Original Paper

Folia Phoniatrica
et Logopaedica

Folia Phoniatr Logop 509-T1
DOI: 10.1159/000XXXXXX

Published online: ■■■

Voice Pathology Detection Based on Short-Term Jitter Estimations in Running Speech

Miltiadis Vasilakis[a, b] Yannis Stylianou[a, b]

[a] Department of Computer Science, University of Crete, ~~■■■~~, and [b] Institute of Computer Science,
Foundation of Research and Technology Hellas (FORTH), ~~■■■■■,~~ Greece

Key Words
Jitter · Short-term jitter · Sustained vowels · Continuous speech · Beat spectrum · Spectral jitter

Abstract
In this paper, we investigate the use of jitter estimation over short time intervals (short-term jitter) for voice pathology detection in the case of running or continuous speech. Short-term jitter estimations are provided by the spectral jitter estimator (SJE), which is based on a mathematical description of the jitter phenomenon. The SJE has been shown to be robust against errors in pitch period estimations, which makes it a good candidate for measuring jitter in continuous speech. On two large databases of sustained vowel recordings from healthy and pathological voices, we suggest a threshold for the SJE for pathology detection based on cross-database validation. Applying that to a database of continuous speech (reading text) from normophonic and dysphonic speakers, a second threshold and new features are suggested for monitoring jitter in continuous speech. Detection performance of the suggested thresholds and features was evaluated using receiver operating characteristic curves and their discriminative efficiency between healthy and pathological voices was judged using the area under the curve index. In terms of area under the curve, the suggested features for reading text provide a discrimination score of about 95%, while the second threshold provides a classification rate of 87.8%. Furthermore, estimated short-term jitter values from reading text were found to confirm the studies showing a 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.

Introduction

One of the most prominent phenomena among those we seek to measure in the context of voice quality assessment is jitter. Jitter may occur during voice production, especially in vowel phonation, and it is defined as small fluctuations in glottal cycle lengths [1, 2]. Jitter and shimmer (amplitude perturbations) over successive speech cycles help give the vowel its naturalness in contrast to constant pitch and amplitude that can result in a machine-like sound. Moreover, jitter (and shimmer) contributes to the voice quality of a speaker. In terms of signal processing, jitter is a form of modulation noise. Specifically, jitter is a modulation of the periodicity of the voice signal. A high degree of jitter results in a voice with roughness that is usually perceived in recordings of pathological voices. Therefore, a reliable estimation of jitter can be used to discriminate between healthy and dysphonic speakers.

KARGER
Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

© 2009 S. Karger AG, Basel
1021–7762/09/0000–0000$26.00/0

Accessible online at:
www.karger.com/fpl

Miltiadis Vasilakis
Department of Computer Science
University of Crete
■■■■■ (Greece)
Tel. ■■■, Fax ■■■, E-Mail mvasilak@csd.uoc.gr

There are also other forms of induced noise that occur, similarly, in voices with a pathological condition, such as additive noise which results in a breathy voice quality. The human ear indeed perceives jitter as noise. Note that humans cannot differentiate between the noise produced by jitter and shimmer, while they are able to differentiate modulation from additive noise [3].

Based on the definition of jitter, many 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 [4–7]. Usually they are applied on recordings of sustained vowels where perturbations are expected to be steady. These methods are based on the estimation of pitch periods and therefore they are sensitive to error in the estimation of pitch period; for a given voiced speech segment, different pitch period estimators will lead to quite different jitter estimates. This means that the suggested methods for jitter estimation are quite unstable. To minimize the variance of jitter estimations, the majority of the methods provides an average measurement, usually over a series of pitch period estimations. Methods based on the averaging of jitter are statistically biased, since it has been found that they underestimate jitter [2]. Also, averaging implies that pitch cycle perturbations are generated by an independent and identically distributed (gaussian) stochastic process. However, it has been shown that there is a correlation between successive values of jitter [8]. Therefore, this correlation should be removed before applying the average operator [2].

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 the limitations of the suggested estimators of jitter [9, 10]. 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 [11, 12]. Processing of continuous speech for voice pathology detection was studied before [10, 12–15]. In Askenfelt and Hammarberg [14], patients read a tale for approximately 40 s, and then 7 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 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) [14]. In Umapathy et al. [10], a time-frequency representation based on matching pursuit decomposition with Gabor time-frequency atoms of various scale factors was used. It was found that the distribution of these scale factors was a potential feature for the discrimination of normal and pathological speech signals. In Fourcin and Abberton [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.

In this paper, we suggest new features for the analysis of continuous speech for voice pathology detection based on short-term measurements of jitter which are robust against the ambiguities of pitch period estimators. Vasilakis and Stylianou [16, 17] presented a novel short-term jitter estimator, referred to as spectral jitter estimator (SJE), that estimates the jitter phenomenon based on a mathematical model. This model transforms the jitter estimation problem from the time domain to the frequency domain showing that jitter leads to beat spectrum. The SJE allows for time-varying measurement with a high local accuracy, as demonstrated on synthetic signals with known jitter. In Vasilakis and Stylianou [16], it was shown how jitter manifests in the magnitude spectrum of a speech frame. Specifically, it was shown that jitter can be estimated by counting the number of intersections between harmonic and subharmonic spectra. Although the SJE uses pitch period information, it was shown that this is not crucial in counting the number of intersections between the harmonic and subharmonic spectra [17]. The performance of the 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 [18] and the Multidimensional Voice Program (MDVP) [19] of KayPENTAX. On two different databases of sustained vowel recordings, the estimates of the SJE were shown to be more correlated with pathology than the estimates by Praat and MDVP. Let us note here that classification between normophonic and dysphonic cases using various features is a necessary step for the evaluation, but by no means sufficient for pathology detection.

As mentioned earlier, it has been shown that methods that produce an average estimate for jitter are statistically biased and actually underestimate jitter [2]. For short-term jitter estimators, however, averaging is not necessary. Actually, the generated sequence of local measurements of jitter can be used to gain further insight into the temporal behavior of jitter, for both healthy and pathological voices. In this paper, we extend the use of the SJE on reading text recordings by suggesting features based on the short-term measurements of jitter as provided by the SJE. For this purpose, we determine a relevant threshold for pathology which leads to high discrimination for normal versus pathological voices, in databases of either sustained vowel recordings or reading text recordings. Using this threshold and based on the time series of local jitter estimations from the SJE, three new features are suggested. It is shown that they are all highly correlated with the existence of pathology while they are ideal for running speech signals. Furthermore, we show examples on how one of these three features can be efficiently used for monitoring the jitter effect in running speech. Please note that by the term 'running speech' in this paper, we mean only voiced segments of the speech signal and not regions with unvoiced consonants or lack of voice.

The paper is organized as follows. In the next section, we present an overview of the SJE by providing the mathematical model that it is based on, and its properties in the time and frequency domain. Thereafter, the procedure for threshold selection for the SJE is developed. In the following section, the application of the SJE on reading text recordings is presented and useful features are proposed that consider the short-term behavior of jitter. Finally, the last section concludes the paper and provides information on future work and possible extensions of this work.

Overview

In this section, a short overview of the SJE and the mathematical model of jitter is presented. More details can be found in Vasilakis and Stylianou [16, 17]. We also summarize the results for discriminating between healthy and dysphonic speakers on two speech databases, using the absolute jitter measurement, as implemented by Praat [18], MDVP [19] and SJE [17].

Spectral Jitter Estimator

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 phenomena, which, when combined appropriately, may produce the observed perturbations. Let's 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. 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 [20]. 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 as [16]

$$g[n]=\sum_{k=-\infty}^{+\infty}\delta\big[n-(2k)P\big]+\sum_{k=-\infty}^{+\infty}\delta\big[n+\varepsilon-(2k+1)P\big] \qquad (1)$$

where P is the pitch period and ε is the pitch deviation, both in samples. In this model, shown in figure 1, ε is the shift that corresponds to jitter. The value of ε can range from 0 (no jitter) to P (pitch halving).

The square of the magnitude spectrum $g[n]$ can be shown to be [17]

$$|G(\omega)|^2=\frac{\omega_0^2}{2}\sum_{k=-\infty}^{+\infty}\left(1+\cos\left[(P-\varepsilon)k\frac{\omega_0}{2}\right]\right)\delta\left(\omega-k\frac{\omega_0}{2}\right) \qquad (2)$$

where

$$\omega_0=\frac{2\pi}{P}$$

is the fundamental frequency in rad. The cosine term inside the sum corresponds to a beat spectrum described by the formula

$$1+\cos\left[(P-\varepsilon)k\frac{\omega_0}{2}\right]=1+\cos(k\pi)\cos\left(k\frac{\varepsilon}{P}\pi\right) \qquad (3)$$

The frequency interval between intersections of the envelope in the beat spectrum is π/ε (rad) and since both cosines in formula 3 have zero phase, we can then locate the intersections at frequencies

$$\omega_k=\left(k+\frac{1}{2}\right)\frac{\pi}{\varepsilon} \qquad (4)$$

with $\omega_k \le \pi$.

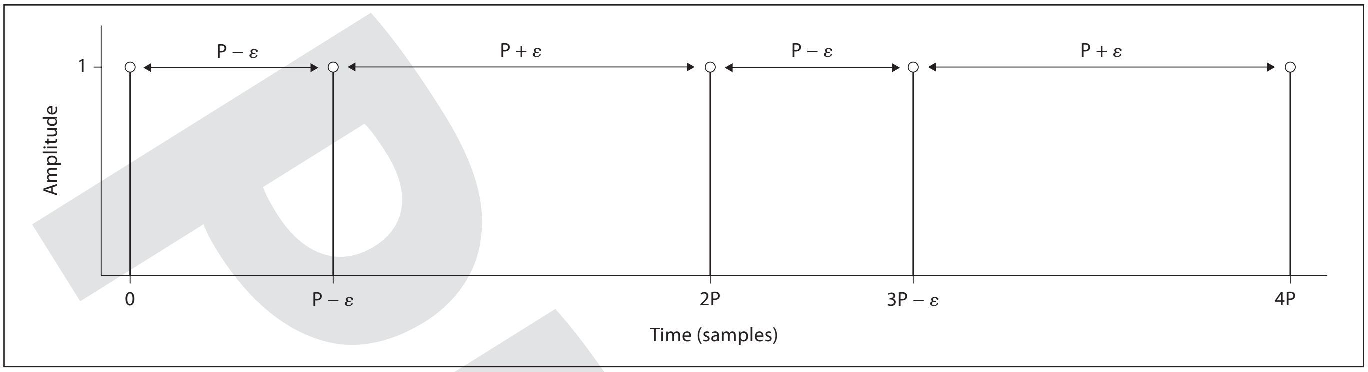


Fig. 1. Jittered impulse train of the two-event model for jitter.

