Perturbation measurements in highly irregular voice signals: Performances/validity of analysis software tools

C. Manfredi, A. Giordano, J. Schoentgen, S. Fraj, L. Bocchi, P.H. Dejonckere

Abstract

In this paper we present results concerning validity of jitter measurement in strongly irregular voice signals (sustained vowels) moderately corrupted by noise. The performance of four tools for voice analysis is compared on synthetic signals as far as fundamental period and jitter estimation are concerned. Synthesised vowels offer the advantage of a perfect control of the amount of jitter put in.

Though implementing the same formula for jitter estimation, the results obtained with these approaches become quite different for increasing jitter. The reason could be searched in the different methods used for the separation of voiced and unvoiced frames as well as for fundamental period estimation.

Results show that all the tools give reliable results up to a jitter level J>15%, that encompasses the maximum value J=12% as obtained by expert raters by visual inspection. Hence, up to this limit, the tools presented here for jitter estimation can give a valid support to clinicians also in term of reproducibility of results and time saving.

For jitter values larger than 15% all programs tend to underestimate the true jitter value, but with large differences among them. Just two methods succeed in estimating jitter values up to and larger than 20% and could thus be better suited for perturbation measure in strongly irregular voice signals.

1. Introduction

Quantitative analysis of irregularity in vocal emissions is usually performed through irregularity measures among which the well known and most used jitter. Jitter is a measure of the periodicity of the signal measuring the changes in fundamental period T0 from cycle to cycle. Of course, “good voices” should have low jitter.

Voice signals can be classified in three categories: type 1 signals are nearly-periodic; type 2 signals contain intermittency, strong subharmonics or modulations; type 3 signals are chaotic or random. Perturbation analysis has considerable utility in the analysis of type 1 signals as perturbation measures, including jitter, assume the presence of a periodical or quasi-periodical signal [1].

A quantitative evaluation of jitter requires a two-step process: in the first step, the algorithm has to assess the presence or absence of periodicity in the signal. A major challenge is thus an effective separation between voiced and unvoiced frames. Only frames classified as “voiced”, i.e. having a periodical structure (type 1 signals), can in fact be processed in the second step. This step involves a correct estimation of the fundamental period (i.e. T0) and subsequently of the amount of jitter. However, currently most of the software tools...
for voice analysis have limitations concerning effective analysis of strongly irregular voices due to low signal stationarity.

The general problem of $T_0$ estimation consists in taking a frame of signal and finding the dominant frequency of repetition. However most real signals are quasi-periodic, may show changes in fundamental frequency over time and can be corrupted by noise. The $T_0$ extraction method directly affects the accuracy of the measures, particularly if several waveform types (with or without formant structure) are under consideration and if noise and modulation are present in the signal. Differences in $T_0$ estimation among different software tools may therefore arise from differences in either of the two steps mentioned above.

As most of the available software tools implement the same formula for jitter estimation (see Eq. (2) in Section 2), the difference among various jitter measures could mainly be due to differences in the technique applied to measure $T_0$. Some programs perform waveform-matching, in which the length of a cycle is determined by looking for the best matching between wave shapes (a “cross-correlation” maximum), while others are based on peak-picking, where the duration of a cycle is determined by measuring the time difference between two locally highest peaks in the waveform. It was shown that the waveform-matching method averages away much of the influence of additive noise, whereas peak-picking is highly sensitive to additive noise [2].

As a practical guideline, perturbation measures less than about 5% have been found to be valid [1]. In the same paper, authors concluded that waveform-matching performs best in signals with a frequency variation below 6% per cycle. Possible reasons why this is so are profoundly explored and discussed in [3].

Several studies have been performed to investigate the inter-program effectiveness of jitter measures. Perturbation measures from CSpeech, Computerized Speech Laboratory, SoundScope, and a hand marking voice analysis system were compared using sustained vowels with mild to severe dysphonia [4]. In this paper poor rank order correlations between programs using similar measures of perturbation were noted. In [5] Karnell et al. compare MDVP, CSpeech and AUDEED on a set of real voice signals showing that jitter measures were poorly correlated. In [6] rating reliability within and across listeners was compared to the reliability of jitter measures produced by several voice analysis systems (CSpeech, SoundScope, Computerized Speech Lab (CSL), and an interactive hand-marking system). Results showed that overall listeners agreed as well as or better than “objective” algorithms. More recently, acoustic voice analysis as calculated by two different analysis systems (Doctor Speech and CSL) were compared [7] on a group of normal voices and were found not comparable in absolute figures, but with identical judgment against normative data. Maryn et al. [8] compare several spectral, cepstral and perturbation measures on tracheo-sphagial voices with poor significance of perturbation measures in this case.

