University of Crete
Department of Computer Science

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

Spectral Based Short-Time Features for Voice Quality Assessment

MSc. Thesis

Miltiadis Vasilakis

Heraklion

March 2009

DEPARTMENT OF COMPUTER SCIENCE

UNIVERSITY OF CRETE

Spectral Based Short-Time Features for Voice Quality Assessment

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

March 27, 2009

Author:

Miltiadis Vasilakis
Department of Computer Science

Board
of enquiry:
Supervisor

Yannis Stylianou
Associate Professor

Member

Athanasios Mouchtaris
Assistant Professor

Member

Apostolos Traganitis
Professor

Accepted by:
Chairman of the
Graduate Studies Committee

Panos Trahanias
Professor

Heraklion, March 2009

Abstract

In the context of voice quality assessment, phoniatricians are aided by the measurement of several phenomena that may reveal the existence of pathology in voice. Of the most prominent among such phenomena are these of jitter and shimmer. Jitter is defined as perturbations of the glottal cycle and shimmer is defined as perturbations of the glottal excitation amplitude. Both phenomena occur during voice production, especially in the case of vowel phonation. Acoustic analysis methods are usually employed to estimate jitter using the radiated speech signal as input. Most of these methods measure jitter in the time domain and are based on pitch period estimation, consequently, they are sensitive to the error of this estimation. Furthermore, the lack of robustness that is exhibited by pitch period estimators, it makes the use of continuous speech recordings as input problematic, and essentially limits jitter measurement to sustained vowel signals. Similarly for shimmer, time domain acoustic analysis methods are usually called to estimate the phenomenon in speech signals, based on estimation of peak amplitude per period. Moreover, these methods, for both phenomena, are affected by averaging and explicit or implicit use of low-pass information. The use of mathematical descriptions for jitter and shimmer, in order to transfer the estimation from the time domain to the frequency domain, may alleviate these problems.

Using a mathematical model that couples two periodic events to achieve the local aperiodicity, allows jitter to be modeled as the shift of one of the two periodic events with respect to the other. Said model, when transformed to the frequency domain, displays interesting spectral trends between the harmonic and subharmonic subspectra. The two spectral parts are shown to form a beat spectrum, with the number of intersections between them directly dependent on the shift related to jitter. This behavior was exploited to develop a short-time Spectral Jitter Estimator (SJE). Experiments with synthetic signals of jittered phonation showed that SJE provides accurate local estimates of jitter. Further evaluation was conducted on two databases of actual sustained vowel recordings from healthy and pathological voices. Comparison with corresponding estimations from the Multi-Dimension Voice Program (MDVP) and the Praat system revealed that SJE outperforms both in normal versus pathological voice discrimination accuracy by at least 4%, as this was judged using Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) index. Examination of the short-time statistics of SJE showed that there is a higher correlation with the existence of pathology in voice, due to the fact that SJE takes into account the full spectrum.

SJE was also shown to be robust against errors in pitch period estimations, which combined with the ability of jitter estimation over short time intervals, deemed SJE a very good candidate for measuring jitter in continuous speech. Through cross-database validation a threshold of pathology for SJE has been determined. By applying this threshold to a database of reading text recordings from normophonic and dysphonic speakers, a second threshold and new features were established, especially for monitoring jitter in continuous speech. In terms of AUC, the suggested features for reading text provide a discrimination score of about 95%, while the second threshold provides a

Classification Rate (CR) of 87.8%. Furthermore, estimated short-time jitter values from reading text were found to confirm the studies showing the decrease of jitter with increasing fundamental frequencies, and the more frequent presence of high jitter values in the case of pathological voices as time increases.

A mathematical model that combines two periodic events, allows also for modeling of shimmer by applying different amplitude deviations on the two events. Again, by transforming the model from the time domain to the frequency domain, notable spectral properties are observed. Using this properties four features indicative of shimmer were created to evaluate the model. Experiments with synthetic shimmered phonation signals, as well as the two afore-mentioned databases of sustained vowel recordings, showed that the model captures correctly the shimmer phenomenon and further development should be pursued.

Acknowledgments

Props to the following:

- My supervisor, professor Yannis Stylianou, for his guidance and encouragement during this work, and in general for the fruitful collaboration we had all these years.
- My colleagues, past and present, at the Multimedia Informatics Laboratory, for creating a pleasant and inspiring environment to work in. Special thanks should go to Yannis Agiomyrgiannakis, Ioannis Pantazis, Andre Holzapfel, Maria Markaki, Christos Tzagarakis, George Kafentzis, Ioannis Sfakianakis, Christina-Alexandra Lionoudaki and Maria Astrinaki.
- csdlista & friends for all the fun times, especially the Heraklion charter.
- Last, but certainly not least, my family, for their undivided aid and support.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Voice quality assessment is an essential diagnostic tool used by phoniatricians to help them in determining the existence of pathological voice. Several methods have been developed to extract measurements that can be used for that purpose. These can be categorized in three main groups:

- Videoendoscopy methods that use an image or video recording of the vocal folds, such as videokymography.
- Electroglottography (EGG) methods that use the measurement of the electrical resistance between two electrodes placed around the neck, through which the vocal fold contact area can be estimated.
- Acoustic analysis methods that use recordings of the radiated speech signal to compute parameters related to pathological voice.

Comparing these three groups of methods, acoustic analysis seems to have several advantages. In particular, acoustic analysis methods have a lower cost of implementation, while they require less time and are non-invasive for the patient when applied. Furthermore, acoustic analysis can produce automatic quantitative results, which, apart from assisting clinical doctors, can be used for unsupervised classification of a voice as pathological or healthy, or even detect specific cases of dysphonia.

Noise is what is mainly perceived as the immediate effect of a pathological condition in voice. The two prominent kinds of noise due to dysphonia are those of additive noise, such as in cases of breathiness, and of modulation noise, such as in cases of roughness. Regarding modulation noise, it can be further categorized, in modulation of the amplitude or modulation of the periodicity, of the voice signal. The case of periodicity modulation is referred to as jitter, while the case of amplitude modulation is referred to as shimmer. Both these phenomena may occur during voice production, especially in vowel phonation. Jitter is defined in that context as small fluctuations in glottal cycle lengths [22] [31], while shimmer is defined as perturbations of the glottal excitation amplitude.

Over successive speech cycles, jitter and shimmer help to give the vowel its naturalness in contrast to constant pitch and amplitude that can result in a machine-like sound. Moreover, the two phenomena contribute to the voice quality of a speaker. Note that humans are not able to differentiate between the noise produced by jitter and shimmer, while they are able to differentiate modulation from additive noise [19]. As it was mentioned above, a high degree of jitter and/or shimmer results in a voice with roughness, that is usually perceived as noise in recordings of pathological voices. Therefore, a reliable estimation of these phenomena can be used to discriminate between normophonic and dysphonic speakers. The actual measurement may take place in the time domain, the frequency domain (magnitude spectrum), or the quefrency domain.

Based on the definition of jitter, many acoustic analysis methods have been proposed for the computation of a value that quantifies the aperiodicity that is introduced to the voice signal on account of jitter [27] [11] [2] [29]. The most common methods are time domain ones that are based on the estimation of a sequence of pitch period values, over a length of time that comprises several periods. This sequence is then used to produce an average value of jitter over that duration. If N is the total number of pitch periods and $u(n)$ is the sequence of pitch period estimates, the definitions of some widely accepted jitter measurements are given below:

- Local jitter is the period-to-period variability of pitch (%)

$$\text{local jitter} = \frac{\dfrac{1}{N-1}\displaystyle\sum_{n=1}^{N-1}|u(n+1)-u(n)|}{\dfrac{1}{N}\displaystyle\sum_{n=1}^{N}u(n)} \tag{1.1}$$

- Absolute jitter is the period-to-period variability of pitch in time

$$\text{absolute jitter} = \frac{1}{N-1}\sum_{n=1}^{N-1}|u(n+1)-u(n)| \tag{1.2}$$

- Relative Average Perturbation (RAP) jitter provides the variability of pitch with a smoothing factor of three periods (%)

$$\text{RAP} = \frac{\dfrac{1}{N-2}\displaystyle\sum_{n=1}^{N-2}\frac{|2u(n+1)-u(n)-u(n+2)|}{3}}{\dfrac{1}{N}\displaystyle\sum_{n=1}^{N}u(n)} \tag{1.3}$$

- Pitch Period Perturbation Quotient (PPQ) provides the variability of pitch with a smoothing factor of five periods (%)

$$\text{PPQ} = \frac{\dfrac{1}{N-4}\displaystyle\sum_{n=1}^{N-4}\frac{|4u(n+2)-u(n)-u(n+1)-u(n+3)-u(n+4)|}{5}}{\dfrac{1}{N}\displaystyle\sum_{n=1}^{N}u(n)} \tag{1.4}$$

Since these methods are based on pitch period estimation, they are sensitive then to the error of this estimation. It is not uncommon for the same voiced speech segment to obtain different fundamental period estimations when using different estimators of the pitch period. This will lead to quite different jitter estimates, making thus the above jitter measurements quite vulnerable to the variability of pitch period estimators. This variability is partially due to the quasi-periodic character of speech even for sustained vowels. A strictly periodic speech signal would have strong frequency components at the integer multiples of the fundamental frequency (the reciprocal of the pitch period) referred to as harmonics. From high frequency resolution speech analysis, results show that voiced speech signals have strong components *near* the harmonic frequencies (but not on the harmonics) usually referred to as inharmonics[25]. At low frequencies ($\leq$ 700 Hz), the speech magnitude spectrum has more of a harmonic structure compared to its structure in higher frequencies. If we consider then only the lower frequencies of a voiced speech segment, this may be closer to being periodic than the corresponding full-band voiced speech segment. Consequently, the methods trying to estimate the pitch period of quasi-periodic speech signals, either explicitly (by low-pass filtering) or implicitly (by ignoring the quasi-periodic character of voiced speech), rely on the low frequencies of speech signals. Furthermore, the average measurement of jitter, over a series of pitch periods, is used by the majority of the methods as a means to to minimize the variance of the jitter estimation. Methods based on the averaging of jitter are statistically biased, since it has been found that they underestimate jitter [31]. Also, averaging implies that pitch cycle perturbations are generated by an independent and identically distributed (Gaussian) stochastic process. It has been shown, however, that there is correlation between successive values of jitter [10]. Therefore, this correlation should be removed before applying the average operator [31].

The choice of applying jitter estimation methods on sustained vowels rather than on continuous (running) speech, is mostly driven by the lack of robustness in the automatic extraction of the fundamental frequency of speech and on the limitations of the suggested estimators of jitter [26] [34]. However, there are arguments in favor of using continuous speech or isolated sentences, such as reading text, for voice pathology detection, since difficulties in abducting or adducting, or asymmetries in the vocal folds, because of pathology, may be revealed during non-stationary areas of speech [20] [12]. Processing of continuous speech for voice pathology detection was studied before, for example in [14] [1] [21] [34] [12]. In [1] patients read a tale for a duration of approximately 40 seconds, and then seven acoustic measures of cycle-to-cycle perturbations in the speech waveform were investigated. It was suggested that the standard deviation of the distribution of the relative frequency differences between consecutive pitch periods provides a useful acoustic measure of waveform perturbations. Since these approaches are based on the pitch period estimation, their accuracy is a function of the accuracy of the pitch period estimators. Given the pseudo-periodic character of voiced speech there is an ambiguity in pitch period estimation and therefore an ambiguity in the estimation of jitter. Moreover, there is no control if the perturbations observed in the speech waveform are due to jitter, or shimmer, or other sources (vocal folds and vocal tract interactions) [1]. In [34], a time-frequency (TF) representation based on the matching pursuit decomposition with Gabor TF atoms of various scale factors was used. It was found that the distribution of these scale factors is a potential feature for discrimination of normal and pathological speech signals. In [12] hearing and phonetic criteria in voice measurement were discussed. Various features were considered taking into account functions of the estimated fundamental frequency and vocal fold closed quotient during connected speech. It was found that these measurements were related both to vocal fold function and to the perceptual attributes of pitch, loudness, and voice quality.

For shimmer, as well, many acoustic analysis methods that quantify the phenomenon have been proposed. Such methods are usually based on peak extraction, in order for a time-series of amplitude values, one for each pitch period, to be established. This series is used to provide an average value of shimmer over a number of several periods. If N is the total number of pitch periods and $a(n)$ is the peak amplitude sequence, the definitions of widely accepted shimmer measurements are given below:

- Local shimmer is the period-to-period variability of the peak-to-peak amplitude (%)

$$\text{local shimmer} = \frac{\frac{1}{N-1}\sum_{n=1}^{N-1}|a(n+1)-a(n)|}{\frac{1}{N}\sum_{n=1}^{N}a(n)} \tag{1.5}$$

- Absolute shimmer is the period-to-period variability of the peak-to-peak amplitude in dB

$$\text{absolute shimmer} = 20\frac{1}{N-1}\sum_{n=1}^{N-1}|\log_{10}[a(n+1)-a(n)]| \tag{1.6}$$

- Amplitude Perturbation Quotient (APQ) provides the variability of the peak-to-peak amplitude with a smoothing factor of 11 periods (%)

$$\text{APQ} = \frac{\dfrac{1}{N-10}\displaystyle\sum_{n=1}^{N-10} \frac{|10a(n+5) - \sum_{k=0}^{4} a(n+k) - \sum_{k=6}^{10} a(n+k)|}{11}}{\dfrac{1}{N}\displaystyle\sum_{n=1}^{N} a(n)} \tag{1.7}$$

These methods exhibit problems similar to those of time-domain jitter methods, regarding averaging, pitch period estimation sensitivity and reliance to low frequencies.

The previous pitch-period based methods for measuring jitter and shimmer can be considered to model the effects of these phenomena using solely their assumed temporal properties. Another approach is to model each phenomenon using its spectral characteristics. In [24], by taking the Fourier series of two periods of the glottal pulse waveform, it was shown how these terms contribute to the different frequencies appearing in the spectrum, for different kinds of perturbations. For jitter (i.e., having two periods of different length) and for shimmer (i.e., having two periods of different amplitude), interesting trends between the harmonic and subharmonic subspectra were noted. Regarding jitter, the pattern persisted in synthetic jittered glottal airflow signals, with either cyclic or random variation of the fundamental frequency. Similarly for shimmer, the respective pattern was also present in synthetic shimmered glottal airflow signals, with either cyclic or random variation of the glottal signal amplitude. However, these models were developed in a heuristic manner, and no mathematical proof was provided.

In this work, we present a new method for the estimation of jitter, based on a mathematical description of the time domain properties of the jitter phenomenon. More specifically, we show how jitter can be described as the combination of two periodic events. When these two events are viewed through a single prism, then they can achieve the local aperiodicity that is jitter. This modeling of jitter as a cyclic process, allows us to identify it quantitatively as the shift of one of the two periodic events with respect to the other. By transforming this model to the frequency domain, it is shown that jitter leads to a beat spectrum defined by the above mentioned shift of the two periodic events. This spectral behavior is in concordance with the afore-mentioned heuristic development and can be used then to indirectly estimate the value of jitter by counting the number of intersections between harmonic and subharmonic subspectra [37]. Based on this we created a novel short-time jitter estimator, referred to as Spectral Jitter Estimator (SJE), that allows for time-varying measurement with a high local accuracy, as it was demonstrated on synthetic phonation signals with known jitter. Although SJE uses pitch period information, it was shown that this is not crucial in counting the number of intersections between the harmonic and subharmonic subspectra [35]. Additionally, by producing a short-time sequence of local jitter values on small intervals, SJE provides estimates without assuming long-term periodicity as in the purely time domain approaches. The performance of SJE in discriminating between normal and pathological voice status was compared to jitter measurements obtained by two established systems for quantitative acoustic assessment of voice quality, namely Praat [4] and Multi-Dimensional Voice Program (MDVP) [9] of KayPENTAX. On two different databases of sustained vowel recordings, the Massachusetts Eye and Ear Infirmary (MEEI) Disordered Voice Database [8] and the Príncipe de Asturias (PdA) Hospital in Alcalá de Henares of Madrid database [13], the estimates of SJE were shown to be more correlated with pathology than the estimates by Praat and MDVP.

As it was mentioned earlier, it has been shown that methods that produce an average estimate for jitter are statistically biased and actually underestimate jitter [31]. For short-time jitter estimators, however, averaging is not necessary. Actually, the generated sequence of local measurements of jitter can be used to gain further insight on the temporal behavior of jitter, for both healthy and pathological voices. For this purpose, we extended the use of SJE on reading text recordings by suggesting new features for analysis of continuous speech, based on the short-time measurements of jitter as provided by SJE. Through cross-database comparison, we determined a relevant threshold for pathology that leads to high discrimination for normal versus pathological voices, in databases of either sustained phonation recordings or reading text recordings. Using this threshold and based on the time-series of local jitter estimations from SJE, three new features have been suggested [36]. It has been shown that all three features have a high correlation with the existence of pathology, while they are ideal for voiced segments of running speech signals. Furthermore, one of these three features can be efficiently used to monitor the jitter effect in running speech.

A mathematical model of the time domain properties of shimmer, as this is achieved by two periodic events, was also examined. The amplitude deviation that characterizes this cyclic shimmer process leads to a frequency domain where two regimes of constant magnitude are observed, one for the harmonic part and another for the subharmonic part [38]. This spectral property, previously noted in [24], can be used to create acoustic analysis features that are related to the shimmer phenomenon. For the evaluation of this model four such features have been developed and tested on synthetic signals and the databases MEEI and PdA.

