Robust Estimation of Jitter in Pathological Voices

Pablo Daniel Agüero, Juan Carlos Tulli, Esteban Lucio Gonzalez, Alejandro Uriz, Juan Garin, and Gonzalo Aranda

Communications Lab - Engineering Faculty - University of Mar del Plata
Buenos Aires - Argentina
pdaguero@fi.mdp.edu.ar

Abstract. *Acoustical analysis of speech using computers has reached an important development in the latest years. The subjective evaluation of a clinician is complemented with an objective measure of relevant parameters of voice. Praat, MDVP and SAV are some examples of software for speech analysis. In this paper we describe an algorithm for the estimation of the fundamental frequency that considers the non-periodic nature of the speech signal under analysis. The experiments show that the use of these estimated f_0 values reduces the errors in perturbation measures of f_0, compared to the errors of other state-of-the-art speech analysis softwares, such as Praat and MDVP.*

Key words: Speech analysis, pitch estimation, pathological voice, JAVA

1 Introduction

In the latest years the acoustical analysis of speech has reached an important development thanks to the progress of computers. The main advantage of computer analysis of speech is the non-invasive assessment of the voice. Furthermore, the evaluation becomes objective through a set of numerical parameters.

The human auditory system is one of the main obstacles in the perceptual diagnostic of voice by the clinician ear. Humans are fundamentally prepared to perceive the voice as a whole, which is particularly advantageous from the point of view of linguistic communication. However, this ability is limited when it is necessary to individualize relevant aspects from a clinical perspective.

It is often difficult to determine the origin of certain anomalies of the voice using a perceptual procedure. For example, Baken et al. [2] show that some aspects of the pitch are more related to resonance frequencies of the vocal tract rather than to the frequency of vibration of vocal chords. The hypernasality of voice can be a consequence of the desynchronization in the timing of velar occlusion instead of an incomplete occlusion. Hence, the same attribute or alteration of the vocal quality may have its origin in different subsystems which can not be easily isolated with the audition of an expert.

In other cases, an adequate perception can not be quantized with the degree of precision of a numerical measure. For example, it is possible to measure the

degree of breathiness of a breathy voice through the corresponding speech parameter, the index of turbulence of voice (or VTI). In this way, the subjective evaluation of a clinician is complemented with an objective measure of relevant parameters of voice. As a consequence, the objectivity of the report is enhanced, and it is possible to measure the degree of progress more accurately.

Validity and reliability of acoustic analysis performed with different tools is affected by many factors. These include microphone type, noise levels, data acquisition system, sampling rate and software used for analysis [7, 8]. Ostensibly, the values of the commonly used frequency and amplitude perturbation measures should not be dependent on the software used to obtain them. Jitter and shimmer, for example, are defined by relatively simple and standardized formulas [3]. The differences observed between numerical values obtained for these measures using different softwares apparently stem from the raw fundamental frequency (f_0) data on which these calculations are based. Despite the basic nature of this parameter, there is no standardized algorithm to calculate f_0, which has been adopted and implemented by all programs.

While different methods for calculating f_0 may yield relatively small differences in the f_0 mean, they may influence the perturbation measures to a far greater extent. This introduces a difficulty for the clinical voice specialist, because different programs which are available for conducting voice analysis could report different values when analyzing identical voice samples. Moreover, it is not clear whether normative data which are presented by specific software (e.g., the data used for the radial graph in Multi- Dimensional Voice Program (MDVP)) are comparable with values obtained in other programs. This possible discrepancy between the results obtained by different programs was previously noticed and addressed by various researchers [12, 7, 18, 10].

Several methods are proposed in the literature to estimate f_0. They may be classified according to the domain where the calculation is performed: time, frequency and quefrency.

The autocorrelation method uses the cross-correlation of the signal with itself to estimate f_0 in the time domain. The fundamental period T_0 ($T_0 = \frac{1}{f_0}$) is determined as the elapsed time between the main maximum $R(0)$ and the first secondary maximum [17].

In addition, an algorithm of optimized autocorrelation may be used to find non-integer periods of pitch. This method proposed by Yohav Medan [15] finds out a rational value of pitch between the integer value given by the autocorrelation method and adjacent values.

Spectral techniques allow to find out the fundamental frequency in the frequency domain by analyzing the principal harmonics of the signal. The method explained by Bagshaw [1] uses the FFT and the spectral multiplication of different spectral harmonics to extract a maximum. Such a maximum corresponds to the fundamental frequency of the signal.

