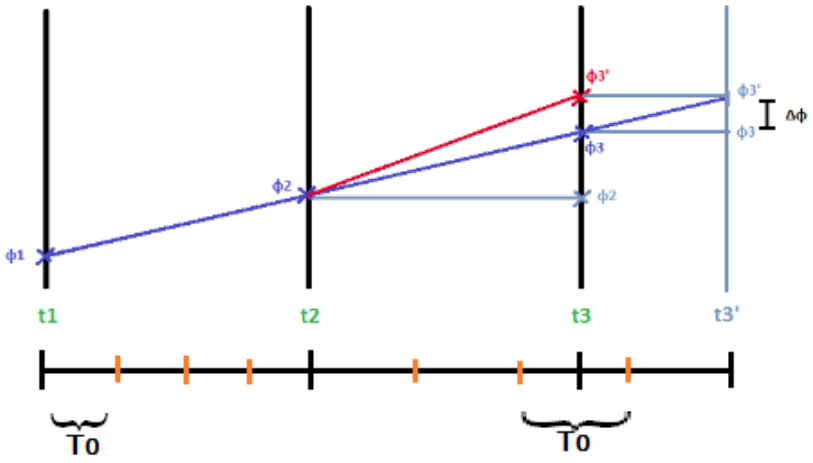


Figure 3.2: Phase Difference

A further step to complete the revealing of the structure of the phase, is to remove the dependency with the harmonic component k, because the variance of RPS will increase for higher frequencies. The resulting equation is the Phase Distortion (PD) [14]:

$$PD_k^i = \underset{k}{\Delta}\widetilde{RPS_k^i} = (\widetilde{\phi}_{k+1}^i - (k+1)\widetilde{\phi}_1^i) - (\widetilde{\phi}_k^i - k\widetilde{\phi}_1^i) = \theta_{k+1}^i - \theta_k^i - \theta_1^i \tag{3.7}$$

The PD represents the system distortion. Features connected to the glottal source can be extracted using the PD without estimating the glottal source.

In Fig. 2.8(b) the instantaneous phase spectrogram of two vowels is illustrated and in Fig. 2.8(a) we can see the corresponding RPS spectrum. In Fig. 2.8(a) the structure of the glottal phase can

easily be distinguished, in contrast to Fig. 2.8(b) (where the structure cannot be easily pointed). In Fig. 3.3 the phase structure is also visible as well as the variations in the glottal source signal.

3.1.1 Phase Distortion for glottal model estimation

Traditional methods for estimating the glottal source parameters use Iterative Adaptive Inverse Filtering and minimum/maximum phase decomposition methods (Complex Cepstrum (CC) and Zero of the Z-Transform (ZZT based methods)). As mentioned in the previous Section, the method that is used in this thesis for detecting voice pathologies doesn't need the estimation glottal or vocal tract characteristics. What is needed is the glottal source shape. This method is the Phase Distortion method that is used for extracting features. The proposed method is much simpler than the others because the features that are extracted by this method are produced only by manipulating the instantaneous phases of the output signal of speech which are computed by aHM. The Eq. 3.7 shows the definition of the PD and its relation to the shape of the glottal source. Phase Distortion is the finite difference of RPS, as it was shown in the previous Section. PD is proposed in [14] and was used for resynthesis. In our case it is introduced for removing the dependency of the harmonic index k in the obtained phase. This is necessary because otherwise the variance of the RPS will be increased in higher frequencies. The PD can also characterize the glottal source without depending on: the duration of the glottal pulse, its excitation amplitude, its position as well as the position of the analysis window and the influence of the minimum phase component of the speech signal (vocal tract filter).
Due to the relation of the PD with the shape of the pulses of the glottal source and the significant information that is produced for the voice variations, PD is an ideal indicator for voice pathology detection. As could be observed from the visualization of the PD, the PD spectrum of the normophonic speaker (Fig. 3.3(a)) appears stable, with stable values across the time and across the harmonics. Unlike in Fig. 3.3(b) the PD spectrum of the dysphonic speaker, varies across the time and across the harmonics it seems to be noisier. So, the variance of the PD is a better voice regularity descriptor than PD itself. In the case of a normophonic speaker, for a sustained phonation the variance of the PD should be close to zero as the shape of the glottal pulse is preserved in time. In the next Section the PDD feature, which is the variance of the PD, is introduced.

3.2 Phase Distortion Deviation: detecting voice irregularities

The proposed feature in this thesis is the Phase Distortion Deviation (PDD). It is obtained from the PD by making its standard deviation, as could be shown in Eq. 3.10. For the analysis-

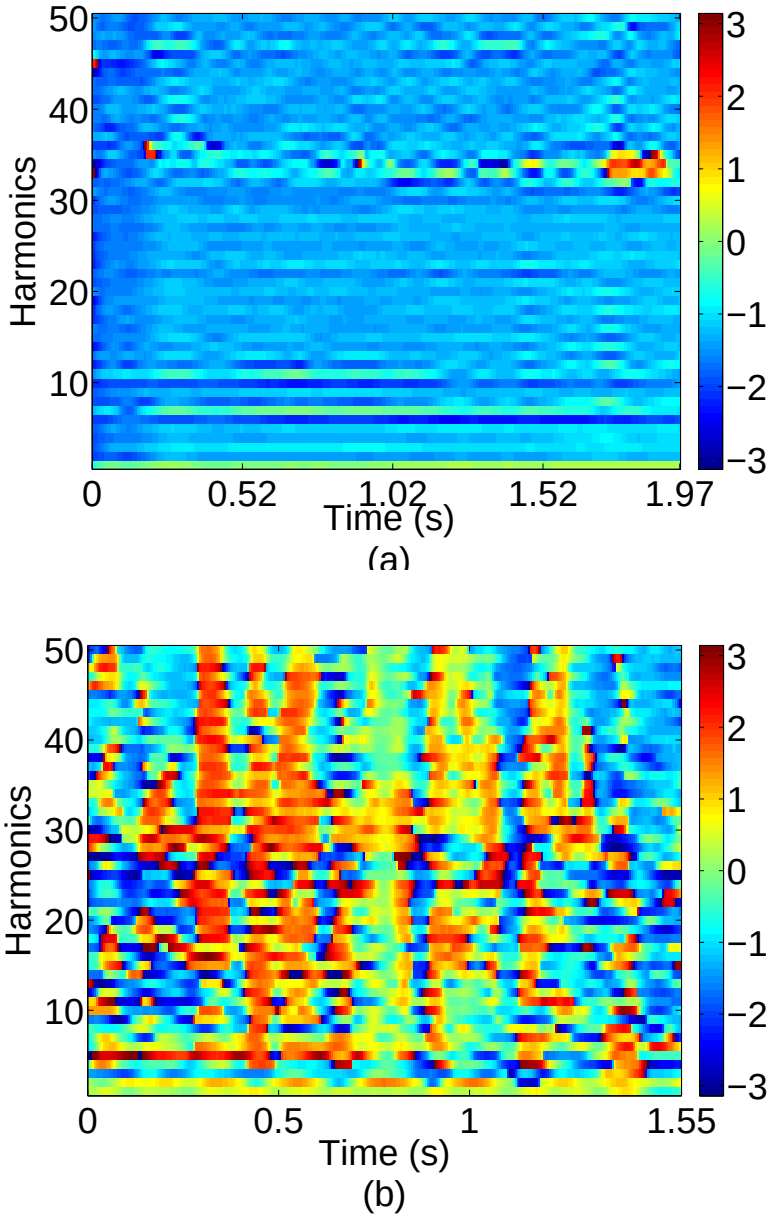


Figure 3.3: PD spectrograms of a normophonic male speaker (Fig. 3.3(a)) and a dysphonic male speaker (Fig. 3.3(b)). Both speakers utter the sustained vowel /a/.

synthesis systems, the PDD is sufficient to carry the entire information for the reconstruction of the signal [44]. For our purpose PDD is used as a feature for measuring the variations of the voice.