The contents of this work are organized as follows. In Chapter 2 the mathematical model SJE is based on and its properties in time and frequency are presented. The algorithm of SJE is also given in this Chapter. The synthetic signals and databases of recordings used in our experiments, as well as the applied evaluation procedures, are described in Chapter 3. The evaluation of SJE takes place in Chapter 4. Specifically, initial validation was conducted using synthetic signals, while the performance of SJE regarding discrimination of normal and pathological voices, compared to jitter estimations from Praat and MDVP, was examined in two databases of sustained phonation recordings. The results from these experiments lead to the establishment of a pathology threshold for SJE. The Chapter is closed with the study of the short-time statistics of SJE and the introduction of three new voice quality assessment features especially for running speech. The mathematical model for shimmer and its evaluation are presented in Chapter 5. Finally, Chapter 6 concludes this work and provides information on future work and possible extensions. Note that parts of this work have been initially presented in the following journals

- M. Vasilakis and Y. Stylianou. Spectral jitter modeling and estimation. Biomedical Signal Processing & Control Special Issue: M&A of Vocal Emissions, to appear. [35]
- M. Vasilakis and Y. Stylianou. Voice pathology detection based on short-time jitter estimations in running speech. Folia Phoniatrica et Logopaedica, to appear. [36]

and the following workshops

- M. Vasilakis and Y. Stylianou. A mathematical model for accurate measurement of jitter. In MAVEBA 2007, pages 710, Florence, Italy, 2007. [37]
- M. Vasilakis and Y. Stylianou. A mathematical model for accurate measurement of shimmer. In 2nd Advanced Voice Function Assessment International Workshop, Aachen, Germany, 2008. [38]
- M. Vasilakis and Y. Stylianou. Spectral jitter estimation revisited. In 3rd Advanced Voice Function Assessment International Workshop, submitted, Madrid, Spain, 2009. [40]
- M. Vasilakis and Y. Stylianou. Novel short-time jitter features for monitoring of running speech. In 3rd Advanced Voice Function Assessment International Workshop, submitted, Madrid, Spain, 2009. [39]

Chapter 2

Spectral Jitter Estimator development

In this chapter a mathematical model of jitter is presented that exhibits notable spectral behavior. This behavior is utilized in order to develop a short-time Spectral Jitter Estimator (SJE). The algorithm of SJE is also given here.

2.1 Mathematical model of jitter

Jitter is defined as cycle-to-cycle perturbations of the glottal cycle lengths, which lead to a local aperiodicity. This kind of perturbations can be modeled and generated by considering two periodic events, which, when combined appropriately, may produce the observed perturbations. Let us consider a mathematical model that describes two periodic events. The local aperiodicity of jitter can be defined then, in relation to these two events, as the shift of one of the two with respect to the other. This shift can be measured to provide us with a quantitative value for jitter [37]. Therefore, a jittered impulse train can be obtained by applying a constant pitch deviation every second impulse, achieving thus a cyclic perturbation that creates the two aforementioned events [28] (pgs. 102-103). We can then model the glottal airflow signal under the presence of jitter as the convolution of the glottal signal over one glottal cycle with such a jittered impulse train. The jittered impulse train can be expressed then as

$$g[n] = \sum_{k=-\infty}^{+\infty} \delta[n-(2k)P] + \sum_{k=-\infty}^{+\infty} \delta[n+\epsilon-(2k+1)P] \tag{2.1}$$

where P is the pitch period and ϵ is the pitch deviation, both in samples. In this model, shown in Fig. 2.1, ϵ is the shift that corresponds to jitter. The value of ϵ can range from 0 (no jitter) to P (pitch halving).

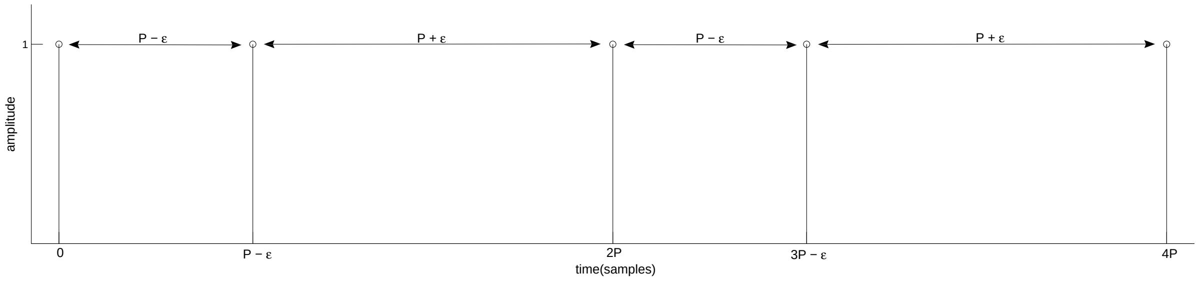

Figure 2.1: Jittered impulse train of the two event model for jitter.

The Fourier transform of the cyclically jittered impulse train (2.1) is

$$
\begin{aligned}
G(\omega) &= \sum_{n=-\infty}^{+\infty} g[n]e^{-j\omega n} \\
&= \sum_{n=-\infty}^{+\infty} \left(\sum_{k=-\infty}^{+\infty} \delta[n-(2k)P] + \sum_{k=-\infty}^{+\infty} \delta[n+\epsilon-(2k+1)P] \right) e^{-j\omega n} \\
&= \sum_{k=-\infty}^{+\infty} e^{-j\omega 2kP} + \sum_{k=-\infty}^{+\infty} e^{-j\omega[(2k+1)P-\epsilon]} \\
&= \left(1+e^{-j\omega(P-\epsilon)}\right) \sum_{k=-\infty}^{+\infty} e^{-j\omega 2kP} \\
&= \left(1+e^{-j\omega(P-\epsilon)}\right) \sum_{k=-\infty}^{+\infty} \frac{2\pi}{2P}\delta\left(\omega - k\frac{2\pi}{2P}\right) \\
&= \left(1+e^{-j\omega(P-\epsilon)}\right) \sum_{k=-\infty}^{+\infty} \frac{\omega_0}{2}\delta\left(\omega - k\frac{\omega_0}{2}\right)
\end{aligned}
\tag{2.2}
$$

where $\omega_0 = \dfrac{2\pi}{P}$ is the fundamental frequency in rad.
The squared of the magnitude spectrum (2.2) is then given by

$$
\begin{aligned}
|G(\omega)|^2 &= G(\omega)G^*(\omega) \\
&= \left(1+e^{-j\omega(P-\epsilon)}\right)\left(1+e^{+j\omega(P-\epsilon)}\right)\left[\sum_{k=-\infty}^{+\infty} \frac{\omega_0}{2}\delta\left(\omega - k\frac{\omega_0}{2}\right)\right]^2 \\
&= \frac{\omega_0^2}{4}\left(1+e^{+j\omega(P-\epsilon)}+e^{-j\omega(P-\epsilon)}+1\right)\sum_{k=-\infty}^{+\infty} \delta\left(\omega - k\frac{\omega_0}{2}\right) \\
&= \frac{\omega_0^2}{2}\left(1+\cos\left[(P-\epsilon)\,\omega\right]\right)\sum_{k=-\infty}^{+\infty} \delta\left(\omega - k\frac{\omega_0}{2}\right) \\
&= \frac{\omega_0^2}{2}\sum_{k=-\infty}^{+\infty}\left(1+\cos\left[(P-\epsilon)\,k\frac{\omega_0}{2}\right]\right)\delta\left(\omega - k\frac{\omega_0}{2}\right)
\end{aligned}
\tag{2.3}
$$

The cosine term inside the sum corresponds to a beat spectrum described by the formula

$$1 + \cos\left[(P - \epsilon)\, k\frac{\omega_0}{2}\right] = 1 + \cos(k\pi)\cos(k\frac{\epsilon}{P}\pi) \tag{2.4}$$

This beat spectrum has a center period of $2\pi/P$ and a deviation period of $2\pi/\epsilon$, both in rad. Therefore, the frequency interval between intersections of the envelope in the beat spectrum is π/ϵ (rad). Since both cosine signals in (2.4) have zero phase, the intersections can be located at frequencies

$$\omega_k = \left(k + \frac{1}{2}\right)\frac{\pi}{\epsilon} \tag{2.5}$$

with $\omega_k \leq \pi$.

From (2.3) the log magnitude spectrum of (2.1) is be shown to be

$$\begin{aligned}
&20\log_{10}|G(\omega)| = 10\log_{10}|G(\omega)|^2\\
&= 10\log_{10}\left(\frac{\omega_0^2}{2}\left(1 + \cos\left[(P-\epsilon)\,\omega\right]\right)\right)\sum_{k=-\infty}^{+\infty}\delta\left(\omega - k\frac{\omega_0}{2}\right)\\
&= 10\log_{10}\left(\frac{\omega_0^2}{2}\left(1 + \cos\left[(P-\epsilon)\,\omega\right]\right)\right)\left[\sum_{l=-\infty,k=2l}^{+\infty}\delta\left(\omega - k\frac{\omega_0}{2}\right) + \sum_{l=-\infty,k=2l+1}^{+\infty}\delta\left(\omega - k\frac{\omega_0}{2}\right)\right]\\
&= 10\log_{10}\left(\frac{\omega_0^2}{2}\left(1 + \cos\left[(P-\epsilon)\,\omega\right]\right)\right)\left[\sum_{l=-\infty}^{+\infty}\delta\left(\omega - l\omega_0\right) + \sum_{l=-\infty}^{+\infty}\delta\left(\omega - (l+\frac{1}{2})\omega_0\right)\right]
\end{aligned} \tag{2.6}$$

Based on (2.6), we can divide the spectrum to a harmonic and a subharmonic part, by sampling this beat spectrum at frequencies $l\omega_0$, which are multiples of the fundamental frequency and at frequencies $(l + 1/2)\omega_0$, which are in between the harmonic locations, respectively, where we remind that $\omega_0 = 2\pi/P$. The harmonic part of the log magnitude spectrum, as it is influenced by jitter, is described then by

$$H(\epsilon, l\omega_0) = 10\log_{10}\left(\frac{\omega_0^2}{2}\left(1 + \cos\left[(P-\epsilon)\, l\omega_0\right]\right)\right),\ l \in \mathbf{N} \tag{2.7}$$

while the subharmonic part of the log magnitude spectrum, that appears because of the existence of jitter, is given by

$$S(\epsilon, (l+\tfrac{1}{2})\omega_0) = 10\log_{10}\left(\frac{\omega_0^2}{2}\left(1 + \cos\left[(P-\epsilon)\,(l+\tfrac{1}{2})\omega_0\right]\right)\right),\ l \in \mathbf{N} \tag{2.8}$$

Examples of the two subspectra, for various values of ϵ, are depicted in Fig. 2.2. It can be observed that the harmonic and subharmonic parts follow a specific pattern, where for a specific value of ϵ the two parts intersect ϵ times. As it was mentioned previously, the locations of the intersections are provided in (2.5). For example, when $\epsilon = 2$ the intersections are located at frequencies $\pi/4$ and $3\pi/4$ (rad), as it is also shown in Fig. 2.2. It is very important to observe that the locations of the intersections in the beat spectrum only depend on ϵ and not on the pitch period (P) of the signal. Hence, by counting the number of intersections between the harmonic and subharmonic subspectra, an estimation of ϵ, and therefore of jitter, can be obtained. It is also interesting to note that this spectral property has been confirmed previously in a heuristic manner for synthetic jittered glottal airflow signals, with either cyclic or random variation of the fundamental frequency [24].

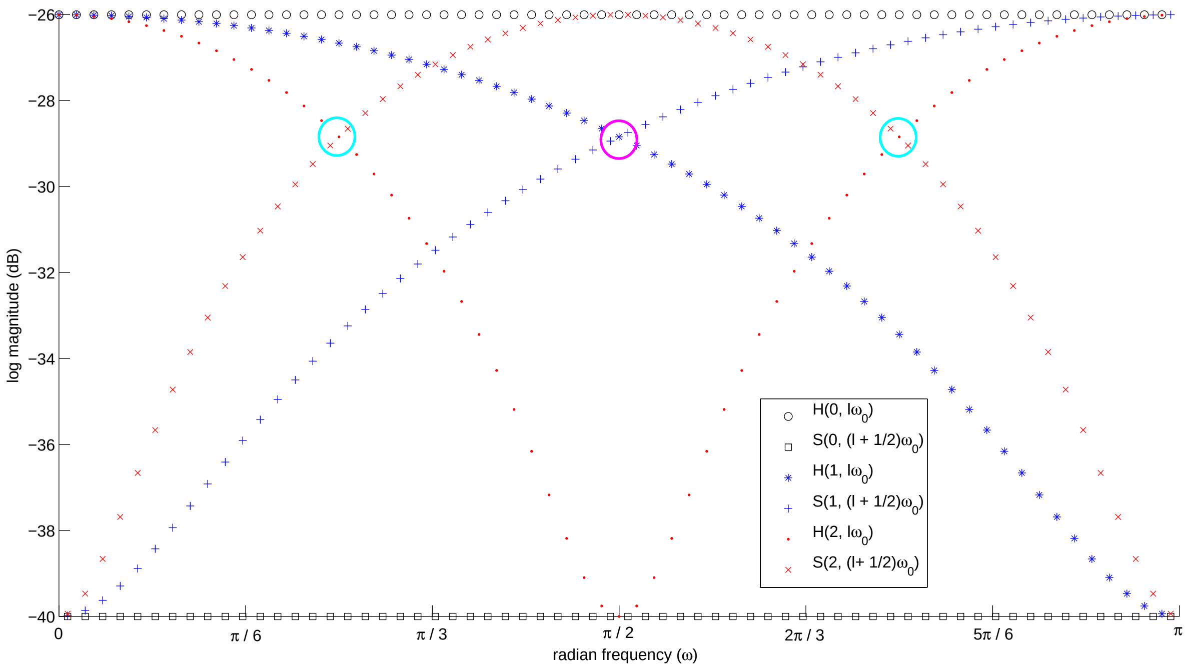

Figure 2.2: Log magnitude spectra of the harmonic and subharmonic parts of the mathematical model for jitter. It is worth to note that the circled intersections between the two parts, reveal each time the value of jitter.

2.2 Spectral Jitter Estimator

If a jittered impulse train, such as in (2.1), is used as the input of a linear system, then the aforementioned spectral structure remains visible also in the output. Therefore, it is expected to observe such a spectral behavior in phonation recordings, whenever jitter is present. Exploiting this fact, a short-time Spectral Jitter Estimator (SJE) has been developed, based on the previously observed properties of the log magnitude spectrum in the presence of jitter [35].

In the mathematical model for jitter it was assumed that the two periodic impulse trains have infinite duration and so a cyclic jitter is modeled. In real speech signals, however, the value of jitter may be modified from period to period. By windowing the signal with a sliding frame this problem is alleviated; we are allowed to examine the signal gradually in time and thus to calculate a sequence of local jitter estimates. Given a phonation recording $\mathbf{s}[n]$ of , with $n = (0 : N - 1)$, the implemented algorithm is presented in pseudo code in Table 2.1. The pitch period is estimated beforehand and we can either use the local value of the pitch period or the average pitch period of the whole signal. Most usually the input signal is a sustained phonation recording of duration 1-2 seconds. In these cases especially, using the average pitch period may provide a more robust result. The size and step of the sliding frame, indicated by the variables L and S, respectively, are chosen to be multiples of the pitch period. Specifically for the frame size, in order to compute the perturbation from one period to the next, at least two periods are required. Because of discontinuities of the time

```
jitter = []
time = []
$index$ = 0
$start$ = 0
pitch = []
pitch = pitch period sequence estimation($\mathbf{s}[n]$)
while $start < N$
    { if the current frame is unvoiced
          $P$ = average(pitch)
      else                                    }  or  {  $P$ = average(pitch)  }
          $P$ = pitch[$index$]
      endif }
    $end$ = $start + L * P - 1$
    frame = $\mathbf{s}[start : end]$
    frame = Hanning window(frame)
    F = $20\log_{10}(|$FFT(frame)$|)$
    H = F[harmonic frequencies]
    S = F[subharmonic frequencies]
    intersections = locations of valid intersections between the two parts(H, S)
    $candidateJitter$ = length(intersections)
    while $candidateJitter > 0$
        if intersections satisfy $candidateJitter$ expected locations then break from the loop
        $candidateJitter$ - -
    endwhile
    jitter[$index$] = $candidateJitter$
    time[$index$] = $(start + end)/2$
    $start$ += $S * P$
    $index$++
endwhile
return jitter, time
```

Table 2.1: The algorithm of the short-time Spectral Jitter Estimator described in pseudo code.

domain signal (end effects), alias frequencies may appear in the computed log magnitude spectrum. In order to avoid this the windowing of the frame is suggested. For this purpose a Hanning window is used. However, with a tapered window like Hanning, the data is distorted. To minimize the effect of this distortion we concluded from our experiments that a window length of three or four periods provides a high enough resolution in the computed spectrum for the estimation to be successful, while the applied Hanning window concentrates on the two middle periods, providing thus the desired short-time precision. The log magnitude spectrum of the Fourier transform of the frame is then computed and using the local pitch period (or the average pitch period) estimate, the magnitude spectrum is split into the harmonic and subharmonic spectra.

By taking into account the two subspectra, the locations of the occurring intersections between them are computed. In order to overcome potential resolution problems in the spectral magnitude, a threshold is used to determine if an intersection is rightfully indicated as such. For any given intersection, if the harmonic and subharmonic parts after its occurrence never reach a difference in amplitude over the threshold, before the next potential intersection, then these two intersections are rejected. The remaining intersections are termed as “valid intersections” in the algorithm. Through experiments with synthetic jittered signals, this threshold has been set to 3 dB. It is worth mentioning that this threshold has been kept constant during subsequent experiments. The valid intersections are further examined, by taking into account the prior knowledge of their expected locations, for each possible jitter value, as these are given in (2.5). In detail, starting with the candidate jitter of the highest possible value, which is the number of all valid intersections, we divide the spectrum in that many equal segments. It is expected that at least one intersection exists in the area around the center of each segment. If this is true, then the candidate jitter value is accepted as our estimation. Otherwise, the candidate jitter is decreased by one and the process is repeated until either the structural requirement is met or zero is reached. This refinement is necessary to suppress any spurious intersections that may arise between the harmonic and subharmonic spectral parts, especially in higher frequencies. In essence this process enhances the intersections by eliminating the spurious ones and grouping neighboring ones in clusters. An example of this enhancement is illustrated in Fig. 2.3. The algorithm of SJE is also shown as a block diagram in Fig. 2.4, while visual examples of its usage on a frame from a synthetic jittered phonation signal and a frame from an actual sustained vowel recording are presented in Figs. 2.5 and 2.6, respectively.