Another method found in the literature is based on the cepstrum transform [19]. The fundamental period of the signal (T_0) may be found in the local

maximum of the quefrency domain. The quefrency of the maximum corresponds to the T_0 of the signal.

The absolute envelope error was also proposed by Yohav Medan [15]. In this case the author uses the mean squared error to find out the T_0 of the signal in the time domain. The T_0 is the optimal T that minimizes the mean squared error e of this expression: $e = \sum_{k=0}^{T} |s[k] - s[k+T]|^2$.

In this paper an algorithm it is proposed for the estimation of the fundamental frequency for people with speech disorders. f_0 is used to calculate several jitter measurements already established in software for clinical speech analysis: relative jitter (jittr), absolute jitter (jitta), relative average perturbation (rap), five-point period perturbation quotient (ppq5) and average absolute difference between consecutive differences between consecutive periods (ddp). The focus of this paper is in the average local jitter.

The local jitter is defined as the absolute difference between consecutive periods, divided by the average period. The average of all local jitter derives in a parameter named jittr (an MDVP definition), and MDVP gives *jittr*=1.040% as a threshold for pathology.

The proposed f_0 estimation algorithm is compared with two state-of-the-art softwares in the calculation of the average local jitter: MDVP (Multi-Dimensional Voice Program) and Praat.

This paper is organized as follows. Section 2 describes two techniques to estimate f_0 used in software for clinical speech analysis: Praat and MDVP. Section 2.3 depicts our f_0 estimation algorithm which is focused in people with speech disorder. An experimental assessment of the methods shown in Section 2 and 2.3 can be found in Section 3. Finally, conclusions and future work are drawn in Section 4.

2 Fundamental frequency estimation and analysis in speech disorders

In this section the algorithms used to estimate the fundamental frequency in two commonly used software for voice analysis are described: MDVP (Section 2.1) and Praat (Section 2.2). At the end of the section, the proposed algorithm used in the voice analysis software of our lab is introduced and explained (Section 2.3).

2.1 Multi-Dimensional Voice Program f_0 estimation

The period-to-period pitch extraction [11] is a classic type of demodulation used for evaluation of voice pathology [13, 14]. However, the irregularity of the disordered voice makes the pitch extraction inaccurate, often impossible. In order to provide reliable data an adaptive time-domain pitch-synchronous method for pitch extraction was proposed by Deliyski [6]. It is used in the software named MDVP, and consists of the following main steps: fundamental frequency (f_0) estimation, f_0 verification, period-to-period f_0 extraction and computation of time domain voice parameters.

f_0 estimation provides preliminary information about the pitch. It is based on short-term autocorrelation analysis with non-linear sgn-coding [16] of the voice signal $x(n)$. Sgn-coding consists of a sign function with a dead region. The width of the dead region is shaped by the parameter K_p.

$$R(\tau) = \sum_{n=0}^{N-\tau-1} x'(n)x'(n+\tau), 0 \leq \tau \leq N/2 \quad (1)$$

where: $x'(i) = 0$ if $P_{min} < x(i) < P_{max}$, $x'(i) = 1$ if $x(i) \geq P_{max}$, and $x'(i) = -1$ if $x(i) \leq P_{min}$, with $P_{max} = K_p A_{max}$ and $P_{min} = K_p A_{min}$.

A_{max} and A_{min} are the global extremes of the voice signal in the current window.

The length of the autocorrelation window is 30ms or 10ms depending on the f_0 extraction range ($67 - 625$Hz or $200 - 1000$Hz). The sampling rate is 50kHz and every window is low-pass filtered at 1800Hz before coding. The value of the coding threshold at this stage of the analysis is $Kp = 0.78$, in order to eliminate the incorrect classification of any f_0 harmonic components as f_0 [5]. The current window is considered to be voiced with period $T_0 = \tau_{max}$ if the global maximum is $R(\tau_{max}) > K_d R(\tau = 0)$, where the voiced/unvoiced threshold value id $K_d = 0.27$ [5].

The f_0 verification procedure is similar to f_0 estimation. The autocorrelation function is computed again for the same windows at $K_p = 0.45$ in order to suppress the influence of sub-harmonic components of f_0. The results are compared to the previous step and the decision about the correct T_0 is made for all windows where a difference is discovered.