Furthermore, the log magnitude spectrum of $g[n]$ can be shown to be

$$20\log_{10}|G(\omega)| = 10\log_{10}\left(\frac{\omega_0^2}{2}\left(1+\cos\left[\left(P-\varepsilon\right)\omega\right]\right)\right) \times\left[\sum_{l=-\infty,k=2l}^{+\infty}\delta\left(\omega-l\omega_0\right)+\sum_{l=-\infty,k=2l+1}^{+\infty}\delta\left(\omega-\left(l+\frac{1}{2}\right)\omega_0\right)\right] \qquad (5)$$

Based on formula 5, we can divide the spectrum into 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. The harmonic part is described by

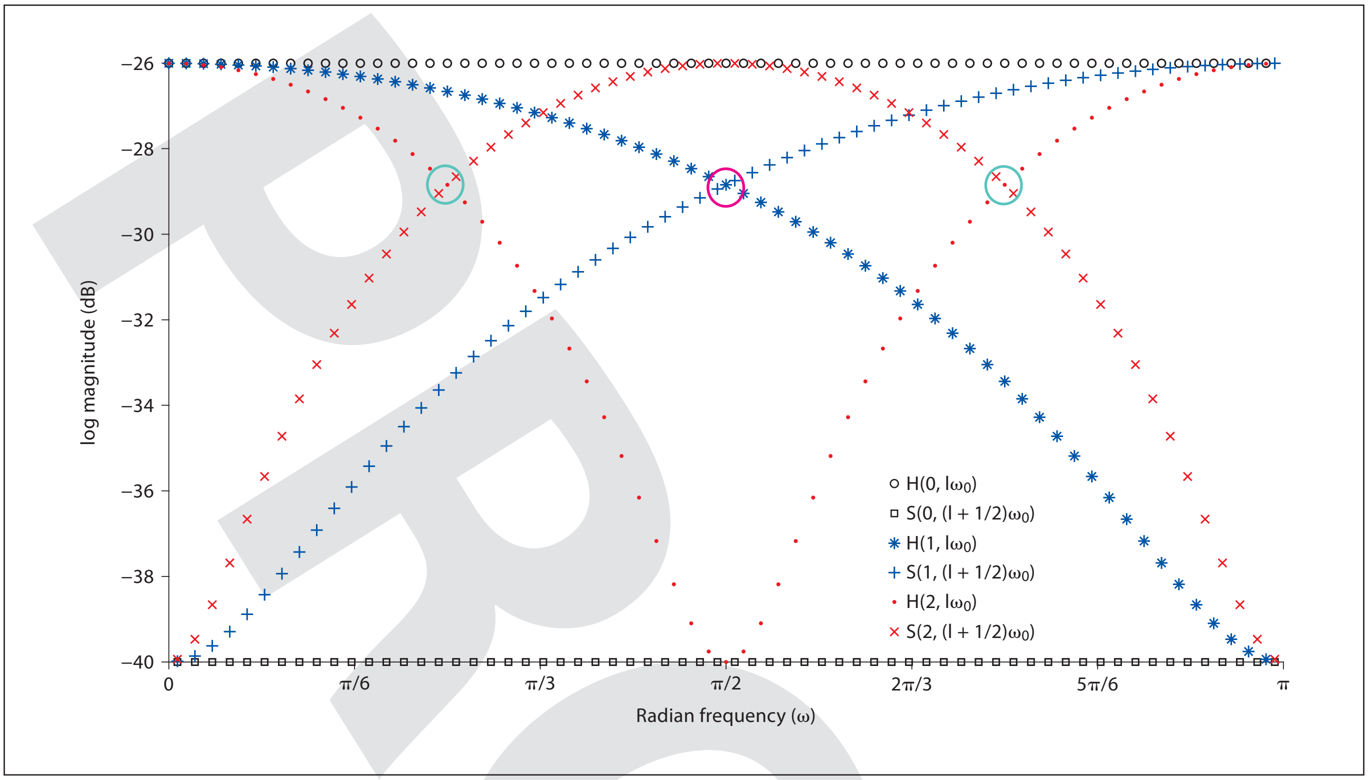
$$H\left(\varepsilon, l\omega_0\right) = 10\log_{10}\left(\frac{\omega_0^2}{2}\left(1+\cos\left[\left(P-\varepsilon\right)l\omega_0\right]\right)\right), l\in\mathbf{N} \qquad (6)$$

and the subharmonic part is given by

$$S\left(\varepsilon,\left(l+\frac{1}{2}\right)\omega_0\right) = 10\log_{10}\left(\frac{\omega_0^2}{2}\left(1+\cos\left[\left(P-\varepsilon\right)\left(l+\frac{1}{2}\right)\omega_0\right]\right)\right), l\in\mathbf{N} \qquad (7)$$

Examples of these two parts for various values of ε are depicted in figure 2. We can observe that the harmonic and subharmonic parts follow a certain pattern, where for a specific value of ε the two parts intersect ε times. As mentioned previously, the locations of the intersections are provided in formula 4. Since jitter, using formula 6 and 7, is estimated in the spectral domain, the jitter estimator is referred to as spectral jitter estimator (SJE). It is 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 [21].

If a jittered impulse train, such as in formula 1, is used as the input of a linear system, then the aforementioned spectral structure remains visible also in the output. Therefore, we expect to observe such a spectral behavior in phonation recordings, whenever jitter is present. Based on this fact, a short-term jitter estimator has been developed in Vasilakis and Stylianou [17]. By applying a sliding window on the signal a sequence of local jitter values is obtained. The size of the window is chosen to be a multiple of the pitch period, usually 3 or 4 times that. 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 signal, especially in cases where we examine sustained phonation recordings. The hop size is one pitch period. We use a Hanning window for analysis, which allows us to avoid the appearance of alias frequencies in the spectrum (because of discontinuities in time), and to also concentrate on the 2 middle periods. The magnitude spectrum of the Fourier transform of the windowed speech segment 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. Next, the total number of intersections between the harmonic and subharmonic spectra is computed. The computed number of intersections is further enhanced by rejecting some intersections and grouping others. First, intersections that may appear due to a lack of resolution are rejected, based on a minimum threshold of difference between the two subspectra after a potential intersection. This threshold has been set through experiments at 3 dB. The remaining accepted intersections may be further elimi-



Color version available online

Fig. 2. log magnitude spectra of the harmonic and subharmonic parts. It is worth noting that the circled intersections between the two parts reveal each time the value of jitter.

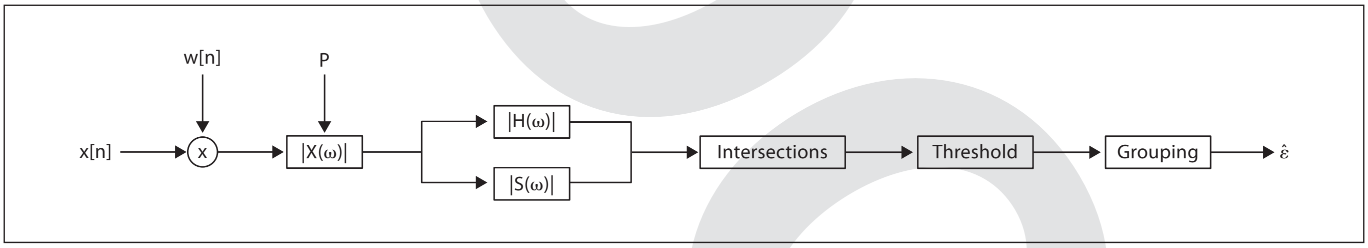
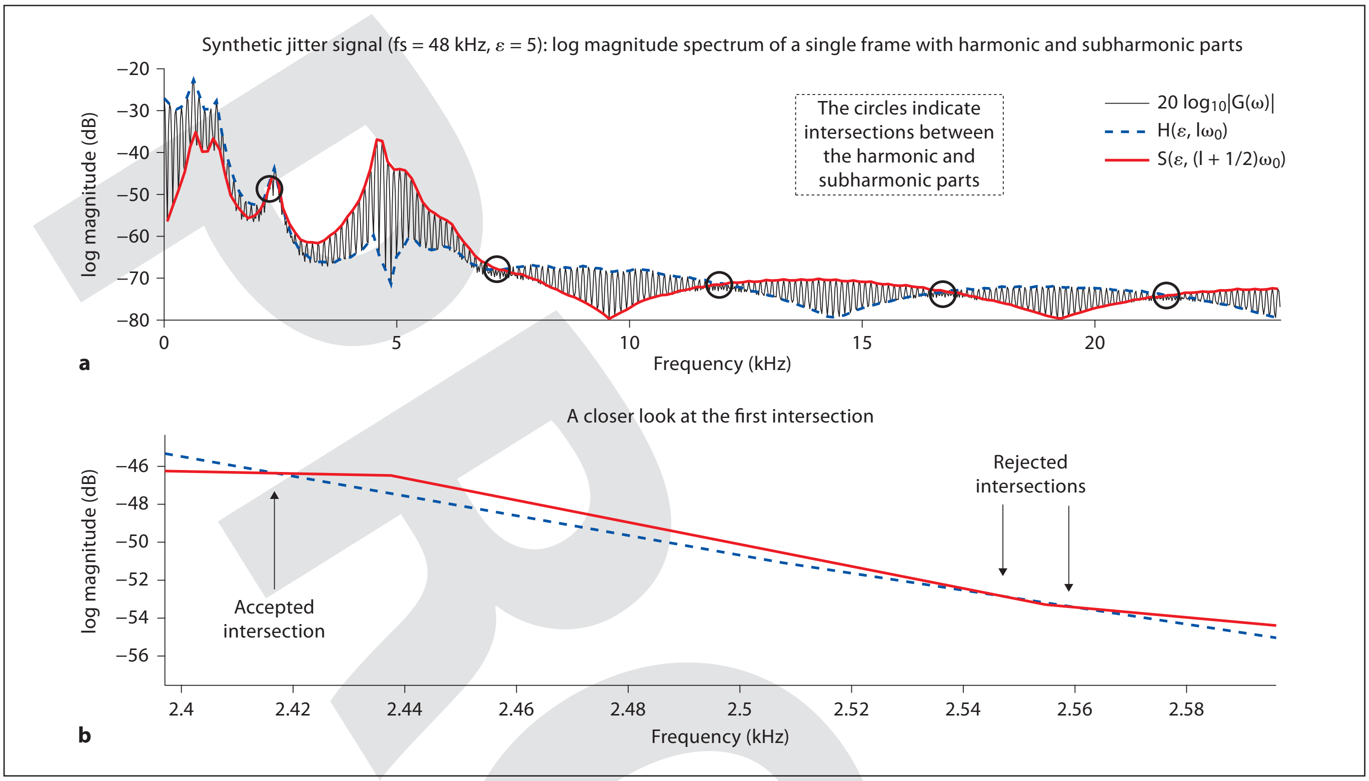