In the present paper, four $T_0$-extraction methods and their software implementations are considered and compared as far as jitter estimation is concerned. The specificity of our work is that it concerns strongly perturbed signals, and that we have an exact reference. In fact, instead of using real data, the comparison is based on synthetic voice signals, whose characteristics are thus perfectly known. Specifically, MDVP (peak-picking) [9], PRAAT (waveform matching) [10], AMPEX (waveform matching) [11] and BioVoice (waveform matching) [12] are considered and tested on synthesized vowel/a with added noise and increasing levels of jitter [13].

Results are compared to those obtained by visual inspection made by three experienced raters who performed manual selection of glottal cycles by means of an option implemented in one of the above mentioned tools. The level of jitter up to which the raters agree in defining boundaries of successive cycles is in fact the level an analysis program at least should reach.

A detailed analysis of the four tools would require knowledge of full details of each algorithm, which is seldom available. However it could be reasonably assumed that the frames classified as voiced have stronger periodicity with respect to unvoiced ones. Hence, with high jitter levels, the programs disregard a large part of frames as they are classified as “unvoiced” due to low periodicity. As a consequence, voice parameters related to irregularities in the signal, such as jitter, are estimated on a subset of the whole signal made up by more regular frames only, thus giving rise to jitter underestimation.

Moreover, different programs use quite different methods to determine whether an irregular part of the signal is voiced or not. A comparison of PRAAT and MDVP points out the differences between the two programs that influence the measurement of the height of the autocorrelation peaks, that is taken as a criterion for voicing: if it is higher than the voicing threshold, the frame is considered voiced, otherwise voiceless. In PRAAT, the standard voicing threshold is higher than in MDVP, which suggests that MDVP tends to regard more frames as voiced than PRAAT. The two programs were shown to have a comparable sensitivity in measuring small jitter values with large difference when noise is added [10,14].

AMPEx and BioVoice perform a slightly different search for voiced frames, which also takes into account voice continuity that is a proximity operator which is intended to overlook small errors in peak positions determined from the autocorrelation function. This approach seems to partially overcome the limitations in MDVP and PRAAT, as described in Section 3 devoted to experimental results.

The next section is devoted to the description of the voice synthesizer program and the four tools considered here: MDVP, PRAAT, AMPEx and BioVoice. The analysis will be restricted to the parameters of interest with this paper, i.e. selection of voiced/unvoiced frames, fundamental frequency and jitter estimation. The section ends with the description of the perceptual rating made by experienced clinicians. Experimental results will be presented and discussed in Section 3.

2. Methods

2.1. Synthesis of hoarse voices

The software tool used here for the synthesis of the disordered voices was developed by some of the authors of this paper [13]. It involves four stages that are:

1. generation of a sinusoidal driving function the instantaneous frequency of which is disturbed to simulate vocal frequency jitter;
2. modelling of the glottal area via a pair of polynomial distortion functions into which the (pseudo-)harmonic driving function is inserted;
3. generation of the volume velocity at the glottis, including acoustic tract-source interactions, via an algebraic model;
4. simulation of the propagation of the acoustic waves in the trachea and vocal tract.

Vocal jitter is simulated by perturbing stochastically the instantaneous frequency of a co-sinusoidal driving function $A \cos[\theta(n)]$.

$$\begin{align*}
\theta(n) &= \theta(n-1) + 2\pi F_0 \Delta + 2\pi b \xi(n) \\
\xi(n) &= \sqrt{\Delta} \begin{cases}
  +1, & p = 0.5 \\
  -1, & p = 0.5
\end{cases}
\end{align*}$$  

where $\theta$ = phase of the driving cosine, $\Delta$ = sampling time step and $F_0$ = typical fundamental frequency that is continually perturbed stochastically with probability $p$ by small positive and negative-going unit pulses the sizes of which are equal to $b$. 

