Using artificial neural networks to solve an inverse problem applied to the estimation of random variables associated to a glottal pulse model in a system to produce voice

Edson Cataldo and João Marcos Meirelles da Silva

Universidade Federal Fluminense,
Telecommunications Department and Graduate Program in Electrical and Telecommunications Engineering
ecataldo@id.uff.br

Abstract. The aim of this paper is to use Artificial Neural Networks (ANNs) to solve a stochastic inverse problem related to a model for the glottal signal used in voice generation. Three parameters of the model should be considered as random variables: the time interval corresponding to a complete glottal cycle (the fundamental period), the time interval corresponding to the open phase of the vocal folds and the time interval corresponding to the closing phase of the vocal folds. For each random variable, the associated probability density function is constructed using the Maximum Entropy Principle. Parameters of these probability density functions should be identified using the ANN designed. At first, realizations of glottal signals are generated based on Monte Carlo Method. Then, features are extracted from the glottal signals obtained. The direct problem is then constructed associating the realizations of the three random variables to the features extracted. An ANN is then designed to solve the proposed inverse problem which maps the three random variables from the voice signals, through the features extracted, and use the ANN to construct its solution. Features are taken as inputs for the designed ANN, which outputs are the random variables. This is a way to validate the Rosenberg model showing that it is possible to fit it to experimental data identifying its parameters by measures. This paper also highlights an interesting application of ANN.

Keywords: inverse problem, artificial neural networks, voice production, glottal signal

1 Introduction

Speech technology has had an essential role in the development of information and communications engineering that has affected extensively all levels of our society. Understanding the human voice production mechanism is, however, not only extremely difficult but also highly challenging due to the fact that humans

are capable of varying extensively the functioning of their vocal organs. A simplified manner to study the functioning of the human speech production mechanism is to categorize speech sounds into three main classes according to the production mechanism: the voiced sounds, which are excited by the fluctuation of the vocal folds; unvoiced sounds, where the sound excitation is turbulent noise; and explosives, which are transient-type sounds made up by abruptly releasing the air flow that has been blocked by, for example, the lips [1].

Modelling voice production is a challenging issue, both on the theoretical and on the application side. The voice apparatus is complex and direct experimental measurements are difficult to obtain particularly because they should be invasive. Almost all the studies in voice production process consider a source-filter model [2]. In the case of voiced sounds, where the vowels are included, the voice source signal (glottal signal) is produced at the glottis, and then filtered and amplified by the vocal tract and further radiated by the mouth. Many glottal source models have been proposed with varying levels of complexity, such as the Rosenberg [3], Liljencrants-Fant (LF) [4], Fujisaki-Ljungqvist (FL) [5], and Rosenberg++ (R++) [6] models. These models were derived from an analysis of physiological measurements. They all share the following common features: they are bell-shaped, positive or null, quasi-periodic, continuous, and differentiable (except at glottal closure in some situations). Nevertheless, they use neither the same parameters nor the same number of parameters.

Here, the main interest is in signal models, assuming a source-filter decomposition and plane-wave propagation, the acoustics of voice production is reduced to an one-dimensional signal processing. The advantage of the signal approach is that the parameters obtained can be linked both to production and to perception. This is the approach used in an unified set of parameters. As compared to present physical models [7], the main potential of signal models is their usefulness in speech voice-quality related studies. Voice quality is mainly due to the characteristics of vocal-fold vibratory movement. Thus, a better understanding of these properties would help to characterize voice quality.

As the idea of this paper is to use an ANN to solve an inverse stochastic problem and identify parameters of a glottal pulse model, the simplest model will be chosen. The Rosenberg model will be considered to generate the glottal signal. The model generates deterministic signals. However, the small random fluctuation in each glottal cycle length is called jitter and it is a way to characterize voice signals even those with pathological characteristics. Typical values of jitter are between 0.1% and 1% of the fundamental period, for the so-called normal voices; that is, without presence of pathologies. The jitter value can be seen as a measure of the irregularity of a quasi-periodic signal and it can be a good indicator of the presence of pathologies such as vocal fold nodules or a vocal fold polyp [8, 9]. Here, the glottal signal to be generated will take into account the presence of jitter. The corresponding stochastic model will be constructed.