Fig. 3. Block diagram of the short-term SJE algorithm.

nated or grouped together, taking into account the prior knowledge of their expected locations, for each possible jitter value, as illustrated in figure 2. This action is required to suppress any spurious crossings that may arise, especially in higher frequencies. More details about the intersection computation and enhancement process are provided in Vasilakis and Stylianou [17]. The algorithm of the SJE is shown as a block diagram in figure 3, while examples of its usage on a frame from a synthetic jitter signal and a frame from an actual speech signal are presented in figures 4 and 5, respectively. The validity of the estimator has been demonstrated also using synthetic jittered signals where the method produced zero error in estimating jitter [17].

Evaluation of the SJE

In this subsection, we provide results from the evaluation of the SJE in the task of discriminating between normophonic and dysphonic speakers using as feature the absolute jitter. Absolute jitter is a widely implemented



Color version available online

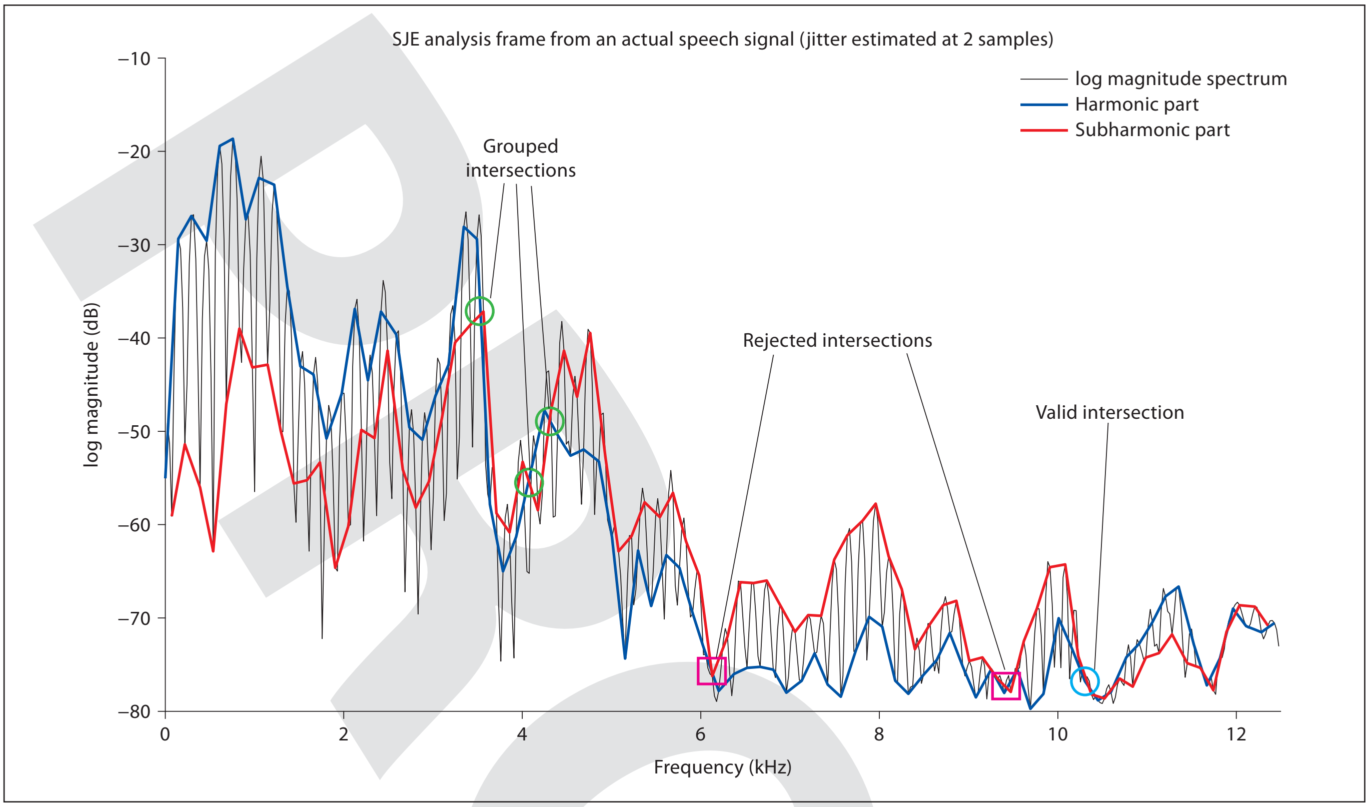
Fig. 4. a Harmonic and subharmonic spectra for a synthetic jitter signal for ε = 5. The detected intersections, after intersection enhancement, are illustrated by circles. The expected structural behavior of jitter in this synthetic example is clearly demonstrated. **b** One example of accepted and one example of rejected intersections.

measurement of the period-to-period variability of pitch in time [4]:

$$\text{Jitt Abs} = \frac{1}{N-1}\sum_{n=1}^{N-1}\left|u(n+1)-u(n)\right| \qquad (8)$$

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 [18] system and the MDVP [19]. Praat implements it as the *Jitter (local, absolute)* function, while MDVP provides the *Jita* analysis parameter. Both systems produce a single jitter estimate for the whole input signal in microseconds. The SJE was compared to the above methods by converting the estimates of ε to microseconds accordingly, and averaging the produced sequence of local values to a single absolute jitter measurement per signal (this averaging was only performed for the purpose of comparison). The discriminative ability of each method on the problem of normal versus pathological classification based on the absolute jitter estimation, can be examined through receiver operating characteristic (ROC) analysis [22]. The ROC curve for each method, that is the true positive rate (TPR) versus false positive rate (FPR) curve, is determined by a variable discrimination threshold. The discriminative efficiency of a method can be 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. Considering the problem of two-class discrimination, such as the normal versus the pathological voices, AUC is an index with analogous discrimination power. AUC is preferred over other measurements of discrimination performance, because it is free from any bias due to the size of each class. The standard error of the AUC index provides information regarding its confidence interval for each case it is applied [23].

The three methods, the absolute jitter estimator as implemented by Praat and MDVP, and the SJE, were applied on two databases of recordings of sustained phonation of /a/ from healthy and pathological voices. The Massachu-



Color version available online

Fig. 5. Harmonic and subharmonic spectra from a frame of an actual sustained phonation recording. SJE results, after intersection enhancement, to an estimate of $\hat{\varepsilon} = 2$.

Table 1. AUC score (standard error) for SJE and the two implementations of absolute jitter by MDVP and Praat, when applied on two databases containing normal and pathological voices (MEEI and PdA)

	AUC, %		
	SJE	MDVP	Praat
MEEI	94.82 (0.92)	90.66 (1.42)	90.47 (1.44)
PdA	84.65 (1.92)	70.65 (2.50)	62.94 (2.67)

setts Eye and Ear Infirmary (MEEI) Disordered Voice Database [24] contains recordings of the sustained vowel /a/. Recordings from 53 subjects with healthy voice and 631 subjects with a wide variety of pathological conditions were used for our comparison experiments. All normal signals in MEEI have a sampling frequency of 50 kHz, while the pathological signals 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 s, while that of the pathological ones from 0.4 to 1.4 s. For SJE in MEEI, a frame size of 4 times the average pitch period, as provided by MDVP, was used in our experiments. The Príncipe de Asturias (PdA) Hospital in Alcalá de Henares of Madrid database [25] was the second database used. PdA consists of recordings of the sustained vowel /a/, with the first and last part of the utterance removed to avoid onset and offset effects. Similar to MEEI, the speech signals were labeled accordingly by clinical doctors. It was found that 238 samples were from normophonic speakers and 201 samples were from dysphonic speakers with a wide range of disorders. All signals in PdA have a sampling frequency of 25 kHz, with 16 bits per sample, and their duration ranges from 1.5 to 4 s. For PdA we again used a frame size of 4 times the average pitch period, this time provided by Praat. The ROC curves of the three methods for MEEI and PdA are depicted in figures

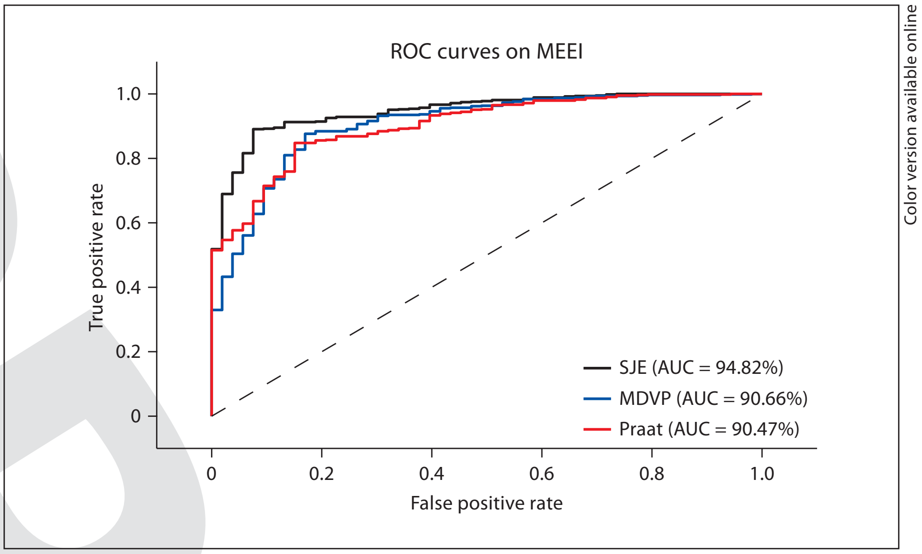


Fig. 6. ROC curves for SJE and absolute jitter, as implemented by MDVP and Praat, for the MEEI database. SJE is more discriminative than the other two methods.