This model simulates frequency jitter by perturbing vocal frequency \( F_0 \) sample by sample by a small amount. Observed frequency jitter is the outcome of these many small disturbances over one glottal cycle. Additional details regarding the simulation of vocal jitter can be found in [13,15–17].

In a second step, the co-sinusoidal driving function takes the place of \( x \) in a pair of polynomials

\[
P(x) = c_0 + c_1 x + c_2 x^2 + c_3 x^3 \ldots \quad \text{and} \quad Q(x) = d_0 + d_1 x + d_2 x^2 + d_3 x^3 \ldots
\]

The weighted sum: \( P[A \cos(\theta)] + A \sin(\theta) \) \( Q[A \cos(\theta)] \) simulates the glottal area function. The polynomial sum outputs a cyclic signal, the cycle lengths of which are identical to the cycle lengths of the harmonic driving functions. When amplitude \( A \) equals 1, the cycle shape is equal to the shape of the template of the glottal area whose Fourier series coefficients have been used to compute polynomial coefficients \( c_i \) and \( d_i \). The template glottal area waveform is a non-symmetric Klatt model, with feeble skewing to the right.

Volume velocity \( \dot{v}_g(t) \) at the glottis is obtained by means of an algebraic model that comprises the glottal area as well as the cross-sections of the tracheal and epi-laryngeal ducts below and above the glottis. In addition, the model involves the amplitudes of the infra and supra-glottal acoustic pressure waves moving towards the glottis, as well as physical constants such as the speed of sound, the density of air and the glottal entry and exit losses. The model simulates the generation of the pulsatile volume velocity at the glottis and takes into account the acoustic interaction between the source and vocal tract as well as trachea [18].

Additive noise owing to turbulence in the vicinity of the glottis, which gives rise to audible breathiness, is simulated by means of low-pass filtered white Gaussian noise that is multiplied by an affine function \( a + b \dot{v}_g(t) \) of the volume velocity. The volume velocity-modulated noise then is delayed and added to the clean volume velocity obtained via the algebraic model described above, which is in agreement with published observed and simulated data [19,20]. Before adding noise to the volume velocity, the square of the volume velocity samples and the square of the noise samples have been summed separately over the signal length. The log-ratio of both sums multiplied by ten is the signal-to-noise ratio in dB at the glottis. Even for signals rated as very breathy, this ratio is > 17 dB. The explanation is that speech sound radiation at the lips favors high over low frequencies, that is, the perceived noise is broadband compared to the glottal noise. The affine function causes the noise to be pulsatile when the glottis is open and zero when the glottis is closed. This means that the glottal noise energy averaged over the signal length is a flawed predictor of perceived breathiness. The feeble delay between noise and clean glottal volume velocity takes into account that the noise is generated above the glottis [13].

Finally, the simulated glottal volume velocity, which is the acoustic excitation signal, is inserted into the vocal tract and tracheal models. Trachea and vocal tract are mimicked by a concatenation of cylindrical ducts of identical lengths, but different cross-sections. The tract cross-section values rest on published data [21,22]. Wall vibration losses are taken into account by inserting an auxiliary tube at each junction [23]. Viscous and heat conduction losses are simulated via digital filters [24]. Lip radiation and losses at the glottis are modelled according to [25,26]. Tracheal losses are taken into account by means of a real attenuation coefficient at the lung end.

### 2.2. MDVP

This is one of the most widely used tool for voice analysis in the biomedical field. MDVP consists of the following main steps: \( F_0 \) estimation, \( F_0 \) verification, \( F_0 \) extraction and computation of voice parameters [9].

\( F_0 \) estimation is based on short-term autocorrelation analysis of the voice signal on frames of 30 ms or 10 ms of length depending on the \( F_0 \) extraction range (67–625 Hz or 200–1000 Hz respectively) at a sampling rate of 50 kHz. A low pass filter at 1800 Hz is applied to every signal frame before coding. A coding threshold \( K_p = 0.78 \) is applied in order to eliminate incorrect classification of \( F_0 \)-harmonic components as \( F_0 \). The current window is considered to be voiced with period \( T_0 = \tau_{\text{max}} \) if the global maximum of the autocorrelation \( R_{\text{max}} = R(\tau_{\text{max}}) \) is larger than \( K_p R(0) \), where \( K_p \) is a properly selected voiced/unvoiced threshold and \( R(0) \) is the value of the autocorrelation \( R(\tau) \) at \( \tau = 0 \).