In the statistical analysis point of view, the data $PD_i(\omega)$ (phases/angles) that are used can be represented as complex numbers of unit magnitude $z = cos(\theta)+isin(\theta) = e^{i\theta}$, where $\theta = PD_i(\omega)$

is the measured angle. Each angle represents a vector of length one in the direction of the angle. The resultant length is denoted with an R [1] and $\overline{R}$ is the mean resultant length. $\overline{R}$ varies between zero and one, so if a value of R is close to one that means that there was little variation in the values of the angles.

The circular standard deviation $\sigma^i_{PD_k}$, or else the PDD value is defined as $v = \sqrt{ln(\frac{1}{\overline{R}}^2)} = \sqrt{-2ln(\overline{R})}$ and $\overline{R} = \frac{1}{N}\sum_{n=1}^{N} z_n$

$$v = \sqrt{ln(\frac{1}{\overline{R}}^2)} = \sqrt{-2ln(\overline{R})} \tag{3.8}$$

$$\overline{R} = \frac{1}{N}\sum_{n=1}^{N} z_n \tag{3.9}$$

Combining the foregoing data the final equation is formed as following:

$$PDD = \sigma^i_{PD_k} = std_i(PD^i_k) = \sqrt{-2ln|\frac{1}{M}\sum_{m\in B_i} e^{jPD^m_k}|} \tag{3.10}$$

The circular standard deviation takes values between zero and infinity. This means that the PDD values of normophonic speakers are low in contrast to these of dysphonic speakers. This definition of the standard deviation is proposed by Fisher [45]. An advantage of this is that for small values of the standard deviation, the circular distribution is standardized, as in the linear case. It is also useful because for a wrapped normal distribution, standard deviation is an estimator of the standard deviation of the underlying normal distribution.

In Fig. 3.4(b) and in Fig. 3.4(a) the PDD spectrogram of a dysphonic and a normophonic speaker are illustrated, respectively, using a small/fixed window of $100ms$ duration. Comparing them to the figures of the PD we can observe that Fig. 3.4(b) and Fig. 3.4(a) are more informative as it shows the time characteristics of the PD for each harmonic. Also, normophonic speakers have PDD which is close to zero, revealing the correlation of PDD with the deviation of the glottal shape in time. Another important point is that using a small window for calculating the PDD the noisiness effect is not very clear in the case of the dysphonic speaker. The small window leads to an underestimation of the standard deviation of the PD. Dysphonic speakers in some time instants have high PD values (peaks) and if we use a big window that high variability in time domain of PD is captured. In order to capture such changes in the entire phoneme the maximum possible window is used, for calculating the PDD (length M in Eq. 3.10). The speech

[1] *Resultant length is the length of a vector that is stretched from the first vector in the origin to the last vector in the end.*

recordings have variable lengths so the window is selected to be proportional to the duration of the speech signal, that is 1/3 of the duration of the signal. This window length is the maximum possible, since there is an order restriction imposed by our zero-phase moving average filter that implements the summation in Eq. 3.10.

The Fig. 3.5(a) and Fig. 3.5(b), that depicts the PDD spectrums of a normophonic and a dysphonic speaker, respectively, show high values if the signal varies in time and in frequency. In these PDD spectrums, it could be observed that in the normophonic speaker's spectrogram the colors are cool hues (values close to zero) which means that normophonics have low variations in their Phase Distortion spectrogram showing that their voice is stable/normal. In contrast, in the dysphonic speaker's spectrogram (Fig. 3.5(b)) the colors are warm hues, which means that the PDD takes high values, thus the Phase Distortion varies a lot and this proves that their voice varies/abnormal.

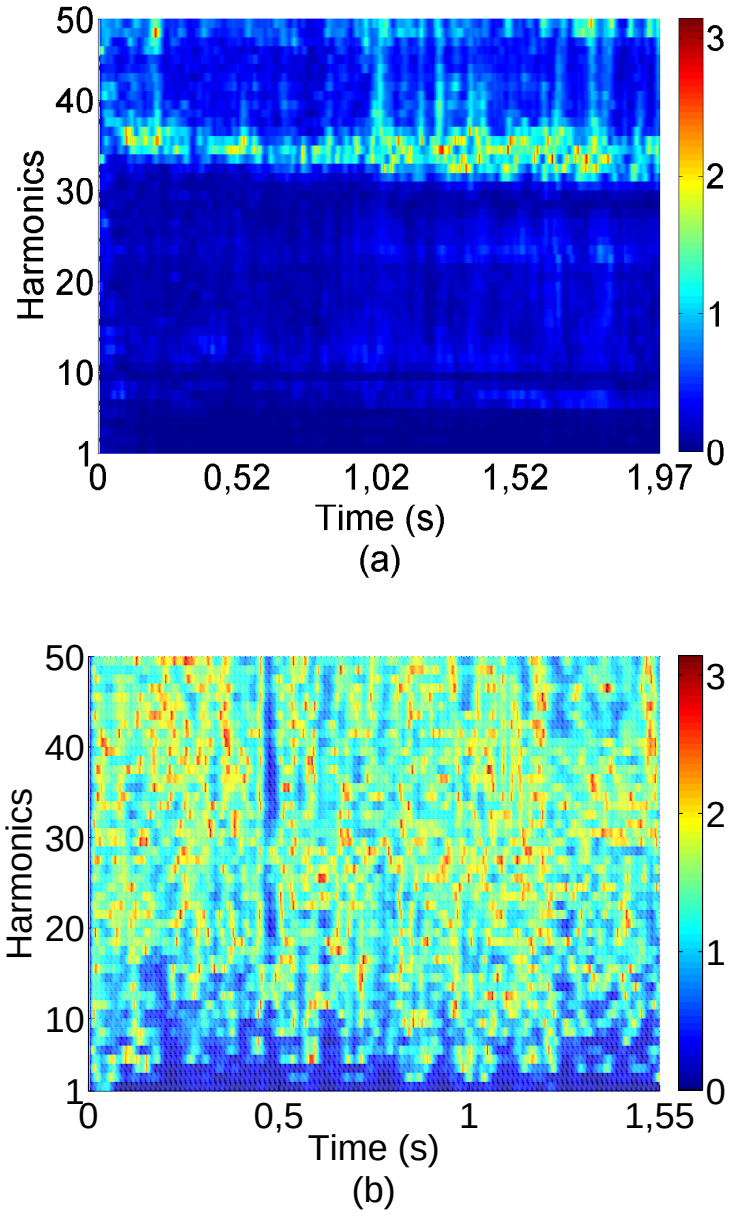


Figure 3.4: PDD spectrograms of a normophonic male speaker (Fig. 3.4(a)) and a dysphonic male speaker (Fig. 3.4(b)) using $100ms$ window length. Both speakers utter the sustained vowel /a/.

