Effect of different jitter-induced glottal pulse shape changes in periodicity perturbation measures

Carlos A. Ferrer¹, Diana Torres¹, Eduardo González¹, José R. Calvo², Eduardo Castillo³

¹ Center of Study on Electronics & Information Technology, Central University of Las Villas, Cuba
² Centre for Application of Advanced Technologies, CENATAV, Cuba.
³ Department of Electrical and Computer Engineering, University of New Brunswick, Canada

{cferrer,dtb,moreira}@uclv.edu.cu, jcalvo@cenatav.co.cu, Eduardo.Calvo@unb.ca

Abstract

Jitter has long been used to describe period instability in voiced speech signals. In spite of this long history of measuring and modeling jitter, the ways the different glottal pulse phases are affected by jitter differ across studies. The models have quite dissimilar implications, and their selection has been rather arbitrary in the literature. This paper describes different choices for modeling jitter within the glottal pulse, and their implications in several speech processing tasks. Based on the discussion presented, any model selection departs from some accredited standpoint in voiced speech processing. Experiments to evaluate the effect of different models on the performance of periodicity perturbation measures are carried out on synthetic signals. The results obtained show large differences across models, demonstrating the need to evaluate each model's fit to actual data.

Index Terms: Jitter, Vocal Quality, Glottal Source.

1. Introduction

Jitter is defined as cycle-to-cycle perturbation in the fundamental frequency (F0) of vocal fold vibration [1][2]. Early findings and measurements of speech cycle aperiodicity were reported since the first decades of the 20th century [3-5]. Measures of jitter have proliferated since the introduction of computers: 15 pages of tabulated algorithms appear in the compilation by Budner [6], with an explicit statement (p. 131) that “the measure is historically the most widely applied” among vocal quality measures. One of the many expressions to measure jitter is shown on equation (1):

\[ \alpha = \frac{1}{N-1} \sum_{n=1}^{N-1} \frac{F(n+1) - F(n)}{0.5(F(n+1) + F(n))} \times 100 \]  

(1)

where \( T(n) \) is the length of the \( n \)th glottal pulse, among the \( N \) pulses analyzed.

The presence of jitter has proven important in both analysis and synthesis. In analysis, small values of jitter are found in normal voices [7], while large values are frequently related to pathological conditions [1][2][8]. In speech synthesis, the introduction of small perturbation to cycle lengths has improved speech naturalness from the inception of the field [9] and later [10], and has allowed researchers to manipulate emotion as well [11]. The introduction of jitter in synthesized voice has additionally been used to explore the perception of acoustic aspects of vocal quality [1][12], approach also suggested in the evaluation of rating scales in that area [13].

Given this broad use of jitter, the way a single glottal pulse is modified by this change in duration should be already known, or at least hypothesized with sound arguments. In this paper the different options are presented and their implications discussed. Experiments are performed to evaluate the practical implications of obtaining measures of periodicity perturbation when one or other model is assumed in synthetic signals.

2. The glottal pulse model

The actual measurement of jitter is generally performed in terms of glottal cycle lengths (in time units), and then converted to frequency. Markers of pulse boundaries were placed manually in the speech waveform by early researchers, following relevant events in the waveform like the rising edge prior to the pulse maximum [3][4][14][15] and automatically later by means of cycle-to-cycle automated pitch detection algorithms (PDA). The latter approach is also frequently oriented towards a specific waveform event, such as pulse maximum or minimum [7], or their previous zero crossing.

In the source-filter model of speech [16], the radiated speech signal \( s(t) \) is assumed to be the convolution of a series of glottal flow pulses \( g(t) \) with the vocal tract impulse response \( h(t) \) and the lip radiation \( r(t) \). The \( r(t) \) is assumed as a derivative operation applied to \( g(t) \), so that the actual convolution is performed between \( g'(t) \) and \( h(t) \):

\[ s(t) = g(t) * h(t) * r(t) \]

\[ s(t) = g(t) * h(t) \]  

(2)

In the absence of a straightforward relationship between the glottal cycle phases and the different markers used to locate the pulse in the radiated speech waveform, it is better to perform all further discussion in terms of the glottal pulse waveform. Figure 1 depicts a schematic representation of the glottal flow waveform \( g(t) \) and its derivative \( g'(t) \) as shown elsewhere [17-22], together with the location of the most commonly described phases. The phases depicted are: opening phase (with length \( T_o \)), closing phase (with length \( T_c \)), open phase (sum of previous phases, \( T_o + T_c \)), closed phase (with length \( T_c \)) and whole pulse (length \( T \)). Several ratios of these phase lengths (i.e. Open Quotient= \( T_o/T \) and Speed Quotient= \( T_c/T \)) have been used in analysis and synthesis of vocal qualities [23].