\( F_0 \) verification: The autocorrelation is computed again for the same windows at \( K_p = 0.45 \) in order to suppress the influence of components sub-harmonic to \( F_0 \). The results are compared to those obtained in the previous step and the decision about the correct \( T_0 \) is made for all windows where a difference between the two estimates is discovered.

\( F_0 \) extraction: It is made on the original signal using a peak-to-peak extraction measurement. It is synchronous with the verified pitch and voiced/unvoiced results computed in the previous steps. A linear 5-points interpolation is applied on the final results in order to increase the resolution, that is necessary for meaningful frequency perturbation measurements.

Jitter is evaluated as follows:

\[
\text{Jitter} = \frac{\left(1/(n-1)\right)\sum_{i=0}^{n-1} |I_i - I_{i+1}|}{(1/n) \cdot \sum_{i=0}^{n-1} |I_i|}
\]

where \( T = 1/F_0 \) is the fundamental period and \( n \) is the number of frames.

### 2.3. PRAAT

PRAAT is a short-term analysis method, hence the analysis is performed for a number of small segments (frames) that are shifted (default length is 0.01 s). The program consists of the following main steps [10]:

Step 1. Pre-processing to remove the side lobe of the Fourier transform of the Hanning window for signal components near the Nyquist frequency.

Step 2. Compute the global absolute peak value of the signal.

Step 3. For every frame, look for at most lag-height pairs that are good pitch candidates for the periodicity of this frame. To find the candidates, the following steps are taken for each frame:

Step 3.1. Take a segment from the signal whose length (the window length) is determined by the lowest fundamental frequency to be detected. The window should be just long enough to contain three periods (for pitch detection) or six periods (for HNR measurements).

Step 3.2. Subtract the local average.

Step 3.3. The first pitch candidate is the unvoiced candidate, which is always present. The strength (local maximum of the normalised autocorrelation) of this candidate is computed with two soft threshold parameters, voicing threshold (Vt) and silence threshold (St). A frame is regarded as locally unvoiced if it has a voicing strength below Vt (whose standard value is 0.45 relative to the maximum possible autocorrelation), or a local peak below St (whose standard value is 0.03 relative to the global maximum amplitude).

Step 3.4. Multiply by a window function, perform a Fast Fourier Transform, square the samples in the frequency domain and perform an inverse Fast Fourier Transform. This gives a sampled
version of the normalized autocorrelation of the windowed signal \( r_n(t) \).
Step 3.5. Divide by the autocorrelation of the window. This gives a sampled version of the autocorrelation of the signal \( r_n(t) \).
Step 3.6. Find the positions and heights of the maxima of the continuous version of \( r_n(t) \) as defined in [10]. The only positions considered for the maxima are those that yield a pitch between Min_Pitch and Max_Pitch. The Max_Pitch parameter should be between Min_Pitch and the Nyquist frequency. The only pitch candidates that are saved are the unvoiced candidate and the voiced candidates with the highest values of the local strength. The locally best candidate in each frame is the one with the highest local strength.
Step 4. As one can have several approximately equally strong candidates in any frame, one launches on these pairs the "global path finder", the aim of which is to minimize the number of incidental voiced-unvoiced decisions and large frequency jumps. For every frame \( n \), \( p_n \) is a number between 1 and the number of candidates or that frame. The values \( |p_n| \leq n \leq \text{number of frames} | \) define a path through the candidates. With every possible path we associate a cost. The globally best path is the path with the lowest cost.

Jitter is evaluated according to Eq. (2).

2.4. AMPEX

The AMPEX algorithm [11] is based on the temporal analysis of the virtual tone components generated by an auditory model that is the pre-processor of the algorithm. The algorithm consists of three major parts.

(1) Evaluate the short-time pseudo-autocorrelation of the individual virtual tone components on frames of 30 ms and accumulate the channel contributions into a global pseudo-autocorrelation function. Only peaks whose height exceeds a threshold \( \delta_a = 0.0125 \) are retained. The choice of this threshold comes from psychoacoustic data on the human perception of pitch.

(2) Apply a pitch extraction algorithm to the pseudo-autocorrelation function obtained in step 1. For each peak that exceeds a threshold \( \delta_a \), a pitch candidate and an evidence factor are generated. Evidence is defined as the amplitude of the autocorrelation function in that frame.