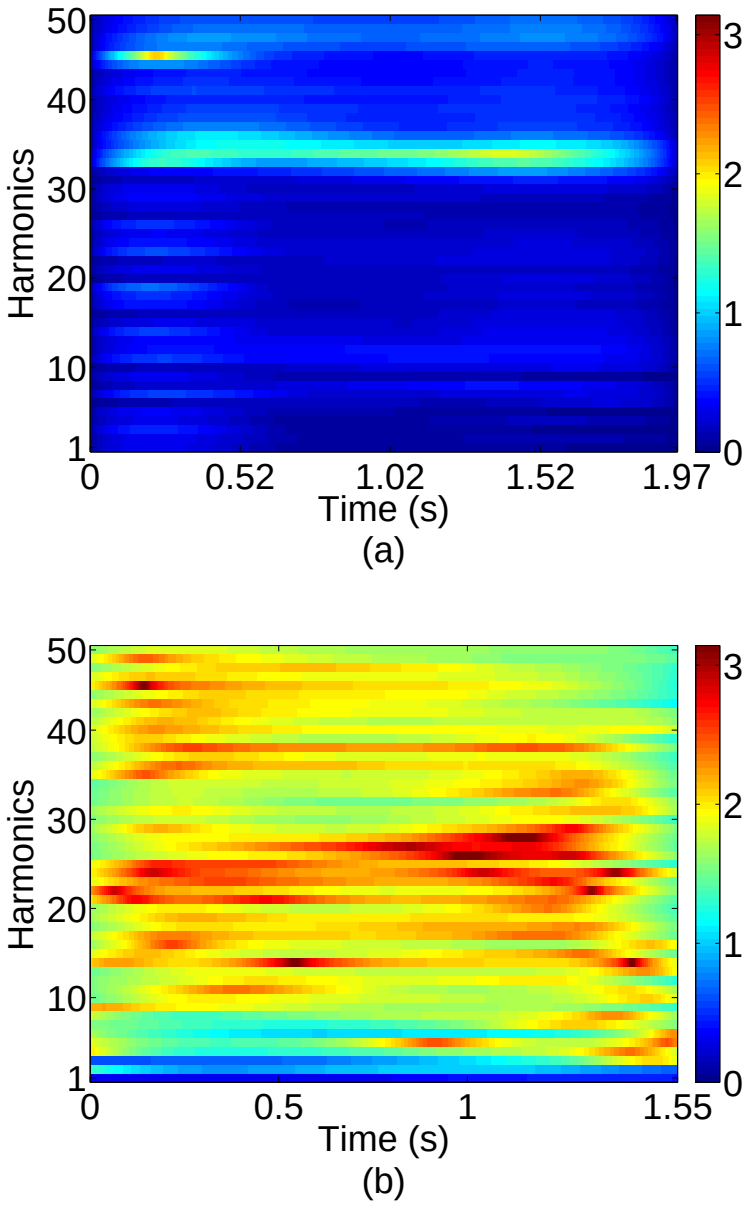


Figure 3.5: PDD spectrograms of a normophonic male speaker (Fig. 3.5(a)) and a dysphonic male speaker (Fig. 3.5(b)) using the maximum possible window length. Both speakers utter the sustained vowel /a/.

Chapter 4

The Regularity Ratio: index of normophonicity

In the previous Chapter a new method for extracting phase-based features of the glottal source is introduced. In this Chapter the proposed method is evaluated on a database consisting of dysphonic speakers who suffer from spasmodic dysphonia and is compared to other objective and subjective metrics. Prior to the evaluation, the disease of SD is thoroughly analysed and then a new metric for quantifying the severity of the dysphonia is proposed. Using this new metric the speakers was ranked according to their degree of dysphonia. This metric can also distinguish normophonic from dyspnonic speakers.

4.1 The disease of Spasmodic Dysphonia

The proposed metric based on phase information is applied to dysphonic speakers who suffer from Spasmodic Dysphonia. In the following paragraphs the SD disease is thoroughly analysed. The data provided below was derived from the [46], a website from the National Spasmodic Dysphonia Association.

Meaning of SD

Spasmodic dysphonia belongs to a family of neurological disorders, called dystonias. A dystonia is a movement disorder that causes muscles to contract and spasm involuntarily. Dystonias can be generalized, affecting the entire body, or focal, affecting only a specific area of the body or group of muscles. Following Parkinson's disease and essential tremor, dystonia is the third most

common movement disorder.

Certain dystonias including SD are task-specific, meaning that the muscles spasm only when they are used for particular actions and not while they are at rest. When a person with SD attempts to speak, involuntary spasms in the tiny muscles of the larynx cause the voice to break up, or sound strained, tight, strangled, breathy or whispery. Improper functioning of any of these muscles results in decreased vocal quality. The spasms often interrupt the sound, squeezing the voice to nothing in the middle of a sentence or dropping it to a whisper. However during other activities such as breathing and swallowing, the larynx functions normally. SD can cause intermittent spasms in any of the muscles that position the vocal cords. The affected muscles determine the exhibited symptom and ultimately the type of SD.

SD Causes

The exact cause of SD remains unknown. A reason for this difficulty in finding what causes SD, is that the spasms do not occur in all types of speech. SD can not be found to unvoiced sounds like "f", "s" where the vocal cords do not require to vibrate. But we can note it in voiced sounds like /a/, /e/ etc, where the vocal cords cause their vibration. In normal voices this vibration is periodic. However, in SD this periodic structure is lost. Therefore, sustained vowels are the most appropriate for voice pathology studies.

Evidence suggests that the problem starts at the base of the brain in the basal ganglia, which regulate involuntary muscle movements. In the SD condition, this nervous system regulator does not function properly and produces incorrect signals, which cause the muscles to contract or relax more than they should or at the wrong time. Genetic factors may put some people at greater risk of developing spasmodic dysphonia, particularly those who have family members with any form of dystonia.

Symptoms of SD

Studies have shown that the symptoms of SD improve or disappear during laughing, crying, yelling, throat clearing, coughing, whispering, and humming. Generally, SD does not affect the emotional aspects of speech. SD tends to affect only normal conversational speech.

SD can start at any time during life, but seems to begin more often when people are middle-

aged. The disorder affects women more often than men. Symptoms usually occur in the absence of any structural abnormality of the larynx, such as nodules, polyps, carcinogens, or inflammation.

Diagnosis of SD

SD has no objective pathology that is evident through x-rays or imaging studies like a CT or MRI scan, nor can a blood test reveal any particular fault. Therefore the best way to diagnose the problem is to find an experienced clinical with a good ear.

Also, there are physical examinations such as looking at the larynx in action, by inserting a rigid endoscope, a straight, narrow medal rod containing a camera through the mouth and toward the back of the throat while the person is saying "eeeee". Another approach to viewing vocal folds involves the use of a flexible endoscope: a very narrow, flexible tube is inserted through one nostril and down through the throat, which allows the doctor to evaluate the movements of the larynx while the person is speaking or singing. But all these methods are painful and not very enjoyable. For these reasons we propose a method to quantify the quality of the voice so that to make the diagnose easier and without any pain. This could help both clinical doctors and patients.

Therapy of SD

SD therapy offered by a speech language pathologist (SPL) involves training the person to alter voicing techniques. For instance, the speech therapist may point out that the patient is producing his or her voice with poor breath support or poor tongue placement in the mouth. Through exercises and practice, the patient can gain better insight into how to speak more efficiently and effectively.

Unfortunately, this approach often produces incremental benefit for the typical SD patient, since SD is a neurologic condition over which the patient has little or no control. Speech therapy is generally seen as a possibly helpful adjunct to other therapies such as botulinum toxin (BTX) injection. It is shown that it helps SD patients who have excess voice strain to "unload" some vocal muscle tension.

In our experiments voiced segments of patients before and after botulinum toxin injection have been used. The databases used for this purpose was provided by Pr. Dejonckere [11] as well as the subjective evaluations of the dysphonic speakers. These databases will be analysed

in more detail in the following Sections.