The strategy to be applied is to construct the stochastic model of the glottal signal, based on the Rosenberg model, considering the three parameters which are the fundamental period, the open interval and the closing interval as random

variables. Then, some features will be extracted of the glottal signal, which will be explained in details later. Estimators of the extracted features will be used as inputs of an Artificial Neural Network whose outputs are the corresponding values of the random variables. The Rosenberg model will be validated being fitted to experimental data. An interesting application of Artificial Neural Network is performed, solving a stochastic inverse problem.

2 The Rosenberg model for the glottal signal

There are many available models to generate glottal signals. In this paper, the model used is the one known as the Rosenberg model [3] which considers the glottal pulse, called U_g, given by:

$$U_g(t) = \begin{cases} \frac{A}{2}\left(1 - cos\left(\pi \frac{t}{T_p}\right)\right) \ , \ 0 \leq t \leq T_p \\ Acos\left(\frac{\pi}{2}\frac{t - T_p}{T_n}\right) \ , \ T_p \leq t \leq T_p + T_n \\ 0 \ , \ T_p + T_n \leq t \leq T_0 \end{cases} \tag{1}$$

where A is a constant related to the amplitude of the glottal pulse, T_p and T_n are parameters related to the opening and closing phase, respectively.

The open phase itself is divided into the opening and closing phases which are defined by the passage of the glottal flow by its maximum. After the return phase, the vocal folds remain closed during the so-called closed phase. With three parameters, the Rosenberg trigonometric model has two separate functions for the opening and closing phases to represent the glottal flow volume velocity.

Fig. 1 shows an example of the glottal pulse using the Rosenberg model.

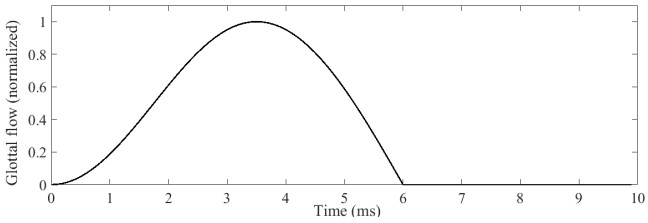

Fig. 1. Glottal pulse Rosenberg model: glottal flow corresponding to one complete glottal cycle.

3 Stochastic model of the glottal signal

The oscillations of the vocal folds are not exactly periodic and the pulses of air, which compose the glottal signal, have not exactly the same time duration. The

small random fluctuation in each glottal cycle length is called jitter and its study is particularly important in different areas related to the voice generation.

One of the first works for quantifying the jitter was proposed by Lieberman [10] who has characterized it by introducing a factor representing all perturbations greater than 0.5 ms. Other preliminary works were based on the calculations of a typical value related to the differences between the lengths of the cycles and their mean values or, more rarely, from the instantaneous frequencies and their mean values. Basically, these works agree with the fact that typical values of the jitter are between 0.1% and 1% of the fundamental period, for the so-called normal voices; that is, without presence of pathologies.

The idea is to consider three main parameters of the glottal signal in the Rosenberg model as random variables and consequently generate jitter. The parameters to be considered will be called the opening time, the closing time and the fundamental period.

Let us consider the duration between two successive times, the first one corresponding to the instant the glottis opens and the second one the instant for which it closes completely. This duration, denoted by T_{fund}, is a random variable, and its inverse is defined as the fundamental frequency that is the random variable $F_{fund} = 1/T_{fund}$.

The second random variable, associated to the opening time, will be denoted by OT, and the third one, associated to the closing time, CT. It is important to say that the three parameters T_p, T_n and T_0 in Eq. 1 will be considered as the random variables T_{fund}, OT and CT, respectively.

To construct the probability density functions (p.d.f.'s) associated to the random variables T_{fund}, OT and CT , the Maximum Entropy Principle is used (see [13, 14]) in the context of the Information theory introduced by [15]. This principle states: *Out of all probability distributions consistent with a given set of available information, choose the one that has maximum uncertainty (entropy).* Details about the available information of each random variable used to construct the probability density functions can be found in [11].