(3) Apply a pitch continuity analysis by computing a final pitch estimate and its evidence \( E_{n_f} \) in frame \( n_f \). \( E_{n_f} \) is then used to make the voiced/unvoiced decision. The final pitch and its evidence for frame \( n \) are derived from the preliminary candidates generated for frames \( n-2 \) to \( n+2 \). The pitch for frame \( n \) is always selected from the set of candidates for that particular frame. The pitches and evidences from the other frames (\( n-2 \), \( n-1 \), \( n+1 \), \( n+2 \)) are used to weight the candidates from frame \( n \) when a continuity between the candidates from frame \( n \) and the considered pitch candidates from the frames (\( n-2 \), \( n-1 \), \( n+1 \), \( n+2 \)) is found. If no candidate prevails, then the pitch and its evidence are set to zero. The continuity operator \( F \) is a function of the pitch period defined as follows:

\[
F(T_{n_f}, T_{n_f-1}) = \begin{cases} 
1 & \text{if } T_{n_f} - T_{n_f-1} < \delta_T \\
0 & \text{otherwise}
\end{cases}
\]

where \( T_{n_f} \) and \( T_{n_f-1} \) is the pitch period of two consecutive frames \( n_f \) and \( n_f-1 \), respectively. \( \delta_T \) is a threshold defined by the user, usually set equal to 0.09 [11]. \( F(T_{n_f}, T_{n_f-1}) \) is thus a proximity operator which is intended to overlook small errors in peak positions as they were determined from the autocorrelation function.

The voiced/unvoiced decision is based on the pitch evidence and on the continuity of the pitch estimates according to the following relationships:

\[
\begin{cases} 
E_{n_f} > \delta_v & \text{then } n_f \text{ is classified as voiced} \\
E_{n_f} < \delta_u & \text{then } n_f \text{ is classified as unvoiced}
\end{cases}
\]

A frame is unvoiced unless the evidence is greater than a fixed threshold \( \delta_v \), and the continuity criterion is satisfied by \( T_{n_f}(n) \) and \( T_{n_f}(n-1) \). According to [12] there are two parameters that need to be optimised: \( \delta_T \) and \( \delta_v \); \( \delta_T \) is not crucial.

The AMPEX algorithm differs from other pitch extraction algorithms which make a frame by frame decision first and use a postprocessor to correct these decisions afterwards. Also, the pitch for a frame is selected from the set of candidates proposed for that frame only.

Jitter: uses the same formula as MDVP (Eq. (2)).

2.5. BioVoice

BioVoice has been designed for the analysis of a wide range of voice signals (newborn cry, adult speech, and singing voice) [12]. It also allows for manual selection of signal frames to help the expert clinician in measuring fundamental period length by visual inspection. The program computes the exact duration between two cursors, respectively positioned at the beginning and at the end of the analysis interval.

BioVoice automatically performs voice analysis according to the following steps:

1. Load wave file and divide it into non overlapping windows of variable length inversely proportional to varying vocal frequency \( F_0 \). Frame length is linked to patient’s gender, as \( F_0 \) depends on size, thickness and tension of the vocal folds and is chosen in the range \( 3F_0/F_{\text{min}} \leq M \leq 3F_0/F_{\text{max}} \), where \( F_0 \) is the signal sampling frequency and \( F_{\text{min}}, F_{\text{max}} \) are respectively the minimum and maximum allowed \( F_0 \) values for the signal under consideration (set to 50–250 Hz for male, 100–350 Hz for female).

Specifically \( F_0 \) is estimated with a two-step procedure. Simple Inverse Filter Tracking (SIFT) is applied first to signal time windows of fixed length \( M = 3F_0/F_{\text{min}} \). From this first step, a raw \( F_0 \) tracking is obtained along with its range of variation \( [F_1, F_2] \), where \( F_1 = \text{lowest } F_0 \) value and \( F_2 = \text{highest } F_0 \) value [7,18,19]. In the second step, \( F_0 \) is estimated inside \( [F_1, F_2] \) with a combined Short-Time Autocorrelation Function (STACF)–Average Magnitude Difference Function (AMDF) approach. For a signal frame, compute the STACF and search for the maximum \( R_{\text{max}} \) and the corresponding cycle length \( T_{\text{max}} \) in the range \( (1/F_{\text{high}}, 1/F_{\text{low}}) \) with \( F_{\text{high}}, F_{\text{low}} \) estimated in the first step. If \( T_{\text{max}} \) falls outside this range, the cycle length is computed again for the same frame by means of the Average Magnitude Difference Function (AMDF):

\[
\text{AMDF}(k) = \sum_{m=k}^{N} |s(m) - s(m-k)|
\]