4.2 Database of normophonic speakers

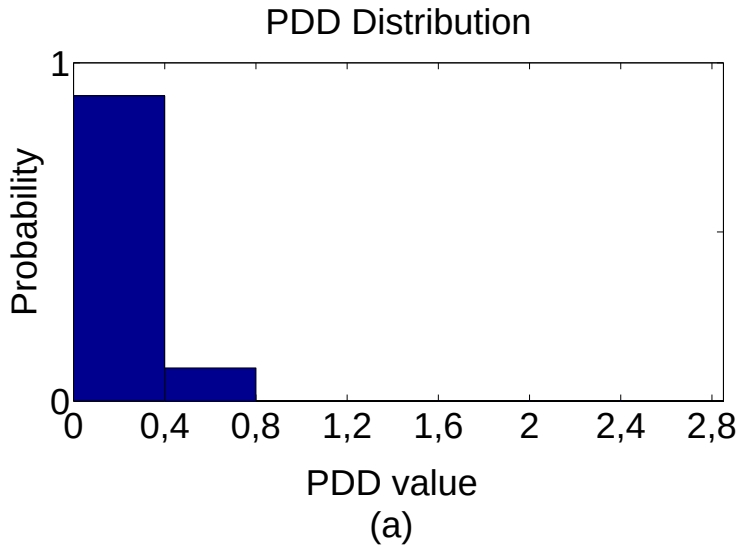
The proposed method is applied on a database of normophonic speakers. The database consists of eleven male and five female healthy subjects, whose age varies between 23 and 45 years old. The recorded sustained vowel was */a/* at 48kHz and then downsampled at 16kHz . The duration of the sustained vowels varies from 2sec to 8*sec* depending on the speaker.

Due to the difficulty of comparing 2-D data, such as PDD data, the histogram of the PDD is used. A histogram is a graphical representation of the distribution of the data. The histogram in the x-axis depicts the different values from the PDD table and in the y-axis the quantity of their occurrence. In Fig. 4.1(a) a normalized histogram is shown, which depicts an averaged histogram of all the normophonic speakers from the database. From the figure it is observed that almost all the values which appear in the normophonic speakers PDD spectrograms vary up to 0.4 and very few exceed this value (concentrated near zero). Due to this observation the area under the value 0.4 is called *normophonic area*.

4.3 Database of dysphonic speakers

The main application of the proposed method in this thesis is implementing the method on pathological voices. The aim is to rank the speakers according to the degree of severity.

For this purpose, a database which consists of speech signals of twenty speakers who suffer from Spasmodic Dysphonia before and after the botulinum toxin injections is used. The sixteen speakers of the database were male and the four speakers were female. For every speaker the sustained vowel of */a/* is used as input in our method.
In Fig. 4.1(b) a histogram of a dysphonic speaker is illustrated. It can be observed that the PDD values are in a great span all over the histogram and only very few values are close to 0. Respectively with the normophonic area, the area over 0.4 is called *dysphonic area*. Considering the above remarks, the value 0.4 is used as a threshold for separating normophonic from dysphonic speakers.



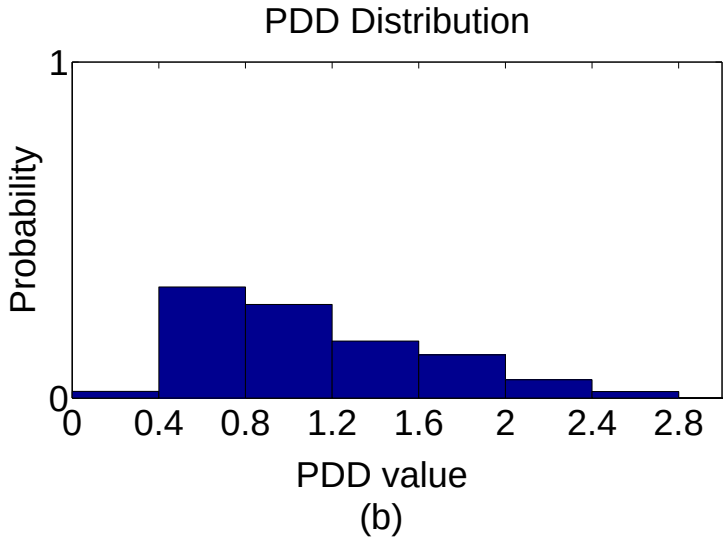


Figure 4.1: Fig. 4.1(a) shows an average histogram of the PDD for the 16 normophonic speakers. Fig. 4.1(b) shows a histogram of PDD of a dysphonic speaker.

4.4 Regularity Ratio metric

Regularity Retio (RR) is the new metric which is proposed in this thesis as well. RR quantifies the normophonicity, how close is the distribution of a histogram to 0. The type of RR is shown in Eq. 4.1, where i is the number of bin, with P denoted the bars probability in the histogram, P_1 is the bar that belongs to the normophonic area and the denominator is the summation of

the rest of the histogram (the summation of the bars which belong to the dysphonic area). If the nominator is grater than the denominator, the RR will be positive, otherwise RR takes negative values(that means that PD varies rapidly). The histogram's bar length is chosen to be 0.4 because we observed that for the majority of the normophonic speakers, when the bar's length was greater than 0.4 the values of the PDD were all in the first bar and so the denominator of the RR metric was 0 and the value of the RR was infinite. Due to this, the 0.4 value is chosen as the maximum possible value for the length of the bar.

$$RR = \log_{10}\left(\frac{P_1}{\sum_{i=2}^{\infty} P_t}\right) \quad , \qquad 1 \leq i < \infty \tag{4.1}$$

Using RR metric speakers can be ranked according to the severity of their disease. The ranking of the database that is used for the experiments together with the RR values are shown in the Table 4.1. The higher the RR value is, the more normophonic the speaker is.

	Subjective Classification OVERALL SEVERITY		**Objective Classification RR**
Knipos	**1.00**	**Roopre**	**-0.7263**
Heupre	**1.00**	**Burpos**	**-0.6527**
Burpre	**0.9231**	**Heupre**	**-0.6160**
Burpos	**0.9231**	**Knipos**	**-0.5777**
Vropre	**0.7692**	**Roopos**	**-0.5476**
Roopos	**0.6923**	**Lulpre**	**-0.4542**
Lulpre	**0.6923**	**Vropre**	**-0.3993**
Roopre	**0.6154**	**Burpre**	**-0.3926**
Stupos	**0.6154**	**Stupre**	**-0.3598**
Stupre	**0.5385**	**Vropos**	**-0.2503**
Vropos	**0.5385**	**Stupos**	**-0.2038**
Plupos	**0.5385**	**Plupos**	**-0.0859**
Esspos	**0.4615**	**Esspos**	**-0.0165**
Heupos	**0.1538**	**Heupos**	**-0.0098**
Lulpos	**0.1538**	**Lulpos**	**1.00**

Table 4.1: Objective evaluation of the dysphonic speakers using the RR metric

4.5 Performance evaluation

The RR is compared with subjective metrics evaluated by medical doctors (Jitter, Tremor and Overall Severity) and an objective metric based on tremor, WMTV.

4.5.1 Subjective metrics

In Chapter 2, some metrics which detect the voice pathologies using the Phase Spectrum or the Amplitude Spectrum were analysed. In this Section are discussed: an objective metric (WMTV) and three subjective metrics (tremor, jitter, overall severity). The proposed method will be compared to them in a next Section.

Jitter is the cycle-to-cycle period perturbation. It is evaluated objectively by [1] using the analysis program Ampex (Auditory Model Based Pitch Extractor) [47]. This program is able to extract the period in irregular signals with background noise and it is efficiently used for substitution voices and for SD. The speech samples are ranked from high to low jitter. The formula that is used for the jitter is Eq. 4.2

$$Jitter = sum\ of\ VE(i) * |T_0(i) - T_0(i-1)| / sum\ of\ VE(i) * T_0(i-1) \tag{4.2}$$

where $T_0 = 1/F_0$, VE = voicing evidence and F_0 = fundamental frequency.