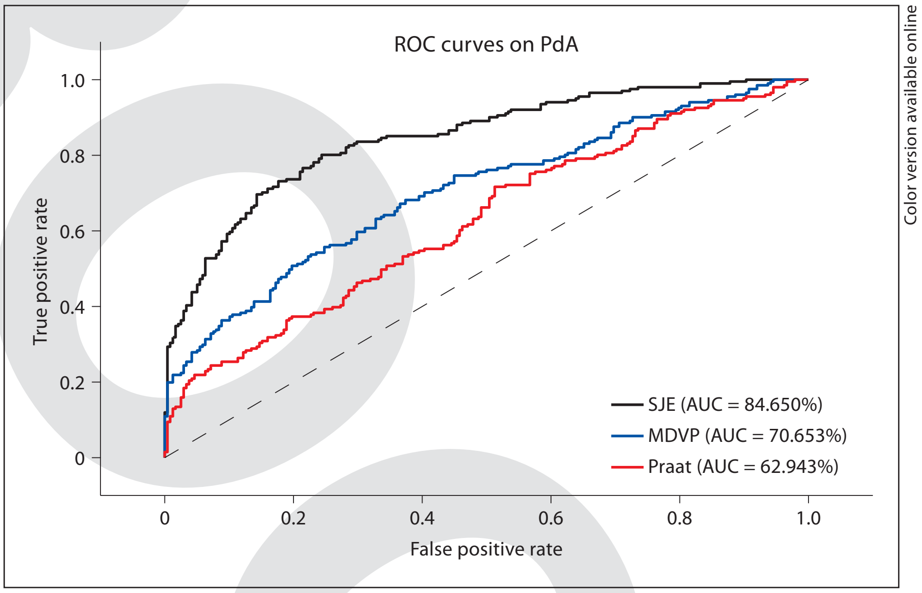


Fig. 7. ROC curves for SJE and absolute jitter, as implemented by MDVP and Praat, for the PdA database. SJE outperforms the other methods in discriminating the normal from the pathological voices.

6 and 7, respectively. In table 1, for each method and for each database, the AUC score and its standard error is provided. It is worth noting that the SJE outperforms both implementations of absolute jitter in discriminating the normal from the pathological voices.

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 the SJE. Since we use two databases, it will be very interesting to examine the consequences of using one database to determine the threshold, and then applying the result to the other database. This will allow 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 the SJE, separately for each database, therefore providing 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 on the MEEI database, the

Table 2. Cross-database evaluation of thresholds determined by SJE in terms of classification rate (CR), and AUC (standard error) for number of frames which are over a threshold *(Over)*, and maximum consecutive frames that are over *(Max Over)* or under *(Max Under)* a threshold

Database	CR, %	*Over* AUC, %	*Max Over* AUC, %	*Max Under* AUC, %
Thr$_{\mathrm{MEEI}}$ 124.24 μs				
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$_{\mathrm{PdA}}$ 161.08 μs				
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)

largest difference is achieved when TPR = 89.07% and FPR = 7.55%, leading to a threshold of 124.24 μs, which we will refer to as Thr$_{\mathrm{MEEI}}$. Similarly, the ROC curve regarding SJE for the PdA database suggests a threshold of 161.08 μs (we will refer to this threshold as Thr$_{\mathrm{PdA}}$), when TPR = 80.10% and FPR = 24.37%. To compare the two thresholds we performed a series of experiments on the two databases. Initially, we measured the classification rate, which is the number of correct detections from both classes divided by the total number of detections, using each threshold.

Since SJE provides us with a short-term sequence of jitter values for each signal, we also calculated three features that make use of the thresholds we presented above. Having in mind that each short-term value corresponds to an analysis frame, then the three features are defined as (1) the percentage of frames that are over the threshold *(Over);* (2) the maximum number of consecutive frames that are over the threshold *(Max Over),* and (3) the maximum number of consecutive frames that are under the threshold *(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 (3 or 4 times the average pitch period), and a fixed hop size was used (hop size was equal to the average pitch period of the signal). Consequently, we calculated the AUC index for these three features for each threshold on each database. All the results are summarized in table 2. Using Thr$_{\mathrm{MEEI}}$ provides, in general, better results than Thr$_{\mathrm{PdA}}$. Given also that it represents a low FPR of 7.55%, we concluded that Thr$_{\mathrm{MEEI}}$ is the preferred value for our following experiments. It is interesting to add that the threshold of 83.20 μs provided by MDVP [26] for its own implementation offers a classification rate of 60.23% for MEEI and 64.46% for PdA, both lower than those provided by Thr$_{\mathrm{MEEI}}$ (89.33% for MEEI and 67.88% for PdA) and 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, Thr$_{\mathrm{MEEI}}$ gives a better classification score for the MEEI database, while in PdA the best classification score is obtained by Thr$_{\mathrm{PdA}}$. For all the experiments conducted hereinafter, the threshold Thr$_{\mathrm{MEEI}}$ (124.236 μs) will be used.

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 aperiodicities 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 longer period of time we may use recordings of reading text. Speakers reading a text with a normal pace are able to breathe occasionally, while in sustained phonation a single intake of breath is involved. This allows us to attain longer recordings for examination and since SJE provides a short-term sequence of jitter estimates, it is ideal for the examination of jitter in running speech signals.

The MEEI Disordered Voice Database, apart from sustained vowel recordings, also includes reading text recordings of the standard text 'The Rainbow Passage'. These recordings are limited to 12 s, usually including up to the two first sentences of the text. For our experiments,

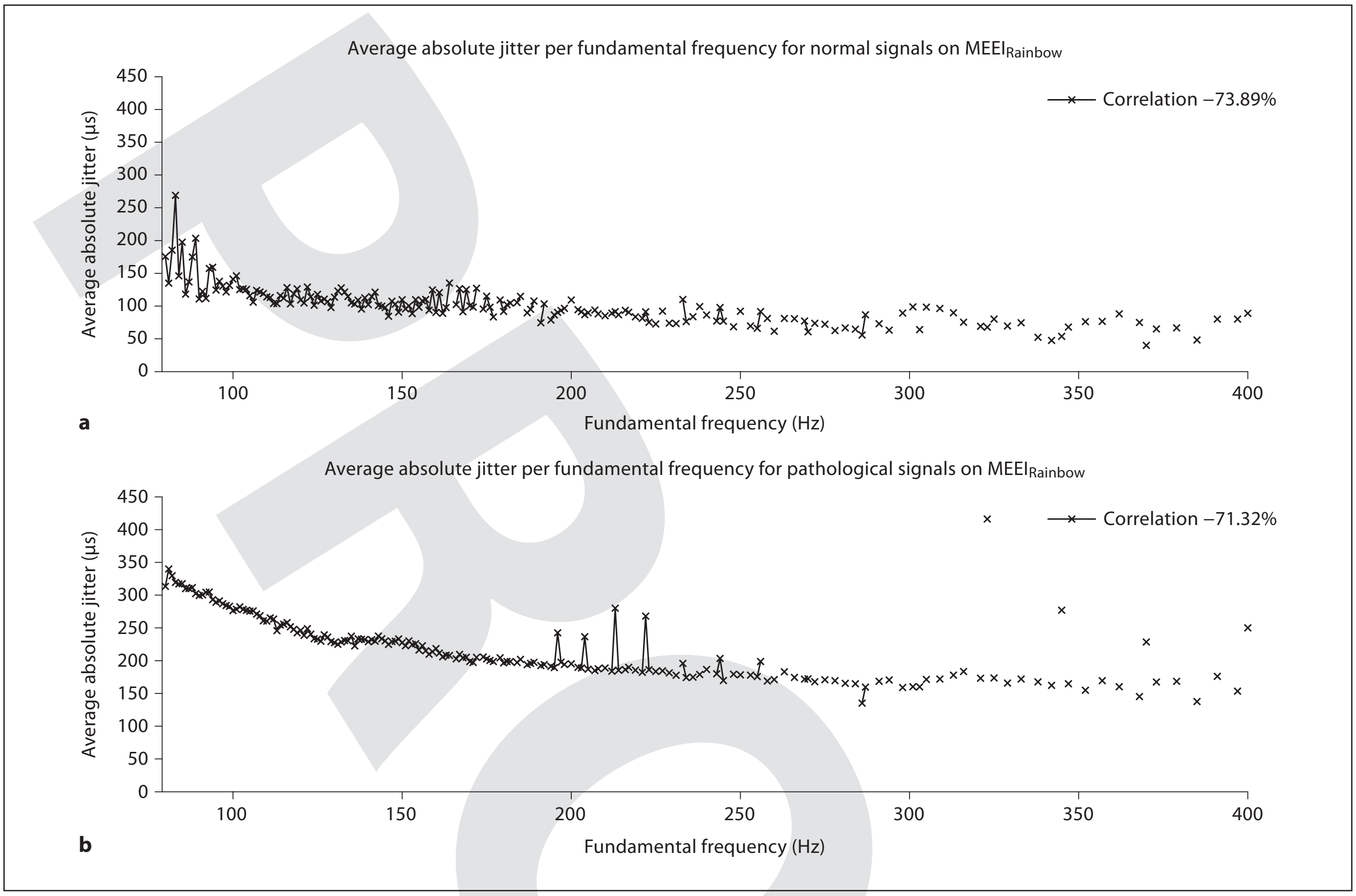


Fig. 8. Average absolute jitter as a function of fundamental frequency, for normal (**a**) and pathological signals (**b**) from MEEI$_{Rainbow}$.