A first problem appears at the very definition of pulse boundaries, since glottal pulse models assume the pulse as starting with the opening phase, and most PDAs are designed to detect the glottal closure instant, for being the most prominent point in \( g'(t) \). This divergence between the
boundaries of modeled glottal pulse and the extracted pitch markers may cause differences in jitter measures depending on how jitter is conceived to modify the glottal pulse, as will be shown in next section.

![Diagram](image)

Figure 1: Typical representation of waveforms and phases of the glottal flow $g(t)$ and its derivative $g'(t)$.

3. Models for jitter effects on pulse phases

We define three options to distort individual phases due to changes in the pulse duration (i.e., jitter). According to their effect on the duration of each pulse, we denote them as: A) the constant warping model, B) the piecewise warping model, and C) the variable warping model. These models are depicted in Figure 2, where the instantaneous relationship between the phases of an original glottal pulse with length $T$ and a jittered pulse with length $T'$ is shown. The piecewise warping model is split in two graphs, with the general case in B1), and a particular case (described in Section 3.2.1) in B2).

![Diagram](image)

Figure 2: Schematic representation of the different models of jitter-induced glottal phase lengths changes. A) Constant warping, B) Piecewise warping and C) Variable warping.

It should be noted that all models have been represented considering the glottal pulse starting point and phases as depicted in Figure 1. Alternative representations exists, like using the glottal closure instant as starting point, or more pulse sub-phases like the return phase in the LF-model [18]. However, the analysis emerging from Figure 2 remains unaffected by these alternative representations. Particular descriptions and considerations on each model follows.

3.1. The constant warping (CW) model

This is a straightforward model to modify the durations of individual pulse phases according to the change in the total pulse length. The glottal pulse contracts or expands as a whole, being the relation between both pulses a straight line with slope equal to $T'/T$, as shown in Figure 2 A). Any original phase duration $T_j$ changes into a $T_j^*$ according to:

$$T_j^* = T_j \cdot \frac{T'}{T} \tag{3}$$

This process is implicit in the figures in [1][2][24], and in the theoretical derivations in [25]. As a convenient result, several glottal-ratios measures remain constant for all pulses.

As drawbacks, there is the already mentioned problem of pulse boundary definitions which causes the pulse lengths measured between adjacent glottal opening instants (as conceived in the glottal models) differ from lengths measured between adjacent glottal closure instants (as measured by most PDAs). This difference hinders the correct evaluation of PDAs performance when this model is assumed, since synthesized and detected contours would differ even for an error-free detection of the event of interest.

A second problem for this model arises from the use of equation 2 in the presence of jitter, since a constant amplitude $g(t)$ produces a variable amplitude (shimmered) $g'(t)$ and, consequently, $s(t)$. This is due to the amplitude modification that the contraction/expansion in $g(t)$ yields when the derivative is applied. Shorter $g(t)$ pulses produce higher amplitude $s(t)$ responses, and vice versa. This fact complicates the estimation of glottal shimmer from radiated signals if this option is assumed.

In spite of being the most popular and straightforward alternative, the rationale for this approach is not fully justified: it functions differently for different pulse starting points.

3.2. The piecewise warping (PW) model

In this type of warping the relationship within phases of different pulses is kept constant, but is different from one phase to another, as shown in Figure 2 B1). The proposal of a model of this kind might seem quite arbitrary, the exception being the one described below.

3.2.1. Only closed-phase warping

In this particular case, the slope of the line relating the open phases of any pair of glottal pulses equals the unity. All the warping ($\pm 1$) required to transform one pulse into the other falls into their closed phases, as shown in Figure 2 B2):

$$T_j^* = T_j + (T' - T) \tag{4}$$

This behavior emerges from a popular synthesis technique: the use of a sequence of jittered Dirac’s impulses for positioning in time a series of a given (constant) individual glottal pulse waveform $g(t)$, or simply using the impulse train as excitations to the vocal tract transfer function $h(t)$ [26].

With this approach all the nonzero phases of glottal pulses are equal. A graphic representation can be seen in [27], p. 328. The use of this approach is frequent in the synthesis of jittered speech to evaluate PDA’s performances [28] or HNR estimators [29]. It is also the underlying model in the first Klat’s synthesizer [30].

One reason favoring this option is its simplicity for synthesis: there is a unique nonzero part of $g(t)$ and $g'(t)$ for all pulses. Another reason is the equivalence of pulse length measures when considering either the opening phase or the
closed shimmer is also simpler in this case, due only to the variable superposition of one pulse with the tails of the previous in $s(t)$.