The obtained short-time sequence of SJE estimations is quantized and consists of integer values that represent the jitter deviation in samples. Other methods that quantify the phenomenon of jitter produce estimates in time units. For comparison purposes the SJE estimates should also be converted from samples to the appropriate time units. Absolute jitter in (1.2) is one of the most common time domain methods. In relation to the mathematical model used for our estimator, absolute jitter provides a value of $2 \times \hat{\epsilon}$, where $\hat{\epsilon}$ is the jitter estimate. Both Praat and MDVP implement this kind of measurement and in fact return results in μs. Converting the samples value

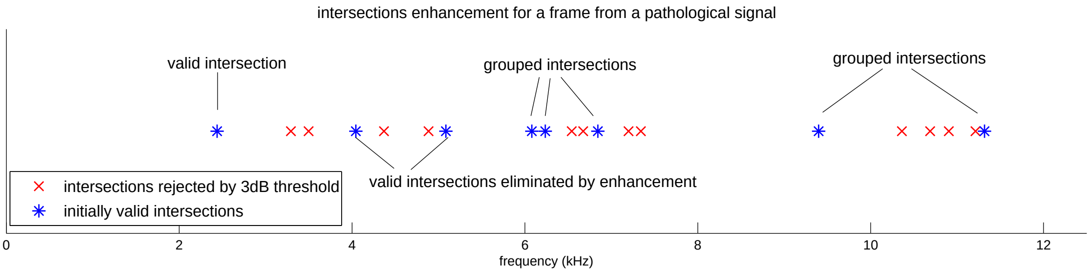

Figure 2.3: An example of intersections enhancement for a frame from a pathological signal. Prior to the enhancement SJE estimates jitter to $\hat{\epsilon} = 8$ samples, while after enhancement it produces a value of $\hat{\epsilon} = 3$ samples.

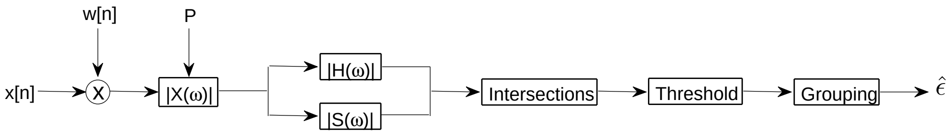

Figure 2.4: Block diagram of the short-time Spectral Jitter Estimator algorithm.

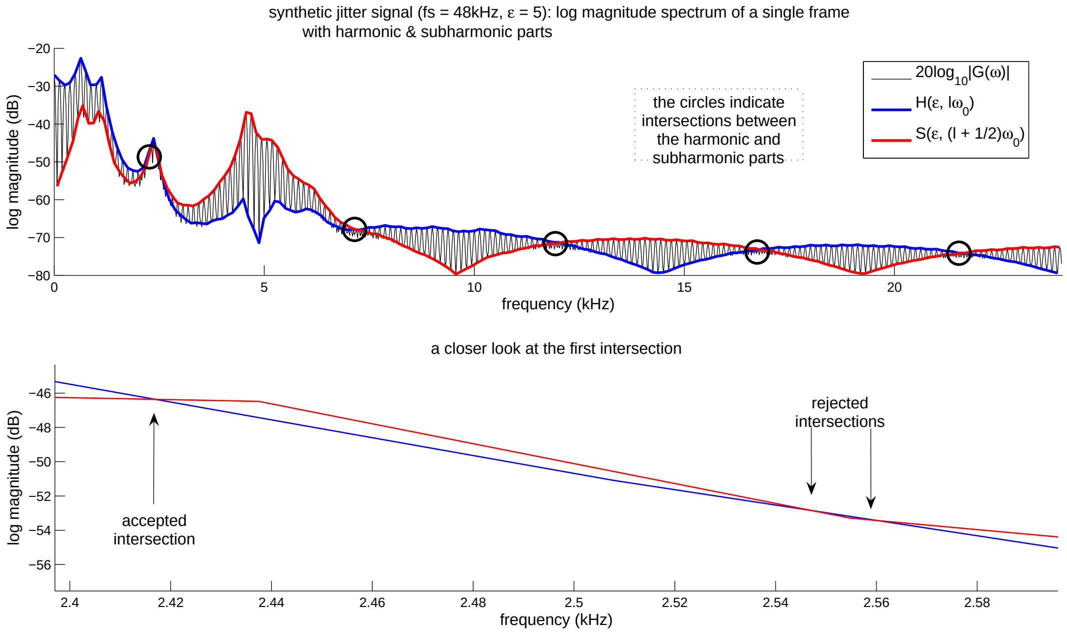

Figure 2.5: (a) Harmonic and subharmonic spectra for a synthetic jitter signal with $\epsilon = 5$. The detected intersections, after the enhancement process, are illustrated by circles. The expected structural behavior of jitter in this synthetic example is clearly demonstrated. (b) One example of a valid intersection and one example of a pair of rejected intersections.

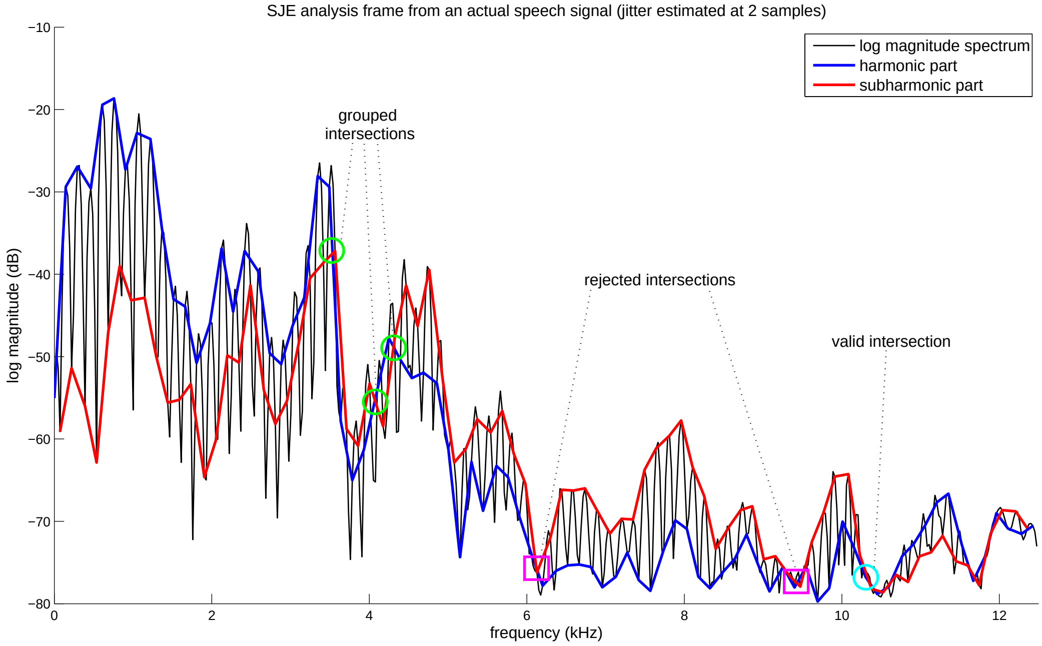

Figure 2.6: Harmonic and subharmonic spectra from a frame of an actual sustained phonation recording. The Spectral Jitter Estimator results, after intersections enhancement, to an estimate of $\hat{\epsilon} = 2$.

accordingly, again a quantized value is computed, with a quantum of

$$\hat{\epsilon}_q(F_s) = \frac{2 \times 10^6}{F_s}(\mu s) \tag{2.9}$$

where F_s is the sampling frequency in Hz. It is evident, that the larger the sampling frequency, the larger the resolution of the measurement. Consider as an example a signal sampled at 50 kHz that is estimated to have a jitter value of 1 sample, which translates to 20 μs in time units (40 μs for the absolute jitter equivalent). If this same signal were down sampled at 25 kHz, then it could be the case that the estimation revealed no jitter at all. However, if an up sampling took place at 100 kHz, then the estimated jitter would be 2 samples, which in time units is again 20 μs. Therefore, a higher sampling frequency would improve the results resolution-wise, only if the information in the signal is not already enough for an accurate estimation.

Chapter 3

Data and evaluation procedures

The signals used for validation and experiments are described in this chapter. They consist of synthetic signals that were created for the purposes of this work and also existing databases of recordings from healthy and pathological voices. The procedures employed for evaluation of the proposed methods are also presented here.

3.1 Synthetic jittered phonation signals

For the initial verification of the validity of the proposed methods, synthetic signals of sustained phonation were created with specific jitter perturbations. For all the synthetic signals, a vocal tract autoregressive model of order 50 was used. The model was created from one period of a sustained phonation recording of the vowel /a/, with average fundamental frequency of 125 Hz. The vocal tract envelope was excited then by jittered impulse trains as these are described in (2.1). Signals were created for sampling frequencies of both 16 and 48 kHz. The value of ϵ varied from 0 samples to 10% of the pitch period, that is up to 13 and 39 samples, for the sampling frequencies of 16 and 48 kHz, respectively. The duration of the signals were set to 1 seconds.

3.2 Databases of recordings

Databases of actual signals from both healthy and pathological speakers were used for further evaluation experiments of the proposed methods. Specifically, two databases of sustained phonation recordings and one database of reading text recordings were used. The signals in all databases have been labeled as normal or pathological by clinical doctors of the institute that created each database.

3.2.1 Massachusetts Eye and Ear Infirmary Disordered Voice Database

The Massachusetts Eye and Ear Infirmary (MEEI) Disordered Voice Database [8] contains recordings of the sustained vowel /a/, from 53 subjects with a healthy voice and 657 subjects with a wide variety of pathological conditions. The subjects were labeled accordingly by clinical doctors. There are also included, for most of the recordings, the acoustic analysis parameters produced by the Multi-Dimensional Voice Program [9] (MDVP). For the purposes of our experiments, the sustained phonation recordings with the MDVP parameters available were selected. Specifically, this concerns all 53 of the healthy voice cases and 631 of the pathological ones. All normal signals have a sampling frequency of 50 kHz, while the pathological signals may have either 25 or 50 kHz, all with 16 bits per sample. In order to avoid potential correlation of the results with the sampling frequency, all 50 kHz signals used were resampled to 25 kHz. The duration of the normal signals ranges from 2 to 3 seconds, while that of the pathological ones from 0.4 to 1.4 seconds. This database will be referred to as “MEEI”.

3.2.2 Príncipe de Asturias database

The database from the Príncipe de Asturias (PdA) Hospital in Alcalá de Henares of Madrid [13] also consists of recordings of the sustained vowel /a/. The stored signals have the first and last part of the utterance removed to avoid onset and offset effects. 238 samples from normophonic speakers and 201 samples from dysphonic speakers, with a wide range of disorders, were used for our experiments. The labeling of the speakers was done from the clinical doctors of the hospital. All signals have a sampling frequency of 25 kHz, with 16 bits per sample, and their duration ranges from 1.5 to 4 seconds. This database will be referred to as “PdA”.

3.2.3 MEEI reading text database

The MEEI Disordered Voice Database, apart from sustained phonation recordings, also includes reading text recordings of the standard text “The Rainbow Passage”. These recordings are limited to 12 seconds, usually including up to the two first sentences of the text, which are the following:
“When the sunlight strikes raindrops in the air, they act as a prism and form a rainbow. The rainbow is a division of white light into many beautiful colors.”

For our experiments 53 signals from healthy voices and 660 signals from pathological voices were used. 683 of these signals have a sampling frequency of 25 kHz (36 normal and 647 pathological) and 30 have a sampling frequency of 10 kHz (17 normal and 13 pathological), all with 16 bits per sample. This database will referred to as “$\text{MEEI}_{\text{Rainbow}}$”.

3.3 Absolute jitter comparison

Absolute jitter is a widely implemented measurement of the period-to-period variability of pitch in time [27]

$$\text{absolute jitter} = \frac{1}{N-1}\sum_{n=1}^{N-1} |u(n+1) - u(n)| \tag{3.1}$$

where N is the total number of pitch periods and $u(n)$ is the pitch period sequence. This type of jitter estimation is implemented by two of the most established systems for acoustic voice quality assessment, the Praat [4] system and the Multi-Dimensional Voice Program (MDVP) [9]. Praat implements it as the *Jitter (local, absolute)* function, while MDVP provides the *Jita* analysis parameter. Both systems produce a single absolute jitter estimate for the whole input signal in μs. In the experiments with actual speech recordings, SJE was compared to the above methods. Since SJE produces a sequence of local estimates in samples, these measurements were first converted to μs accordingly, and then the average value of the new sequence was computed for the purpose of comparison.

3.4 Receiver Operating Characteristic analysis

The ability of a method to discriminate samples, from a given set, that belong to two different classes, based solely on a single score that is provided by the method for each sample, can be examined through Receiver Operating Characteristic (ROC) analysis [7]. The ROC curve of the method, that is the True Positive Rate (TPR) vs. False Positive Rate (FPR) curve, is determined by considering a variable discrimination threshold. This curve describes in essence the performance of all possible binary classifiers based on the examined method. The discriminative efficiency of a method can be then summarized in an accuracy index referred to as Area Under the Curve (AUC), which is the area under the ROC curve produced for the method. For the specific problem of two class discrimination, such as normal versus pathological voices, AUC is an index that is analogous of discrimination power. AUC is preferred over other measurements of discrimination performance, because it is free from any bias due to the size of the set of each class. The standard error of the AUC index provides additional information regarding its confidence interval [15].

Chapter 4

Spectral Jitter Estimator evaluation

The validity and performance of the Spectral Jitter Estimator (SJE) was evaluated using synthetic signals and databases of actual speech recordings. The results of the performed experiments are presented in this chapter. SJE has been initially validated using synthetic signals. Comparison with the Praat system and the Multi-Dimension Voice Program (MDVP) on the MEEI and PdA databases have shown that SJE is more discriminant than the other methods. Building on these results, a threshold for pathology has been established for SJE, through cross-database examination. Furthermore, since SJE is able to provide short-local estimations, short-time statistics of those have been also examined. Finally, experiments with reading text recordings from $\mathrm{MEEI}_{\mathrm{Rainbow}}$ have been conducted that led to the creation of new features especially for running speech.

4.1 Synthetic Signals

Applying the short-time Spectral Jitter Estimator (SJE) on synthetic signals, with prior knowledge of the actual pitch period, the results did confirm our theoretical observations. The structural pattern of the jittered impulse train is maintained on the final signal and thus the exact measurement of jitter, in a short-time fashion, is possible. The previously mentioned Fig. 2.5 shows a frame from a synthetic jittered phonation signal with sampling frequency of 48 kHz and pitch deviation of 5 samples (ϵ = 5). The number of counted intersections indeed corresponds to the value of ϵ, while two false intersections were correctly rejected using the 3dB threshold.

To assess the legitimacy of the synthetic signals used, and consequently that of SJE, Praat [4] was used as a reference system, comparing the absolute jitter estimates of the two methods. We remind that the measurement provided by Praat is a single estimate for the whole signal, while SJE computes a sequence of short-time values. Therefore, for the purpose of comparison the average value of this short-time sequence was used. The average value is also doubled and then converted from samples to μs, so that it is analogous to the absolute jitter units from Praat.

The error difference, between the theoretical jitter value and the estimates from SJE and Praat, for all the synthetic jitter signals, is presented in Fig. 4.1. As SJE has given precise local values, the zero error was to be expected. The results were the same using a frame size of either three or four times the pitch period. The error difference of Praat is in the order of some μs, for almost all different values of ϵ. For three cases at 48 kHz, Praat determined the signals as unvoiced and thus it did not provide a jitter measurement. Given that for the 16 kHz case the estimate quantum is 125 μs, and for the 48 kHz case it is 41.7 μs, the error produced by Praat can be considered negligible, and so the validity of the synthetic signals can be ascertained.

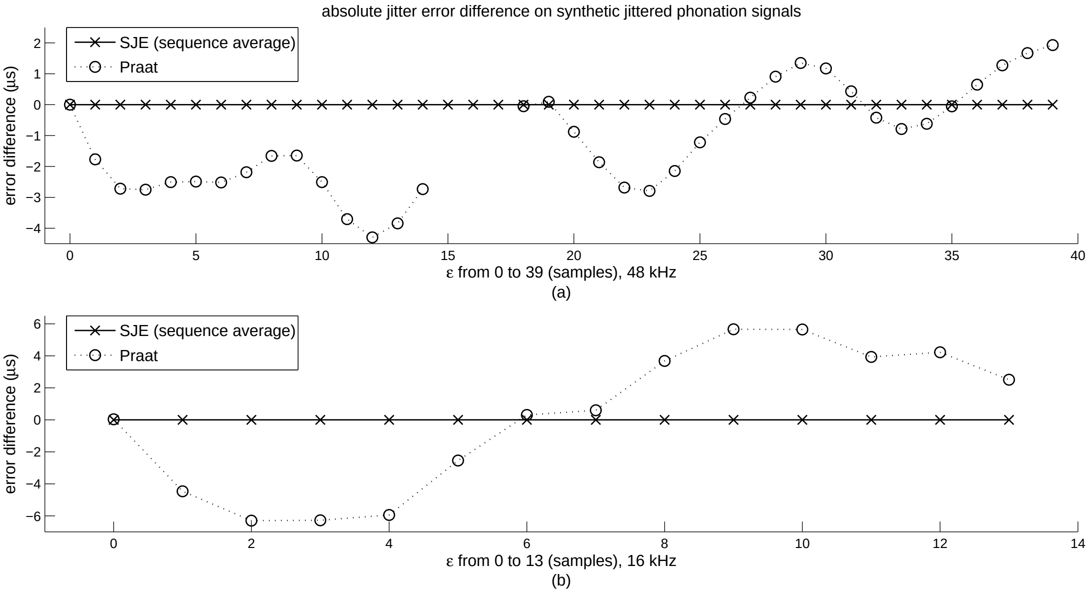

Figure 4.1: Absolute jitter error difference between theoretical value and estimation on synthetic jittered phonation signals with sampling frequency of (a) 48 kHz and (b) 16 kHz. The minimal error in the estimates of Praat verifies the validity of SJE, which presents zero error.