53 signals from healthy voices and 660 signals from pathological voices were used. 683 of these signals have a sampling frequency of 25 kHz and 30 have a sampling frequency of 10 kHz, all with 16 bits per sample. We will refer to this database as MEEI$_{Rainbow}$. 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 [27] for each recording in MEEI$_{Rainbow}$. To eliminate any onset and offset effects in the voiced areas, we disregarded any voiced frames that do not have at least two voiced neighboring frames in each direction, along of course with the unvoiced frames. For the remaining voiced frames, referred to hereafter as *valid frames,* the short-term jitter estimator SJE was used to measure the local absolute jitter value, using a window of 4 times the local pitch period. Next, the terms *frame* and *valid frame* will be used interchangeably unless explicitly specified.

An initial examination of the local estimates from SJE shows that these are in accordance with documented statistical behavior. It is expected that on average jitter decreases with increasing fundamental frequencies [28–31]. 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 as a whole. In figure 8, 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 estimates of jitter was used to calculate several features that reflect the average and short-term behavior of jitter.

The average value of jitter is only computed here for comparison purposes. Specifically, if $j(n)$ is the aforesaid

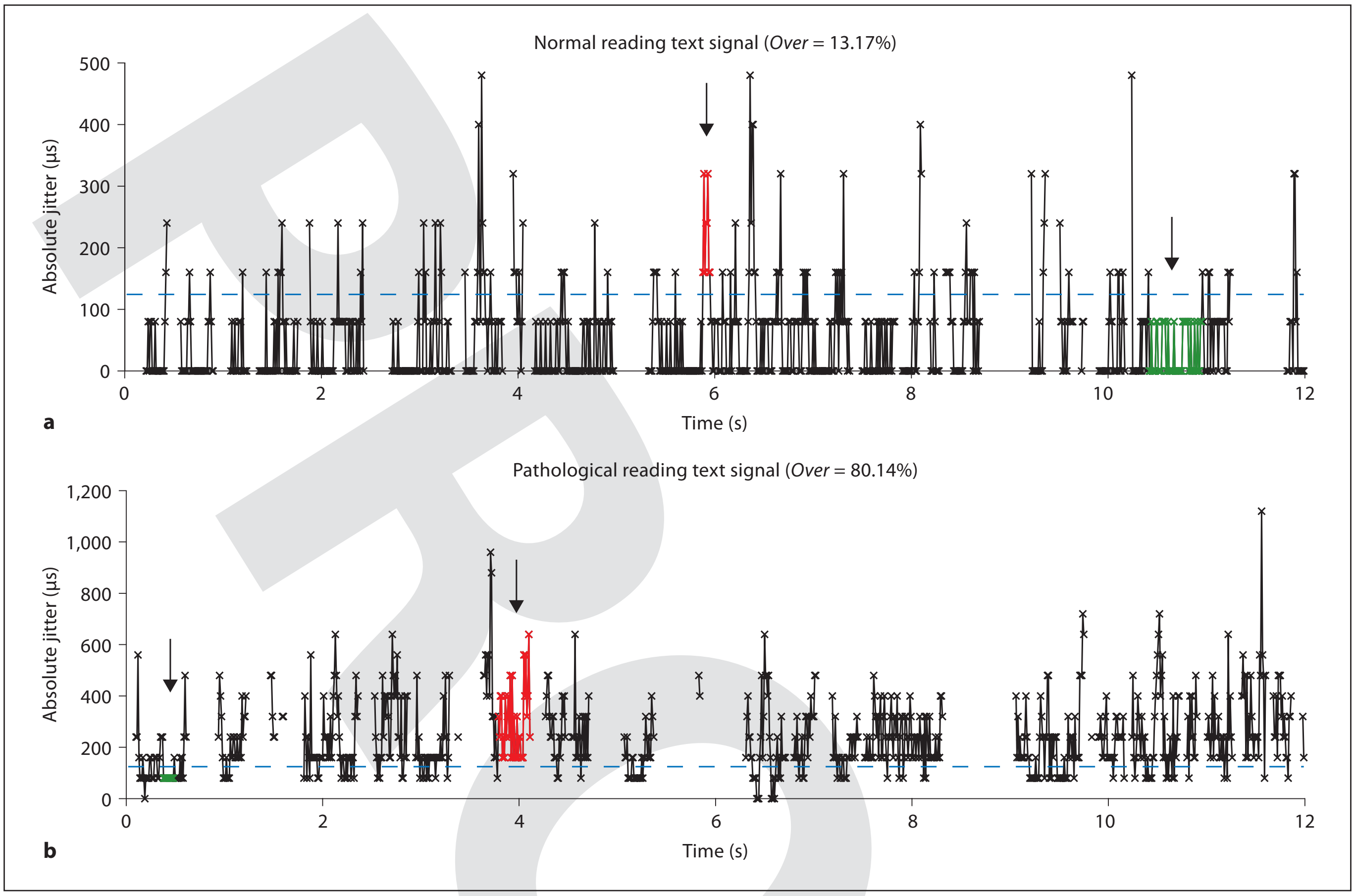


Fig. 9. The pathology threshold $\mathrm{Thr}_{\mathrm{MEEI}}$, indicated by the dashed line, as applied on the short-term SJE estimates of two signals from $\mathrm{MEEI}_{\mathrm{Rainbow}}$, one normal (**a**) and one pathological (**b**). Arrows indicate the maximum number of contiguous frames that are below or above the threshold for each signal.

sequence with length *N*, then for each signal the following features were computed.

- The average absolute jitter from all valid frames which will be referred to as *Jit Mean*.

$$\text{Jit Mean} = \frac{\sum_{n=1}^{N} j(n)}{N} \; (\mu\text{s})$$

- The percentage of valid frames that have an absolute jitter value over $\mathrm{Thr}_{\mathrm{MEEI}}$, which will be referred to as *Over*.

$$\text{Over} = 100 \frac{\left|\left\{ j(n) : j(n) > \mathrm{Thr}_{\mathrm{MEEI}} \right\}\right|}{N} \; (\%)$$

where $|A|$ denotes the cardinality of *A*, or otherwise the number of elements in the *A* set.

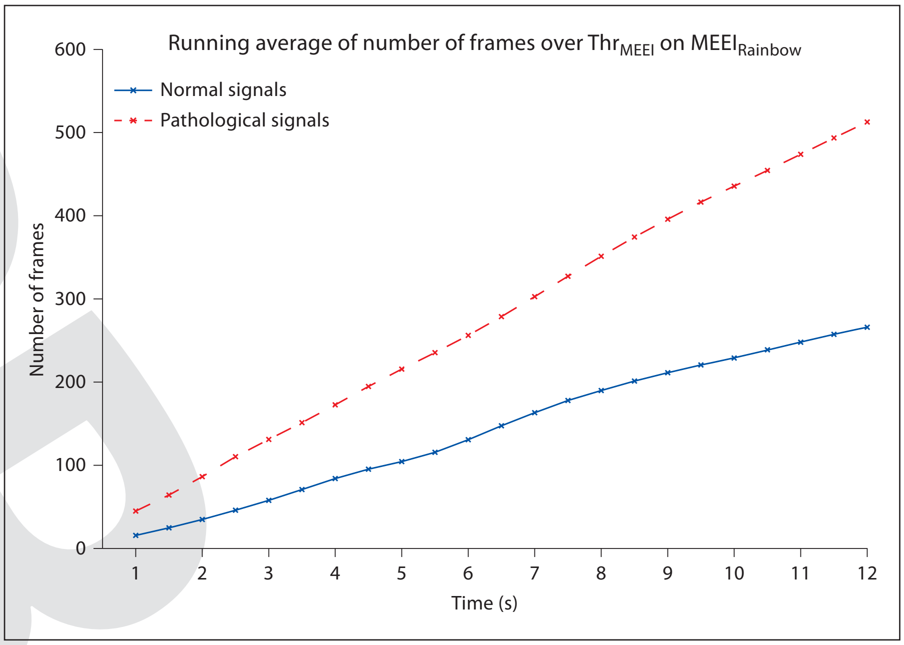
- The maximum number of consecutive valid frames that have an absolute jitter value over $\mathrm{Thr}_{\mathrm{MEEI}}$, which will be referred to as *Max Over*.

$$\text{Max Over} = \max(|\{ j(n) : j(n) > \mathrm{Thr}_{\mathrm{MEEI}} \text{ and consecutive frames}\}|) \text{ (scalar)}$$

- The maximum number of consecutive valid frames that have an absolute jitter value under $\mathrm{Thr}_{\mathrm{MEEI}}$, which will be referred to as *Max Under*.

$$\text{Max Under} = \max(|\{ j(n) : j(n) \le \mathrm{Thr}_{\mathrm{MEEI}} \text{ and consecutive frames}\}|) \text{ (scalar)}$$

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-term absolute jitter estimations for two signals from $\mathrm{MEEI}_{\mathrm{Rainbow}}$, one normal and one patholog-



Color version available online

Fig. 10. Running average number of frames over Thr$_{MEEI}$, for the normal (solid line) and pathological class (dashed line) on MEEI$_{Rainbow}$. The horizontal axis denotes the length of the analysis window in seconds.