For a periodic signal with period \( T \), this function is expected to have a strong minimum at \( k = T \). The pitch period is thus estimated as the one corresponding to \( \text{min}(\text{AMDF}) \). The choice of the AMDF instead of ACF is due to the non-stationarity of the signals under study that was shown to often cause misestimating of the true signal periodicity, with enhanced results with respect to ACF [27–29].

2. Delete unvoiced frames from the signal and obtain a news audio file made up of voiced frames only, named the "voiced signal". In case of long sentences (several minutes) this considerably reduces computation time. This step is based on a revised version of the pitch continuity function in the AMPEX algorithm (see previous section).
From the “voiced signal” file a set of basic objective parameters is estimated: \( F_0 \), jitter, noise, formants, as well as specific parameters for each class of signals. In this paper we will describe \( F_0 \) and jitter estimation only.

For jitter estimation BioVoice implements the same formula (2) as with MDVP.

### 2.6. Visual inspection

A relevant information about the perturbation up to which computing jitter at least makes sense is obtained by exploring the ability of human visual perception in recognizing cycle patterns in synthetic signals. This is the level an analysis program should reach at least. To this aim, three experienced raters participated in the perceptual experiment, all three having more than 10 years experience in voice pathology, particularly voice analysis. They had to define the duration of each cycle on a visual display by positioning cursors on a 19 in. computer monitor with a definition of 1280 × 1024 pixels, using BioVoice. All raters reported that they base their cursor positioning mainly on peak positions. For each vowel sample, the cursor was positioned 40 times in order to define 39 cycle durations, from the 11th to the 49th Voice onset as well as end of phonation were thus excluded from the rating. Picking the dominant (upper or lower) peak appeared to be the basic starting point, but with possible preference for an alternative peak as soon as the duration-difference with the previous cycle exceeded 25%, and quasi-systematically if this duration-difference exceeded 50%.

### 3. Experimental results

Synthetic signals were generated using the synthesiser described in Section 2.1, adding a moderate level of pulsatile low-pass filtered white Gaussian noise, which was found to best mimic voice quality of typical clinical patients with dysphonia. The frequency modulation is simulated by means of small pulses that disturb sample by sample the instantaneous frequency of the harmonic driving function. The size of the pulses is fixed and their sign (plus or minus) is assigned stochastically. The small instantaneous frequency disturbances add up to cycle length perturbations. Because the sign of the sample-by-sample perturbations of the instantaneous vocal frequency is stochastic, the observed cycle lengths vary from one cycle to the next as well as from one vowel sound to the other, even when the magnitude of the parameter that fixes the size of the instantaneous frequency perturbations is constant.

Thirteen levels of increasing jitter are considered, corresponding to increasing values of parameter \( b \) in Eq. (1) from 0.315 to 5.0 ("jitter put in" in Table 1).

The relative cycle length jitter is summarized by means of the ratio of the average of the magnitudes of the inter-cycle length differences and the average cycle length.

For the current experiment a mild level of noise corresponding to 29 dB signal-to-noise ratio at the glottis was chosen, after a preliminary auditory-perceptual trial. It was considered by the perceptual raters as approximately corresponding to a B1 score, on the traditional GRBAS-scale [30].

Each synthetic voice signal has an average fundamental frequency of 100 Hz, and comprises 200 cycles, with gradual on- and offset. The sampling frequency is 22.050 Hz.

Concerning the sampling frequency \( F_s \), notice that its default values are quite different among the four approaches: 50 kHz with MDVP, 22 kHz with PRAAT, 44 kHz with BioVoice, and 20 kHz with AMPEX. In [31] the impact of \( F_s \) from 44.1 kHz to 10 kHz was explored by analyzing a large number of measures of fundamental frequency, jitter, and shimmer. In particular it was found that the recommended and critical \( F_s \) for acoustic voice analysis is above 26 kHz and 12 kHz for MDVP and PRAAT, respectively. Thus, voice samples captured above 26 kHz can be used for data analysis without significant bias in the results. In our case data are sampled at 22 kHz which is a value reasonably close to the lower limit mentioned above. Recall that for MDVP the default sampling rate is \( F_s = 50 \) kHz but other sampling values are allowed in the range 25 kHz ≤ \( F_s \) ≤ 81.92 kHz. Hence with MDVP signals were resampled at \( F_s = 25 \) kHz. PRAAT and BioVoice allow working with any \( F_s \) thus the original value was kept, with Ampex signals were resampled at the allowed \( F_s = 20 \) kHz.