4.2 Databases of recordings

Further evaluation of Spectral Jitter Estimator (SJE) was performed using actual sustained phonation recordings from two databases. SJE was compared to absolute jitter measurements provided by Praat and MDVP. The initial pitch estimation required by SJE was taken from four sources:

- average of pitch period sequence computed by the YIN [5] pitch estimator ($\mathrm{P}_{\mathrm{YIN(average)}}$)
- local pitch period again estimated by YIN ($\mathrm{P}_{\mathrm{YIN(local)}}$)
- average pitch period estimated by Praat ($\mathrm{P}_{\mathrm{Praat(average)}}$)
- average pitch period estimated by MDVP ($\mathrm{P}_{\mathrm{MDVP(average)}}$)

We remind that for SJE the average value of the produced short-time sequence is used.

4.2.1 Massachusetts Eye and Ear Infirmary Disordered Voice Database

The Massachusetts Eye and Ear Infirmary (MEEI) Disordered Voice Database was used for experiments with actual recordings from healthy and pathological voices. The distributions of the absolute jitter measurement from the three different methods, for normal and pathological signals, are presented in Fig. 4.2. For SJE, the case depicted is the one where $P_{\text{MDVP(average)}}$ is taken as the initial pitch estimation, with a frame size of four times that. All methods have similar distributions, with slight differences in the produced values. Specifically, Praat has a smaller range for normal signals, while MDVP has twice the range of the other methods for pathological signals. This difference can be attributed to the fact that Praat is less sensitive to additive noise than MDVP [3]. In Fig. 4.3 the absolute jitter estimates from the three methods for each signal in MEEI are shown (the estimates are plotted after sorting the jitter estimations by MDVP). SJE gives average values which are in general larger than the estimates from MDVP and Praat. This is partly explained by the fact that SJE produces quantized values, described in (2.9), but the main reason is the different nature of the three methods. Although in theory all three methods measure absolute jitter, the implementations of MDVP and Praat depend too much on the notion of periodicity, while MDVP even takes steps to alleviate quasi-periodicity of speech through low-pass filtering [6]. Therefore, either explicitly or implicitly, these methods work on low-pass information and don't capture the full range of the jitter phenomenon. SJE, on the other hand, looks for the structural effects of jitter on the full spectrum.

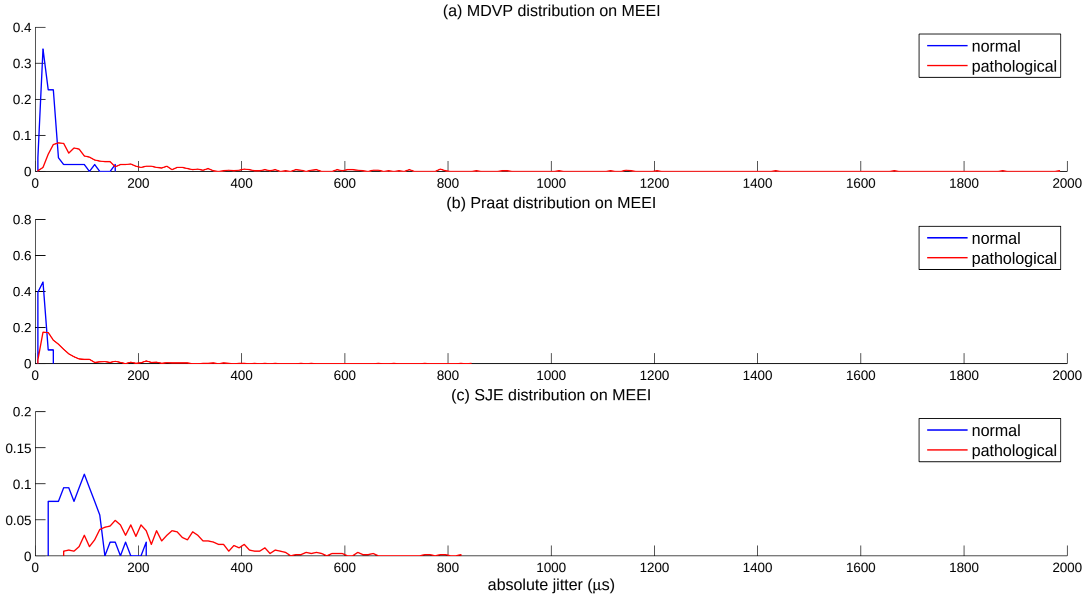

Figure 4.2: Distributions of the absolute jitter estimates for normal and pathological signals in MEEI, from three methods (a) MDVP, (b) Praat and (c) SJE. Praat has a smaller range of estimates for normal signals, while MDVP produces a larger range for pathological ones.

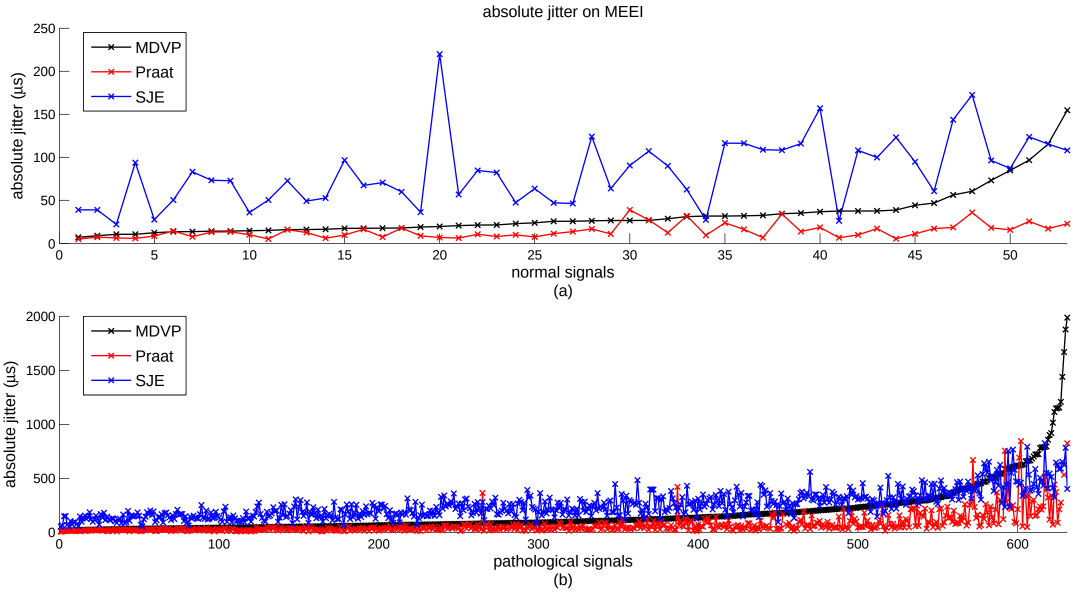

Figure 4.3: Absolute jitter estimates from three methods for (a) normal and (b) pathological signals in MEEI. The values are sorted by the estimation of MDVP for convenience. The SJE average is in general larger than the estimates of other absolute jitter methods.

Since a direct comparison between the results of the different methods is of no use, the Receiver Operating Characteristic (ROC) analysis of their ability to discriminate healthy from pathological voices allows for an indirect comparison. The ROC curves for the three methods in contest are portrayed in Fig. 4.4. Notice that SJE using four different pitch estimators has a steady performance in classification, as it is shown by the very similar ROC curves and Area Under the Curve (AUC) indexes. This provides proof that SJE is robust enough to perform consistently without relying so much to the initial pitch period estimation. Additionally, with an improvement of nearly 4% in the AUC index, our method is indeed more discriminant by both MDVP and Praat. When a frame size of three times the pitch period is used for SJE, the performance in discrimination doesn't change significantly. The AUC score and its standard error, for all the cases tested, are given in Table 4.1.

AUC (standard error) % for MDVP on MEEI				90.66 (1.42)
AUC (standard error) % for Praat on MEEI				90.47 (1.44)
AUC (standard error) % for SJE on MEEI				
Frame size	$P_{\text{MDVP(average)}}$	$P_{\text{Praat(average)}}$	$P_{\text{YIN(average)}}$	$P_{\text{YIN(local)}}$
Three times	94.73 (0.93)	93.09 (1.13)	94.68 (0.94)	91.70 (1.30)
Four times	94.82 (0.92)	93.17 (1.12)	94.77 (0.92)	91.37 (1.34)

Table 4.1: AUC score (and standard error) in % for all absolute jitter methods tested on MEEI. For SJE, cases with four different pitch estimators and two different frame sizes are presented.

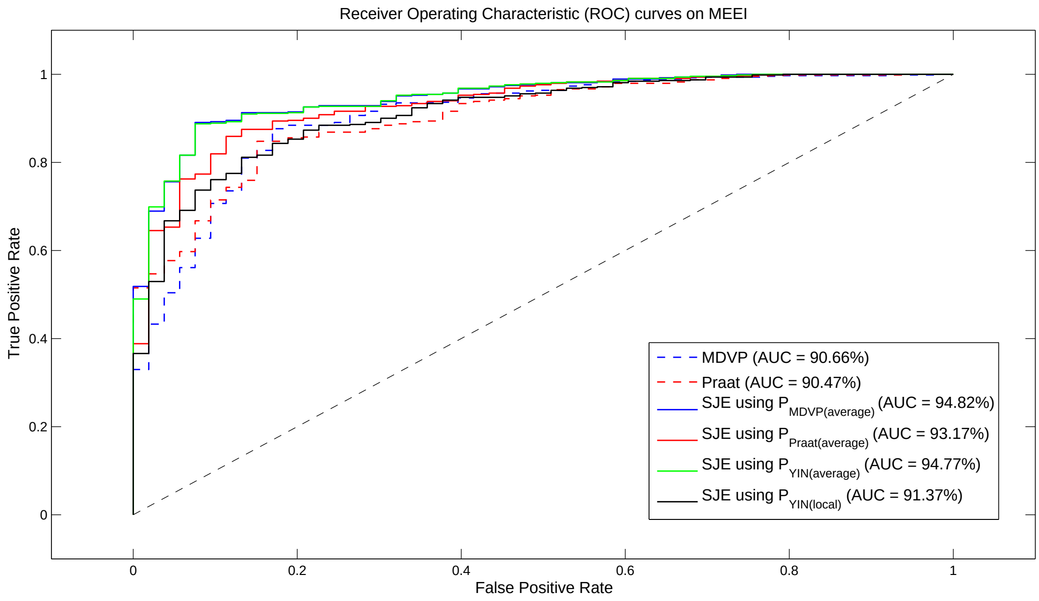

Figure 4.4: ROC curves for the absolute jitter methods tested on MEEI. SJE is the most discriminant among the three methods compared.

4.2.2 Príncipe de Asturias database

Additional experiments with actual recordings were performed using the Príncipe de Asturias (PdA) database. The distributions of the absolute jitter measurements for the two groups of signals are illustrated in Fig. 4.5. For SJE, the case shown is the one where $P_{\text{Praat(average)}}$ is used for the initial pitch estimation, with a frame size of four times that. Similarly to the results of the previous experiments, Praat has a smaller range of estimations for normal signals, while MDVP gives a larger range for pathological signals. It is worth mentioning that SJE provides absolute jitter measurements with a smaller overlap between the two classes as compared to the other two estimators. In Fig. 4.6, the estimations of jitter per signal from the three estimators are depicted (again the signals are sorted by their corresponding estimations from MDVP). It can be observed that for PdA, same as for MEEI, SJE provides larger values for the estimation of jitter.

When it comes to the discrimination ability between healthy and pathological voices, SJE is the most discriminant for this database as well. In Fig. 4.7 the Receiver Operating Characteristic (ROC) curves for classification of normal vs. pathological signals are depicted, using the absolute jitter results from Praat, MDVP, and SJE. For SJE four cases are presented with different initial pitch estimations. Praat is by far the least discriminant, while the SJE average is better by both Praat and MDVP in all cases. Using Praat's own pitch estimate our method provides an improvement in the area under the curve (AUC) index of more than 20% from Praat, and almost 14% from MDVP, while for the other pitch estimators the AUC index is also quite high. If we instead use a frame size of three times the pitch period, then SJE performs more or less the same. In Table 4.2 the AUC

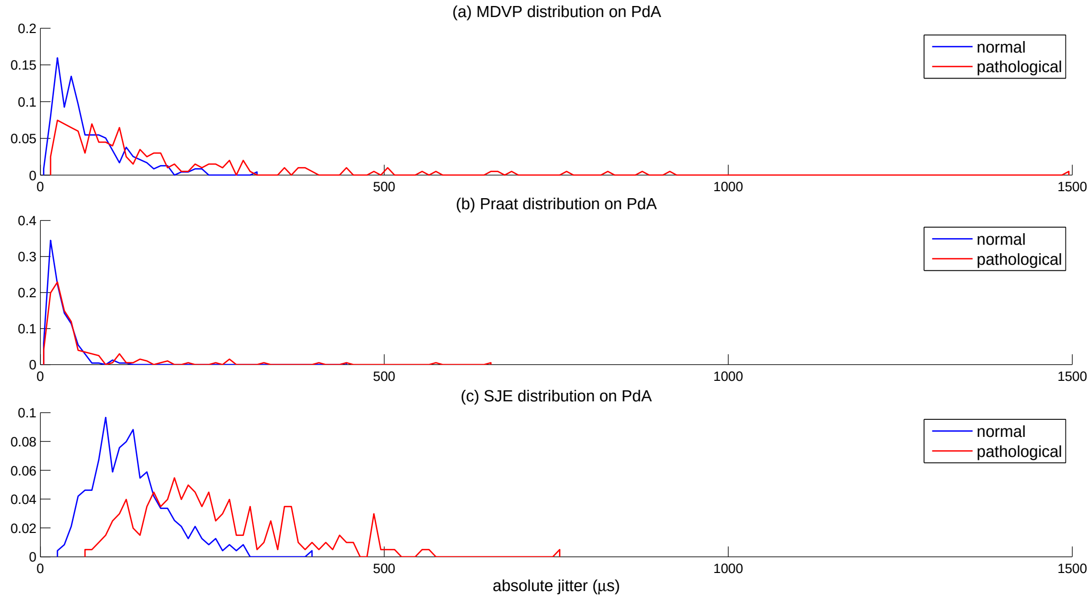

Figure 4.5: Distributions of the absolute jitter estimates for normal and pathological signals in PdA, from three methods (a) MDVP, (b) Praat and (c) SJE. The overlap between the distributions of the estimates for the normal and pathological signals is significantly smaller for SJE.

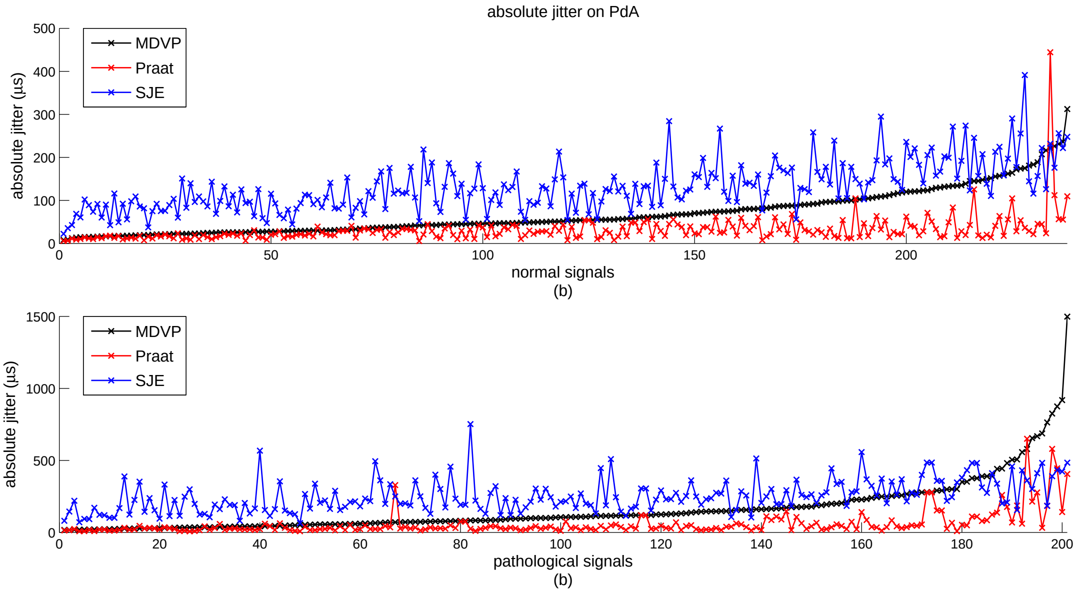

Figure 4.6: Absolute jitter estimates from three methods for (a) normal and (b) pathological signals in PdA. The values are sorted by the estimation of MDVP for convenience. The SJE average is again larger than the estimates of the other absolute jitter methods.

score and its standard error for all different experiments are shown.

AUC (standard error) % for MDVP on PdA				70.65 (2.50)
AUC (standard error) % for Praat on PdA				62.94 (2.67)
AUC (standard error) % for SJE on PdA				
Frame size	$P_{\text{MDVP(average)}}$	$P_{\text{Praat(average)}}$	$P_{\text{YIN(average)}}$	$P_{\text{YIN(local)}}$
Three times	79.73 (2.17)	84.10 (1.95)	78.46 (2.23)	77.76 (2.26)
Four times	79.71 (2.17)	84.65 (1.92)	78.13 (2.24)	77.61 (2.26)

Table 4.2: AUC score (and standard error) in % for all absolute jitter methods tested on PdA. For SJE, cases with four different pitch estimators and two different frame sizes are presented.

