# EVALUATION OF A NEW MODEL FOR VOWELS SYNTHESIS WITH PERTURBATIONS IN ACOUSTIC PARAMETERS

Gabriel A. Alzamendi†, Gastón Schlotthauer†, Hugo L. Rufiner‡ and María E. Torres†‡♮

† *Laboratorio de Señales y Dinámicas no Lineales (FI-UNER) - CONICET*
‡ *Centro de I+D en Señales, Sistemas e Inteligencia Computacional (FICH-UNL) - CONICET*
♮ *metorres@santafe-conicet.gov.ar*



***Abstract*— Voice signal contains intrinsic irregularities which become more evident in the presence of pathologies. The acoustic parameters are very useful for the clinical assessment of voice and pathologies detection. Most existing voice models handle irregularities as additive noise and not as information carriers. In this work, a new model is proposed allowing to generate synthesized voices with previously selected acoustical parameters shimmer and jitter. Artificial voices are generated from a glottal source signal, obtained by conveniently disturbing amplitudes and periods, and then filtered using an auto-regressive linear filter. Models were developed for amplitude and period perturbations based on statistical methods. Several signals were generated and the performance of the model was analyzed. The quality of synthesized voices was evaluated using an objective quality measurement. The obtained jitter and shimmer values mostly agreed with the theoretically predicted values. These results suggest that this model is useful for artificial voices generation.**

***Keywords*— irregular-voice, voice synthesis model, acoustic parameters, jitter, shimmer.**

## I. INTRODUCTION

Study and modeling of voice generation mechanisms cover diverse scientific fields and demand interdisciplinary points of view, due to the complexity and diversity of the involved elements. The main issues are the analysis of the anatomical structures and the phenomena involved in the speech processes, considering dynamical behaviors and structural relationships.

Applications of speech models include methods for speaker recognition, techniques to improve the quality of artificial voices, strategies applied to man–machine interfaces and a variety of techniques for modeling, conditioning, synthesis, compression and transmission of speech signals. The different models developed to analyze and imitate the process of voice generation differ on the employed strategies and methods, and depend on the considered application. Recently, voice models have been applied to the study and synthesis of pathological voices. This made possible to develop a knowledge and understanding of the etiologies and alterations that can be found in different voice disorders (Schlotthauer, 2010; Torres *et al.*, 2009). Moreover, it has been demonstrated that even normal voices present intrinsic irregularities and that they are responsible of the degree of perceived naturalness (Baken and Orlikoff, 2000; Schlotthauer, 2010).

Acoustical parameters are usually employed in the practice of clinical medicine. Added to perceptual analysis and specific tests, they allow specialists to characterize the voice of an individual and to determine the presence of pathologies (Schlotthauer, 2010). *Shimmer* and *jitter* are the parameters most frequently used for quantifying instantaneous alterations in amplitude and frequency, respectively. It has been proved that these parameters are useful to characterize different types of voice and are sensitive to voice disorders (Baken and Orlikoff, 2000; Brockmann *et al.*, 2011; Velasco García *et al.*, 2011).

The purpose of this study is to propose and develop a simple voice synthesis model based on acoustical parameters of interest in the voice clinical practice. Particularly, here the focus is centered in the quantities *shimmer* and *jitter*, considering both healthy and pathological voices. For the validation of the proposed model two approaches are followed. First, the quantities estimated from the synthesized signals will be contrasted with the theoretical values. Also, the synthesized speech quality will be evaluated by means of an objective quality measure, with high correlation with psycho–acoustic perception.

This article is organized as follows: in Sec. II. the proposed model is introduced, the methodology is presented and the used materials are detailed. In Sec. III. the experimental results are shown and discussed. In Sec. IV. the conclusions and future works are presented.

## II. MATERIALS Y METHODS

In this work, we propose a voice synthesis method based on the *source-filter* speech production model.

This approach possesses a simple theoretical framework and has proved to be useful in a variety of applications (Proakis *et al.*, 1993). The model is inspired on the physiology of the voice system and the phonation process, where the air flow coming from the lungs is modified by the actions of the vocal folds producing regular pulses, called *Glottal Pulses* (GP). Those are acoustically transmitted through the *Vocal Tract* (VT), giving as result the proper voice signal (Rufiner, 2009). Each component of our model is detailed in the following sections.