Table 1 shows the actual jitter % ("jitter put in") corresponding to the 13 levels of jitter.

A first test has been performed to check BioVoice’s capability to estimate \( F_0 \) in three different synthesized signals with \( F_0 = 100 \) Hz, 120 Hz and 140 Hz, respectively, without jitter (\( J = 0 \)% and without noise. Results have shown the perfect correspondence between the original and the estimated \( F_0 \) values in all cases (0 Hz STD and 0% Jitter).

A second test concerned the capability of BioVoice to estimate the fundamental period \( T_0 \) for different values of \( F_0 \), moderate level of additional low-pass filtered white Gaussian noise and increasing \( J \). Figs. 1 and 2 show the correspondence between the cycle length \( T_0 = 1/F_0 \) of the synthesized signal and the one estimated with BioVoice for \( F_0 = 100 \) Hz and jitter level 5 and 9 respectively. Notwithstanding the high jitter values that exceed the upper limit of reliability of 6% as indicated in [1], BioVoice succeeds in correctly estimating \( T_0 \) in both cases with mean cycle length and STD almost equal to the synthetic ones. Specifically for jitter level 5 BioVoice estimates a mean \( T_0 \) value equal to 10.06 ms against \( T_0 = 10.09 \) ms for the reference one, with a difference of ±0.03 ms (Fig. 1). With jitter level 9 the estimates are \( T_0 = 9.763 \) ms with BioVoice against \( T_0 = 10.232 \) ms for the original signal (Fig. 2). Given the high amount of jitter, a difference of ±0.21 ms appears acceptable.

Finally the four algorithms (MDVP, PRAAT, AMPEX and BioVoice) have been compared on the whole set of synthetic signals reported in Table 1.

From Fig. 3 it can be seen that all programs behave quite well from the lowest jitter value (level 1, \( b=0.315 \), corresponding to \( J = 2.795 \)) up to level 6 (\( b=1.89, J=14.3 \)), the upper value being just above the maximum one (level 5, \( b=1.575, J=12 \)) that could be handled via visual inspection by three experienced raters who performed manual selection of cycles as described in Section 2.4. The mean value among the three raters is also shown in the figure ("visual inspection").

Above a \( b \) value of 1.89 (\( J=14.3 \)), all programs apart from BioVoice and AMPEX tend to dramatically under-estimate the true jitter value. This is particularly evident for MDVP and PRAAT, that also fail to evaluate \( J \) for parameter \( b \) above 4 and 3.45 respectively (\( J=29.7 \% \) and \( J=23.5 \% \)). Biovoice give valid estimates at least up to \( b=3.15 \) (\( J=21 \)), while AMPEX largely underestimates above \( b=2.205 \) (\( J=16.6 \)). This could be in part due to the unreliability of the auditory model implemented. However, both programs

---

**Table 1**

<table>
<thead>
<tr>
<th>Level</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model param.</td>
<td>0.315</td>
<td>0.63</td>
<td>0.945</td>
<td>1.26</td>
<td>1.575</td>
<td>1.89</td>
<td>2.205</td>
<td>2.52</td>
<td>3.15</td>
<td>3.45</td>
<td>4.0</td>
<td>4.5</td>
<td>5.0</td>
</tr>
</tbody>
</table>
Fig. 1. Comparison between $T_0 = 1/F_0$ obtained with the original synthesized signal and estimated with BioVoice for jitter level 5 ($J \approx 12\%$).

Fig. 2. Comparison between $T_0 = 1/F_0$ obtained with the original synthesized signal and estimated with BioVoice for jitter level 9 ($J \approx 21.2\%$).

Fig. 3. Jitter estimated by different approaches. Comparison of MDVP, PRAAT, AMPEX, BioVoice and visual inspection with theoretical values, for moderate noisy synthetic signals with jitter increasing from 2.795% to 37% (jitter put in).
are capable of evaluating jitter \( J \) up to the maximum level with an increasing trend, though underestimating it.