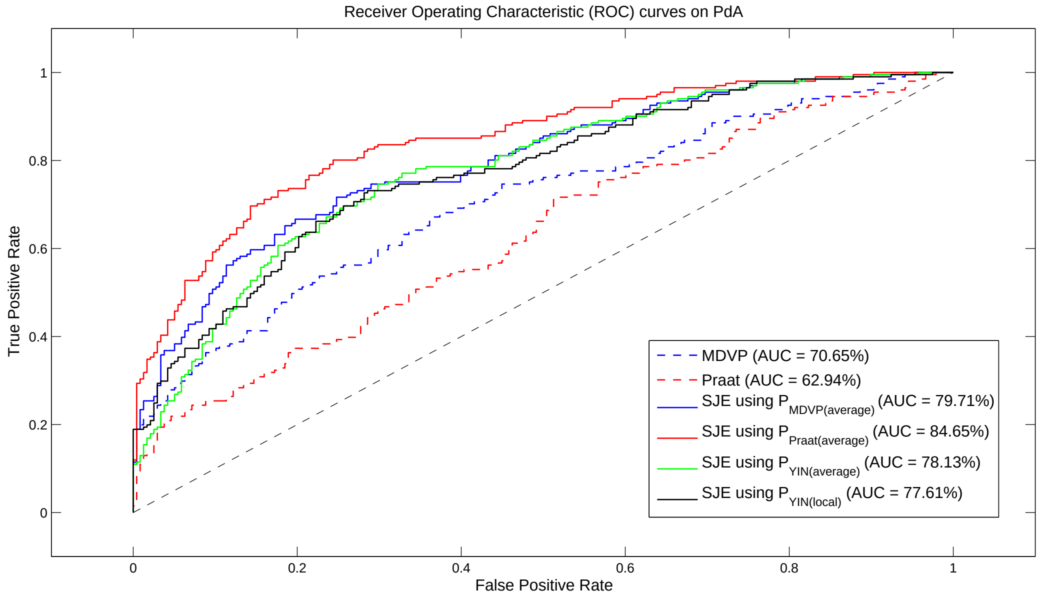

Figure 4.7: ROC curves for the absolute jitter methods tested on PdA. SJE is the most discriminant among the three methods compared for this database as well.

4.2.3 Intersections enhancement impact

It is interesting to examine what the estimates of SJE would be like without the enhancement of intersections described in section 2.2. In Figs. 4.8 and 4.9 the absolute jitter estimates from the application of SJE on MEEI and PdA, respectively, with and without enhancement are presented (the estimates are plotted after sorting the corresponding jitter estimations by MDVP, similarly to Figs. 4.3 and 4.6). The estimates of SJE without enhancement are significantly larger than those after the enhancement process. However, even with these flawed results, the AUC score is very high, specifically 96.64% with a 0.68% standard error for MEEI, and 83.33% with a 2.00% standard error for PdA. The reason is that while the spurious intersections occur in both healthy and pathological voices, they tend to appear more frequently in the pathological ones.

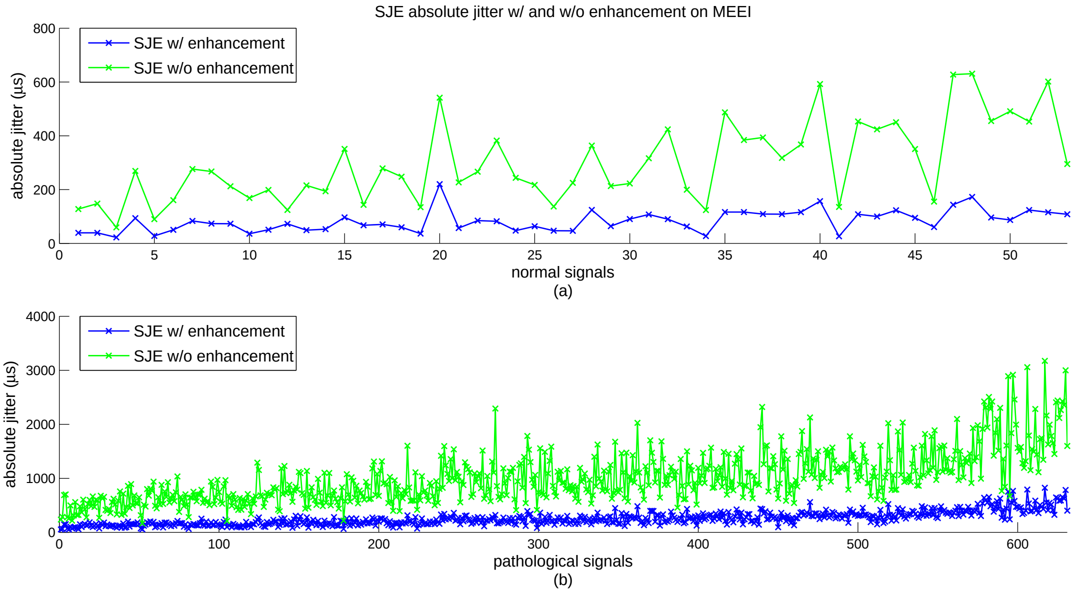

Figure 4.8: Absolute jitter estimates from SJE with and without enhancement, for (a) normal and (b) pathological signals in MEEI. The values are sorted by the corresponding estimation of MDVP for convenience.

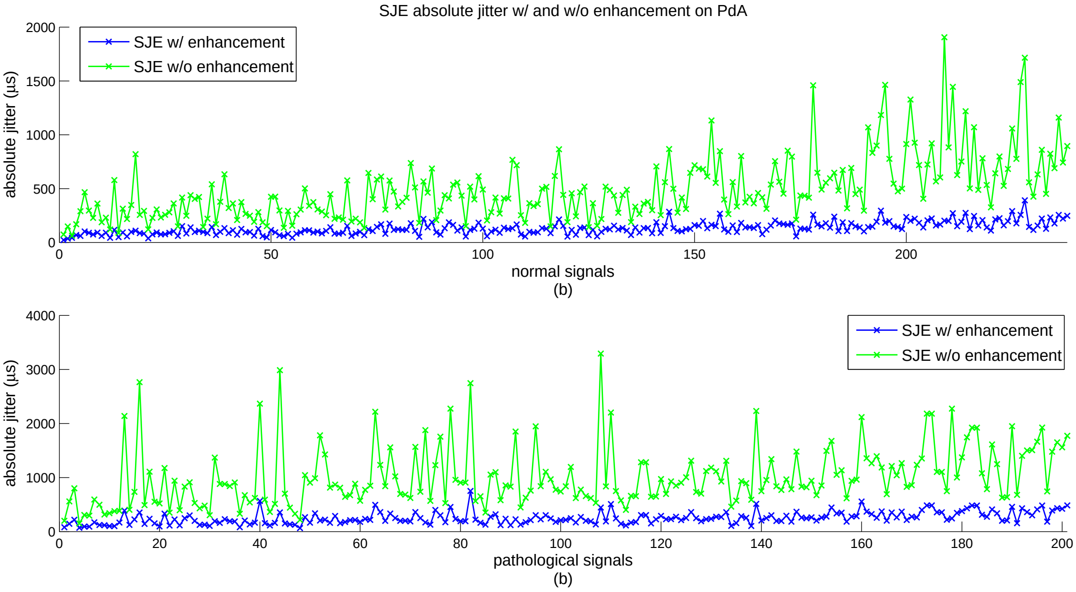

Figure 4.9: Absolute jitter estimates from SJE with and without enhancement, for (a) normal and (b) pathological signals in PdA. The values are sorted by the corresponding estimation of MDVP for convenience.

4.3 Pathology Threshold for SJE

We can take advantage of the results of the experiments presented in the previous section, in order to suggest a threshold for pathology when using SJE. Since two databases have been used, it is interesting to examine the consequences of using one database to determine the threshold, and then apply the result to the other database. This allows us to perform cross-database study comparisons which is quite rare in the literature of voice pathology detection. A threshold can be determined by taking into account the ROC curve for SJE, separately for each database, providing therefore two thresholds, one per database. Given the ROC curve, the discrimination instance that provides the best classifier is the one where the difference of the TPR to the FPR is the largest. For the case of SJE (with $\mathrm{P}_{\mathrm{MDVP(average)}}$ as the pitch period estimation) on the MEEI database , the largest difference is achieved when TPR=89.07% and FPR=7.55%, leading to a threshold of 124.24 (μs), which we will referred to as "$\mathrm{Thr}_{\mathrm{MEEI}}$". Similarly, the ROC curve regarding SJE (with $\mathrm{P}_{\mathrm{Praat(average)}}$ as the pitch period estimation) for the PdA database suggests a threshold of 161.08 (μs), when TPR=80.10% and FPR=24.37%. This threshold which will be referred to as "$\mathrm{Thr}_{\mathrm{PdA}}$". To compare the two thresholds a series of experiments were performed on the two databases. Initially, the classification rate (CR) was measured, which is the number of correct detections from both classes divided by the total number of detections, using each threshold.

Since SJE provides a short-time sequence of jitter values for each signal, three new features, that make use of the thresholds presented above, were also calculated. Having in mind that each short-time value corresponds to an analysis frame, then the three features are defined as

- the percentage of frames that are over the threshold,
  referred to as "Over",
- the maximum number of consecutive frames that are over the threshold,
  referred to as "Max Over", and
- the maximum number of consecutive frames that are under the threshold,
  referred to as "Max Under".

The three features are based on frames rather than time, since for each signal all frames were equal in size, because the analysis window per signal was determined by the average pitch period of the signal, as four times this value, and also a fixed step size was used, equal one time the same value. Consequently, the AUC index for these three features, for each threshold and for each database, was calculated. All the results are summarized in Table 4.3. It is interesting to add that the threshold of 83.20μs provided by MDVP [6] for its own implementation, offers a classification rate of 60.23% for MEEI and 64.46% for PdA, both lower than those provided by $\mathrm{Thr}_{\mathrm{MEEI}}$ (89.33% for MEEI and 67.88% for PdA) and $\mathrm{Thr}_{\mathrm{PdA}}$ (75.15% for MEEI and 77.68% for PdA) in the case of SJE. As it was expected, the threshold which was defined in a specific database provides the best classification score for that database. Therefore, $\mathrm{Thr}_{\mathrm{MEEI}}$ gives a better classification score for the MEEI database, while in PdA the best classification score is obtained by $\mathrm{Thr}_{\mathrm{PdA}}$. However, using $\mathrm{Thr}_{\mathrm{MEEI}}$ provides, in general, better results than $\mathrm{Thr}_{\mathrm{PdA}}$. Given also that it represents a low FPR of 7.55%, for all the following experiments $\mathrm{Thr}_{\mathrm{MEEI}}$ has been used.

Thr_{MEEI} 124.24 (μs)				
database	CR %	Over AUC %	Max Over AUC %	Max Under AUC %
MEEI	89.33	94.52 (0.96)	82.97 (2.22)	96.48 (0.70)
PdA	67.88	83.93 (1.96)	81.98 (2.07)	79.62 (2.18)
Thr_{PdA} 161.08 (μs)				
database	CR %	Over AUC %	Max Over AUC %	Max Under AUC %
MEEI	75.15	92.79 (1.17)	81.44 (2.36)	97.50 (0.55)
PdA	77.68	83.86 (1.97)	79.14 (2.20)	81.28 (2.10)

Table 4.3: Cross-database evaluation of thresholds determined by SJE in terms of Classification Rate (CR), and AUC with its standard error for number of frames which are over a threshold (Over), maximum consecutive frames that are over a threshold (Max Over), and maximum consecutive frames that are under a threshold (Max Under).

4.4 Short-time statistics

The local estimates produced by SJE have interesting statistical properties. These may be examined to gain insightful information on the difference between healthy and pathological voices, or even between different kinds of pathological disorders. Since all signals used in the experiments are of the same sampling frequency 25 kHz, local absolute jitter value in μs can be used for comparison between the two databases. In Fig. 4.10 and 4.11 the average distribution of the short-time SJE estimation for the MEEI and PdA databases, respectively, reveal a certain consistency for the two cases of normal and pathological signals. The distribution of the estimates for pathological signals has a larger variance, with a peak around 200 μs, while that for normal signals concentrates on small values of absolute jitter, with a peak around 100 μs.

Regarding the temporal behavior of jitter, a study of the transitions of the local jitter value from one frame to the next is of interest. The average distribution of the transition from one short-time SJE estimate to the next, for normal and pathological signals, from the MEEI and PdA databases, are shown in Fig. 4.12. The main diagonal in the case of healthy voices indicates that areas where jitter has a stable very small value, mostly from 0 to 2 samples of jitter, dominate the short-time sequence. Pathological signals seem to also contain such areas, where the local jitter value is constant, between 1 and 4 samples of jitter In general, for normal voices the variance of the distribution is small, with a peak located near the origin. For pathological voices the distribution has a noticeably larger variance, including more steep transitions in the value of jitter from frame to frame.

4.5 Reading Text Experiments

Jitter analysis is preferably performed on sustained vowels, because during phonation the radiated speech signal is expected to be quasi-periodic and therefore in the presence of jitter the aperiodic-

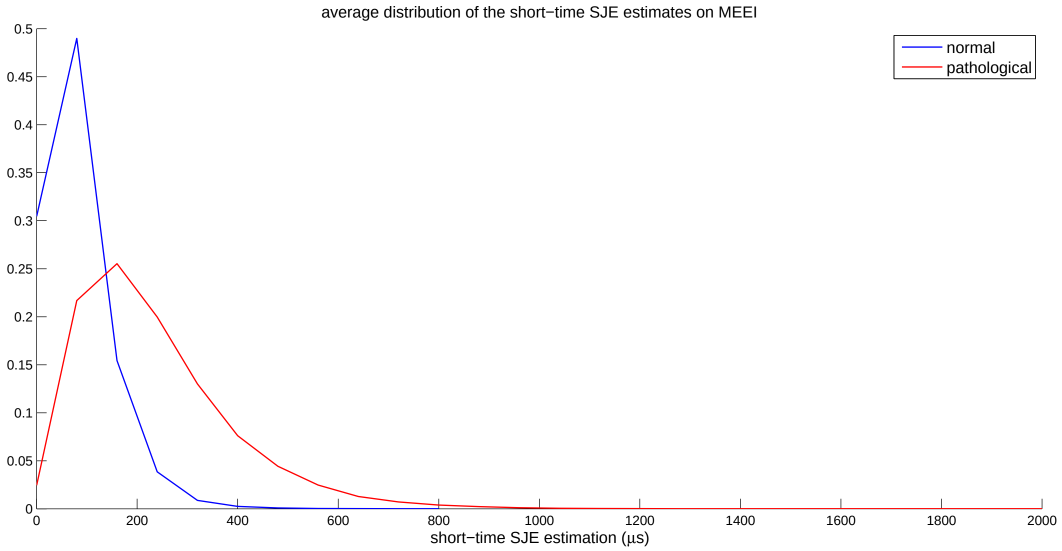

Figure 4.10: The average distribution of the short-time SJE estimates, for normal and pathological signals on MEEI, reveals that pathological samples have a larger variance.

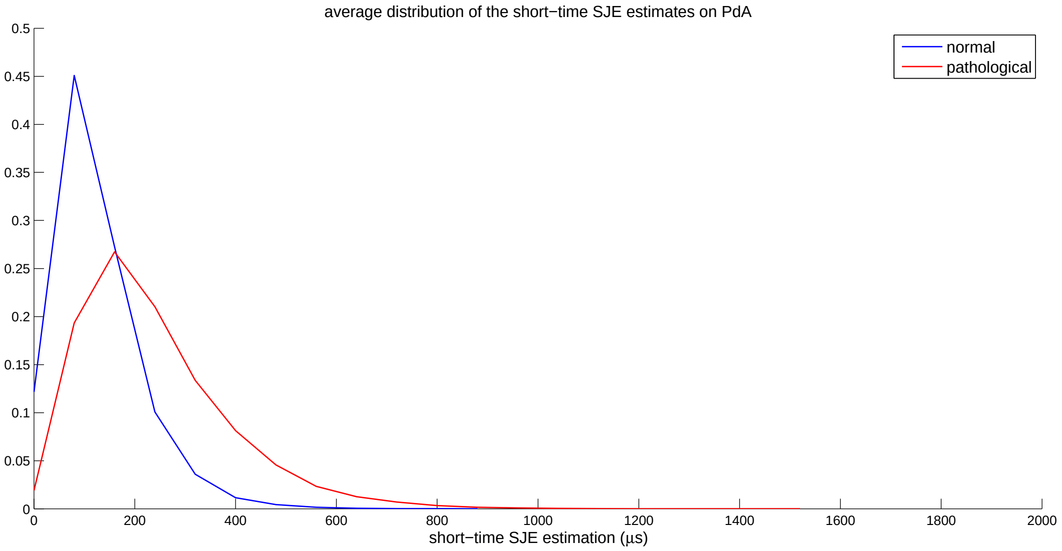

Figure 4.11: The average distribution of the short-time SJE estimates, for normal and pathological signals on PdA, is similar to that of MEEI, indicating the consistency of the short-time statistics.

ities that occur are more easily perceived. However, sustained phonation recordings are limited by nature to a small duration. After the first few seconds of voicing, pathological speakers may feel discomfort, while even healthy speakers may not be able to maintain a steady voice. To consider the behavior of jitter for a larger period of time recordings of reading text may be used. Speakers reading a text with a normal pace are able to breath occasionally, while in sustained phonation a single intake of breath is involved. This allows us to attain longer recordings for examination and

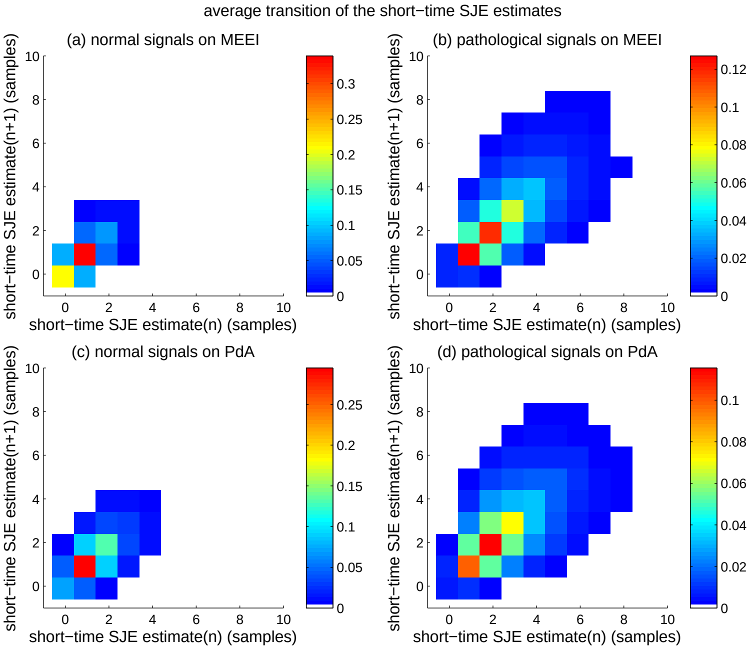

Figure 4.12: Average distribution of the transition of the short-time SJE estimates for (a) normal and (b) pathological signals on MEEI, and (c) normal and (d) pathological signals on PdA. Healthy voices are mostly characterized by large areas of constant small jitter, while the pathological ones have a larger variance in the difference from one frame to the next.

since SJE provides a short-time sequence of jitter estimates, it is ideal for the examination of jitter in running speech signals.