### A. Glottal source

The morphology of the *Glottal Source* (GS), considered in the model, depends on the speech sounds to be analyzed or generated. In particular, only the synthesis of sustained vowels will be considered in this work. The vowels exhibit a regular morphology and, in case of healthy voices, a semi-periodic behavior. These emissions are the most widely used in acoustic tests. Considering these properties, here we propose to generate the GS from an impulse train with variable amplitude and period, represented by:

$$u[n] = \sum_{i=1}^{I} A_i \; \delta \left[ n - \sum_{j=1}^{i} P_j \right], \qquad (1)$$

where $A_i$ and $P_j$ are respectively the amplitude and period of each impulse (Proakis *et al.*, 1993). The value $1/P_j$ determines the *instantaneous frequency* ($F_0$) of $j$-th impulse. The main advantages of the representation of GS as in Eq. (1) are that it allows: *i*) to achieve the regularity and periodicity needed for the application, and *ii*) to freely modify the values $A_i$ and $P_j$, thus introducing controlled alterations in the speech signal. Ruinskiy and Lavner (2008) used similar perturbations such as random noise added to the GS in voice synthesis processes. Here we propose to generate artificial voices from statistical models of *jitter* and *shimmer* parameters. Therefore, it is needed an appropriate connection between these quantities and the parameters of GS.

The measure of the amplitude perturbation is usually called *shimmer*. This quantity takes into account instantaneous alterations in the amplitude of a voice signal, considering two consecutive pulses (Baken and Orlikoff, 2000). The *shimmer factor* ($shimmer_{\%}$) is computed as follows:

$$shimmer_{\%} = 100 \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} |A_{i+1} - A_i|}{\frac{1}{N} \sum_{i=1}^{N} A_i}, \qquad (2)$$

where $A_i$ is the amplitude of the $i-th$ pulse and $N$ is the number of pulses in the signal.

Period perturbation measures in the voice signal receive the name of *jitter* (Baken and Orlikoff, 2000). Among all existing quantities, the most widely used is the *jitter factor*, given by:

$$jitter_{\%} = 100 \frac{\frac{1}{N-1} \sum_{j=1}^{N-1} |P_{j+1} - P_j|}{\frac{1}{N} \sum_{j=1}^{N} P_j}, \qquad (3)$$

where $P_j$ is the period of the $j-th$ pulse and $N$ is the number of pulses.

Variations in amplitude and period in the GS pulses are assumed statistically independent. This fact allows to use Eq. (1). Furthermore, we consider that the series $A_i$ and $P_j$ obey a Gaussian behavior with probability density functions $\mathcal{N}(A_0, \sigma_A)$ and $\mathcal{N}(P_0, \sigma_P)$ respectively, where the quantities $A_0$ and $P_0$ are the corresponding mean values, and $\sigma_A$ and $\sigma_P$ are their standard deviations. These hypothesis have been previously used both in the analysis of the voice signal dynamics (Titze, 1995) and in classification of healthy and pathological voices (Torres *et al.*, 2009).

From the density functions, the series $\Delta A_i = A_{i+1} - A_i$ and $\Delta P_j = P_{j+1} - P_j$, with $i, j = 1, \ldots, N-1$, were generated. These series possess probability distribution functions given by $\mathcal{N}(0, \sqrt{2}\,\sigma_A)$ and $\mathcal{N}(0, \sqrt{2}\,\sigma_P)$ respectively. Therefore, the absolute values series $|\Delta A_i| = |A_{i+1} - A_i|$ obeys a hemi-Gaussian behavior and possesses a probability density function given by:

$$\begin{cases} \mathcal{N}(0, \sqrt{2}\,\sigma_A), & \text{if } |\Delta A_i| = 0; \\ 2\,\mathcal{N}(0, \sqrt{2}\,\sigma_A), & \text{if } |\Delta A_i| > 0; \\ 0, & \text{in other case.} \end{cases} \qquad (4)$$