Tremor is the rhythmic change in pitch and loudness. It is subjectively estimated by medical doctors and the speech samples are ranked from high to low tremor value.

Overall Severity of Spasmodic Dysphonia is evaluated by doctors.

4.5.2 Objective metrics

WMTV is the acronyms of the **W**eighted **M**ean **T**remor **V**alue. This is an objective tremor metric and is proposed in [23]. This metric quantifies the severity of spasmodic dysphonia and is based on tremor estimation on the speech signal. The vocal tremor consists of the involuntary modulations of the frequency and/or the amplitude in sustained phonation. The pathological tremor can be caused by diseases like Parkinson and essential tremor. There is also physiological tremor, the natural stochastic modulations in the interval of [2,25]Hz with low amplitude. The

acoustic vocal tremor attributes are how fast (modulation frequency) and how strong (modulation level) are the modulations. This method (WMTV) uses an AM-FM decomposition algorithm based on the adaptive time-varying quasi-harmonic model. The modulation frequency $\frac{1}{2\pi}\frac{d\psi(t)}{dt}$ and the modulation level, $m(t)$, are used for calculating the metric. These parameters are estimated by the FM and AM component, respectively.
Also some of the objective metrics (**HNR, HRF, H1H2**) and the Jitter metric which were described in Chapter 2 are compared with the subjective overall severity of SD metric. The Jitter objective metric is computed by using the Praat [1].

4.5.3 Correlations and Evaluation Results

The doctor's evaluation for the patients of our database (subjective evaluation) and the corresponding classification using the WMTV method (objective evaluation) are shown in the following Table 4.2. The patients have been ranked before and after the surgical procedure. In the end of the name of each one in the table there is a syllable *pre* or *pos*: *pre* means that the value is before the surgery and *pos* is the value after the surgery. The values in the Table has been normalized to 1.

To measure the correlation between the proposed method with the above metrics, Spearman and Pearson correlation method was used. Correlation measures the strength of association between two variables. The Pearson correlation method is the most widely used. It measures the strength of the linear relationship between normally distributed variables. The Pearson correlation between A and B is shown in Eq. 4.3.

$$\rho_{A,B} = \frac{Cov(A,B)}{\sigma_A \sigma_B} \tag{4.3}$$

where A is the first set's values, B the second set's values, the correlation is denoted with a ρ, σ_A, σ_B is the standard deviation of A and B, respectively and $Cov(A,B)$ is the covariance of A and B which are defined in Eq. 4.4

$$Cov(A,B) = \frac{1}{n-1}\sum_{t=1}^{n}(r_a - \overline{r_a})(r_b - \overline{r_b}) \tag{4.4}$$

where r_a, r_b are the return for series A, B in period t_i respectively, $\overline{r_a}, \overline{r_b}$ are the arithmetic mean for A and B and n is the number of periods.

[1] Praat is a free scientific computer software package for the analysis of speech in phonetics.

	Subjective Classification				Objective Classification
	TREMOR	JITTER	OVERALL SEVERITY		WMTV
Burpre	1.00	0.6769	0.9231	Heupre	1.00
Burpos	0.94	0.5335	0.9231	Burpos	0.91
Roopre	0.82	0.7934	0.6154	Burpre	0.85
Stupre	0.71	0.6058	0.5385	Roopos	0.68
Roopos	0.59	0.7235	0.6923	Roopre	0.56
Vropre	0.53	0.8301	0.7692	Lulpre	0.39
Vropos	0.47	0.7588	0.5385	Vropre	0.37
Heupre	0.41	1.00	1.00	Stupre	0.33
Knipos	0.41	0.8265	1.00	Vropos	0.30
Lulpre	0.24	0.8852	0.6923	Esspos	0.24
Plupos	0.12	0.5603	0.5385	Plupos	0.20
Esspos	0.06	0.5065	0.4615	Knipos	0.18
Heupos	0.06	0.5790	0.1538	Stupos	0.18
Lulpos	0.06	0.1663	0.1538	Heupos	0.15
Stupos	0.0	0.5166	0.6154	Lulpos	0.11

Table 4.2: Subjective and Objective evaluation of the dysphonic speakers

The Spearman rank correlation method makes no assumption about the distribution of the data. This method first takes the return data and assigns a rank number to each return. It doesn't matter if the highest return gets the lowest rank or vice versa. If the ranks of the data of the two series are the same there is obviously a very strong relationship between the two series. The Pearson correlation won't necessarily show this because it shows the strength of the linear relationship and the relationship might not be linear. The Pearson correlation is good for measuring the strength of the linear relationship between two normally distributed data sets. The Spearman correlation method is better for data that are not distributed normally. The Spearman is computed from Eq. 4.5.

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)} \tag{4.5}$$

where $d_i = x_i - y_i$ is the difference between ranks. Spearman assesses how well the relationship between two variables can be described using a monotonic function. The perfect Spearman correlation of +1 or -1 occurs when each of the variables is a perfect monotone function of the other. In both tables (Table 4.3 and Table 4.4) there are two values in each cell. The value in the parentheses is the $p - value$ which is the probability of getting a correlation as large as the

observed value by random chance, when the true correlation is zero. If $P(i,j)$ is small, e.g less than 0.05, then the correlation $R(i,j)$ is significant. The other value in the cell is the correlation coefficient.

The Pearson correlation method is applied for comparing the proposed method with the other metrics. This method returns high significance results in the correlation between all the subjective metrics and the proposed metric RR as it can be seen in the Table 4.3 and Table 4.4. This entails two main conclusions: firstly the data of RR set is linear correlated with the subjective metrics (jitter, tremor and overall severity) and secondly the RR metric can capture a variety of voice abnormalities using the PDD-feature. Another statistical method Spearman is applied for making the same comparisons, because Spearman is a more general correlator and gives the monotonicity level. For these reasons it is expected that Spearman will give even better results than Pearson and indeed was what we expected. The results are shown in the Table 4.3 and Table 4.4.

	Jitter (Ampex)		Tremor		Overall Severity	
	S	P	S	P	S	P
RR	-0.65 (0.0082)	-0.78 (0.0006)	-0.68 (0.005)	-0.70 (0.0037)	-0.81 (0.003)	-0.82 (0.0002)
WMTV	0.50 (0.0585)	0.44 (0.0991)	0.75 (0.0012)	0.72 (0.0024)	0.67 (0.0066)	0.68 (0.0053)

Table 4.3: Pearson's (P) and Spearman's (S) correlation coefficient of the ranking of RR and WMTV with the ranking of subjective evaluations (Tremor, Overall severity) and objective evaluations (Jitter) provided by [1]. 15 speakers are ranked.

	Jitter (Ampex)		Tremor		Overall Severity	
	S	P	S	P	S	P
RR	-0.62 (0.0033)	-0.71 (0.0004)	-0.59 (0.0067)	-0.64 (0.0023)	-0.82 (<0.0001)	-0.82 (<0.0001)

Table 4.4: Pearson's (P) and Spearman's (S) correlation coefficient of the ranking of RR with the ranking of subjective evaluations (Tremor, Overall severity) and objective evaluations (Jitter) provided by [1]. All speakers are ranked.

In the above tables (Table 4.3 and Table 4.4) it can be observed that RR metric is negatively correlated to the other metrics. This is actually correct because the more normophonic the speaker is, the higher the RR value gets, but for all the other metrics with which RR is compared, the result is the opposite (the more normophonic the speaker is, the lower the metric value gets). In Table 4.3, the correlations of the subjective metrics with the WMTV metric is positive correlated for the reason that WMTV takes higher values if the severity is high like the subjective metrics.