Table 2 shows the mean value and the standard deviation over all the 13 levels of the jitter estimated by the four tools and by visual inspection made by the three raters, as compared to the jitter put in. Though underestimating the true jitter, BioVoice performs better than the other tools. Visual inspection overestimates jitter.

A statistical comparison between the four programs and visual inspection would be desirable concerning the estimated cycle lengths for increasing jitter. However, cycle length was not available for MDVP and AMPEX. Praat allows to obtain the “glottal pulses” showing their position on the waveform by means of markers from which the cycle length can be evaluated. However it was found that results rapidly deteriorate with increasing jitter, as the number of identified cycles decreases from 97.5% to 5% and the number of outliers increases correspondingly, thus making results unreliable. Instead, the percentage of identified cycles is around 100% with both original synthetic data, BioVoice and visual inspection, independently of jitter. Thus MDVP, Praat and Ampex were not included in the comparison. Fig. 4 shows the coefficient of variation of the cycle lengths obtained for the 13 jitter levels for the original signals (“put in”), BioVoice and visual inspection. The figure shows that the results obtained with BioVoice are close to the corresponding ones with the original data, while visual inspection gives slightly overestimated and more spreading values.

4. Discussion and conclusions

Perturbation measures, including jitter, assume the presence of a periodical or quasi-periodical signal as only frames classified as “voiced”, i.e. having a periodical structure, can be processed in order to obtain an estimate of the amount of jitter in that frame.

Differences among software tools may therefore arise from differences in either of the two steps of voiced/unvoiced separation and fundamental period estimation. For instance, it can be observed that at high jitter levels, the MDVP and Praat programs tend to classify a large number of frames as “unvoiced” thus excluding them from the analysis step. Hence, when a large part of the frames is classified as unvoiced, the ones having larger variations in cycle lengths most probably belong to this class.

Although a detailed analysis would require accessing the internal details of the MDVP algorithm, it can be assumed that the frames classified as voiced have stronger periodicity with respect to unvoiced ones. Voiced frames, on the contrary, include the ones where the jitter is lower. As the estimate includes only voiced frames, this lead to an underestimation of the jitter, as it can be indeed observed in the experimental results.

As concerns the second step relative to jitter estimation, it has already been pointed out that the peak-to-peak estimate, used e.g. in MDVP program is valid up to jitter values approximately around 5% [17]. From the reported results, both peak-to-peak and waveform-matching seem to perform well up to at least \( J = 10\% \) with some more bias for AMPEX.

As all the approaches use the same formula for jitter estimation, and the level of added noise is relatively moderate, the differences must be due to the different techniques for voiced/unvoiced selection and pitch estimation. From the results reported here the best performance is achieved with BioVoice, thanks to the variable frame length for analysis linked to the locally varying signal characteristics and the modified AMPEX method for the selection of voiced/unvoiced frames and a two-step robust \( F_0 \) estimation.

Future work will be devoted to testing the proposed approaches on synthetic signals with increasing noise levels, mixed noise and jitter, as well as real pathological voices.

Acknowledgments

The authors greatly acknowledge COST Action 2103 and Ente Cassa di Risparmio di Firenze Project n. 2008.1586 for their contribution to this project.

---

**Fig. 4.** Boxplot of the coefficient of variation of the cycle length estimated with BioVoice and visual inspection compared to the original data (put in) for all levels of jitter.

**Table 2**

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jitter put in</td>
<td>18.15</td>
<td>11.008</td>
</tr>
<tr>
<td>MDVP</td>
<td>9.691</td>
<td>3.704</td>
</tr>
<tr>
<td>Praat</td>
<td>7.773</td>
<td>3.790</td>
</tr>
<tr>
<td>Ampex</td>
<td>11.285</td>
<td>6.198</td>
</tr>
<tr>
<td>BioVoice</td>
<td>14.907</td>
<td>8.875</td>
</tr>
<tr>
<td>Visual inspection</td>
<td>23.093</td>
<td>12.016</td>
</tr>
</tbody>
</table>

The authors greatly acknowledge COST Action 2103 and Ente Cassa di Risparmio di Firenze Project n. 2008.1586 for their contribution to this project.
References


