Using Waveform Matching Techniques in the Measurement of Shimmer in Voiced Signals

Carlos Ferrer¹, María E. Hernández-Díaz¹, Eduardo González¹

¹Center for Studies on Electronics and Information Technologies, Central University of Las Villas (UCLV), Cuba
cferrer@uclv.edu.cu, mariae@uclv.edu.cu, moreira@uclv.edu.cu

Abstract
In this work several approaches of amplitude contours estimation for shimmer measurement are analyzed and compared. The approaches covered incorporate a waveform matching procedure proposed in this work, based on existent least squares measures. The experimental comparisons evaluate each method’s sensitivity to periodicity perturbations like jitter, shimmer, and noise, as well as their combination. The waveform matching technique shows a better overall performance than the other methods.

Index Terms: shimmer, vocal quality, waveform matching

1. Introduction
Amplitude perturbations (shimmer) are, together with frequency perturbations (jitter) and noise measures (HNR or SNR), the most widely used objective measures of vocal quality [1][2]. Despite their widespread usage, there are ambiguities in the very definition of the terms which can lead to different results in the experimental measures obtained for the same set of signals. For instance, jitter stands for differences in glottal pulse durations, but the boundaries between pulses are arbitrarily assigned to significant events (glottal closure, maximum peak, etc.) in the signal. In the absence of a good reason not to think so, all the measures obtained are assumed to be equivalent.

Shimmer is an even more ambiguous term to measure. The problem of setting the pulse boundaries still holds, and another one arises: what is the “amplitude” of the pulse? There are many possible answers to this question: peak value (positive, negative, difference, rectified), pulse RMS, rectified DC level, and others [3][4]. Although maximum peak is the most common approach, measures involving the pulse as a whole, like RMS, could arguably be more robust to noise. To the authors’ knowledge, there is no comprehensive analysis and comparison of the different choices for the estimation of the pulse amplitude.

In this work, several reported methods’ weaknesses are exposed, and a novel method is proposed based on waveform matching techniques [5][6]. Besides, an experiment involving synthetic voice signals is carried out to evaluate and compare their performances in the accurate estimation of amplitude contours, in the presence of a wide range of shimmer, jitter, and noise.

The next sections are structured as follows: Section 2 is devoted to the description of each method addressed in this work and to comment on their theoretical strengths and weaknesses. In Section 3 the experiments performed to evaluate the methods are described. The results are shown in Section 4, along with their discussion. The summary of the findings in this work is made in Section 5.

2. Amplitude measures and shimmer
Most of the many shimmer measures devised (see [2] for a survey) depart from a detected sequence of pulse amplitudes. The approaches to determine this amplitude contour can be classified in three main groups: peak (instant) amplitude values, averaged (integral) amplitude values and a final group based on waveform-matching techniques. Description and comments on each group follow.

2.1. Peak amplitude values
This group comprises all the amplitude contours determined by measuring instantaneous peak values in the waveform. Maximum peak, minimum peak, maximum absolute peak, amplitude difference between maximum and minimum in the pulse, are examples of definitions of pulse amplitude that fall in this group.

This type of measure is the most frequently used, both in researchers’ software [3] and in the commercially available analysis software like Kay Elemetrics’ MDVP [7] used by researchers [8]. The known sensitivity of peak values to noise is an obvious drawback of this approach, but its simplicity favors its use.

A factor influencing the robustness of this group of measures is the Peak-to-RMS value of the voiced waveform. The more prominent the peak, the more insensitive to the underlying noise.

2.2. Averaged amplitude values
This group includes measures working over all the duration of the pulse, like RMS [9] value and rectified DC value. These measures have been used in the belief that using the whole pulse can reduce the influence of noise in the amplitude estimate.

A simple analysis shows that noise has an important bias effect in the amplitude estimates based in RMS. For stationary uncorrelated noise, the RMS value of noise is added as a constant to the signal’s RMS amplitude contour. The effect in the measured contour is that the relative magnitude of the fluctuation (shimmer) is reduced with respect to the contour’s mean value. For a zero mean noise, DC based measurements are free of this drawback.

2.3. Waveform-matching techniques
Waveform matching techniques between adjacent pulses have been successfully used in the task of pitch determination [5][6][10][11]. The amplitude relationship between the pulses being matched is obtained as a by-product of these techniques. In [5] a least squared error (LSE) approach yields that for a pair of pulses \( x_1 \) and \( x_2 \) modeled by:
\[ x_1(t) = p(t) + e_1(t) \]  
\[ x_2(t) = a_1 p(t) + e_2(t) \]  

(1)

the amplitude relationship between the pulses, given that noise realizations are uncorrelated and stationary, is

\[ a_1 = \sqrt{1 + R^2} \quad R = \frac{(x_1 - x_2) - (x_1 - x_1)}{2(x_1 - x_1)} \]  

(2)

where the (,) denote the scalar product operation.