The difference of the two tables (Table 4.3 and Table 4.4) is that Table 4.3 consists of the metrics correlation but only in 15 of 20 speakers of the dysphonic database. This happens because WMTV depends on the duration of the phoneme, so the signal has to be more than 1.5 *sec* for extracting the features needed in WMTV. The AQHM algorithm uses an autocorrelation method to estimate the first and last frame of the analysed signal. As a result it fails to capture the signal for these frames. Therefore, these frames are omitted from the analysis. However, this imposes a limitation on the duration of the phoneme. Phonemes cannot be shorter than $1.5sec$. Subjects who suffered from spasmodic dysphonia found it hard to speak and thus their vowels had a duration less than 1.5 sec. However this criterion especially for dysphonic speakers is very significant. Because dysphonic speakers cannot hold their voice for a long period of time and so the samples are not enough for WMTV method. The rest 5 speakers of the database have a shorter duration than $2ms$ and they were not used in Table 4.3. One of the advantages of the proposed method is that it is independent of the duration of the speech. For this reason in Table 4.4 the whole database of the dysphonic speakers is used and RR metric is compared again with the subjective metrics. It can be seen that the results of the correlations are a bit better than in Table 4.3. Other observations are that RR outperforms the WMTV metric by 14% on the overall severity, RR is significantly correlated with tremor and overall severity, WMTV is also significantly correlated with them but WMTV is not correlated with jitter whereas RR is. From the correlation between jitter and WMTV, the WMTV metric has $p-value = 0.0585$ (using Spearman) or $p-value = 0.0991$(using Pearson) which means that it is not significantly correlated since $p-value > 0.05$. It can be observed from both of the tables that $p-values$ for the correlation between the RR metric and the subjective metrics are very low. Which means that the proposed method is significantly correlated with the three of the subjective metrics. Also, comparing the RR metric with the WMTV (Table 4.3), the RR has higher correlation values ($r-values$) with all the subjective metrics than the WMTV. This means that RR detects more voice pathologies and more accurately.

Finally, the correlations between the overall severity and the rest of the objective metrics (HRF, H1-H2,HNR and Jitter(Praat)) were made. All of them had high $p-values$ ($p-value > 0.05$) both for Spearman and Pearson correlations which means that the correlation of these objective metrics are not significant. So, these metrics are not ideal to characterize the overall severity of SD. In Table 4.5 the correlations are illustrated. Only the $p-value$ of the Pearson correlation of H1-H2 with the overall severity is significant (0.0327) but the $r-value$ is not high (0.55). For calculating the ranking of the metrics H1-H2 and HRF the implementations of the COVAREP were used [48]. HNR ranking was made using the implementation of the algorithm which described in Chapter 2.

	Overall Severity of SD	
	S	P
HRF	-0.16 (0.5596)	-0.36 (0.1884)
H1-H2	0.48 (0.0709)	0.55 (0.0327)
HNR	-0.25 (0.3613)	-0.38 (0.1614)
Jitter (Praat)	0.25 (0.3755)	0.12 (0.6805)

Table 4.5: Pearson's (P) and Spearman's (S) correlation coefficient of the ranking of HRF,H1-H2,HNR and Jitter from Praat with the ranking of the subjective evaluation of the Overall severity of SD.

Chapter 5

Conclusions and future work

In this thesis we propose a new method for estimating voice pathologies using the phase information of the glottal source. The new metric introduced, called Regularity Ratio, quantifies the severity of voice pathology.

A well-known proverb says, *"Prevention is better than the cure"*. The aim of this thesis is to find an algorithm that early detects voice disorders, so that possible future diseases connected to vocal abnormalities can be prevented.
To sum up, the innovative phase-based method that is introduced, extracts the features from the glottal source signal. Previous methods use more complex techniques for extracting features from the glottal source from the amplitude spectrum of the output/speech signal. Classical computation for glottal excitations estimated from natural speech is often problematic due to formant ripple (i.e the fluctuating component embedded in the glottal excitation due to incomplete cancelling of formants by the inverse filter). As it has been shown in Chapter 2, the amplitude spectrum is not sufficient for capturing the glottal source characteristics which are connected to the voice abnormalities studied. In this thesis, the instantaneous characteristics (amplitude, phase, frequency) are extracted firstly by using an adaptive harmonic model on the speech signal. Then, from the instantaneous phases through mathematical formulas the glottal shape arises. Using this proposed method there is no need to manipulate with complex inverse filtering methods and in addition the problem with the wrapping of the phase is solved by using these formulas. Thereafter, (from the glottal shape where the voice abnormalities are connected), features for measuring the variabilities in speech are created (PDD features). These features without estimating explicitly the glottal waveform, are correlated with the glottal shape and therefore with voice. PDD is free from the linear phase influence and the phase contributions of the vocal tract and therefore can characterise the regularity of the glottal source. The advantage of using the

PDD is that it alleviates the necessity of detecting the Glottal Closure Instants (GCI) and the accurate estimation of the glottal source. From these 2-D PDD features it is difficult to quantify voice disorders and to evaluate our method. For this reason the RR metric is introduced (Chapter 4). The RR metric is extracted from the histograms of the PDD in normophonic speakers and then is evaluated in a database of dysphonic speakers with Spasmodic Dysphonia in a ranking problem of the severity of the pathology. Our metric is compared with subjective and objective rankings using other metrics. PDD-ranking is highly correlated with the subjective ranking from the medical doctors and extracts better results compared to an other objective metric, WMTV.

A future work can be the application of this method to other databases which contain other vocal diseases. It is expected that the results would be satisfactory because in Chapter 4 it was shown that the proposed method is correlated with more than one acoustic features (jitter, tremor). Therefore, it is expected that this method will be able to detect several voice abnormalities. An other expansion of this work may be the classification between dysphonic and normophonic speakers. In this thesis no classification was performed but as it was shown in Chapter 4, the normophonic speaker can be distinguished from the dysphonic speakers so the classification should not be difficult.

Voice quality in speech processing research usually refers to the perceived degree of the characteristics like breathiness, creakiness and loudness. Voice quality variations are considered to be mainly due to the variations in the phonation process (production of the glottal excitation signal) at the glottis. The voice quality is important for emotional/expressive speech synthesis since it gives as much information about the state of the speaker as does the prosodic information. From a signal processing point of view, voice quality variations mainly correspond to variations in spectral tilt, in the relative amount of aperiodic components in speech, and in some spectral variations in the low frequency part of the spectrum (variations in the glottal formant frequency (Fg), in the first formant bandwidth etc.)(Fig. 5.1). So, in the future the proposed method could be applied also for expressive speech.

One last concept for future work is applying the proposed method for detecting voiced and unvoiced segments of speech. In Fig. 5.2 it can be observed the amplitude envelope and the corresponding PDD spectrogram of the voiced/unvoiced segments.

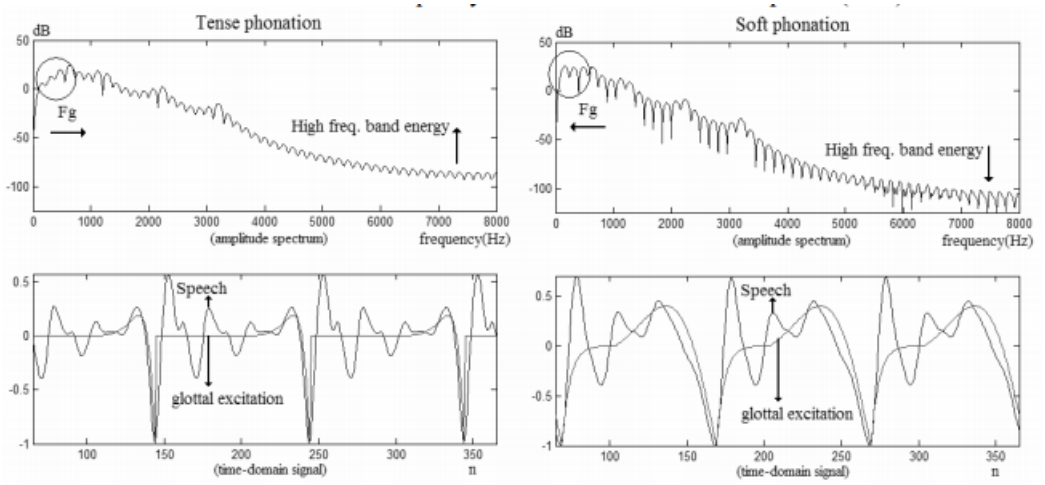


Figure 5.1: Spectral variations due to variations in phonation. Top figures show the magnitude spectrum of the speech signals and bottom figures show the time domain signals for glottal excitation and speech.

