A MATHEMATICAL MODEL FOR ACCURATE MEASUREMENT OF JITTER

Miltiadis Vasilakis\(^1\,\,^2\) and Yannis Stylianou\(^1\,\,^2\,*\)
\(^1\)Department of Computer Science, University of Crete, Hellas
\(^2\)Institute of Computer Science, Foundation of Research and Technology Hellas (FORTH)

Abstract: Jitter is a fundamental metric of voice quality. The majority of jitter estimators produce an average value over a duration of several pitch periods. This paper proposes a method for short-time jitter measurement, based on a mathematical model which describes the coupling of two periodic phenomena. The movement of one of the two periodic phenomena with respect to the other is what is considered as jitter and what the proposed method measures. Through tests with synthetic jitter signals it has been verified that the suggested method provides accurate local estimates of jitter. Further evaluation was conducted on actual normal and pathological voice signals from the Massachusetts Eye and Ear Infirmary (MEEI) Disorders Voice Database. Compared with corresponding parameters from the Multi-Dimension Voice Program (MDVP) and the Praat system, the proposed method outperformed both in normal vs. pathological voice discrimination.

Keywords: Jitter, short-time, pathological voice.

I. INTRODUCTION

Evaluation of voice quality is an essential diagnostic aid for the assessment of pathological voice. Methods based on acoustic analysis have several advantages. In comparison to methods such as videoendoscopy or electroglottography (EGG), they cost less, require less time and are non-invasive for the patient. Furthermore, acoustic analysis can produce automatic quantitative results, which, apart from assisting clinical doctors, can be exploited for unsupervised classification of a voice as pathological or normal, or even detect specific cases of dysphonia.

The main effect of a pathological condition, as we perceive it, is noise. The parameters produced by acoustic analysis for voice quality, usually quantify the presence of this aperiodic component; mainly additive noise, such as in cases of breathiness, or modulation noise, such as in cases of roughness. Further regarding modulation noise, this can be detected either in frequency, called jitter, or in amplitude, called shimmer. Jitter is defined as perturbations of the glottal source signal that occur during phonation and affect the glottal pitch period. The measurement of jitter can be performed by using the radiated speech signal, or by using measurements of glottal conductivity through (EGG). The computation may take place in the time domain, in the frequency domain (magnitude spectrum), or using cepstrum.

Several methods have been proposed for the computation of quantitative values for jitter. Time domain methods are usually based on pitch period measurements that are used to estimate an average value of jitter, over a number of several periods. If \( N \) is the total number of pitch periods and \( u(n) \) is the pitch period sequence, the definitions of widely accepted jitter measurements are given below.

Local jitter is the period-to-period variability of pitch \((\%)
\[
\frac{1}{N-1} \sum_{n=1}^{N-1} |u(n+1) - u(n)|
\] (1)

Absolute jitter is the period-to-period variability of pitch in time
\[
\frac{1}{N-1} \sum_{n=1}^{N-1} |u(n+1) - u(n)|
\] (2)

Relative Average Perturbation (RAP) jitter provides the variability of pitch with a smoothing factor of 3 periods \((\%)
\[
\frac{1}{N-2} \sum_{n=1}^{N-2} \frac{2u(n+1) - u(n) - u(n+2)}{3} \frac{1}{N} \sum_{n=1}^{N} u(n)
\] (3)

Pitch Period Perturbation Quotient (PPQ) provides the variability of pitch with a smoothing factor of 5 periods \((\%)
\[
\frac{1}{N-4} \sum_{n=1}^{N-4} \frac{4u(n+2) - u(n) - u(n+1) - u(n+3) - u(n+4)}{5} \frac{1}{N} \sum_{n=1}^{N} u(n)
\] (4)

The pitfalls with such techniques is that they heavily rely on a periodicity that doesn’t actually exist in speech, while some methods specifically provide a jitter value that is a percentage of that notion of periodicity. In order to overcome this problem (the existence of non-periodicity), a standard solution is to perform a low pass filtering before pitch estimation, which solution essentially destroys the
details of the speech signal; it reduces the effect of non-periodicity, which is however what we would like to measure.