A period-to-period f_0 extraction is made on the original signal x(n) using a peak-to-peak detection method. It is synchronous with the verified pitch and voiced/unvoiced results computed in the previous steps. A linear 5-point interpolation is applied on the final period-to-period f_0 data in order to increase the resolution. This increased resolution is necessary for meaningful frequency perturbation measurements. The peak-to-peak amplitude is also extracted for every period.

2.2 Praat f_0 estimation

The estimation of the f_0 using Praat involves several steps, as shown in the paper of Boersma [4]. The first step removes the sidelobe of the Fourier transform of the Hanning window for signal components near the Nyquist frequency. It is performed through upsampling and spectral manipulation of the components near the Nyquist frequency.

In steps two and three a set of candidate periods per frame are calculated, using a voicing threshold to avoid spurious peaks not related to periodicity. A silence threshold is also used to detect voiceless frames.

In the last step one optimal candidate for each frame is found through a dynamic programming algorithm named Viterbi decoding. The best candidates

are those in the path that minimizes the joint cost: transition cost plus autocorrelation cost.

For stationary signals, the global path finder can easily remove all local octave errors, even if they comprise as many as 40% of all the locally best candidates, because the correct candidates will be almost as strong as the incorrectly chosen candidates. For most dynamically changing signals, the global path finder can still cope easily with 10% of local octave errors.

2.3 SAV f_0 estimation: the proposal of this paper

SAV (Software for Analysis of Voice) is a free software written in Java by our group, which may be used in multiple platforms, such as Windows and Linux. The main goal of SAV is the development of a free software that may be used by voice specialists and to provide channels of communication with the users to receive feedback for future improvements or bug correction.

Nowadays SAV is a software in beta version. Therefore, the authors are open to critics and suggestions to improve it in a daily fashion. SAV must be evaluated, calibrated and tested before being used by specialists to trace the treatment of patients.

In this paper we describe an algorithm for the estimation of the fundamental frequency that considers the non-periodic nature of the speech signal under analysis. In SAV these f_0 values are used to calculate several of the f_0 perturbation measures already mentioned in the Introduction.

The definition of the autocorrelation or the absolute envelope error methods do not include the fact that two consecutive periods may be different. In a non-pathological voice it is not an important problem, because a steady pronunciation of a vowel will have a very low jitter. Therefore, consecutive periods may be assumed as equal.

However, people with pathologies will have high values of jitter, and the assumption of consecutive periods with equal duration is inadequate. It is necessary to reformulate the definition of autocorrelation and absolute envelope error to fulfill the requirements of pathological speech signals.

The autocorrelation may be rewritten for signal with jitter as shown in Equation 2. T_1 and T_2 are the consecutive periods under analysis. The argument $\left[\frac{T_2-1}{T_1-1}k + T_1\right]$ is a linear warping function to consider an elongation and variations of the open and closed phases in the second period of the glottal excitation.

$$R(T1, T2) = \frac{1}{T_1} \sum_{k=0}^{T_1-1} s\left(k\right) s\left(\left[\frac{T_2-1}{T_1-1}k + T_1\right]\right) \tag{2}$$

In the case of the absolute envelope error algorithm the formulation is also modified to include a warping function for the second period T_2. In this case it is also included the argument $\left[\frac{T_2-1}{T_1-1}k + T_1\right]$ for a proportional elongation of T_2.

$$e(T1,T2) = \frac{1}{T_1}\sum_{k=0}^{T_1-1}\left(s\left(k\right) - s\left(\left[\frac{T_2-1}{T_1-1}k + T_1\right]\right)\right)^2 \quad (3)$$

In the experiments of the Section 3 it is used only the absolute envelope error algorithm, because it is the one that is implemented in SAV.

3 Experiments

The experiments were conducted using a set of speech signals with different degrees of jitter. The audio files have a sampling frequency of 44100Hz and 16 bits. The signal files only contain voice, without any additional information, such as an electroglottograph channel. The parameter that is evaluated in the experiments is the estimated average local jitter, using the algorithms described in Section 2.

In order to have experimental results under controlled conditions, synthetic voices with different degrees of average local jitter were generated. The glottal source used in the experiments corresponds to the Liljencrants-Fant glottal model [9], as shown in Equations 4 and 5.

$$g(t) = E_0 e^{\alpha t}\sin(\omega_g t), 0 \leq t \leq T_e \quad (4)$$

$$g(t) = -\frac{E_e}{\epsilon T_a}[e^{(-\epsilon(t-T_e))} - e^{(-\epsilon(T_c-T_e))}], T_e \leq t \leq T_c \leq T_0 \quad (5)$$

The synthetic glottal source with the desired jitter is filtered using a set of linear predictor coefficients (LPC) estimated from a real voice without any pathology. The resulting waveform has a known average local jitter that can be used as a reference in the experiments.