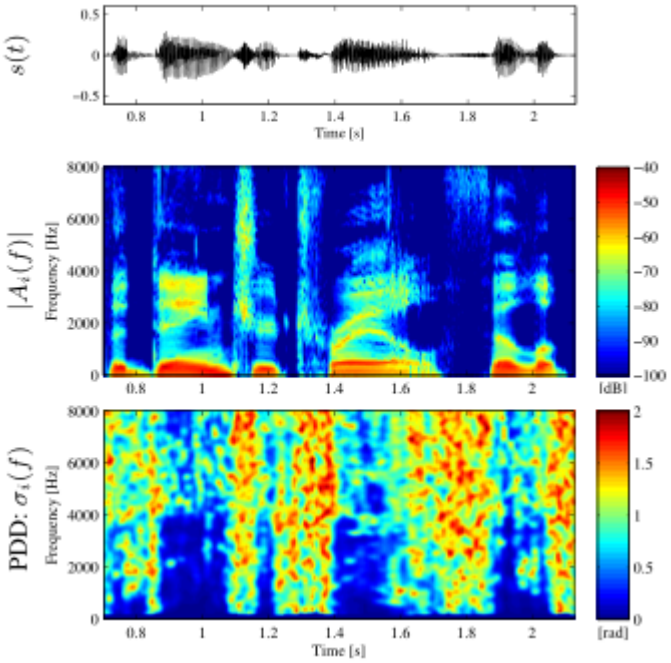


Figure 5.2: In the above spectrograms is represented the amplitude of the vocal tract and the PDD of the $s(t)$ signal.

Appendix

Appendix A

Adaptive Iterative Refinement for aHM

Algorithm

Create a sequence of time t_i according to $f_0(t)$.
Initiate each $f_0^i = f_0(t_i)$
Initiate each K^i using $f_{corr} = 20Hz$ and $K = \lfloor 0.5N_w/|f_{corr}| \rfloor$
while $\exists i$ such as $f_0^i K^i < f_s/2$ **do**
    for each anchor c **do**
        Create a segment of 3 periods around t_c using f_0^c
        Compute $\phi_0(t)$ and interp. of all f_0^i
        Compute LS solution (a_k^c, b_k^c) of a aQHM
        Compute df_k and $f_{corr} = mean(df_k/k)$
        Correct ${f'}_0^c = f_0^c + f_{corr}$
        if $f_0^c K^c < f_s/2$ **then**
            Update $K^c = \lfloor 0.5N_w/|f_{corr}| \rfloor$
        end if
    end for
    Set $f_0^i = {f'}_0^i \forall i$
end while

Appendix B

Data provided by medical doctor

In the following Table is shown the subjective evaluations of the speakers from the database which is used for the experiments in this thesis. The database and the subjective evaluations were provided by the Pr. Dejonckere from Utrecht University.

	PVS	PVF	AVE	PVFU	JIT	Jc	VL90	Tmax	Overall perc. rating	Rating tremor
Burpos	84.758	96.260	0.893	2.698	9.026	6.892	0	3.290	6	8
Burpre	46.453	84.906	0.877	9.586	11.452	5.976	1.510	1.580	6	8.5
Dijpos	29.686	62.154	0.678	10.864	16.284	8.438	0.28	0.410	8	7
Dijpre	29.341	64.725	0.697	13.965	17.643	8.384	0.26	0.92	7.5	7
Esspos	45.656	96.301	0.920	1.72	8.569	7.597	1.230	1.100	3	0.5
Esspre	24.780	63.361	0.8	11.376	15.577	10.476	0.6	2.45	7	4.5
Heupos	84.469	96.447	0.886	2.802	9.795	7.054	0	2.25	1	0.5
Heupre	49.692	87.934	0.813	8.662	16.918	10.582	5.01	2.22	6.5	3.5
Knipos	22.355	75.896	0.739	7.853	13.982	8.646	0.330	1.61	6.5	3.5
Knipre	27.239	67.884	0.738	10.962	16.815	10.986	0.560	0.660	7	3
Lulpos	89.603	99.110	0.982	0.549	2.814	2.417	5.01	7.670	1	0.5
Lulpre	40.873	91.075	0.886	9.556	14.976	9.416	5.01	1.930	4.5	2
Plupos	49.843	92.857	0.8	4	9.479	5.646	5.01	0.880	3.5	1
Plupre	20.505	72.519	0.74	17.133	18.434	10.589	0.430	0.510	6	4.5
Roopos	71.968	95.553	0.889	5.062	12.241	6.652	1.770	1.85	4.5	5
Roopre	53.012	92.199	0.798	6.274	13.422	7.053	1.640	1.220	4	7
Stupos	60.534	96.799	0.878	4.556	8.739	4.321	2.160	1.410	4	0
Stupre	50.972	91.414	0.874	5.103	10.249	7.113	1.530	1.480	3.5	6
Vropos	40.975	81.332	0.869	8.641	12.838	6.528	1.410	2.110	3.5	4
Vropre	43.239	86.379	0.861	10.499	14.043	8.029	1.400	2.440	5	4.5

Table B.1: Subjective Evaluations

PVF/PVS: PVF is the proportion of voiced frames and depends on the pauses appearing in speech. Also the PVS, the proportion of voiced speech frames is computed, thus considering inly frames that are classified as speech in the first step of the analysis. Since pauses and weak sounds are typically unvoiced, PVS will typically be larger than PVF. For vowels it should be expected that $PVS = 100\%$: The better the voice, the highest the percentages.

AVE: The average voicing evidence. The more regular(periodic) the voice frames are, the higher the AVE will be.

VL90 parameter: The 90th percentile of the voicing length duration. The voicing length is defined as the number of consecutive voiced frames found in the data. Phonatory breaks reduce this parameter.

JIT and JITc: The cycle-to-cycle period perturbation and the corrected cycle-to-cycle period perturbation. It is evaluated objectively by [1] using Ampex software [47] and the speech samples are ranked from high to low jitter. Better voices show limited jitter.

PVFU: The percentage of frames with "unreliable" F_0 is considered as a second F_0-instability factor. Frequency shifts make F_0 unreliable.

Tmax: The maximum length of speech without pause.

Overall perc. rating: The overall severity of the SD of the speech samples in a descending order as evaluated by the doctors.

Rating Tremor: The rythmic change in pitch and loudness. The speech samples are ranked from high to low tremor value.

Bibliography

[1] P.Dejonkere and C.Manfredi. Long-term follow-up of patients with spasmodic dysphonia repeatedly treated with botulinum toxin injections. *International Journal of Phonosurgery and Laryngology*, 1(2):57–60, July–December 2011.

[2] Paavo Alku. Glottal inverse filtering analysis of human voice production - a review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana, Indian Academy of Sciences*, 36:623–650, 2011.