It can be shown that the expected value of $|\Delta A_i|$ is determined by:

$$\mathcal{E}\left\{ |\Delta A_i| \right\} = \int_0^{\infty} \frac{2\,|\Delta A_i| \exp\left( \frac{-|\Delta A_i|^2}{4\,\sigma_A^2} \right)}{(4\,\pi\,\sigma_A^2)^{1/2}} = \frac{2\,\sigma_A}{\sqrt{\pi}}. \qquad (5)$$

It is known that $\left\{ \frac{1}{N-1} \sum_{i=1}^{N-1} |A_{i+1} - A_i| \right\}$ converges to $E\{|\Delta A_i|\}$ and that $\left\{ \frac{1}{N} \sum_{i=1}^{N} A_i \right\}$ converges to $A_0$, under the condition $N \to \infty$. Finally, replacing Eq. (5) into the Eq. (2) $\sigma_A$ is obtained:

$$\sigma_A = \frac{\sqrt{\pi}\, A_0\, shimmer_{\%}}{200}. \qquad (6)$$

In a similar way, it can be demonstrated that:

$$\sigma_P = \frac{\sqrt{\pi}\, P_0\, jitter_{\%}}{200}. \qquad (7)$$

Eqs. (6) and (7) allow to synthesize vowels with desired values of $shimmer_{\%}$, $jitter_{\%}$, $A_0$, and $P_0$. To this end, the series $A_i$ and $P_j$ must be generated as random Gaussian noise with mean $A_0$ and $P_0$ and standard deviation $\sigma_A$ and $\sigma_P$, respectively.

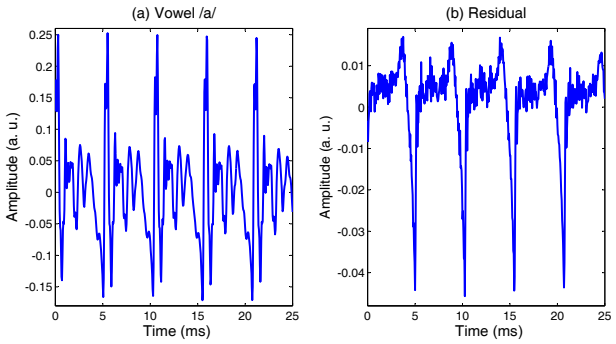


Figure 1: Voice signal corresponding to a healthy voice from a male person ($F_0 = 189.295$ Hz, $jitter_{\%} = 0.269\%$ and $shimmer_{\%} = 1.826\%$). a) Sustained vowel */a/*, b) Residual.

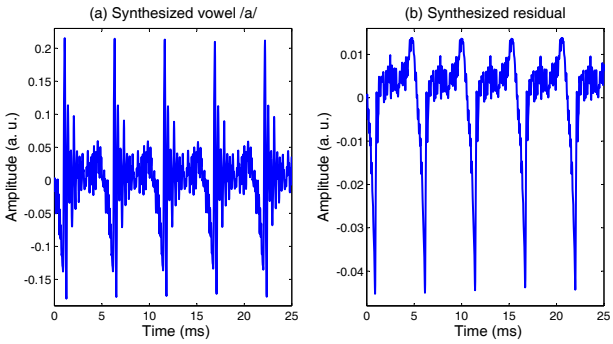


Figure 2: Synthesized voice signal corresponding to a healthy voice from a male person ($F_0 = 189.295$ Hz, $jitter_{\%} = 0.269\%$, $shimmer_{\%} = 1.826\%$ and $F_s = 50$ kHz). a) Sustained vowel */a/*, b) Residual.

### B. Vocal tract

The filtering properties of the VT can be represented by an auto-regressive model, where the voice signal at a particular instant depends on its past values and the current value of the GS (Proakis *et al.*, 1993; Rufiner, 2009). It can be expressed by the following difference equation:

$$s[n] = -\sum_{k=1}^{K} a_k \, s[n-k] + G \, u[n], \qquad (8)$$

where $s[n]$ represents the voice signal, $u[n]$ corresponds to GS, $a_k$ are the *linear prediction* (LP) coefficients and $G$ is a constant.