Table 3. AUC score (standard error) for the four features based on the SJE short-term sequence in the MEEI$_{Rainbow}$ database that contains recordings of reading text from normal and pathological voices

AUC of features on MEEI$_{Rainbow}$ 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)

ical, are illustrated in figure 9. In the same figure, the threshold Thr$_{MEEI}$ is also depicted (dashed line). It is worth noting 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 case of pathology, while only 13% of the valid frames are above the same threshold for the normal signal. The *Max Over* and *Max Under* intervals for each signal are also indicated by arrows in figure 9. Specifically, for this example of pathological 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$_{MEEI}$ correctly separates the majority of the local jitter estimates. The AUC indices for the aforementioned features are given in table 3. Note that all cases show very good discriminative ability with an AUC index over 90%.

Since we have based the above features on a sequence of short-term jitter estimations, we are able to examine their gradual development in time, in terms of value and discrimination. In what follows, when we apply a feature gradually in time using a sliding analysis window of fixed size, we will refer to it 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 extent visualize the above results regarding the AUC scores, we computed the running average number of frames that are over the threshold Thr$_{MEEI}$ for normal and pathological voices by analyzing the entire MEEI$_{Rainbow}$ database. As an example, let's consider a running analysis window of 2 s for the case of the normal class of speakers. Then, for each recording in this class we compute the number of frames that are over the threshold in the current analysis window (from 0 to 2 s). If *L* is the number of recordings, obviously we compute *L* values. The running average of number of frames over the threshold for the current window is obtained by computing the average of these *L* values. Then, the running window is increased by 0.5 s (covering now the time interval between 0 and 2.5 s) and the corresponding running average is again computed. This

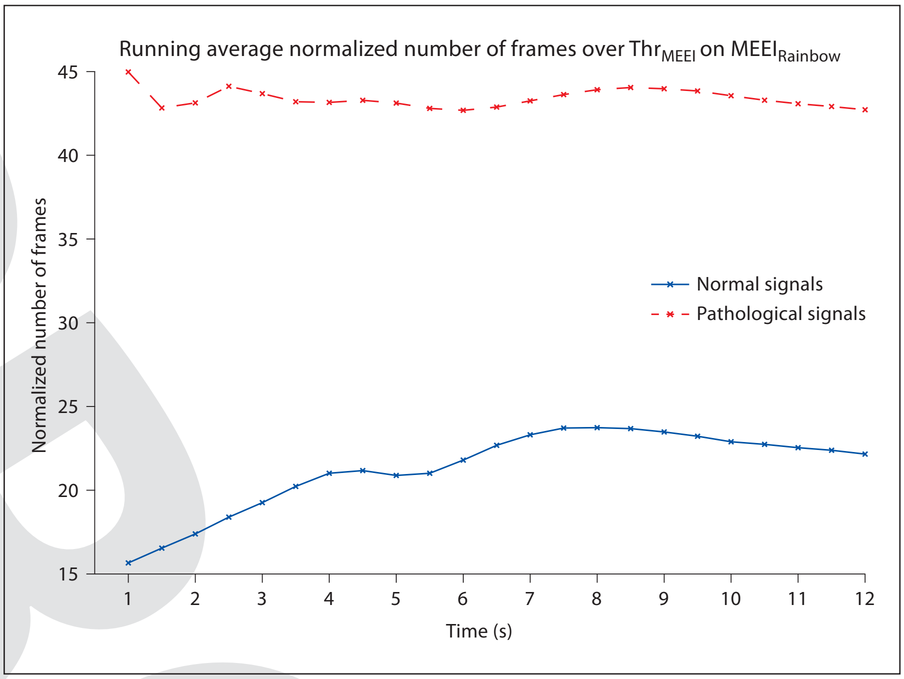


Fig. 11. Normalized running average number of frames over $\mathrm{Thr}_{\mathrm{MEEI}}$, for normal (solid line) and pathological speakers (dashed line) on $\mathrm{MEEI}_{\mathrm{Rainbow}}$. The horizontal axis denotes the length of the analysis window in seconds.

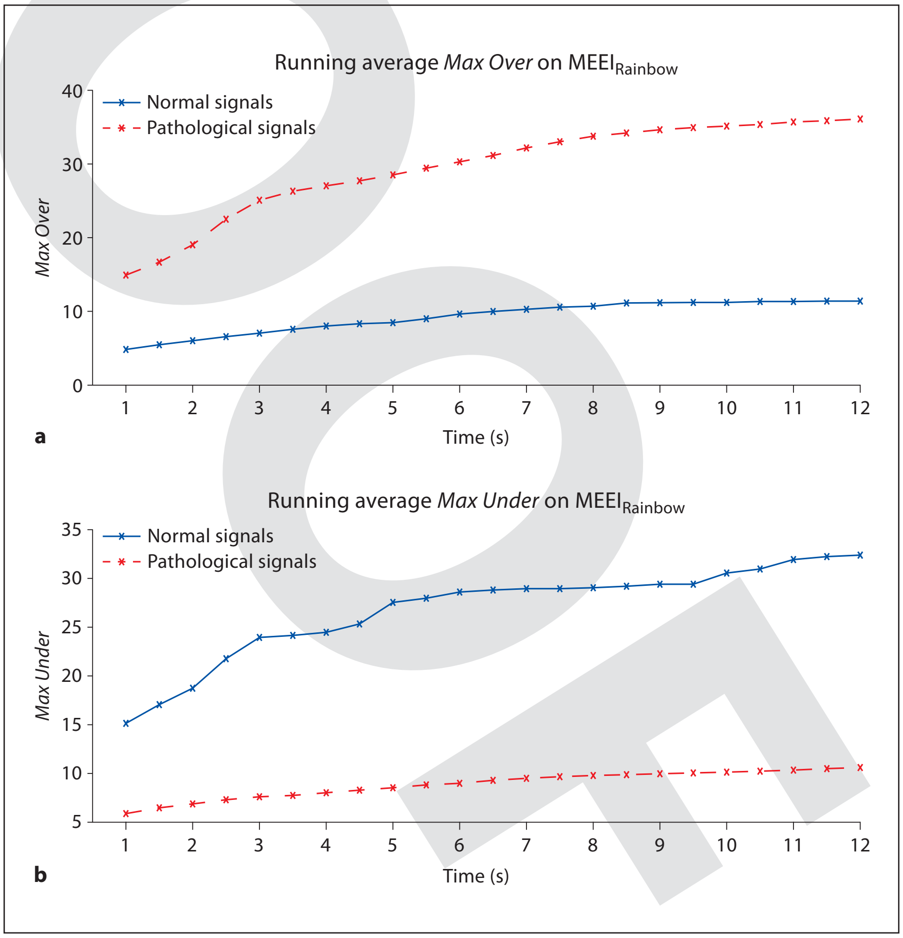


Fig. 12. The running average *Max Over* (**a**) and *Max Under* values (**b**) for the normal signals (solid line) and the pathological signals (dashed line) from $\mathrm{MEEI}_{\mathrm{Rainbow}}$.

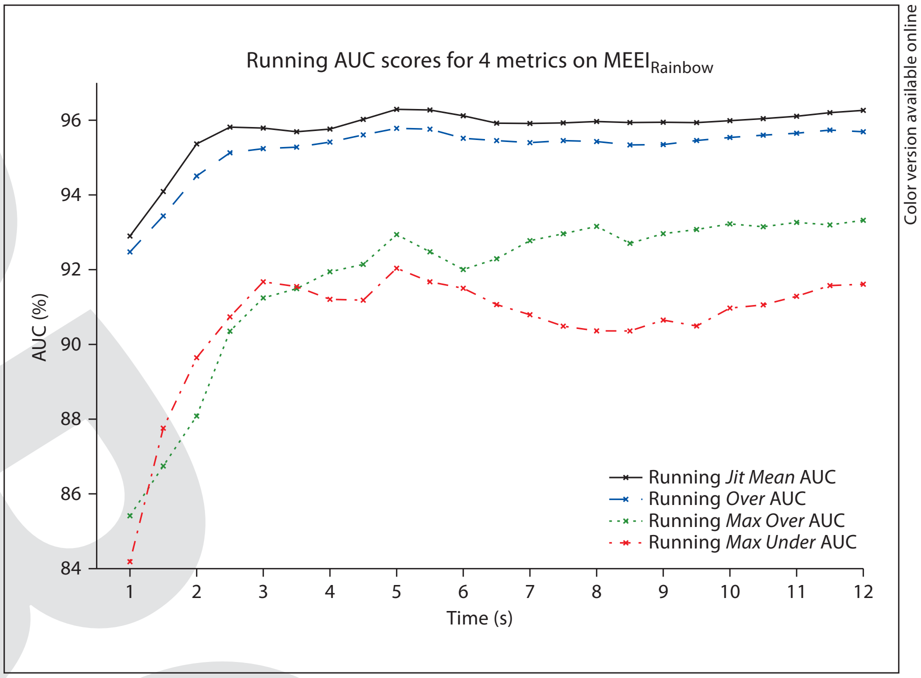


Fig. 13. The running AUC score of the four features applied on the $\mathrm{MEEI_{Rainbow}}$ database. The *Jit Mean* and *Over* features are quite discriminative even from the third second.

procedure is repeated until the running window spans the whole duration of the signal (12 s).

In figure 10, the running average of frames that are above the threshold $\mathrm{Thr_{MEEI}}$ is depicted for a running analysis window from 1 to 12 s, for the normal voices (solid line) and for the pathological voices (dashed line). This running average is equivalent to an average accumulator of the number of pathological frames. Therefore, as 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 interestingly, the increase rate of the pathological class is much higher than the corresponding increase rate of the normal one. In figure 11, 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 s (considering both voiced and unvoiced). Therefore, the numbers shown in the ordinate axis of figure 11 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 (2–3 s, for example 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 signals the corresponding number of frames remains about the same (45%).