An alternative to calculating an average value for jitter, is that of short-time tracking. A sequence of jitter values on small intervals can be more precise without assuming long-term periodicity and may even provide better insight on the evolution of pathological voices. In this work we suggest the use of a mathematical model that enables us to combine two periodical phenomena, in order to achieve the local aperiodicity. Based on that, we identify jitter as the movement of one of the two periodic phenomena with respect to the other. This movement is exactly what we try to measure. Using such a model we are able to calculate the value of short-time jitter with high precision. Comparison was made with the corresponding jitter measurements provided by PRAAT [1] and Multi-Dimensional Voice Program (MDVP) [2] of Kay-Pentax, on the database Massachusetts Eye and Ear Infirmary (MEEI) Disordered Voice Database [3].

The paper is organized as follows. In section II we present the mathematical model we propose and the method we derived from it to measure short-time values of jitter. The conducted experiments and their results are presented in section III. Section IV concludes the paper.

II. METHOD

![Figure 1: Glottal impulse train of the proposed jitter model.](image)

Jitter may be expressed as a perturbation on the glottal excitation impulse train. A simple mathematical model can be obtained by considering a cyclic perturbation, with pitch deviation of a constant value, applied every second impulse [4]. The glottal impulse train can be expressed as

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

where \( P \) is the pitch period and \( \epsilon \) is the pitch deviation, both in samples. This model, shown in Fig. 1, realizes the combination of two periodic phenomena and \( \epsilon \) is the movement that corresponds to the local aperiodicity of jitter and therefore the value we should seek to measure. The value of \( \epsilon \) can range from 0 (no jitter) to \( P \) (pitch halving).

The power spectrum of the impulse train can be shown to be

\[ |P(\omega)|^2 = 2(1 + \cos[(\epsilon - P)\omega]) \left[ \sum_{k=-\infty}^{+\infty} \frac{2\pi}{P}\delta(\omega - k\frac{2\pi}{P}) \right]^2 \]

\[ = 2(1 + \cos[(\epsilon - P)\omega]) \sum_{l=-\infty, k=2l}^{+\infty} \frac{\pi^2}{P^2}\delta(\omega - l\frac{2\pi}{P}) + \sum_{l=-\infty, k=2l+1}^{+\infty} \frac{\pi^2}{P^2}\delta(\omega - l\frac{2\pi}{P} - \frac{\pi}{P}) \]

The last part can be written as

\[ |P(\omega)|^2 = H(\epsilon, \omega) + S(\epsilon, \omega) \]

where \( H(\epsilon, \omega) \) is the influenced by jitter harmonic part of the power spectrum, while \( S(\epsilon, \omega) \) is the subharmonic part that appears because of the jitter.

The two power spectra for various values of \( \epsilon \) are depicted in Fig. 2. We observe that the harmonic and subharmonic parts for a certain value of \( \epsilon \) crossover that many
times. The structure remains the same also on the output from a linear system when the input is the impulse train \( p[n] \).

Based on this perceived structure of power spectra a short-time jitter estimator has been developed. Initially, for a given speech signal, a pitch estimation takes place that provides us with a temporal sequence of the pitch period. A sliding frame is used to allow us to examine the signal gradually in time. The size of the frame can be a period. A sliding frame is used to allow us to examine the signal gradually in time. The size of the frame can be either fixed to 4 times the average pitch period, or variable to 4 times the local pitch period. The frame step used is accordingly either one average pitch period, or one local pitch period. A hanning window is then applied to the frame and the power spectrum is computed. The size of the Discrete Fourier Transform is that of the smallest power of 2 that is closest to the length of the frame. From the power spectrum, the harmonic and subharmonic parts are taken into account, and by counting the number of crossings between them, the jitter value of the current frame is estimated. In order to overcome potential spectrum resolution problems, a threshold is used to determine if a crossing has occurred. If the harmonic and subharmonic parts, after a candidate crossing, never reach a difference over the threshold value, before the next potential crossing, then it is not regarded as one. Through testing, the threshold value has been set to 3dB. In the end a short-time jitter sequence with integer values (i.e. in samples) is obtained. Taking into account the sampling frequency of the signal the sequence is converted to μsec. It is evident, that the larger the sampling frequency, the larger the resolution of the measurement.