The system behavior in the frequency domain can be analyzed by taking the $\mathcal{Z}$-transform on both sides of Eq. (8). As a result, the system transfer function is given by:

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 + \sum_{k=1}^{K} a_k \, z^{-k}} = \frac{G}{A(z)}, \qquad (9)$$

where $S(z) = \mathcal{Z}\{s[n]\}$, $U(z) = \mathcal{Z}\{u[n]\}$ and $A(z) = 1 + \sum_{k=1}^{K} a_k \, z^{-k}$.

The result of filtering a speech signal using the inverse VT filter is defined as *prediction error* or *residual*. It is considered that this residual represents the actual behavior of the GS, and it has been proved that its application to voice synthesis improves the acoustical and perceptual properties of artificial signals (Proakis *et al.*, 1993).

### C. Database

The used database (MEEI, 2009) consists of sustained vowels */a/* phonated by 53 healthy speakers and by 654 speakers with disturbed voices due to a variety of pathologies. These signals were used to obtain the LP coefficients needed for VT modeling, according to Sec.B. Each signal was accompanied by a detailed information gathered from a variety of test and opinions of professionals. Table 1 shows mean, maximum and minimum of $shimmer_{\%}$ and $jitter_{\%}$ of the analyzed population. It can be noted that the pathological voices present higher values and greater dispersions in the acoustic parameters.

In this application 22 LP coefficients were used. In Fig. 1.a, 25 ms of a real vowel signal and in Fig. 1.b its residual are displayed, corresponding to a healthy voice from a male individual.

### D. Synthesized signals

An impulse train was generated with unitary amplitude and fundamental period $P_0$, with $P_0 = 1/F_0$. Amplitude and period of each impulse were modified independently according to Sec. A. in order to obtain the set $jitter_{\%}$ and $shimmer_{\%}$ values. The GS was obtained as result of the convolution between the disturbed impulse train and a residual period. This procedure improved the naturalness of the synthesized signal. Finally, each artificial vowel was obtained from the VT filter, as explained in Secs. B. and C. In Fig. 2.a, 25 ms of a synthesized vowel with a *sampling frequency* $F_s = 50$ kHz is shown and in Fig. 2.b its residual is displayed. For comparison purposes, signals were synthesized estimating the VT filter from the sustained vowels records, and using the parameters $F_0$, $jitter_{\%}$ and $shimmer_{\%}$ reported in the previously described database.

A collection of signals was synthesized taking different $jitter_{\%}$ and $shimmer_{\%}$ values into the ranges $0.00 \leq jitter_{\%} \leq 3.00$ and $0.00 \leq shimmer_{\%} \leq 5.00$ respectively, with step size of 0.05. These ranges were chosen according to the maximum and minimum values of each acoustical parameter for healthy voices, taken from the database (see Table 1). Hereafter, these parameters are referred as the theoretical $jitter_{\%}$ and $shimmer_{\%}$ values.

Table 1: Mean, maximum and minimum of $shimmer_{\%}$ and $jitter_{\%}$ of healthy and pathological voices corresponding to the analyzed database.

| Population | Acoustical parameters | Mean ± SD | Maximum | Minimum |
|---|---|---|---|---|
| Healthy voices | $shimmer_{\%}$ | 2.205 ± 0.924 | 4.802 | 0.963 |
| | $jitter_{\%}$ | 0.615 ± 0.437 | 2.529 | 0.175 |
| Pathological voices | $shimmer_{\%}$ | 7.103 ± 5.027 | 31.296 | 1.230 |
| | $jitter_{\%}$ | 2.539 ± 2.838 | 21.322 | 0.212 |

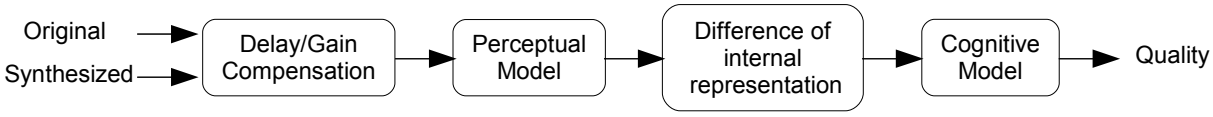


Figure 3: Flow chart of different steps in the PESQ method.