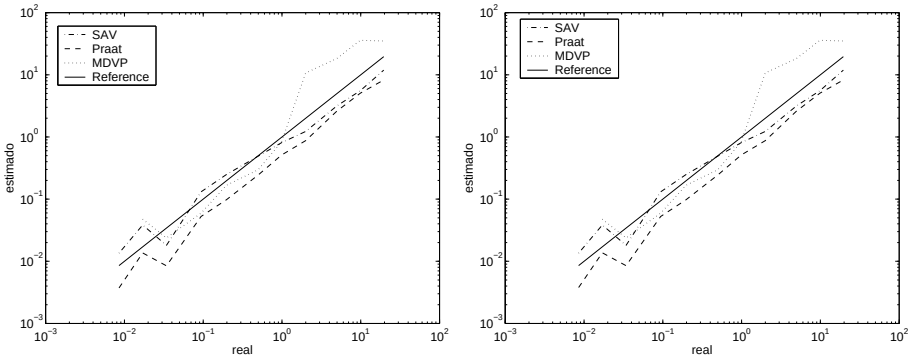
The average local jitter simulated covers a wide range of values from normal until pathological voice: 0.01%, 0.02%, 0.05%, 0.1%, 0.2%, 0.5%, 1.0%, 2.0%, 5.0%, 10.0% and 20%. These values of jitter are only goals. The synthetic voices attain only a similar value with respect to the goal.

3.1 Experimental results

The first experiment was conducted with 22 synthetic voices with known average local jitter and signal-to-noise ratio (SNR=40dB and SNR=20dB) (11 simulated values of jitter for each SNR condition). Figure 1 shows that the average local jitter estimated with SAV has a smaller difference with the reference value than the values estimated using Praat and MDVP, both for SNR=40dB and SNR=20dB.

Low values of jitter are detected more accurately by Praat, mainly for values below 0.1%. After that level of jitter, the detection curve of the algorithm used in SAV has a better approximation to the real jitter value of the reference.

The performance of the algorithm proposed by Deliyski [6] is not as accurate as SAV and Praat for all the range of jitter values in the simulation.



SNR 40dB SNR 20dB

Fig. 1. Average local jitter estimated with SNR=20dB and SNR=40dB for synthetic speech signal with known average local jitter

Another experiment was also performed using ten runs of voices with simulated average local jitter and two different SNR conditions, 20dB and 40dB (220 synthetic voices). In this way it is possible to study the average performance of each algorithm with respect to the reference jitter.

Figure 2 shows the experimental results of the ten runs. The solid curve corresponds to the absolute error of the algorithm used in SAV. This algorithm has the lower mean absolute error for all the range from 0.1% to 20%.

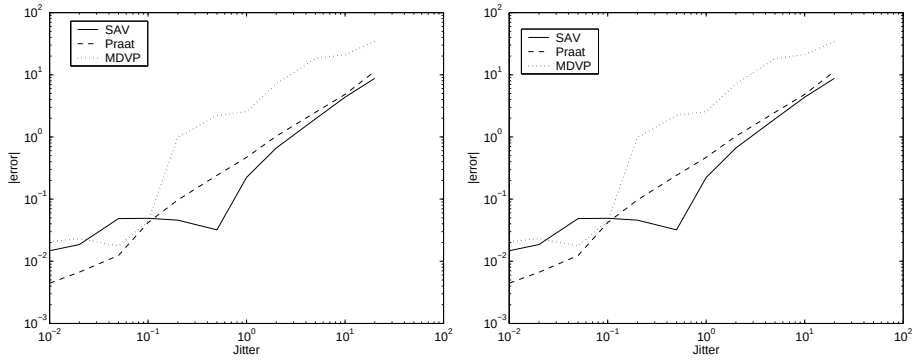
These results are consistent with those obtained in the previous experiment: Praat has a superior performance than SAV only for small values of jitter, below 0.1%. The performance of the algorithm proposed by Deliyski has a worse performance that Praat and SAV for all the range of values.