Using a 10 ms interval from frame to frame, an autocorrelation-based pitch estimator was employed to determine the local pitch period of all voiced frames [33], for each recording in the $\text{MEEI}_{\text{Rainbow}}$ database. To eliminate any onset and offset effects in the voiced areas, any voiced frames that don't have at least two voiced neighboring frames in each direction were disregarded, along of course with the unvoiced frames. For the remaining voiced frames, referred to as "valid frames", the short-time Spectral Jitter Estimator (SJE) was used to measure the local absolute jitter value, using a frame size of four times the local pitch period.

An initial examination of the local estimates from SJE shows that these are in concordance with documented statistical behavior. It is expected that on average jitter decreases with increasing fundamental frequencies [16] [18] [17] [30]. We verified this expectation by calculating the correlation coefficient between estimated jitter and fundamental frequency, with confidence intervals at 95%. Specifically, we found a correlation of -73.89% for the normal signals, -71.32% for the pathological signals, and -84.33% for the database in whole. In Fig. 4.13 the average absolute jitter per fundamental frequency, for frequencies between 80 and 400 Hz, is illustrated, for the two classes of normal and pathological voices.

The sequence of local SJE estimates used to calculate several features that reflect the average

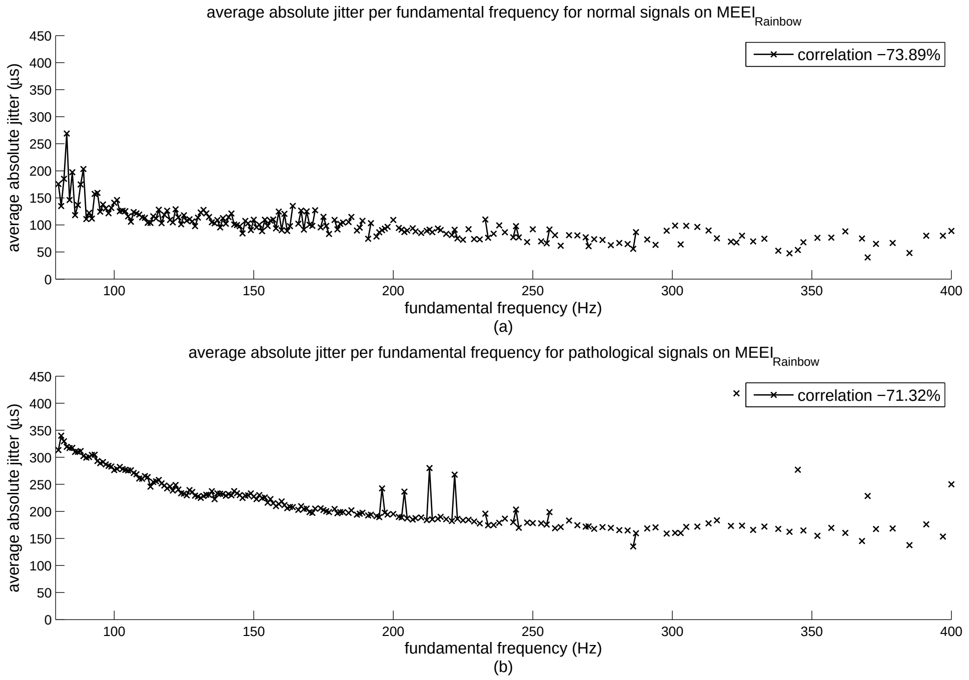

Figure 4.13: Average absolute jitter from SJE as a function of fundamental frequency, for (a) normal and (b) pathological signals from $\text{MEEI}_{\text{Rainbow}}$.

and short-time behavior of jitter. The average value of absolute jitter is only computed here for comparison purposes. Specifically, if $j(n)$ is the aforesaid sequence with length N, then for each signal the following features were computed:

- The average absolute jitter from all valid frames which, referred to as "Jit Mean".
$$\text{Jit Mean} = \frac{\sum\limits_{n=1}^{N} j(n)}{N} \ (\mu s)$$

- The percentage of valid frames that have an absolute jitter value over Thr_{MEEI}, referred to as "*Over*".
$$\textit{Over} = 100\frac{|\{j(n) : j(n) > \text{Thr}_{\text{MEEI}}\}|}{N} \ (\%)$$
where $|A|$ denotes the cardinality of A, or otherwise the number of elements in the A set.

- The maximum number of consecutive valid frames that have an absolute jitter value over Thr_{MEEI}, referred to as "*Max Over*".
$\textit{Max Over} = \max(|\{j(n) : j(n) > \text{Thr}_{\text{MEEI}} \text{ and consecutive frames}\}|)$ (scalar)

- The maximum number of consecutive valid frames that have an absolute jitter value under Thr_{MEEI}, referred to as "*Max Under*".
$\textit{Max Under} = \max(|\{j(n) : j(n) \leq \text{Thr}_{\text{MEEI}} \text{ and consecutive frames}\}|)$ (scalar)

Note that there is no need to convert values that represent a number of frames to time units, because the fixed interval used from frame to frame makes them equivalent. The short-time absolute jitter estimation for two signals from $\text{MEEI}_{\text{Rainbow}}$, one normal and one pathological, are illustrated in Fig. 4.14. In the same figure, the threshold Thr_{MEEI} is also depicted (dashed line). It is worth observing the number of frames that are over this threshold in the case of the pathological signal compared to the corresponding number of frames for the normal signal. More than 80% of the valid frames are over the threshold in the case of pathology, while only 13% of the valid frames are above the same threshold for the normal case. The *Max Over* and *Max Under* intervals for each signal are also indicated in Fig. 4.14 by arrows. Specifically, for this example of pathologic voice, 11 consecutive valid frames are under the threshold, while 33 consecutive valid frames are above the threshold. It is evident that the suggested threshold Thr_{MEEI} does separate correctly the majority of the local jitter estimates. The AUC indexes for the aforementioned features are given in Table 4.4. All cases show very good discriminant ability with an AUC index over 90%.

AUC (standard error) % of features on $\text{MEEI}_{\text{Rainbow}}$ using Thr_{MEEI}			
Jit Mean	*Over*	*Max Over*	*Max Under*
96.26 (0.72)	95.69 (0.80)	93.32 (1.10)	91.61 (1.30)

Table 4.4: AUC score in % for the four features based on the SJE short-time sequence, on the classification of reading text recordings from $\text{MEEI}_{\text{Rainbow}}$ to normal and pathological voices.

Since we have based the above features on a sequence of short-time jitter estimations, we are able to examine their gradual development in time, in terms of value and discrimination. In the following, when we apply a feature gradually in time using a sliding analysis window of fixed size, we will refer to it then as “local”. If instead we apply a feature using an analysis window that starts from the origin and its duration is gradually extended up to the current time instant, then we will refer to it as “running”.

To further investigate, and to some extend visualize the above results regarding the AUC scores, the running average number of frames that are over the threshold Thr_{MEEI} for normal and pathological voices was computed by analyzing all the $\text{MEEI}_{\text{Rainbow}}$ database. To understand the function of the running average consider as an example a running analysis window of 2 seconds duration for the case of the normal class of speakers. Then, for each recording in this class the number of frames that are over the threshold in the current analysis window, from 0 to 2 seconds, is counted. If L is the number of recordings, obviously L values are computed. The running average number of frames over the threshold for the current window is obtained by calculating the average of these L values. After that, the running window is increased by 0.5 seconds, covering now the time interval from 0 to 2.5 seconds, and the corresponding running average is again computed. This procedure is repeated until the running window spans the whole duration of the signal (12 seconds). In Fig. 4.15 the running average of frames that are above the threshold Thr_{MEEI} are depicted, for both the nor-

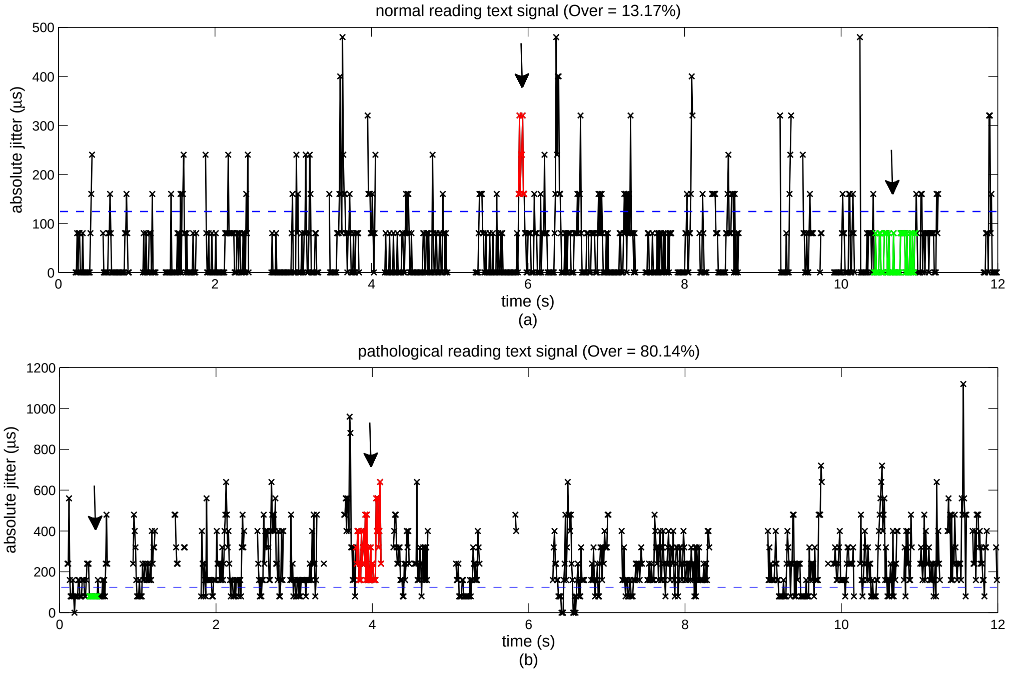

Figure 4.14: Short-time absolute jitter estimates from SJE for (a) one normal and (b) one pathological signal from $\mathrm{MEEI}_{\mathrm{Rainbow}}$. Notice how the pathology threshold $\mathrm{Thr}_{\mathrm{MEEI}}$, indicated by the dashed line, is applied on the local SJE estimates. Arrows indicate the maximum number of consecutive frames that are below or above the threshold for each signal.

mal and pathological voices. The running analysis window starts from 1 second and reaches up to 12 seconds. In fact, this running average is equivalent to an average accumulator of the number of pathological frames. Therefore, as it is expected the computed values are monotonously increasing. It is worth noting that for all running windows, the values computed for the pathological voices are always higher than the values for the normal voices. More interesting, the increase rate of the pathological class is much higher than the corresponding increase rate of the normal one. In Fig. 4.16, the normalized per analysis duration running averages are depicted. Since the hop size of the jitter estimation is constant and equal to 10 ms, it means that there are 100 frames per 1 second, considering both voiced and unvoiced frames. Hence, the numbers shown in the ordinate axis of Fig. 4.16 can be interpreted as percent. We observe that on average a bit less than 25% of frames in the normal signals may be above the threshold for pathology, while for pathological signals about 45% of the frames may be above the threshold. Considering short running windows, for example 2-3 seconds in the case of short phonation, the number of frames that are above the threshold are reduced (just above 15%) in the case of normal voices, while for the pathological voices the corresponding number of frames remains about the same (45%).

In Fig. 4.17 the running average *Max Over* and *Max Under* values, for both normal and pathological signals, are depicted. As it was explained before, we calculated the average *Max Over* and

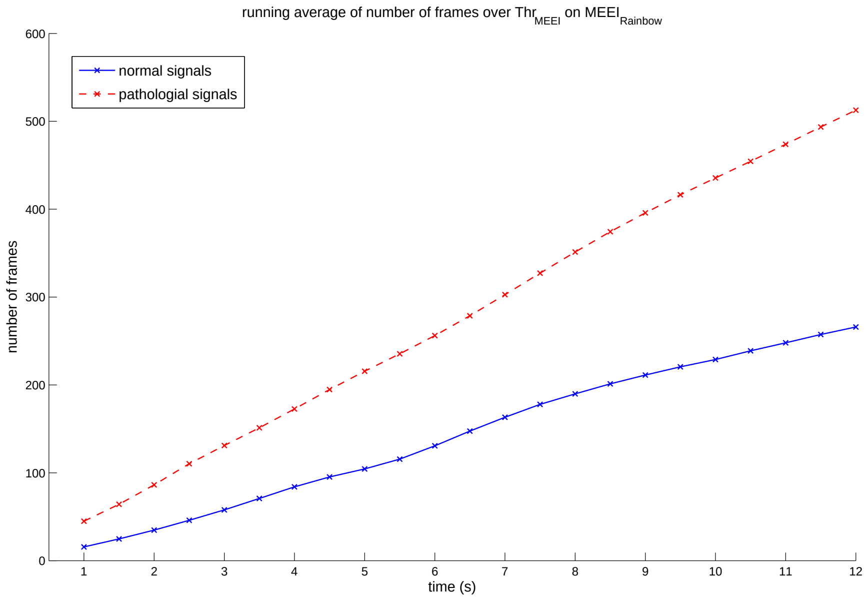

Figure 4.15: Running average number of frames over Thr$_{\mathrm{MEEI}}$, for normal and pathological signals on MEEI$_{\mathrm{Rainbow}}$. The horizontal axis denotes the length of the analysis window in seconds.

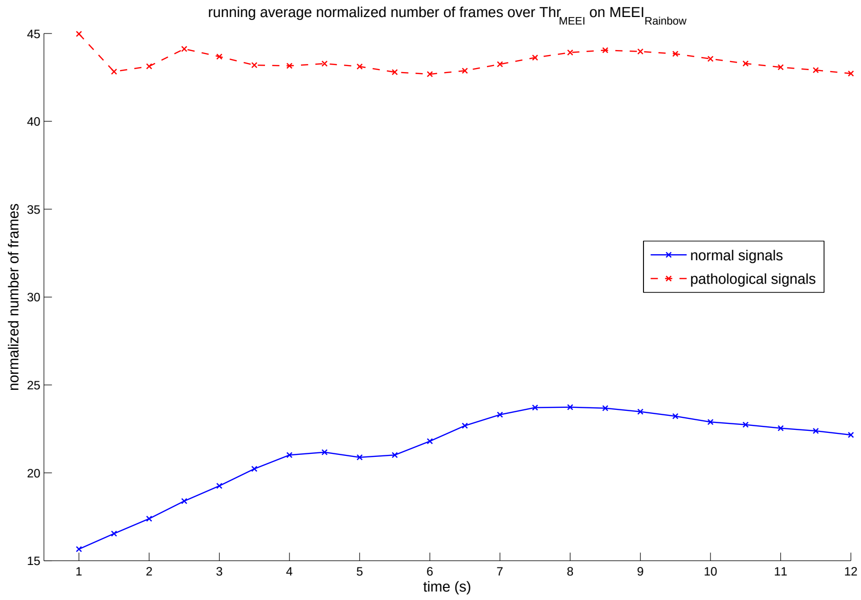

Figure 4.16: Normalized running average number of frames over Thr$_{\mathrm{MEEI}}$, for normal and pathological signals on MEEI$_{\mathrm{Rainbow}}$. The horizontal axis denotes the length of the analysis window in seconds.

the average *Max Under* for the two classes of recordings, starting from the first second and incrementing by half a second, until the full 12 seconds length was reached. It can be observed that for normal signals the related *Max Under* feature rises with a higher rate than *Max Over*. For pathological signals similar behavior is noted for the *Max Over* value, which increases more rapidly than the corresponding *Max Under* value. While the other two values (*Max Over* for normal and *Max Under* for pathological voices) also rise in the first seconds, they do so with a smaller rate (than *Max Over* for pathological and *Max Under* for normal voices), and they both stabilize after the 8 seconds mark. In a similar fashion, the running AUC scores of the four features are presented in Fig. 4.18. *Jit Mean* and *Over* reach stability very early while they are quite high from the beginning. *Max Over* and *Max Under* on the other hand start with a lower AUC and fluctuate more, while they follow closely the trend of the average pathological *Max Over* and the average normal *Max Under* in Fig. 4.17, respectively.

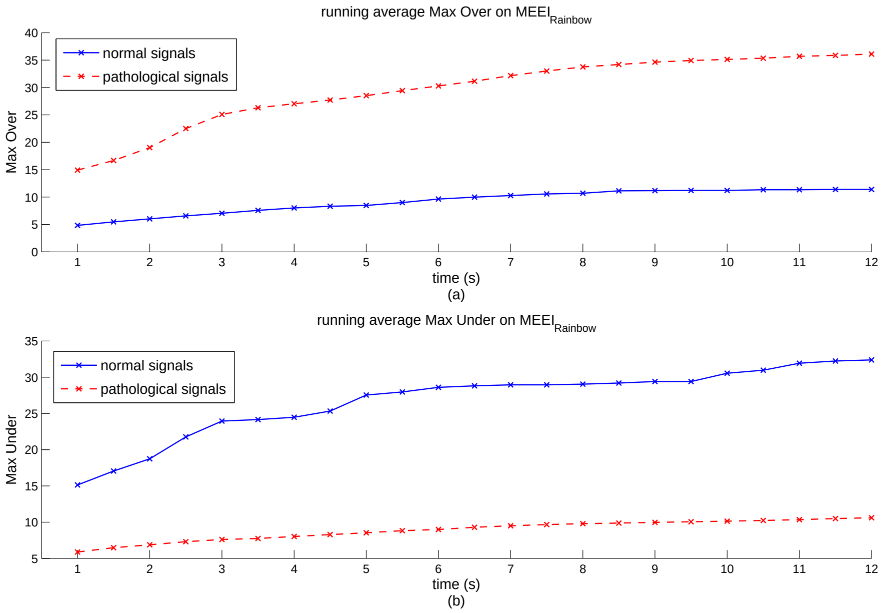

Figure 4.17: Running average of the (a) *Max Over* and (b) *Max Under* values for normal and pathological signals on $\mathrm{MEEI}_{\mathrm{Rainbow}}$.