### E. Perceptual evaluation of speech quality

In order to evaluate the perceptual quality of the synthesized voices, an objective measure called *perceptual evaluation of speech quality* (PESQ) was used. This measure is defined in the standard ITU P.862 as an objective method designed to evaluate the quality of speech transmitted over communication channels (ITU-T, 2001). It has been extensively studied and it has shown a high correlation with subjective quality measures in different situations (Kokkinakis and Loizou, 2011).

In an attempt to imitate human perception, this method employs several levels of analysis. The first step consists in gain/delay compensation. Next, a transformation to a perceptual domain is performed and a density distortion is obtained from the difference between the signal to be analyzed and a reference signal. In the final step, cognitive models are applied (see Fig. 3).

The used PESQ algorithm takes values in the range $[1, 4.5]$ given by nonlinear regression with subjective tests (Hu and Loizou, 2008; available online at `http://www.utdallas.edu/~loizou/speech/software.htm`). Based on healthy voices of the database, 53 sustained vowels were synthesized and the PESQ values were obtained, using real signals as reference. For each signal, the used values of $F_0$, $jitter_{\%}$ and $shimmer_{\%}$ were extracted from database information.

## III. RESULTS

In this section, the performance of the proposed model and the perceptual quality of the synthesized vowels are analyzed, considering different values of the acoustic parameters *jitter* and *shimmer*.

Figs. 1 and 2 display a real and a synthesized signal respectively. Figs. 1.b and 2.b show that the behavior of both residual signals were very similar. Moreover, small oscillations can be noted in the real residual amplitude corresponding to intrinsic irregularities in the speech system, but these oscillations are not observed in the synthesized residual. On the contrary, the morphology of the synthesized vowel considerably differs from the real one (see Figs. 1.a and 2.a), but it is not the case for its regularity. This occurs because the VT filter only provides an approximated model of the spectral behavior of the vocal tract, and consequently the obtained signal is not an exact copy of the real voice (Proakis *et al.*, 1993). The pre-emphasis filter and the convolution between the impulse train and the residual improve the quality of the synthesized signal.

In order to analyze the model performance, a set of sustained vowels were synthesized using $F_s = 50$ kHz. For each signal, $jitter_{\%}$ and $shimmer_{\%}$ quantities were obtained from Eqs. (2) and (3), respectively. Signals were grouped according to the theoretical value of each acoustical parameter separately and statistical measures were obtained in each group. In Fig. 4.a and 4.b the estimated values of $shimmer_{\%}$ and $jitter_{\%}$ are respectively represented in function of their theoretical values. It is presented in solid blue line the mean for each group, in solid gray line the standard deviation and in dotted red line the theoretical value. The estimated correlation coefficient was 0.999986 for $shimmer_{\%}$ and 0.999939 for $jitter_{\%}$. In both cases, it can be seen that the estimated parameters cannot be distinguished from the theoretical values over most of the analyzed range. Furthermore, it can be seen that the dispersion of each acoustic parameter increases with higher perturbation values.

In particular for $jitter_{\%}$, the proposed model slightly differs from the theoretical behavior for values less than 0.2. This phenomenon can be appreciated in the zoom at Fig. 4.b. In order to explain this behavior, we assume that it is due to the discrete nature of synthesized signals. Observe that, in Eq. (3), for small

sinc(i) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
G. A. Alzamendi, G. Schlotthauer, H. L. Rufiner & M. E. Torres; "Evaluation of a new model for vowels synthesis with perturbations in acoustic parameters"
Latin American Applied Research, 2012.

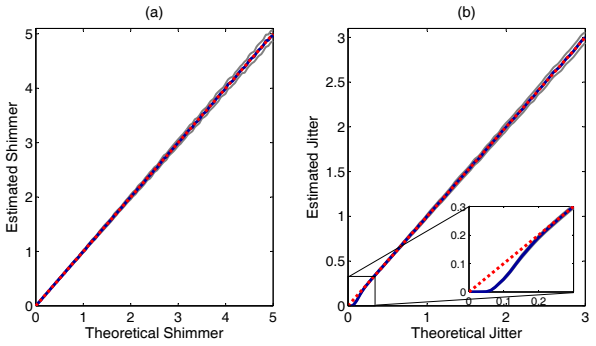


Figure 4: Estimated acoustical parameters in function of the theoretical values. It is presented in solid blue line the mean, in solid gray line the standard deviation and in dotted red line the theoretical value. (a) $shimmer_{\%}$, (b) $jitter_{\%}$.