On the other hand, in [6] an amplitude relationship is obtained using a similar model for the pulses, except that their separations and lengths are equal and denoted as \( \tau \) and varied over a chosen search range. The minimization of a cost function given by:

\[ J(\tau) = \frac{(d, d)}{(x_2, x_2)} \quad d = x_2 - a_1 x_1 \]  

(3)

results in an amplitude relationship as follows:

\[ a_1 = \frac{(x_2, x_1)}{(x_1, x_1)} \]  

(4)

A closer examination of (4) reveals that it tends to underestimate the value of relative amplitude when uncorrelated noise is present. Uncorrelated noise energy is present in the denominator of (4), but not in the numerator, causing a non linear underestimation effect. This is the reason to prefer the use of (2) in this work.

A method to obtain an amplitude contour using (2) can be conceived, by fixing \( x_1 \) to a certain pulse in the signal (i.e. the first), and substituting \( x_2 \) by all the subsequent pulses. The resulting amplitude contour would contain the relative amplitudes of all the pulses with respect to the first \( (x_1) \). This scaled version of the actual amplitude contour is sufficient for most shimmer measures. This method, proposed in this work, is included in the evaluation and comparison performed in the following sections.

3. Synthetic Signals and Experiments

3.1. Signal generation

Simulated vowels were generated according to the method used in [6][10] and [12], where the speech signal \( s(t) \) is obtained as the convolution of two signals, the vocal tract impulse response \( h(t) \) and an excitation impulse train \( i(t) \):

\[ h(t) = \sum_{n} a(t_n) \delta(t - t_n) \]

\[ i(t) = \sum_{n} a(t_n) \delta(t - t_n) \]

\[ h(t) = \sum_{n} k_{mn} e^{-\lambda t} \cos(2\pi f_m t) \quad (t \geq 0) \]  

(5)

this convolution results in:

\[ s(t) = \sum_{n} a(t_n) \delta(t - t_n) \]  

(6)

where \( t_n \) is the time instant of the \( n \)-th excitation and \( a \) is its amplitude. The terms \( k_{mn} \), \( b_m \) and \( f_m \) represent the amplitude, bandwidth and central frequency, respectively, of the M resonators (formants) used to model the vocal tract.

The values of \( k_{mn} \), \( b_m \) and \( f_m \) used for these resonators are the same as in [6][10] and [12], corresponding to a vowel /a/ and \( M=5 \).

The use of equations (5) and (6) to synthesize the vowel yields an easy way to vary the amount of jitter by controlling \( t_m \), while \( a \) can be used to vary shimmer, and the HNR can be controlled by adding noise to \( s(t) \). The sampling frequency \( F_s \) was set to 22.05 kHz, and the mean value of fundamental period \( T_0 \) is \( t_m=1/150 \) seconds ('mean' fundamental frequency \( F_0=150 \) Hz, \( t_m=147 \) samples). These values for \( F_s \) and \( F_0 \) are the same than in [10] and [12]. For simplicity reasons, only the closest two precedent impulses to the current '\( r \)' were considered in the synthesis of (6), since \( h(t) \) decays to negligible values for \( t > 3t_m \). In any case, a certain amount of interference between two consecutive pulses is introduced. The length of the generated speech signals was set to two seconds, which gives, together with the value of \( t_m \) used, an average of 300 pitch pulses on each signal.

3.2. Experiments

Four methods for the amplitude contour determination were evaluated: Main Positive Peak, RMS, Rectified DC level, and the Waveform Matching approach described in Section 2, based in the use of (2). The set of synthetic signals used are generated according to the procedure in [12]. The description is reproduced here for convenience. Four types of signals were generated, according to the presence of jitter, shimmer, additive noise, or a combination of the three periodicity perturbations.

Jittered Signals: the amplitude factor of equation (6) is left constant \( (a=1) \) and no noise is added to the signal \( s(t) \). The impulse excitation instants are obtained as \( t_m = t_{m-1} + t_m + u(n) \), where \( u(n) \) is a random real value uniformly distributed in the interval \( \pm u_m \). The time difference between two adjacent excitation instants will have then a uniform probability distribution in the range \( t_m \pm u_m \). The average pitch duration in samples is \( t_m=147 \). Different values for \( u_m \) were used, varying from 0 to 35, in steps of 5 samples, corresponding to values from 0 to 23.8% of \( t_m \) in steps of 3.4%. According to [6] jitter almost never exceeds the 25% of \( t_m \).