On the other hand, glottal measures like the open quotient are variable from pulse to pulse, and if glottal pulses are considered an oscillatory phenomena rather than a triggered response then the CW model could be considered closer to reality than this one.

3.3. The variable warping (VW) model

In this model the correspondence among jittered pulses is not required to be constant even within a given glottal phase, as shown in Figure 2 C). This model points towards the notion that the vocal fold vibration frequency (F0) is continuously varying [27]. As such, techniques involving an instantaneous frequency [31] (IF) contour have been addressed. Linked to this category are the shaping functions for the glottal waveform in [21], the several perturbations to vocal fold vibrations described in [32], the time normalization in [33], and the IF determination in [27]. A physiologically-based model like the Integral Pulse Frequency Modulation (IPFM) [34] could be added to the group of tools involving a continuous frequency contour. The synthesis in [35][36] also introduces random fluctuations on a sample-to-sample basis.

The definition of jitter, and the need of determining pulse markers/boundaries, makes jitter measures intrinsically different from measures obtained from IF curves representing the fundamental frequency F0 (see [27] for some methods to obtain these IF curves in speech signals). In jitter measurement methods F0 is a discrete-time function, with a value for each pulse as a whole. In IF curves, F0 is a continuous-time function, varying even within the pulse. Assuming this model makes the traditional measurement of jitter meaningless, with more relevance in measures of the IF variability.

A strong drawback for this model is the difficulty to estimate IF. Since the instantaneous frequency only exists for signals with a single sinusoidal component, the voice fundamental frequency must be extracted free of harmonics. As Roark describes in [27], the filtering process tends to suppress the very variability which is meant to be measured.

4. Evaluation procedure

Different metrics of perturbation (i.e. pulse maximum or minimum, positive or negative zero crossing) have reported different results for the same signals [2][27][37][38]. Hence, it has been recommended that the metric used should be reported along with the measure’s value [2][27][38]. However, to what extent the use of either one or other model described in Section 3 can cause a given metric to differ has not been explored. In this work we have evaluated the effect, in the measurement of jitter and shimmer, of using the CW or the PW models is performed. The VW model is not included for being the most complex and less related to jitter measures.

4.1. Signals used

Synthetic signals with seven levels of mean jitter were used, in steps of 3.4% reaching 23.8%. These values are the same used in [39][40][41][42], reaching the jitter limits reported for pathological voices (~25%) [43]. For each level of perturbation, signals comprising 300 jittered pulses were generated, with $Fs=22050$ Hz and $F0=150$Hz.

Individual jittered pulse lengths were obtained as $\tilde{T} = T + J$, being $J$ a uniformly distributed random variable in the range ± L, with L being the mean level of jitter mentioned before (increasing in steps of 3.4%). Individual glottal pulses were generated following Rosenberg’s type B (polynomial) model, with $T_0=0.337$, and $T_d=0.099$, the values producing the most natural-sounding signals [17]. The derivative of this signal was passed through a vocal tract transfer function with the same parameters as in [28][39-43], corresponding to a vowel /a/, to obtain the radiated $s(t)$.

4.2. Measures evaluated

To assess the influence of using the CW or the PW model in the estimation of jitter, two measures were evaluated. A first one is the absolute normalized averaged difference between reference ($T_{det}(n)$) and detected ($T_{det}(n)$) pulse lengths:

$$\beta = \frac{1}{N} \sum_{n} \frac{|T_{ref}(n) - T_{det}(n)|}{T_{ref}(n)} * 100$$

The second is the number of Gross Errors (GE) which counts the number of times the differences between contours is above certain threshold:

$$|T_{ref}(n) - T_{det}(n)| > t_0$$

In this case $\omega=3\%$, as in [42]. Dividing by N, GE can be expressed in %. The difference between jitter values as given by equation (1) for both contours, $T_{ref}$ and $T_{det}$ is not used due to the expressed criticism in [28][42][44]. To avoid the effect of detected pulse misalignment, the procedure described in [42] is applied to $T_{ref}(n)$ and $T_{det}(n)$ and corrected versions of $\beta$ (called $\beta^{c}$) and GE are obtained. Two well-known cycle-to-cycle PDAs were used to obtain the detected $T_{det}$ contours: the least mean squared PDA by Milenkovick [45] and the super-resolution PDA by Median [43].