It is important to observe that the jitter detected with SAV is more precise in the range of values near the threshold for pathology: 1.040%. This result is remarkable, because it is important that a voice analysis software has a precise detection of jitter in the range that is important to trace the treatment of patients: 0.1% to 20.0%.

4 Conclusions and future work

In this paper it was made a set of comparative experiments to study the performance of the algorithm used in SAV (Software for Analysis of Voice) with respect to two state-of-the-art softwares for clinical speech analysis: Praat and MDVP.

Experimental results shown a lower estimation error for voices with simulated average local jitter, mainly in the range of values near the threshold for pathology: 1.040%. Such achievement is remarkable, because voice analysis software must have a precise detection of the jitter in the range that is important to trace the treatment of patients: 0.1% to 20.0%.



SNR 40dB SNR 20dB

Fig. 2. Absolute error of the average local jitter estimated with SNR=20dB and SNR=40dB for synthetic speech signal with respect to the simulated average local jitter

Future work will focus in those aspects that are not already taken into account by our algorithm, such as important variations of open and closed phase times.

References

1. Bagshaw, P.: Automatic prosodic analysis for computer aided pronunciation teaching. Ph.D. thesis (1994)
2. Baken, R., Orlikoff, R.: Clinical measurement of speech and voice. In: Second Edition. San Diego, CA: Singular Publishing Group (2000)
3. Baken, R.: Clinical measurement of speech and voice. In: Allyn and Bacon, Needham Heights, MA (1987)
4. Boersma, P.: Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proceeding of Institute of Phonetic Sciences. vol. 17, pp. 97–110 (1993)
5. Deliyski, D.: Investigation of the autocorrelation function characteristics in pathologic voice signal analysis. In: International Conference on Statistical Theory of Communications. p. 17 (1988)
6. Deliyski, D.: Acoustic model and evaluation of pathological voice production. In: Proceedings of Eurospeech'93. pp. 1969–1972 (1993)
7. Deliyski, D., Shaw, H., Evans, M.: Influence of sampling rate on accuracy and reliability of acoustic voice analysis. In: Logopedics, Phoniatrics, Vocology. vol. 30, pp. 55–62 (2005)
8. Deliyski, D., Shaw, H., Evans, M.: Regression tree approach to studying factors influencing acoustic voice analysis. In: Folia Phoniatrica et Logopaedica. vol. 58, pp. 274–288 (2006)
9. Fant, G., Liljencrants, J., Lin, Q.: A four parameter model of glottal flow. In: STL-QPSR. vol. 26, pp. 1–13 (1985)

10. Godino-Llorente, J., Osma-Ruiz, V., Saenz-Lechon, N.: Acoustic analysis of voice using wpcvox: a comparative study with multi dimensional voice program. In: European archives of otorhinolaryngology. vol. 265, pp. 465–476 (2008)
11. Hess, W.: Pitch determination of speech signals. In: Springer (1983)
12. Karnell, M., Hall, K., Landahl, K.: Comparison of fundamental frequency and perturbation measures among three analysis systems. In: Journal of Voice. vol. 9, pp. 383–393 (1995)
13. Koike, Y.: Application of some acoustic measures for the evaluation of laryngeal dysfunction. In: Studia Phonologica. vol. VII, pp. 17–23 (1973)
14. Koike, Y., Takahashi, H., Calcaterra, T.: Acoustic measures for detecting laryngeal pathology. In: Acta OtoLaryngological. vol. 84, pp. 105–117 (1977)
15. Medan, Y., Yair, E., Chazan, D.: Super resolution pitch determination of speech signals. In: IEEE Transactions on Signal Processing. vol. 39, pp. 40–48 (1991)
16. Rabiner, L.: On the use of autocorrelation analysis for pitch detection. In: IEEE Transactions on Acoustics, Speech, and Signal Processing. vol. 25, pp. 24–33 (1977)
17. Samad, S., Hussain, A., Fah, L.: Pitch detection of speech signals using the cross-correlation technique. In: Proceedings of TENCON 2000. pp. 283–286 (2000)
18. Smith, I., Ceuppens, P., Bodt, M.D.: A comparative study of acoustic voice measurements by means of dr. speech and computarized speech lab. In: Journal of Voice. vol. 19, pp. 187–196 (2005)
19. Wang, F., Yip, P.: Cepstrum analysis using discrete trigonometric transforms. In: IEEE Transactions on Signal Processing. vol. 39, pp. 538–541 (1991)