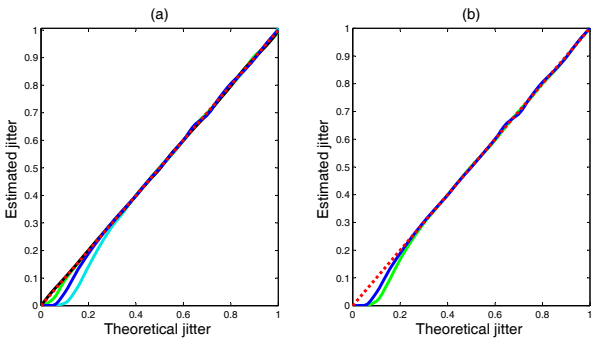


Figure 5: (a) $jitter_{\%}$ of signals with $F_0 = 189$ Hz and different $F_s$ values: 35 (cyan), 50 (blue), 75 (green) and 100 kHz (black). (b) $jitter_{\%}$ of signal with $F_s = 50$ kHz and different $F_0$ values: 189 (blue) and 230 Hz (green.)

values of $jitter_{\%}$ the estimation of $|P_{j+1} - P_j|$ is difficult, because the capacity to resolve as different two consecutive periods depends exclusively on the sampling frequency. Therefore, at low values of $jitter_{\%}$, the impulse periods are very similar. For this reason, jitter values will be underestimated. So, at certain point it will no longer be possible to distinguish differences between consecutive periods, and the estimated value of jitter will tend to zero.

In order to confirm this hypothesis, the experiment was repeated by setting $F_0 = 189$ Hz and varying $F_s$ in the synthesized signals. In Fig. 5.a the $jitter_{\%}$ performance of synthesized voices is shown, using $F_s = 35$ (cyan line), 50 (blue line), 75 (green line) and 100 kHz (black line). The theoretical values are displayed in dotted red line. As it can be seen, the performance of the proposed method improves with higher values of $F_s$, but simultaneously the computational costs increases significantly.

From Eq. (3), it can be seen that $jitter_{\%}$ also depends on the mean period $P_0$ of the analyzed voice. When $F_0$ increases then $P_0$ decreases, and as a consequence, the estimated value moves away from the theoretical value for a wider range of $jitter_{\%}$. This phenomenon was experimentally confirmed, as can be appreciated in Fig. 5.b. Two sets of artificial voices were synthesized, one of them using $F_0 = 189$ Hz (solid blue line) and another with $F_0 = 230$ Hz (solid green line). The model performance was analyzed and it was contrasted with the theoretical one (dotted red line). The signals of these two sets correspond to typical male and female healthy voices, respectively. All signals were synthesized using $F_s = 50$ kHz.

The performance at low values of $jitter_{\%}$ could be considered as a model deficiency. Nevertheless, it can be observed from Table 1 that the minimum value of jitter in the database is 0.175. An analysis of the database reveals that most of the values are greater than this critical value. Therefore, we can assume that the proposed model can be applied to synthesize both healthy and pathological voices (considering $jitter_{\%}$ values greater than 0.2). If the synthesis of signals with $jitter_{\%}$ less than 0.2 is required, then an appropriate $F_s$ must be chosen.

PESQ values were obtained from artificial signals in order to evaluate its perceptual quality, considering real vowels as references. This algorithm was designed to analyze continuous speech signals and, in this situation, speech information is usually considered to remain stable for periods of $20 - 30$ ms. For this reason, rectangular windows of 2500 samples were used in this work. Sustained vowels of long duration present intrinsic oscillations which could influence PESQ. In Fig. 6 a histogram of obtained PESQ values is shown. It can be seen that most synthesized signals have a PESQ value greater than 3.0, with a mean of 3.9 and a standard deviation of 0.4. This result suggests that the perceptual quality of synthesized signals using the proposed model is considerably high.