[3] G. S. Ohm. Uber die definition des tones, nebst daran geknüfter theorie der sirene und ähnlicher tonbildender vorichtungen. *Ann. Phys. Chem.*, 59:513–565, 1843.

[4] H. L. F. von Helmholtz. On the sensations of tone (english translation by a.j. ellis). Longmans, Green and Co., London, 1912 (original wsork published 1875).

[5] R.D. Patterson. A pulse ribbon model of monaural phase perception. *J Acoust. Soc. Am.*, 82:1560–1586, 1987.

[6] M. R. Schroeder and H. W. Strube. Flat-spectrum speech. *J Acoust. Soc. Am.*, 79(5), 1985.

[7] Li Liu, Jialong He, and GG'Onther Palm. Effects of phase on the perception of intervocalic stop consonants. *Speech Communication*, 22(4):403–417, 1997.

[8] Pobloth H. and Kleijn W.B. On phase perception in speech. *IEEE*, 1:29–32, 1999.

[9] Banno, Hideki, Takeda, and F. K. Itakura. A study on perceptual distance measure for phase spectrum of stimuli. *IEEE*, 5:3297–3300, 2001.

[10] Paavo Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Commun.*, 11(2–3):109–118, 1992.

[11] T. Drugman, B. Bozkurt, and T. Dutoit. Complex cepstrum-based decomposition of speech for glottal source estimation. *Proc. Interspeech*, 2009.

[12] Y. Stylianou. Removing linear phase mismatches in concatenative speech synthesis. *IEEE Trans. Speech Audio Process*, 9(3):232–239, 2001.

[13] S.P.Lipshitz, M.Pocock, and J.Vanderkooy. On the audibility of midrange phase distortion in audio systems. *J.Audio Eng. Soc.*, 30(9):580–595, 1982.

[14] G.Degottex, A.Roebel, and X.Rodet. Function of phase-distortion for glottal model estimation. *in Proc.IEEE Int.Conf.on Acoustics, Speech, and Signal Processing(ICASSP)*, pages 4608–4611, 2011.

[15] J. M. Tribolet. A new phase unwrapping algorithm. *IEEE Trans. Acoust. Speech, Signal Process.*, pages 170–177, 1977.

[16] D. C. Ghiglia and M. D. Pritt. *Two-Dimensional Phase Unwrapping: Theory, Algorithms and Software*. Wiley, 1998.

[17] T. Drugman, T. Dubuisson, and T. Dutoit. Phase-based information for voice pathology detection. *ICASSP*, pages 4612–4615, 2011.

[18] H.Banno, K.Takeda, and F.Itakura. The effect of group delay spectrum on timbre. *Acoustical Science and Technology*, 23(1):1–9, 2002.

[19] H.A.Murthy and B.Yegnanarayana. Speech processing using group delay functions. *Elsevier Signal Processing*, 22:259–267, 1991.

[20] R.Smits and B.Yegnanarayana. Determination of instants of significant excitation in speech using group delay function. *IEEE Trans. SpeechAudio Processing*, 3:325–333, 1995.

[21] G.Degottex and Y.Stylianou. Analysis and synthesis of speech using an adaptive full-band harmonic model. *IEEE Trans. on AudioSpeech and Lang. Proc.*, 21(10):2085–2095, 2013.

[22] A.Fourcin and M.Ptok. Closing and opening phase variability in dysphonia. 2003.

[23] M.Koutsogiannaki, Y.Pantazis, Y.Stylianou, and P.Dejonkere. Tremor in speakers with spasmodic dysphonia. *MAVEBA*, 2011.

[24] M. Koutsogiannaki, O. Simantiraki, G. Degottex, and Y. Stylianou. The importance of phase on voice quality assessment. *interspeech*, 2014.

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

[26] Peter J. Murphy and Olatunji O. Akande. Noise estimation in voice signals using short-term cepstral analysis. *J. Acoust. Soc. Am.*, 121(3):1679–1690, 2007.

[27] E. Yumoto and Wilbur J.Gould. Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of Acoustic Society of America*, 71(6), 1982.

[28] Yolanda D.Heman-Ackah, Deirde D. Michael, and George S. Goding. The relationship between cepstral peak prominence and selected parameters of dysphonia. *Journal of Voice*, 6(1):20–27, 2002.

[29] B. Radish Kumar, Jayashree S. Bhat, and Neitica Prasad. Cepstral analysis of voice in persons with vocal nodules. *Journal of Voice*, 24(6):651–653, 2009.

[30] M. Fernandes, F. E. Resende Mattioli, E. Afonso Lamounier, and A. Oliveira Andrade. Assessment of laryngeal disorders through the global energy of speech. *IEEE Latin America Transactions*, 9(7):982–990, 2011.

[31] D. G. Childers and C. K. Lee. Vocal quality factors:analysis synthesis, and perception. *J. Acoust. Soc. Amer.*, 90(5):2394–2410, 1991.

[32] D. Klatt and L. Klatt. Analysis, synthesis and perception of voice quality variations among female and male talkers. *J. Acoust. Soc. Amer.*, 87:820–857, 1990.

[33] Y. Agiomyrgiannakis and Y. Stylianou. Wrapped gaussian mixture models for modeling and high-rate quantization of the phase data of speech. *IEEE Trans. on Audio, Speech and Lang. Proc.*, 17(4):775–786, 2009.

[34] A. Oppenheim, G. Kopec, and J. Tribolet. Signal analysis by homomorphic prediction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24(4):327–322, 1976.

[35] B. Doval, C. Alessandro, and N. Henrich. The voice source as a causal/anticausal linear filter. *VOQUAL,Geneva*, 2003.

[36] B. Bozkurt, B. Doval, C. Alessandro, and T. Dutoit. Zeros of z-transform representation with application to source-filter separation on speech. *IEEE signal processing letters*, 12(4), 2005.

[37] I. Saratxaga, I. Hernaez, M. Pucher, and I. Sainz. Perceptual importance of the phase related information in speech. *Proc. Interspeech. ISCA*, 2012.

[38] K. Paliwal and L. Alsteris. Usefulness of phase spectrum in human speech perception. *Proc. Eurospeech, Geneva, Switzerland*, pages 2117–2120, 2003.

[39] P. D. Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga. Evaluation of speaker verification security and detection of hmm-based synthetic speech. *Audio, Speech and Language Pocessing, IEEE transactions*, 20(8):2280–2290, 2012.

[40] H.Kawahara, I.Masuda-Katsuse, and A. de Cheveigne. Restructuring speech representation using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction:possible role of a repetitive structure in sounds. *Speech Comminication*, 27(3–4):187–207, 1999.

[41] R.McAulay and T.Quatieri. Speech analysis/synthesis based on a sinousoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):744–754, 1986.

[42] Y. Pantazis, O.Rosec, and Y.Stylianou. Adaptive am-fm signal decomposition with application to speech analysis. *IEEE Trans. Audio, Speech, Lang. Process.*, 19(2):290–300, 2010.

[43] A.V Oppenheim and R.Schafer. *Digital Signal Processing*. Prentice Hall, 2nd edition, 1978.

[44] G. Degottex and D. Erro. A measure of phase randomness for the harmonic model in speech synthesis. *interspeech*, 2014.

[45] N.Fisher. *Statistical Analysis of Circular Data*. Cambridge University Press, Oct. 1995.

[46] National spasmodic dysphonia association @ONLINE.

[47] L. V. Immerseel and J. Martens. Pitch and voiced/unvoiced determination with an auditory model. *J. Acoust. Soc. Am.*, 90(5):2394–2410, 1991.

[48] G.Degottex, J.Kane, T.Drugman, T.Raitio, and S.Schererr. Covarep - a collaborative voice analysis repository for speech technologies. *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2014.