The influence of using the CW or the PW model in the estimation of shimmer was evaluated according to two metrics: the maximum peak in the pulse and the pulse RMS. The mean absolute error between detected $a(n)$ and reference $a(n)$ amplitude contours was used as the performance measure. Since each method yields the contour in a different scale, contours are normalized to their mean values ($\bar{a}_g$ and $\bar{a}_r$, respectively) before the subtraction [40]:

$$Error = \frac{1}{N} \sum_{n} \left| \frac{a(n)}{\bar{a}_g} - \frac{a(n)}{\bar{a}_r} \right|$$

Here the fact that $a(n)$ is constant has been used to obtain the simplification in the last expression of equation (7).

5. Results and Discussion

Regarding jitter, measures are reported for both PDAs (denoted “Mil” and “Med”, according to the authors). For each level of jitter present, results for 100 runs were averaged.

In the case of signals synthesized according to the CW model, the pulse boundaries searched-for by the PDA do not follow the boundaries used on the synthesis process. To give a fair evaluation of PDAs performances, the results on the CW model have been reported for two versions of $T_{ref}$: the one following the opening instants (denoted CWo) and the one following closure instants (denoted CWC). The gross errors GE(%) are shown in Table 1, while $\beta^{c}$ is plotted in Figure 3.
Table 1. GE(%) made departing from different models.

<table>
<thead>
<tr>
<th>PDA</th>
<th>Jitter level (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.4</td>
</tr>
<tr>
<td>Mil CWo</td>
<td>0.54</td>
</tr>
<tr>
<td>Med CWo</td>
<td>0.41</td>
</tr>
<tr>
<td>Mil CWc</td>
<td>-</td>
</tr>
<tr>
<td>Med CWc</td>
<td>-</td>
</tr>
<tr>
<td>Mil PW</td>
<td>-</td>
</tr>
<tr>
<td>Med PW</td>
<td>-</td>
</tr>
</tbody>
</table>

From results in Table 1, the first obvious outcome is that PDAs can follow closure instants for both models without many GE. At worst, only 2.45% of the pulses are measured inaccurately, which occurs for “Mil” in the most distorted signals, for the most complex model. Both PDAs are unable to follow opening-instants $T_{o/p}$ contours (CWo). A second result is the remarkable superiority of “Med” PDA in following the closure instants ($T_{c/p}$ based on PW and CWc).

Results for $\beta_c$ confirm the findings above. In particular, “Med” can follow $T_{c/p}$ contours resulting from both the PW and CWc models with errors below 0.5% even in the most distorted signals.

The results for the measurements of shimmer are shown on Figure 4. All measures were performed on pulses located following glottal closure instants, hence CWc timings were not used. Again, two results are noteworthy. First, CW synthesis model poses more difficulty than PW in the subsequent measurement of shimmer. This is in correspondence with the results on jitter. Second, “Peak” and “RMS” behave opposite on each model: “RMS” is better in CW, while “Peak” is (slightly) better in PW. For the CW model there is a significant jitter-induced shimmer caused by derivation of $g(t)$. This effect results smaller in terms of the pulse’s RMS than the pulse peak value: on RMS the higher peak value is compensated to some extent with the shorter pulse duration. In the PW model jitter-induced shimmer only occurs due to the variable superposition of adjacent pulse tails. This apparently has (slightly) less effect on the pulse peak value than on its RMS.

6. Conclusions

Several conflicting views related to the modeling of the glottal pulse in the presence of jitter have been pointed out:

The jitter-induced warping of glottal pulses: Several models were described, being the CW and a particular case of PW the most easily connected to jitter.

The pulse’s starting point problem: This arises from the frequently differing views between synthesis and analysis. The choice of any given point can be seen as arbitrary.

The jitter-induced shimmer: There are two ways in which jitter creates shimmer: through the derivative applied over time-warped glottal pulses, and due to the variable overlap between pulse responses. The first one is by far more significant, as shown by the results on shimmer measures.

According to the results obtained, PDAs’ can’t be expected to perform well following CWo-based $T_{o/p}$ contours. Glottal pulse opening instants usually compose the gold standard against which PDAs’ are evaluated. This could help to explain why it is widely accepted that perturbation measures are unreliable.

Further research is required in establishing which model (CW, PW, VW) better fits actual vocal fold dynamics. Direct measurement techniques like High Speed Videendoscopy, Videokymography, or even Electrogloslography, might help to determine how the different phases of the glottal pulse change in the presence of jitter. If the VW model, not evaluated in this paper, were to emerge as the most realistic option, a deep change in current frequency variability measures should be considered.

Meanwhile, care must be taken when interpreting results obtained using only one of the models of jitter-induced pulse modification. Readers should be aware that other models exist, which could yield quite different results.

7. References