Shimmered Signals: the amplitude \( a(n) \) of the impulses used to generate \( s(t) \) are obtained as: \( a(n)=1+v(n) \), where \( v(n) \) is a random real value, uniformly distributed in the interval \( \pm v_m \). The values of \( v_m \) used were twice the values of \( u_m \) in the previous experiment, measured in percent of the unaltered amplitude \( a=1 \), going from 0 to 47.6% in steps of 6.8. This relationship was chosen to keep the ratio of usually accepted limits of jitter (25%) [6] and shimmer (50%) [1]. The temporal separation between the pulses was kept constant, equal to \( t_m \). No noise was added to \( s(t) \).

Additive Noise Only: a clean signal \( s(n) \), obtained for constant values of \( a=1 \) and \( t_m=1+tm+u_m \), was contaminated with additive white gaussian noise \( e(n) \), such that the ratio of the variances of \( s(n) \) and \( e(n) \) met an intended value of HNR in dB. The values of HNR used (in dB) were \( \infty, 22, 18, 15, 12, 8, 5 \) and 2.

Combined Perturbations: all the parameters \( (u_m, v_m \) and HNR) were varied simultaneously, in ascending order of perturbation, in the same amounts as in the three previous experiments. In this way, eight different cases are obtained, ranging from the perfectly periodic \( s(t) \) to the most distorted waveform with \( u_m=23.8% \), \( v_m=47.6% \) and HNR=2 dB.
3.3. Comparison procedure

Each method’s ability to detect accurately the amplitude contour was tested using the signals described above, with known amplitude contours. The mean absolute error between detected \( a_d \) and reference \( a_r \) 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 \( \bar{a}_d \) and \( \bar{a}_r \) respectively before the subtraction:

\[
\text{Error} = \frac{\sum_{i=1}^{N} |a_d(i) - a_r(i)|}{N}
\]  

(7)

Emphasis must be made in that no perturbation (shimmer) measure can be used to evaluate a contour detector, since two completely different contours can yield the same value of perturbation. This fact has already been pointed out in [10] for the equivalent case of jitter and separation contours.

4. Results and Discussion

The results of ten realizations of the random variables involved were obtained for each perturbation level, and the average is shown in Figures 1-4, for the cases of additive noise, shimmer, jitter and combined perturbations, respectively. The methods have been denoted \text{Peak} (for the main positive peak in the pulse), \text{RMS}, \text{DC} and \text{LSE} (for the waveform matching approach, based in a Least Squared Error criterion).

Figure 1: Mean absolute error of the amplitude estimation methods in the presence of noise

The results vs. additive noise (Figure 1) show the known sensitivity of \text{Peak} to noise, while the apparent better performance of \text{RMS} at higher noise levels can be due to its described tendency (Section 2) to flatten the \( a_d \) contour as the noise level increases.

Figure 2: Mean absolute error of the amplitude estimation methods in the presence of shimmer

Figure 2 shows that the presence of shimmer causes greater distortion in overall measures than in peak values, at least for this particular source-filter combination. It must be highlighted that the Peak-to-RMS ratio of the generated voiced signal is rather high, although not uncommon in natural voices. The \text{DC} approach is remarkably worse than the others.

Figure 3: Mean absolute error of the amplitude estimation methods in the presence of jitter

In the presence of jitter (Figure 3) the results of the four methods keep a similar order of performance, with \text{DC} the worst and \text{Peak} the better. The error values are higher than in the shimmer case. There is a knee-shaped behavior of the curves around the value of 15% maximum jitter, most noticeably in \text{Peak} and \text{LSE} approaches. The reason rests in the particular selected formant frequencies, but it does not change the methods’ performance order.

Figure 4: Mean absolute error of the amplitude estimation methods in the presence of noise and shimmer

In the presence of noise and shimmer the results of the four methods keep a similar order of performance, with \text{DC} the worst and \text{Peak} the better. The error values are higher than in the shimmer case. The methods’ performance order remains similar to that observed in the presence of noise alone.
Figure 4: Mean absolute error of the amplitude estimation methods in the presence of combined perturbations

The performance of the different methods in the presence of combined perturbations (Figure 4) show that there is a significant increment in the mean absolute error, greater than what could be expected by adding the obtained errors when facing single perturbations. It is a sign that methods performance worsens when different types of perturbation are present. It can also be noticed that the LMS method proposed, based on the waveform matching technique in [5], is the one with better overall performance.

5. Conclusions

The waveform-matching techniques, with a demonstrated performance in pitch determination, can produce more accurate shimmer contours than current approaches. Peak amplitude methods, ranking second in overall performance, are an acceptable choice given their simple implementation. However, the Peak method results in this work can be above the average in real life conditions, due to the high Peak-to-RMS ratio of the synthetic signals used. RMS and DC measures were the worst to perform in the experiments, in spite of dealing with the pulse as a whole.

A more comprehensive study is desirable, covering different synthesis procedures, vocal tract configurations, noise types, etc., in order to define the actual advantages of the waveform matching approach.

6. Acknowledgements

This work was supported partially by the Flemish VLIR-UOS Programme for Institutional University Cooperation (IUC).

7. References