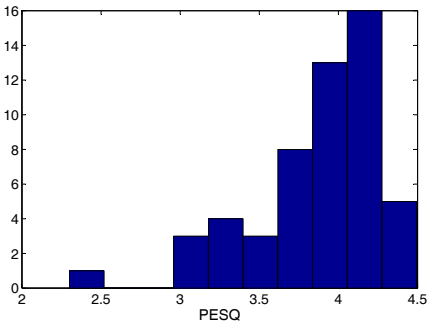


Figure 6: Histogram of PESQ values obtained from the synthesized vowels, using the real signals as references.

## IV. CONCLUSION AND FUTURE WORK

In this article a model for generation of artificial voices with controlled perturbations was proposed. This model differs from previous models by including acoustical parameters *shimmer* and *jitter* in the synthesis process, which are of great interest in the voice clinical practice.

For this purpose, a set of rules for amplitude and period modification in the GS were developed taking into account statistical models. Based on real voice signals, the model was applied to the synthesis of sustained vowels for a wide range of *shimmer* and *jitter*.

It was shown that synthesized signals have a behavior similar to the real vowels and the obtained parameters correspond with the theoretical values, over most of the range of interest. Furthermore, the capability of the proposed model to generate high quality artificial voices was confirmed with the obtained PESQ values.

Future works in this area include the application of this model to synthesize pathological voices, the incorporation of additional acoustic parameters of clinical interest, and the use of advanced signal processing techniques to analyze the synthesized voices.

## ACKNOWLEDGMENTS

This work was supported by ANPCyT, CONICET and UNER. The authors thank Dra. María C. Jackson Menaldi from Lakeshore Ear, Nose and Throat Center, St. Clair Shores (USA), and Wayne State University, Detroit (USA), for her important comments.

## REFERENCES

Baken, R.J. and R.F. Orlikoff, *Clinical measurement of speech and voice*, Singular Thomson Learning, San Diego (2000).

Brockmann, M., M.J. Drinnan, C. Storck, and P.N. Carding, "Reliable jitter and shimmer measurements in voice clinics: The relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task," *J Voice*, **25**,44-53 (2011).

Hu, Y. and P.C. Loizou, "Evaluation of objective quality measures for speech enhancement," *Audio, Speech, and Language Processing, IEEE Transactions on*, **16**, 229-238 (2008).

ITU-T P.862, *Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*, 2001.

Kokkinakis, K. and P.C. Loizou, "Evaluation of objective measures for quality assessment of reverberant speech," *Proc ICASSP 2011*, Prague, Czech Republic, 2420-2423 (2011).

MEEI Massachusetts Eye and Ear Infirmary Voice and Speech Lab, *Disordered voice database*, model 4337 (2009).

Proakis, J.G., J.H.L. Hansen and J.R. Deller, *Discrete-Time Processing of Speech Signals*, Macmillan, New York (1993).

Rufiner,H.L., *Análisis y modelado digital de la voz. Técnicas recientes y aplicaciones.* Ediciones UNL, Santa Fe (2009).

Ruinskiy, D. and Y. Lavner, "Stochastic models of pitch jitter and amplitude shimmer for voice modification", *Proc IEEEI 2008*, Beijing, China, 489-493 (2008).

Schlotthauer, G., *Análisis de señales con descomposición empírica en modos y aplicaciones a la señal de voz*, Tesis de doctorado en ingeniería, Universidad Nacional del Litoral, Santa Fe (2010).

Titze, I.R., *Workshop on acoustic voice analysis: summary statement*, Technical report, National Center for Voice and Speech, Denver, USA (1995).

Torres, M.E., G. Schlotthauer, H.L. Rufiner and M.C. Jackson-Menaldi, "Empirical mode decomposition. Spectral properties in normal and pathological voices", *Proc IFMBE 2009*, Munich, Germany, 252-255 (2009).

Velasco García, M.J., I. Cobeta, G. Martín, H. Alonso-Navarro, and F.J. Jimenez-Jimenez, "Acoustic analysis of voice in Huntington's disease patients," *J Voice*, **25**, 208-217 (2011).