III. EXPERIMENTS AND RESULTS

In order to verify the validity of the proposed method, in theory and in practice, experiments were carried out with both synthetic and actual pathological voice signals. The actual signals were taken from the MEEI Disordered Voice Database [3].

A. Synthetic Signals

The synthetic signals were created using glottal impulse trains as described in (5). These were used to excite an AR model of order 50, extracted from a sustained recording of vowel /a/, with an average fundamental frequency of 125Hz. This was done for sampling frequencies of 16 and 48kHz, and for \( \epsilon \) values from 0 to 10% of each pitch period. The duration of the signals were set to 1sec.

Using a fixed frame size, with knowledge of the actual pitch period, we did confirm our observations. The structure of the glottal excitation was maintained on the final signal and exact measurement of the short-time jitter was possible. Fig. 3 shows the power spectrum of a frame of the synthetic signal, with sampling frequency 48kHz and \( \epsilon = 5 \). The crossings counted correspond to the jitter movement, while two false crossings are correctly rejected.

To verify our results, we used as a reference the Praat [1] system. The absolute jitter (2) measurement as it is implemented in Praat [Jitter (local, absolute)] was used. Since our method calculates a sequence of short-time values, we used for comparison the average jitter value. Having in mind Fig. 1, absolute jitter (2) would return a jitter value of \( 2 \times \epsilon \), while the average value we measure is \( 1 \times \epsilon \). To do an analogous comparison we use double the average jitter value.

The error difference between the actual jitter value and the results of our method and Praat are presented in Fig. 4, for 16 and 48kHz. The proposed method has zero error, while the error difference of Praat is of the order of some μseconds for all \( \epsilon \) values, except three cases in the 48kHz, where Praat determined the signals as unvoiced and didn’t return jitter measurements.

B. MEEI Disordered Voice Database

The MEEI Disordered Voice Database contains sustained vowel and reading text samples, from 53 subjects with normal voice and 657 subjects with a wide variety of pathological conditions. Also included for most of the signals were the acoustic analysis parameters produced by the Multi-Dimensional Voice Program (MDVP) [2]. For the purpose of our experiments, all 53 of the normal voice samples and 632 of the pathological voice samples were used, and specifically the sustained recordings of vowel /a/. The excluded pathological voice samples were the ones that lacked the MDVP parameters. The sampling frequency of the selected signals were originally either 25 or 50kHz, with the normal voice ones only of 50kHz. To
avoid potential correlation of the results with sampling frequency, all signals used in this paper were resampled to 25kHz.

For the pitch estimation required by the proposed method, YIN [5] was used with default parameters. Both fixed and variable frame size experiments took place. The computed short-term jitter sequence, for each sample, was averaged and doubled, to provide a single jitter measurement. The MDVP Jita parameter and the Praat Jitter (local, absolute) value, both implementations of (2), were used for comparison. Receiver Operating Characteristic (ROC) curves for the four measurements in contest are portrayed in Fig. 5. The proposed method outperforms in discrimination, between pathological and normal samples, both MDVP and Praat, with the fixed frame case having slightly better results over the variable frame case.

IV. CONCLUSION

We proposed a method for short-time jitter evaluation, based on a mathematical model of two periodic phenomena. The experiments conducted with synthetic signals verified that the method produces accurate local estimates of jitter. Regarding pathological voice classification, it was shown that the average value of the proposed method is more discriminant than two standard implementations of absolute jitter (2) measurement, namely MDVP and Praat.

The fact that the proposed method allows us to see the behavior of local jitter in time, is something that we plan to examine in depth. Knowledge of the gradual development of jitter, apart from being of use in voice quality evaluation, it may also be useful in automatic pathological condition discovery.

Jitter also contributes to the appearance of noise in the spectrum. This presents problems regarding the computation of a Harmonics to Noise Ratio (HNR) estimate. Identifying in the magnitude spectrum the noise induced by jitter, may provide a more accurate HNR value [6]. The points where the crossings between the harmonic and subharmonic parts of the power spectrum occur, are good candidates for deciding which parts of the spectrum noise should not be considered additive but structural.

REFERENCES