Among the short-time features examined (*Over*, *Max Over*, and *Max Under*), the *Over* feature, that is the percentage of frames with a local absolute jitter value over the $\mathrm{Thr}_{\mathrm{MEEI}}$ threshold, has the best performance regarding discrimination. As it is also shown in Fig. 4.18, this is true even for signals of a small duration. Based on these results, it was investigated if *Over* could be used to establish another threshold for pathology, especially for recordings of reading text. Specifically, given that a threshold of pathology for SJE estimates is already selected (i.e., $\mathrm{Thr}_{\mathrm{MEEI}}$), another threshold for pathology could be set by computing the minimum value of *Over* that is required to indicate a speech segment as pathological. In this way, it is possible to monitor the jitter estimations during continuous speech (i.e., spontaneous speech). The FPR, TPR and threshold that correspond to the best classifier of the *Over* feature, as this evolves over time, are illustrated in Fig. 4.19. For

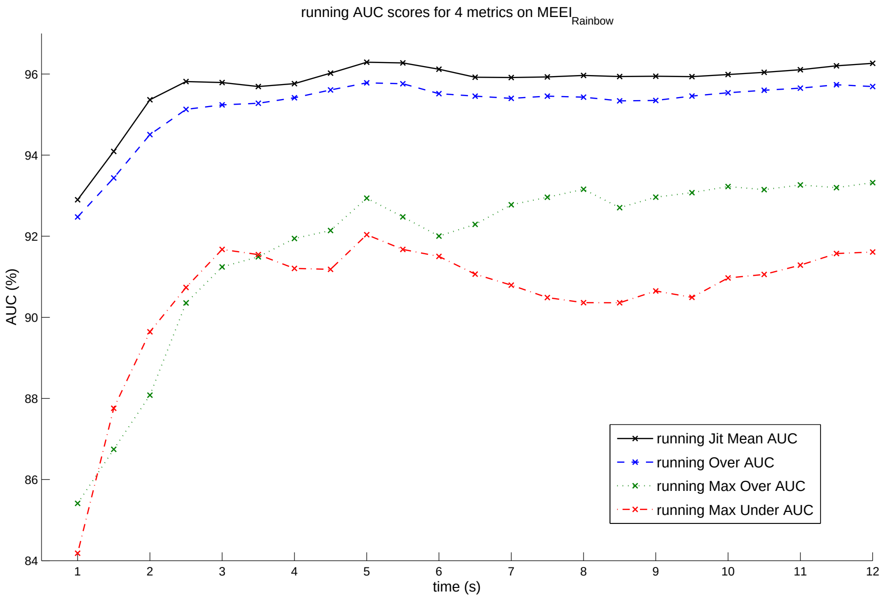

Figure 4.18: Running AUC score of the four features applied on MEEI$_{\text{Rainbow}}$ database. The *Jit Mean* and *Over* features are quite discriminant even from the third second.

example, for a running analysis window of 1.5 seconds, the best classifier, that is the one having the highest distance between FPR and TPR, has a threshold of about 48% which corresponds to an TPR of 80% and an FPR of 4%. In the last 3 seconds, FPR settles on 7.55% and TPR around 89.5%, while the threshold for *Over* ranges from 47.5% to 50%. We propose the use of 45% as a maximum threshold of *Over* for normophonia and 50% as a minimum threshold of *Over* for dysphonia. These limits, will be denoted by Thr$_{\text{Over}}$. The region in between should be considered as an indeterminate area that indicates the need for further information regarding voice quality assessment. When Thr$_{\text{Over}}$ is applied to MEEI$_{\text{Rainbow}}$ a classification rate of 87.80% is gained, with an additional 3.65% (26 files in total, 24 pathologic and 2 normal) classified in the indeterminate area.

An example of the potential use of Thr$_{\text{Over}}$ is presented in Figs. 4.20 and 4.21, for one normal and one pathological signal. The running *Over* percentage for the two signals is shown in Fig. 4.20, while in Fig. 4.21 the corresponding local *Over* percentage is illustrated. The local *Over* is computed using a sliding window of 1 second duration shifted by half a second. For the particular normophonic signal, while the running *Over* feature is under Thr$_{\text{Over}}$ almost exclusively (Fig. 4.20(a)), in the local *Over* plot, it exceeds the threshold of pathology in some intervals (Fig. 4.21(a)). Nonetheless, it does remain in the normal region by majority. For the pathological signal the remarks are alike. While it is clearly in the pathological region from early on regarding the running *Over* feature (Fig. 4.20(b)), in the local *Over* estimates, it lies under the normal threshold for a few intervals only (Fig. 4.21(b)). Hence, for the running *Over*, a recording of at least

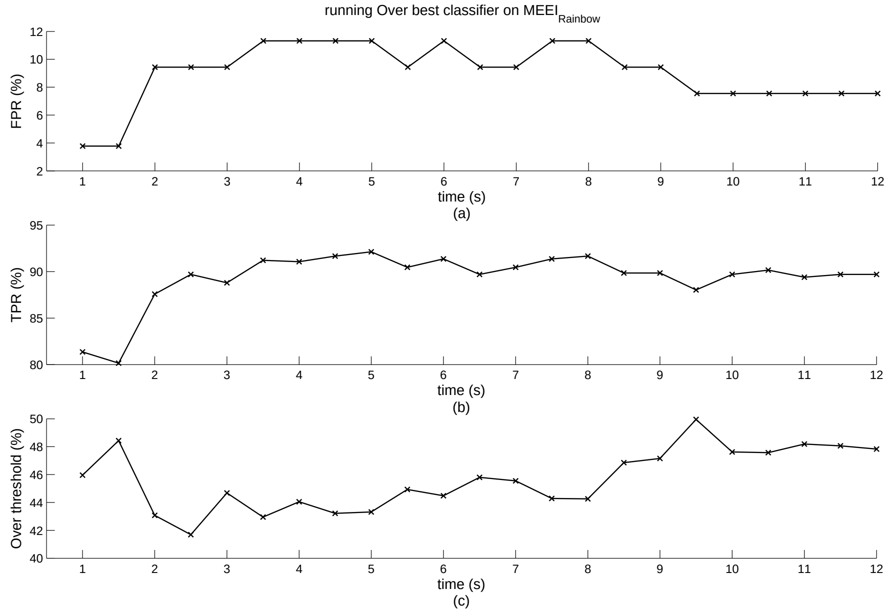

Figure 4.19: (a) FPR, (b) TPR and (c) optimum threshold for the best classifier based on the *Over* feature, when applied on MEEI$_{\text{Rainbow}}$, as a function of time.. It is reminded that FPR and TPR stand for False Positive Rate and True Positive Rate, respectively.

several seconds should be used, so that there are sufficient statistics for the estimation to converge to a specific region without a doubt. Similarly, when we consider the local *Over* feature, we should use an interval of adequate length. It is worth mentioning that the fluctuation of local *Over* feature as shown in Fig. 4.21 corresponds to intervals where there is a short rest of phonation. Therefore, just after these areas the local *Over* feature tends to decrease.

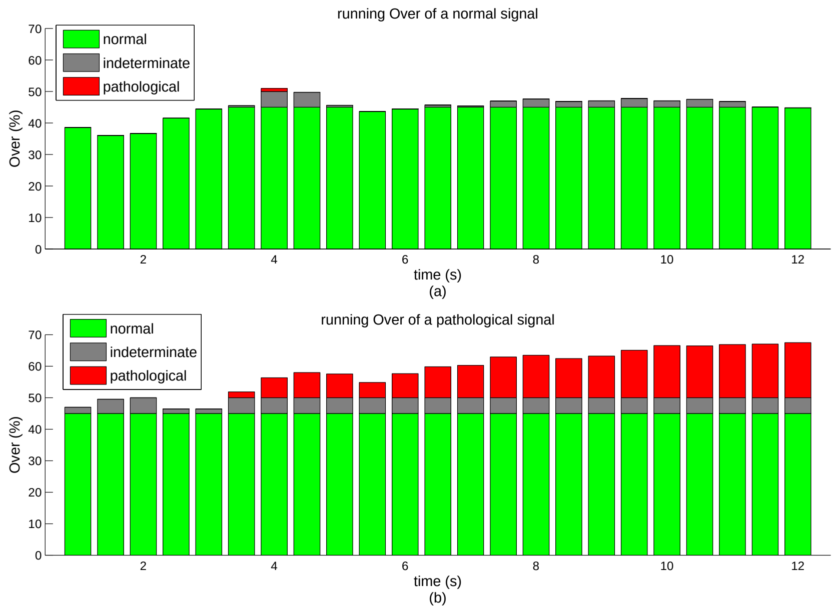

Figure 4.20: Running *Over* estimate example for (a) one normal and (b) one pathological reading text signal from $\text{MEEI}_{\text{Rainbow}}$. Given enough time the estimate settles in the normal or the pathological region.

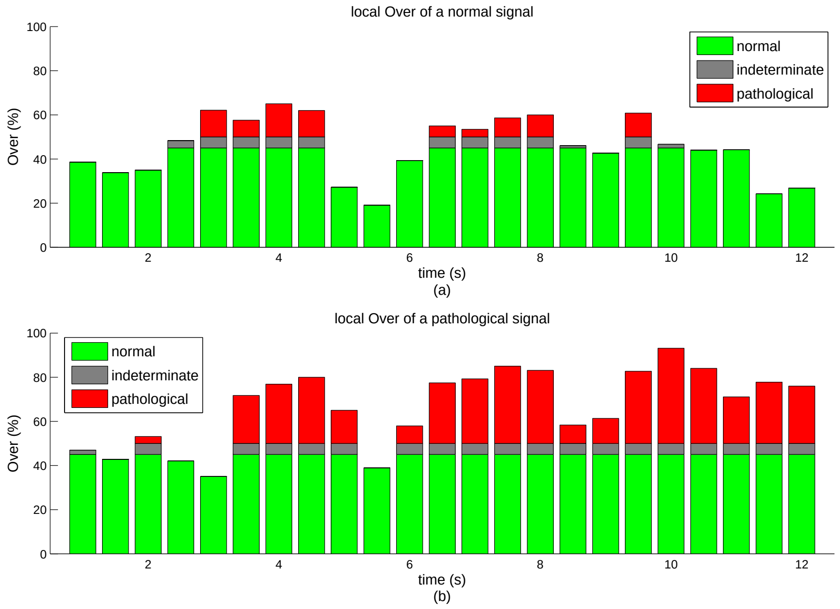

Figure 4.21: Local *Over* estimate example for one normal and one pathological reading text recording from $\text{MEEI}_{\text{Rainbow}}$. The local *Over* value is computed using a sliding window of 1 second duration shifted by half second.

Chapter 5

Shimmer modeling

In this chapter a mathematical model of shimmer is presented, that, like the model suggested for jitter, has interesting spectral properties. Based on these properties four features for the detection of pathology are developed and used as a means of evaluation of the model.

5.1 Mathematical model of shimmer

Shimmer may be modelled on an impulse train that consists of two periodic events, as a perturbation of the amplitude of the pulses. A simple mathematical model may be obtained by considering a cyclic perturbation, with a deviation of a constant value, applied alternately to either increase or decrease the amplitude of each pulse [28](pgs.102-103). The shimmered impulse train can be expressed then as

$$g_s[n] = A(1+\Delta)\sum_{k=-\infty}^{+\infty}\delta[n-(2k)P] + A(1-\Delta)\sum_{k=-\infty}^{+\infty}\delta[n-(2k+1)P] \qquad (5.1)$$

where A is an amplitude modifier, P is the pitch period in samples, and Δ is the shimmer amplitude deviation. This model, depicted in Fig. 5.1, combines two periodic events with Δ being the ratio that characterizes the local aperiodicity of shimmer. The value of Δ can range from 0 (no shimmer) to 1 (pitch halving). By convolution of the glottal signal over one glottal cycle with such a shimmered impulse train, the glottal airflow signal under the presence of shimmer can be modelled.

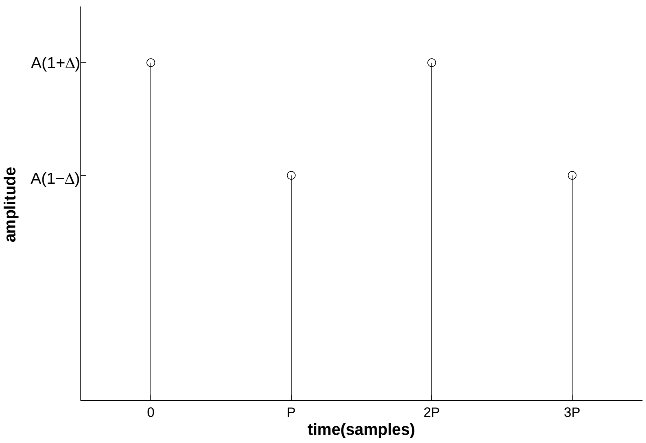

Figure 5.1: Shimmered impulse train of the two event model for shimmer.

The Fourier transform of the cyclically shimmered impulse train (5.1) is

$$\begin{aligned}
G_s(\omega) &= \sum_{n=-\infty}^{+\infty} g_s[n]e^{-j\omega n} \\
&= \sum_{n=-\infty}^{+\infty} \left(A(1+\Delta) \sum_{k=-\infty}^{+\infty} \delta[n-(2k)P] + A(1-\Delta) \sum_{k=-\infty}^{+\infty} \delta[n-(2k+1)P] \right) e^{-j\omega n} \\
&= A(1+\Delta) \sum_{k=-\infty}^{+\infty} e^{-j\omega 2kP} + A(1-\Delta) \sum_{k=-\infty}^{+\infty} e^{-j\omega(2k+1)P} \\
&= A\left[(1+\Delta)+(1-\Delta)e^{-j\omega P}\right] \sum_{k=-\infty}^{+\infty} e^{-j\omega 2kP} \\
&= A\left[(1+\Delta)+(1-\Delta)e^{-j\omega P}\right] \sum_{k=-\infty}^{+\infty} \frac{2\pi}{2P}\delta\left(\omega - k\frac{2\pi}{2P}\right) \\
&= A\left[(1+\Delta)+(1-\Delta)e^{-j2\pi\frac{\omega}{\omega_0}}\right] \sum_{k=-\infty}^{+\infty} \frac{\omega_0}{2}\delta\left(\omega - k\frac{\omega_0}{2}\right) \\
&= A\left[(1+\Delta)+(1-\Delta)e^{-j2\pi\frac{\omega}{\omega_0}}\right] \frac{\omega_0}{2} \left[\sum_{l=-\infty,k=2l}^{+\infty} \delta\left(\omega - k\frac{\omega_0}{2}\right) + \sum_{l=-\infty,k=2l+1}^{+\infty} \delta\left(\omega - k\frac{\omega_0}{2}\right) \right] \\
&= A\left[(1+\Delta)+(1-\Delta)e^{-j2\pi\frac{\omega}{\omega_0}}\right] \frac{\omega_0}{2} \left[\sum_{l=-\infty}^{+\infty} \delta\left(\omega - l\omega_0\right) + \sum_{l=-\infty}^{+\infty} \delta\left(\omega - (l+\frac{1}{2})\omega_0\right) \right]
\end{aligned} \tag{5.2}$$

where $\omega_0 = \dfrac{2\pi}{P}$ is the fundamental frequency in rad.

Similarly with the case for jitter, we can use (5.2) to divide the spectrum to a harmonic and a subharmonic part, by sampling at harmonic frequencies $l\omega_0$ and at subharmonic frequencies $(l+1/2)\omega_0$, respectively. The harmonic part of the spectrum, as it is influenced by shimmer, is described then by

$$\begin{aligned} H_s(\Delta, l\omega_0) &= A\left[(1+\Delta)+(1-\Delta)e^{-j2\pi\frac{l\omega_0}{\omega_0}}\right]\frac{\omega_0}{2} \\ &= \frac{A\omega_0}{2}\left[(1+\Delta)+(1-\Delta)e^{-j2l\pi}\right] \\ &= \frac{A\omega_0}{2}\left[1+\Delta+1-\Delta\right] \\ &= A\omega_0,\, l \in \mathbf{N} \end{aligned} \tag{5.3}$$

while the subharmonic part of the spectrum, that appears because of the existence of shimmer, is given by

$$\begin{aligned} S_s(\Delta, (l+\tfrac{1}{2})\omega_0) &= A\left[(1+\Delta)+(1-\Delta)e^{-j2\pi\frac{(l+\frac{1}{2})\omega_0}{\omega_0}}\right]\frac{\omega_0}{2} \\ &= \frac{A\omega_0}{2}\left[(1+\Delta)+(1-\Delta)e^{-j(2l\pi+\pi)}\right] \\ &= \frac{A\omega_0}{2}\left[1+\Delta-1+\Delta\right] \\ &= A\omega_0\Delta,\, l \in \mathbf{N} \end{aligned} \tag{5.4}$$

The magnitude of the two spectral parts, for various values of deviation Δ, are illustrated in figure 5.2. Both the harmonic and subharmonic parts have a specific level of amplitude, with the former being constant, irregardless of the value of Δ, and the latter being analogous to the percentage of shimmer induced. This observation has been confirmed previously, through heuristic analysis for synthetic shimmered glottal airflow signals, with either cyclic or random variation of the glottal amplitude [24]. It is evident that the ratio of the magnitude of a subharmonic frequency to that of a harmonic frequency is equal to the value of Δ.

5.2 Model evaluation

Unlike the case with the mathematical model for jitter, when speech signals under the effect of shimmer are considered, the aforesaid spectral behavior of the model for shimmer can not be measured efficiently. If the shimmered impulse train in (5.1) is regarded as glottal excitation, then this excitation is used as the source signal to a filter that is formed by the convolution of the glottal signal with the vocal tract. The spectral envelope of this filter, since it is not of constant magnitude, modifies the relation between the harmonic and subharmonic spectra. It would be possible to estimate the glottal excitation magnitude spectrum through inverse filtering, however, to achieve an accurate estimation is a difficult task that is beyond the scope of this work.