In figure 12, 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 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-second length is 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-second mark. In a similar fashion, the running AUC scores of the four features are presented in figure 13. *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 figure 12, respectively.

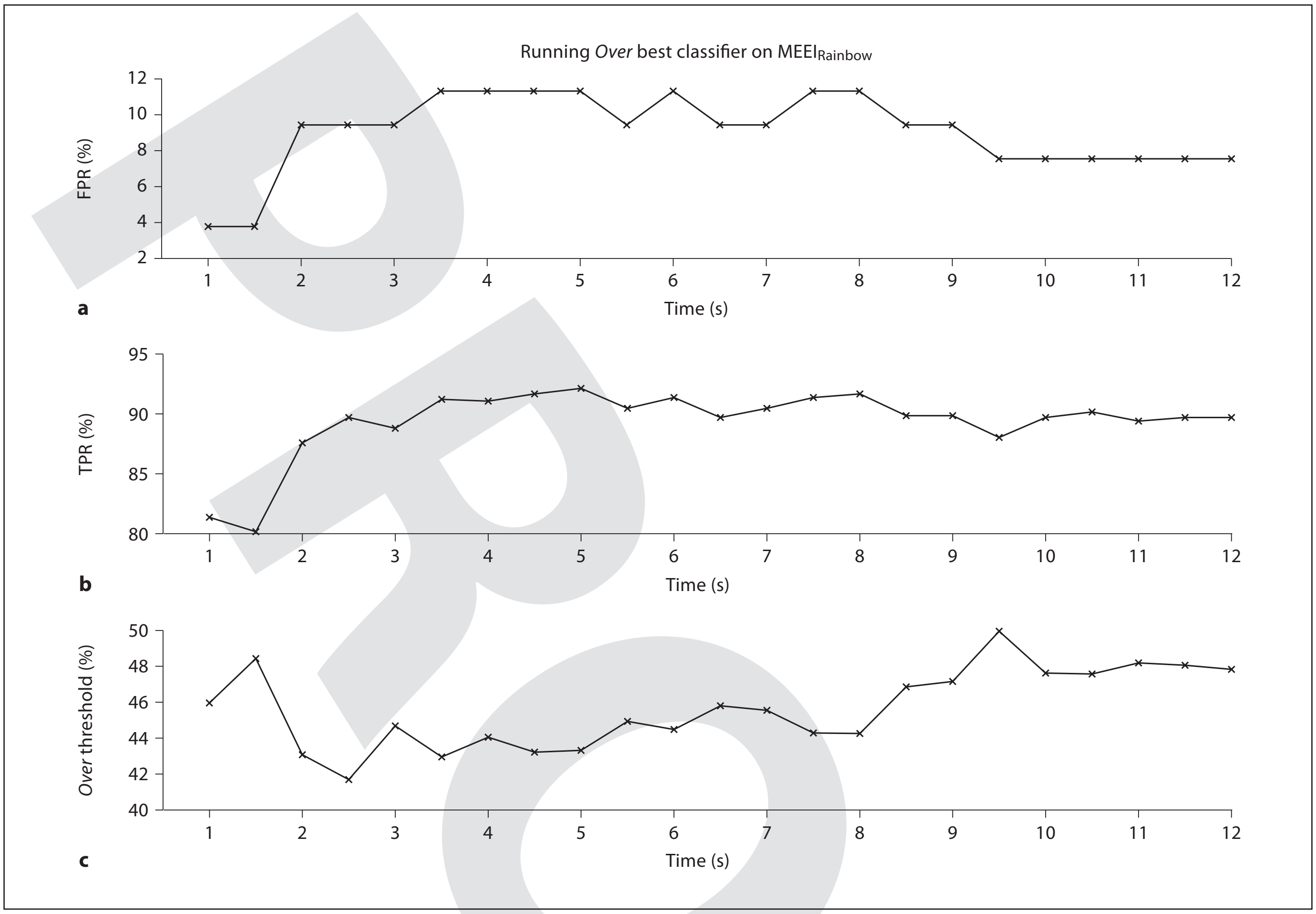
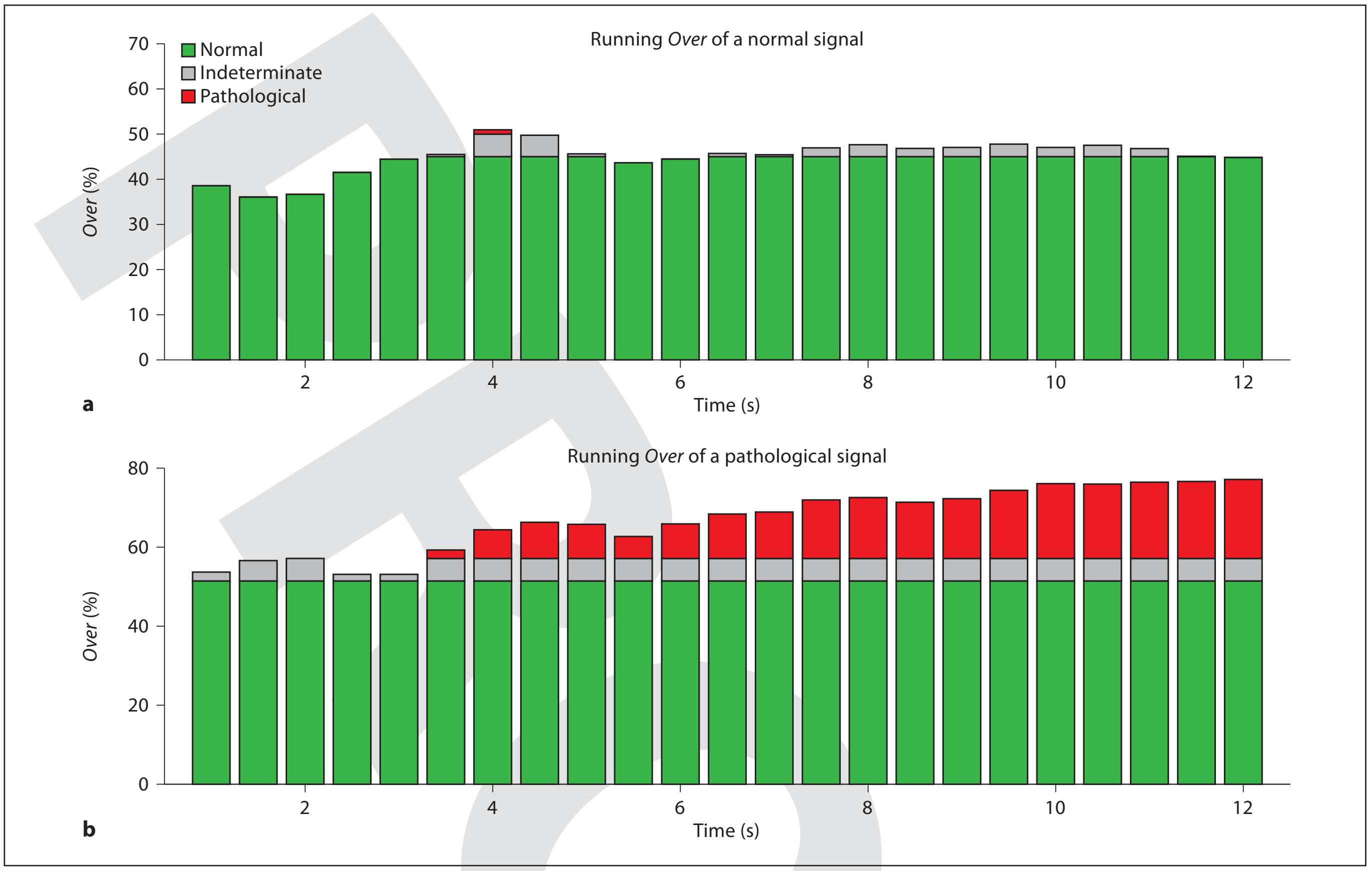


Fig. 14. FPR (**a**), TPR (**b**) and optimum threshold (**c**) for the *Over* feature as a function of time in the $\mathrm{MEEI}_{\mathrm{Rainbow}}$ database.

Among the short-term features we 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 figure 13, this is true even for signals of a small duration. Based on these results we 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, we will be able 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 figure 14. For example, for a running analysis window of 1.5 s, the best classifier (we recall that this is defined as the one having the highest distance between FPR and TPR) has a threshold of about 48% which corresponds to a TPR of 80% and an FPR of 4%. In the last 3 s, 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 $\mathrm{Thr}_{\mathrm{Over}}$. The region in between should be considered as an indeterminate area that indicates the need for further information regarding voice quality assessment. When we apply $\mathrm{Thr}_{\mathrm{Over}}$ to $\mathrm{MEEI}_{\mathrm{Rainbow}}$ we have a classification rate of 87.80%, with an additional 3.65% (26 files in total, 24 pathological and 2 normal) classified in the indeterminate (gray) area.



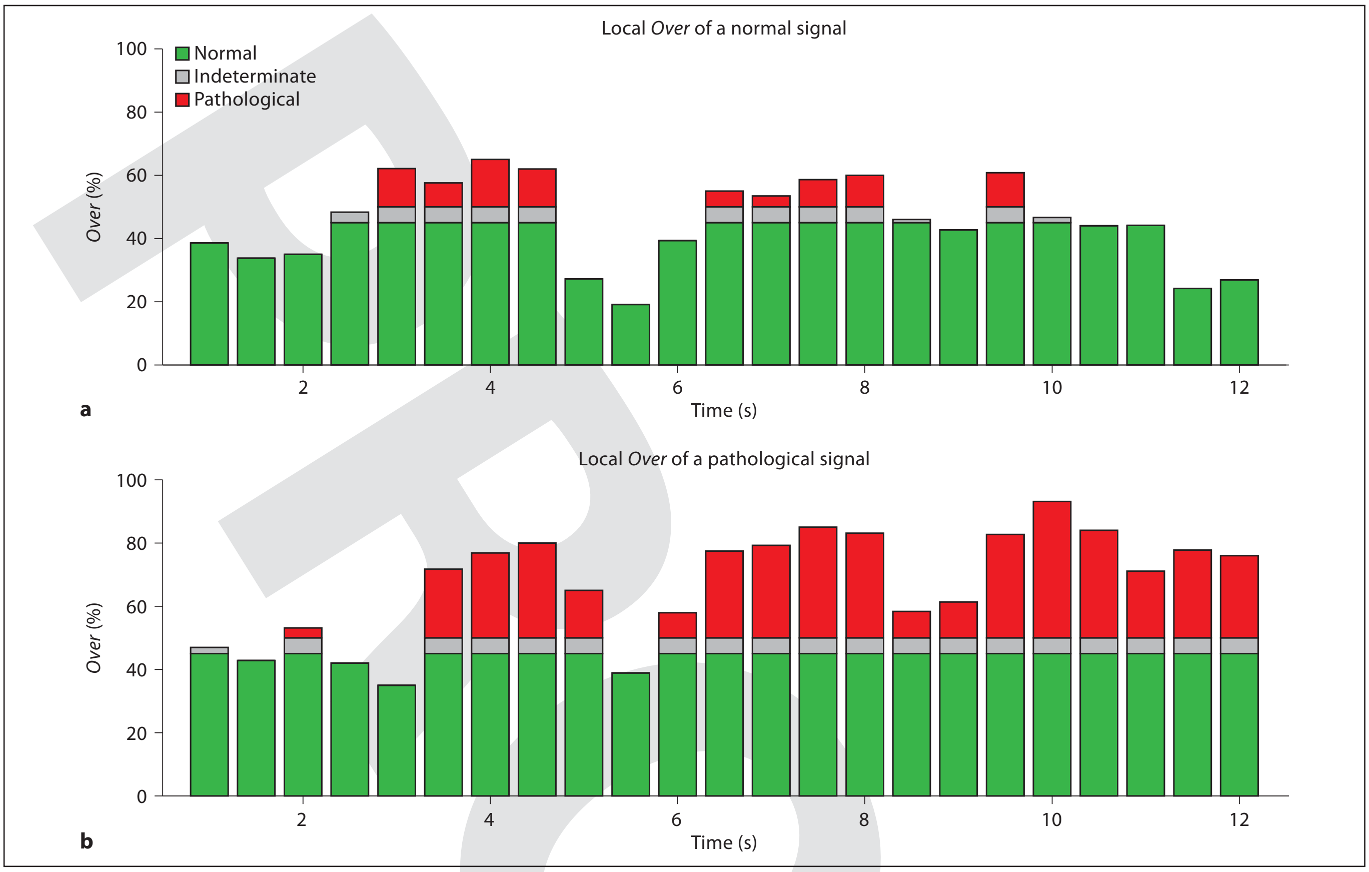
Color version available online

Fig. 15. The running *Over* estimate for a normal (**a**) and a pathological reading text signal (**b**). Given enough time the estimate settles in the normal or the pathological region.

An example of the potential use of Thr_{Over} is presented in figures 15 and 16, for one normal and one pathological signal. The running *Over* percentage for the two signals is shown in figure 15, while in figure 16 the corresponding local *Over* percentage is illustrated. The local *Over* is computed using a sliding window of 1 s shifted by half a second. For the particular normophonic signal, while the running *Over* feature is under Thr_{Over} almost exclusively (fig. 15a), in the local *Over* plot, it exceeds the threshold of pathology in some intervals (fig. 16a). Nonetheless, it does mostly remain in the normal region. 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. 15b), in the local *Over* estimates, it lies under the normal threshold for a few intervals only (fig. 16b). Hence, for the running *Over,* a recording of at least 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 figure 16 corresponds to intervals where there is a short rest of phonation. Therefore, just after these areas the local *Over* feature tends to decrease.

Conclusion

In this work, we expanded on the previously developed SJE by determining a relevant threshold for pathology and also by applying SJE on reading text recordings. Firstly, the suggested Thr_{MEEI} threshold results in high discrimination for normal versus pathological voices, in databases of either sustained vowel recordings or reading text recordings. Using this threshold and based on the time series of local jitter estimations from SJE, we introduced three new features that are highly correlated with



Color version available online

Fig. 16. The local *Over* value computed using a sliding window of 1 s duration shifted by 0.5 s, for a normal (**a**) and a pathological signal (**b**).

the existence of pathology and therefore they can be considered as good candidates for monitoring running 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).* Among them, the *Over* feature provided the best classification score. Furthermore, we established thresholds for the *Over* feature, $\mathrm{Thr}_{\mathrm{Over}}$, that can be used especially for monitoring the jitter effect in running speech.

Several statistical properties of jitter have been documented in the past. In this paper, 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-term 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 Endo and Kasuya [8] and in Schoentgen and de Guchteneere [32], the jitter time series is modeled as an autoregressive 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 [8]. Therefore, it is straightforward to apply similar time series modeling techniques to the short-term jitter sequence estimated by the SJE.

It will also be of great importance to test the suggested features and thresholds in signals recorded before and after successful therapy. Finally, it will be interesting to test the suggested ideas on databases with vocal loading.

Acknowledgments

The authors would like to thank Juan Ignacio Godino-Llorente of the Department of Circuits and Systems Engineering, Universidad Politécnica de Madrid, for the availability of the PdA database.

References

1 Lieberman P: Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. J Acoust Soc Am 1963; 35:344–353.

2 Schoentgen J, De Guchteneere R: Time series analysis of jitter. J Phon 1995;23:189–201.

3 Kreiman J, Gerratt BR: Perception of aperiodicity in pathological voice. J Acoust Soc Am 2005;117:2201–2211.

4 Pinto NB, Titze IR: Unification of perturbation measures in speech signals. J Acoust Soc Am 1990;87:1278–1289.

5 Feijoo S, Hernández-Espinosa C: Short-term stability measures for the evaluation of vocal quality. J Speech Hear Res 1990;33:324–334.

6 Baken RJ, Orlikoff RF: Clinical Measurement of Speech and Voice, ed 2. San Diego, Singular Publishing Group, 1999.

7 Rosa M, Pereira JC, Grellet M: Adaptive estimation of residue signal for voice pathology diagnosis. IEEE Trans Biomed Eng 2000;47: 96–104.

8 Endo Y, Kasuya H: A stochastic model of fundamental period perturbation and its application to perception of pathological voice quality. 4th Int Conf Spoken Lang Processing, Philadelphia, 1996, pp 772–775.

9 Parsa V, Jamieson DG: Acoustic discrimination of pathological voice: sustained vowels versus continuous speech. J Speech Lang Hear Res 2001;44:327–339.

10 Umapathy K, Krishnan S, Parsa V, Jamieson DG: Discrimination of pathological voices using time-frequency approach. IEEE Trans Biomed Eng 2005;52:421–430.

11 De Krom G: Acoustic correlates of breathiness and roughness: experiments on voice quality; thesis, Utrecht Institute of Linguistics OTS, Utrecht, 1994.

12 Fourcin A, Abberton E: Hearing and phonetic criteria in voice measurement: clinical applications. Logoped Phoniatr Vocol 2008;33: 35–48.

13 Gubrynowitz R, Mikiel W, Zarnecki P: An acoustic method for the evaluation of the state of the larynx source in cases involving pathological changes. Arch Acoust 1980;5: 3–30.

14 Askenfelt A, Hammarberg B: Speech waveform perturbation analysis revisited. Speech Transm Lab Q Prog Status Rep 1981;22:49–68.

15 Laver J, Hiller S, Mackenzie J, Rooney E: An acoustic screening system for the detection of laryngeal pathology. J Phon 1986;14:517–524.

16 Vasilakis M, Stylianou Y: A mathematical model for accurate measurement of jitter. MAVEBA 2007, Florence, 2007, pp 7–10.

17 Vasilakis M, Stylianou Y: Spectral jitter modeling and estimation. Biomed Signal Process Control, in press.

18 Boersma P, Weenink D: Praat: Doing Phonetics by Computer. Version 5.0.22. 2008.

19 Kay Elemetrics: Multidimensional Voice Program (MDVP). 2007.

20 Quatieri TF: Discrete-Time Speech Signal Processing: Principles and Practice. Upper Saddle River, Prentice Hall, 2002.

21 Murphy PJ: Spectral characterization of jitter, shimmer, and additive noise in synthetically generated voice signals. J Acoust Soc Am 2000;107:978–988.

22 Egan JP: Signal Detection Theory and ROC Analysis. New York, Academic Press, 1975.

23 Hanley JA, McNeil BJ: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology 1982;143: 29–36.

24 Kay Elemetrics. Disordered Voice Database (Version 1.03). 1994.

25 Godino-Llorente JI, Osma-Ruiz V, Sáenz-Lechón N, Cobeta-Marco I, González-Herranz R, Ramírez-Calvo C: Acoustic analysis of voice using WPCVox: a comparative study with Multi Dimensional Voice Program. Eur Arch Otolaryngol 2008;265:465–476.

26 Deliyski DD: Acoustic model and evaluation of pathological voice production. Eurospeech ’93, Berlin, 1993, pp 1969–1972.

27 Stylianou Y: Harmonic plus noise models for speech, combined with statistical methods, for speech and speaker modification; PhD thesis, Ecole Nationale Supérieure des Télécommunications, Paris, 1996.

28 Hollien H, Michel J, Doherty ET: A method for analysing vocal jitter in sustained phonation. J Phon 1973;1:85–91.

29 Koike Y, Takahashi H, Calcaterra TC: Acoustic measures for detecting laryngeal pathology. Acta Otolaryngol 1977;84:105–117.

30 Horii Y: Fundamental frequency perturbation observed in sustained phonation. J Speech Hear Res 1979;22:5–19.

31 Schoentgen J: Stochastic models of jitter. J Acoust Soc Am 2000;109:1631–1650.

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