An indirect evaluation of the model was performed by examination of four features that are based on approximations of the above spectral properties. The four features are the following:

- The ratio of the first subharmonic to the DC component, referred to as "Par$_0$"

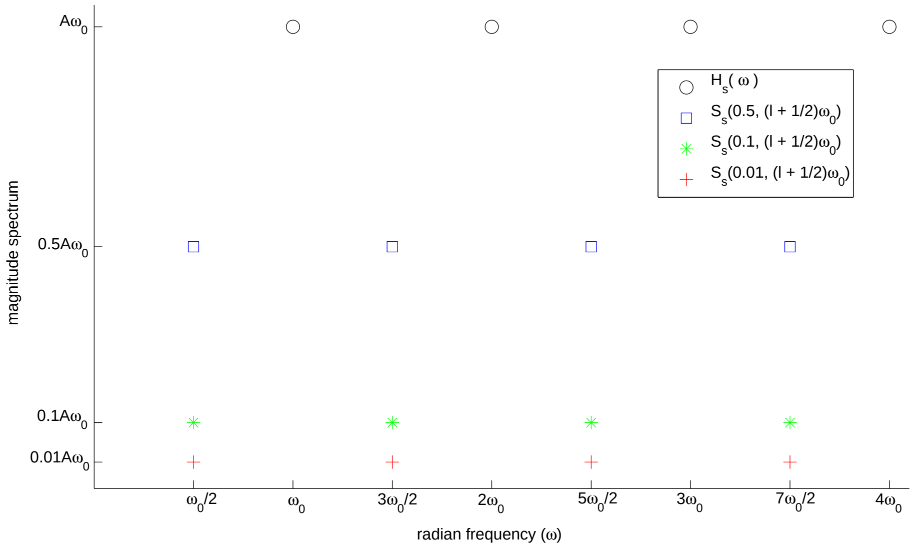

Figure 5.2: Magnitude spectra of the harmonic and subharmonic parts of the mathematical model for shimmer. The harmonic part is the same for all values of Δ, while the subharmonic part is affected accordingly.

$\mathrm{Par}_0 = S_s(\frac{1}{2}\omega_0)/H_s(0)$

- The ratio of the first subharmonic to the first harmonic, referred to as "Par$_1$"
  $\mathrm{Par}_1 = S_s(\frac{1}{2}\omega_0)/H_s(\omega_0)$

- The ratio of the average subharmonic to the average harmonic including the DC component, referred to as "Rat$_0$"
  $\mathrm{Rat}_0 = \overline{S_s((l+\frac{1}{2})\omega_0)}/\overline{H_s(l\omega_0)},\ l \in \mathbf{N}$

- The ratio of the average subharmonic to the average harmonic excluding the DC component, referred to as "Rat$_1$"
  $\mathrm{Rat}_1 = \overline{S_s((l-\frac{1}{2})\omega_0)}/\overline{H_s(l\omega_0)},\ l \in \mathbf{N}^*$

These features were computed for synthetic shimmered phonation signals using a sliding frame, in a fashion similar to that of the Spectral Jitter Estimator (SJE) described in Section 2.2. The synthetic signals were created by excitation of the vocal tract envelope mentioned in Section 3.1, with shimmered impulse trains as these are described in (5.1). Signals were created for sampling frequencies of both 16 and 48 kHz. The value of Δ varied from 0% to 100% with a step of 0.1%, while the duration of the signals were set to 1 seconds. The correlation coefficient between the value of Δ and the average of each feature for an input synthetic signal was found to be 100%, with confidence intervals at 95%, for all four features.

The four features for shimmer were also examined in the two databases of sustained vowel recordings, MEEI and PdA. The estimation was done using a sliding Hanning window, with frame size four times the average pitch period of the signal and hop size one time that. The average pitch period estimation was taken from Praat ($P_{\text{Praat(average)}}$) and MDVP ($P_{\text{MDVP(average)}}$) The average of the time-series generated for a feature, for each signal, was used to perform ROC analysis. The AUC indexes for the four features, using the two different pitch period estimators, on MEEI and PdA, are given in Tables 5.1 and 5.2, respectively. The AUC scores of the local shimmer (1) implementations of Praat (*Shimmer (local)*) and MDVP (*Shim*) are also given for reference.

AUC (standard error) % for MDVP on MEEI		92.58 (1.19)
AUC (standard error) % for Praat on MEEI		90.63 (1.42)
AUC (standard error) % for four features on MEEI		
Feature	$P_{\text{MDVP(average)}}$	$P_{\text{Praat(average)}}$
Par$_0$	89.31 (1.57)	89.14 (1.59)
Par$_1$	86.57 (1.86)	86.36 (1.88)
Rat$_0$	88.46 (1.66)	86.14 (1.91)
Rat$_1$	88.57 (1.65)	86.29 (1.89)

Table 5.1: AUC score (and standard error) in % for all shimmer features tested on MEEI.

AUC (standard error) % for MDVP on PdA		74.65 (2.38)
AUC (standard error) % for Praat on PdA		73.87 (2.40)
AUC (standard error) % for four features on PdA		
Feature	$P_{\text{MDVP(average)}}$	$P_{\text{Praat(average)}}$
Par$_0$	65.91 (2.62)	69.81 (2.53)
Par$_1$	72.55 (2.45)	79.05 (2.20)
Rat$_0$	70.40 (2.51)	79.25 (2.19)
Rat$_1$	70.33 (2.51)	79.34 (2.19)

Table 5.2: AUC score (and standard error) in % for all shimmer features tested on PdA.

Chapter 6

Conclusions

In this work we proposed the use of a mathematical description for modeling the jitter phenomenon, when that is present in voice production. This model transforms the jitter estimation problem from the time domain to the frequency domain and led us to the development of the short-time Spectral Jitter Estimator (SJE). Experiments conducted with synthetic jittered phonation signals verified that SJE produces accurate local estimates of jitter. Comparison of the method with equivalent and widely adopted measurements of jitter, namely the implementations of Praat and MDVP for absolute jitter (3.1), showed that SJE is more discriminant in the classification of normal versus pathological sustained vowel recordings. We expanded on these results by determining through a cross-database study a relevant threshold for pathology and also by applying SJE on reading text recordings. The determined $\mathrm{Thr}_{\mathrm{MEEI}}$ threshold results indeed in high discrimination for healthy versus pathological voices, in databases of either sustained vowel recordings or reading text recordings. These results, in addition to the statistical data gathered from local short-time measurements, clearly show that SJE produces estimates that are highly correlated with the pathological nature of jitter.

Based on the establishment of the $\mathrm{Thr}_{\mathrm{MEEI}}$ threshold and on the time-series of local jitter estimations from SJE, we also introduced three new features that have high correlation with the existence of pathology and therefore can be considered as good candidates for use with continuous speech signals. Specifically, these are the percentage of frames above $\mathrm{Thr}_{\mathrm{MEEI}}$ (*Over*), the maximum number of consecutive frames that are above $\mathrm{Thr}_{\mathrm{MEEI}}$ (*Max Over*), and the maximum number of consecutive frames that are below $\mathrm{Thr}_{\mathrm{MEEI}}$ (*Max Under*). Moreover, we determined thresholds for the *Over* feature, $\mathrm{Thr}_{\mathrm{Over}}$, that can be used especially for monitoring the jitter effect in running speech. A potential beneficiary of this monitoring ability can be vocal loading estimation. This is the estimation of the stress inflicted on speech organs after sustained periods of voicing, which is mostly of interest for healthy speakers that perform a vocal occupation, such as teachers, singers, etc.

The jitter phenomenon also contributes to the appearance of noise in the spectrum. This may have implications in the estimation of a Harmonics to Noise Ratio (HNR). Based on the work presented in Chapter 2 regarding the spectral properties of jitter, we can identify in the magnitude spectrum the noise induced by jitter; if this taken into consideration before HNR estimation, then it may lead to a more accurate HNR value [23]. Indeed, the frequency points where the intersections between the harmonic and subharmonic parts of the spectrum occur, are good candidates for deciding which parts of the spectral noise should not be considered as additive but as of the structural kind.

Several statistical properties of jitter have been documented in the past, as well. In Chapter 4 we examined and verified the behavior of local jitter as a function of fundamental frequency. Other interesting properties could be examined in the future using the SJE short-time measurements. One such property is that jitter in adjacent periods is correlated and thus, present time jitter could be predictable from past values [32]. In [10] and in [32], the jitter time-series is modeled as an Auto Regressive (AR) process. Following that, it is shown that the frequency and bandwidth of the pole of the envelope is related to the rate of pathology perceived in the examined signal [10]. Therefore, it is straightforward to apply similar time-series modeling techniques to the short-time jitter sequence estimated by SJE.

The short-time nature of SJE can lead to the extraction of multi-dimensional parameters that may be of use in automatic pathological condition detection. Distributions of the time-series of the short-time SJE estimations may be used for the classification of new unlabeled signals, through a similarity measure between the distribution of the new signal and these of already labeled recordings. It will also be of great importance to test the proposed method, features and thresholds in signals recorded before and after successful therapy. It is expected that, after treatment, the measurements will show if not values under the threshold of pathology, at least significant reductions compared to the ones before.

We also proposed a mathematical model for shimmer that was used to create four features related with the spectral properties of the phenomenon. Although accurate estimation of shimmer was not achieved, the four parameters nonetheless exhibited good discrimination abilities, performing in the same level with the implementations of Praat and MDVP for the local shimmer (1) parameter. Inverse filtering or spectral envelop estimation could be used to approximate the glottal excitation spectrum, by dividing the magnitude spectrum of a given speech segment with an estimated envelope, in order to improve the accuracy of these features.

Appendix A

Discrete Dirac comb properties

The Kronecker delta function is the discrete analogue of the continuous Dirac function and is defined as

$$\delta[n] = \begin{cases} 1, \text{ if } n = 0 \\ 0, \text{ if } n \neq 0 \end{cases} \tag{A.1}$$

where n is the discrete time index.
It is obvious that for a time shift n_0

$$\delta[n - n_0] = \begin{cases} 1, \text{ if } n = n_0 \\ 0, \text{ if } n \neq n_0 \end{cases} \tag{A.2}$$

Consequently we have

$$(\delta[n - n_0])^2 = \delta[n - n_0] \tag{A.3}$$

The Kronecker delta has also the so-called sifting property, that is for a function $x[n]$

$$\sum_{k=-\infty}^{+\infty} x[k]\delta[n - n_0] = x[n_0] \tag{A.4}$$

The discrete Dirac comb is a periodic impulse train that can be defined as

$$\delta_T[n] = \sum_{k=-\infty}^{+\infty} \delta[n - kT] \tag{A.5}$$

Using the properties of the Kronecker delta we can show that the Fourier transform of the discrete

Dirac comb is itself a Dirac comb. Specifically

$$\begin{aligned}\Delta_T(\omega) &= \sum_{n=-\infty}^{+\infty} \delta_T[n] e^{-j\omega n} \\ &= \sum_{n=-\infty}^{+\infty} \left(\sum_{k=-\infty}^{+\infty} \delta[n-kT] \right) e^{-j\omega n} \\ &= \sum_{k=-\infty}^{+\infty} e^{-j\omega kT} \\ &= \sum_{k=-\infty}^{+\infty} \frac{2\pi}{T} \delta \left(\omega - k\frac{2\pi}{T} \right) \\ &= \frac{2\pi}{T} \delta_{2\pi/T}(\omega)\end{aligned} \tag{A.6}$$

where ω is the frequency in rad.

Note also that

$$(\delta_T[n])^2 = \delta_T[n] \tag{A.7}$$

and that if y is a function of the combed $x[k]$, then

$$y\left(x[k]\delta_T[n]\right) = y\left(x[k]\right)\delta_T[n] \tag{A.8}$$

Bibliography

[1] A. Askenfelt and B. Hammarberg. Speech waveform perturbation analysis revisited. Speech Tansmission Laboratory - Quartely Progress and Status Report, 22(4):49–68, 1981.

[2] R.J. Baken and R.F. Orlikoff. Clinical Measurement of Speech and Voice, 2nd edn. Singular Publishing Group, 1999.

[3] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In Institute of Phonetic Sciences, volume 17, pages 97–110, University of Amsterdam, 1993.

[4] P. Boersma and D. Weenink. Praat: doing phonetics by computer (Version 5.0.22) [Computer Program], 2008.

[5] A. de Cheveigne and H. Kawahara. YIN, A fundamental frequency estimator for speech and music. Journal of the Acoustical Society of America, 111(4):1917–1930, 2002.

[6] D.D. Deliyski. Acoustic model and evaluation of pathological voice production. In EUROSPEECH'93, pages 1969–1972, Berlin, 1993.

[7] J.P. Egan. Signal Detection Theory and ROC Analysis. Academic Press, 1975.

[8] Kay Elemetrics. Disordered Voice Database (Version 1.03), 1994.

[9] Kay Elemetrics. Multi-Dimensional Voice Program (MDVP) [Computer Program], 2007.

[10] Y. Endo and H. Kasuya. A stochastic model of fundamental period perturbation and its application to perception of pathological voice quality. In 4th International Conference on Spoken Language Processing, pages 772–775, Philadelphia, 1996.

[11] S. Feijoo and C. Hernández-Espinosa. Short-term stability measures for the evaluation of vocal quality. Journal of Speech and Hearing Research, 33:324–334, 1990.

[12] A. Fourcin and E. Abberton. Hearing and phonetic criteria in voice measurement: Clinical applications. Logopedics Phoniatrics Vocology, pages 1–14, Apr 2007.

[13] J.I. Godino-Llorente, V. Osma-Ruiz, N. Sáenz-Lechón, I. Cobeta-Marco, R. González-Herranz, and C. Ramírez-Calvo. Acoustic analysis of voice using WPCVox: a comparative study with Multi Dimensional Voice Program. European Archives of Otolaringology, 265(4):465–476, 2008.

[14] R. Gubrynowitz, W. Mikiel, and P. Zarnecki. An acoustic method for the evaluation of the state of the larynx source in cases involving pathological changes. Archives of Acoustics, 5(1):3–30, 1980.

[15] J.A. Hanley and B.J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143:29–36, 1982.

[16] H. Hollien, J. Michel, and E.T. Doherty. A method for analysing vocal jitter in sustained phonation. Journal of Phonetics, 1:85–91, 1973.

[17] Y. Horii. Fundamental frequency perturbation observed in sustained phonation. Journal of Speech and Hearing Research, 22:5–19, 1979.

[18] Y. Koike, H. Takahashi, and T.C. Calcaterra. Acoustic meausures for detecting laryngeal pathology. Acta Oto-Laryngol, 84:105–117, 1977.

[19] J. Kreiman and B.R. Gerratt. Perception of aperiodicity in pathological voice. Journal of the Acoustical Society of America, 117(4):2201–2211, 2005.

[20] G. De Krom. Acoustic correlates of breathiness and roughness: Experiments on voice quality. Dissertation, Utrecht Institute of Linguistics OTS, Utrecht, Netherlands, 1994.

[21] J. Laver, S. Hiller, J. Mackenzie, and E. Rooney. An acoustic screening system for the detection of laryngeal pathology. Journal of Phonetics, 14:517–524, 1986.

[22] P. Lieberman. Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. Journal of the Acoustical Society of America, 35:344–353, 1963.

[23] P. J. Murphy. Perturbation-free measurement of the harmonics-to-noise ratio in voice signals using pitch synchronous harmonic analysis. Journal of the Acoustical Society of America, 105(5):2866–2881, 1999.

[24] P.J. Murphy. Spectral characterization of jitter, shimmer, and additive noise in synthetically generated voice signals. Journal of the Acoustical Society of America, 107(2):978–988, 2000.

[25] Y. Pantazis, O. Rosec, and Y. Stylianou. On the properties of a time-varying quasi-harmonic model of speech. In Interspeech 2008, pages 1044–1047, Brisbane, 2008.

[26] V. Parsa and D.G. Jamieson. Acoustic discrimination of pathological voice: Sustained vowels versus continuous speech. Journal of Speech, Language, and Hearing Research, 44:327–339, 2001.

[27] N.B. Pinto and I.R. Titze. Unification of perturbation measures in speech signals. Journal of the Acoustical Society of America, 87(3):1278–1289, 1990.

[28] T.F. Quatieri. Discrete-Time Speech Signal Processing. Prentice Hall, 2002.

[29] M. Rosa, J.C. Pereira, and M. Grellet. Adaptive estimation of residue signal for voice pathology diagnosis. IEEE Transactions on Biomededical Engineering, 47(1):96–104, Jan 2000.

[30] J. Schoentgen. Stochastic models of jitter. Journal of the Acoustical Society of America, 109(4):1631–1650, 2000.

[31] J. Schoentgen and R. De Guchteneere. Time series analysis of jitter. Journal of Phonetics, 23:189–201, 1995.

[32] J. Schoentgen and R. De Guchteneere. Predictable and random components of jitter. Speech Communication, 21:255–272, 1997.

[33] Y. Stylianou. Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification. PhD thesis, Ecole Nationale Supèrieure des Télécommunications, Paris, France, Jan 1996.

[34] K. Umapathy, S. Krishnan, V. Parsa, and D.G. Jamieson. Discrimination of pathological voices using time-frequency approach. IEEE Transactions on Biomededical Engineering, 52(3):421–430, 2005.

[35] M. Vasilakis and Y. Stylianou. Spectral jitter modeling and estimation. Biomedical Signal Processing & Control Special Issue: M&A of Vocal Emissions, to appear.

[36] M. Vasilakis and Y. Stylianou. Voice pathology detection based on short-time jitter estimations in running speech. Folia Phoniatrica et Logopaedica, to appear.

[37] M. Vasilakis and Y. Stylianou. A mathematical model for accurate measurement of jitter. In MAVEBA 2007, pages 7–10, Florence, Italy, 2007.

[38] M. Vasilakis and Y. Stylianou. A mathematical model for accurate measurement of shimmer. In 2nd Advanced Voice Function Assessment International Workshop, Aachen, Germany, 2008.

[39] M. Vasilakis and Y. Stylianou. Novel short-time jitter features for monitoring of running speech. In 3rd Advanced Voice Function Assessment International Workshop, Madrid, Spain, 2009.

[40] M. Vasilakis and Y. Stylianou. Spectral jitter estimation revisited. In 3rd Advanced Voice Function Assessment International Workshop, Madrid, Spain, 2009.