The construction of the p.d.f.'s of the three random variables will follow the same procedure. Let Y be each one of these random variables, T_{fund}, OT and CT, whose support is $]0, +\infty[$. The probability density function $p_Y(y)$ of the random variable Y has to verify the constraints given by Eqs. 2, 3 and 4:

$$\int_{-\infty}^{+\infty} p_Y(y)\, dy = 1 \quad , \tag{2}$$

$$\int_{-\infty}^{+\infty} y\, p_Y(q)\, dy = \underline{Y}, \tag{3}$$

$$\int_{-\infty}^{+\infty} ln(y)\, p_Y(y)\, dy = c\,. \tag{4}$$

in which c is an unknown positive constant.
Applying the Maximum Entropy Principle yields the p.d.f. given by Eq. 5:

$$p_Y(y) = \mathbf{1}_{]0,+\infty[}(y) \frac{1}{\underline{Y}} \left(\frac{1}{\delta_Y^2}\right)^{\frac{1}{\delta_Y^2}} \times$$

$$\times \frac{1}{\Gamma\left(1/\delta_Y^2\right)} \left(\frac{y}{\underline{Y}}\right)^{\frac{1}{\delta_Y^2}-1} exp\left(-\frac{y}{\delta_Y^2 \underline{Y}}\right) \tag{5}$$

where the positive parameter $\delta_Y = \sigma_Y / \underline{Y}$ is the relative deviation of the random variable Y such that $\delta_Y < 1/\sqrt{2}$ and where σ_Y is the standard deviation of Y. From Eq. 5, it can be proved that Y is a second-order random variable and that $E\{1/Y^2\} < +\infty$.

Clearly, there is a mapping $\mathcal{L}$ such that the glottal signal u_g at time t can be written as

$$u_g(t) = \mathcal{L}(t; T_{Fund}, OT, CT, \delta) \ , \tag{6}$$

When the random variables T_{fund}, OT e CT are considered the glottal signal is then a stochastic process U_g such that

$$U_g(t) = \mathcal{L}(t; T_{fund}, OT, CT, \delta) \ , \tag{7}$$

where δ is the dispersion parameters considered for the three random variables.

It is assumed that this stochastic process can locally be modelled as being stationary and ergodic (see, for instance, [12]). For each realization of (T_{fund}, OT, CT, δ) a glottal pulse is generated and a realization of the stochastic process corresponding to the glottal signal is composed by several glottal pulses.

4 Features extracted

After generating the glottal signal some parameters can be extracted and they will be used to train the Artificial Neural Network designed, which architecture will be discussed later. The parameters of the glottal signal can provide information to examine their importance in biomedical applications. But, here, they can help to fit the Rosenberg model.

In this paper, the features extracted from the voice signals are: MFCCs coefficients, measures of jitter, dH12 and HRF. Each one of these parameters will be discussed in the following.

MFCCs (Mel Frequency Cepstrum Coeficients) The MFCC [16] is a representation defined as the real cepstrum of a windowed short-time signal derived from the fast Fourier transform of the speech signal. In the MFCC, a nonlinear frequency scale is used, which approximates the behavior of the

auditory system. The discrete cosine transform of the real logarithm of the short-time energy spectrum expressed on this nonlinear frequency scale is called the MFCC. Fifteen (15) MFCCs were extracted from the glottal signals and used as entries for the ANN designed.

Jitter measures There are different types of measures for jitter, listed below [17]:

(i) *Absolute jitter*. It is the cycle-to-cycle variation of the fundamental frequency, *i.e*, the average absolute difference between consecutive periods, in seconds, expressed as

$$\mathrm{Jit}_{\mathrm{abs}} = \frac{1}{N-1}\sum_{i-1}^{N-1} \mid T_i - T_{i+1} \mid , \tag{8}$$

in which T_i are the lengths of each glottal cycle and N is the number of periods considered.

(ii) *Local jitter*. It is the average absolute difference between consecutive periods, divided by the average period, and given by

$$\mathrm{Jit}_{\mathrm{loc}} = \frac{\dfrac{1}{N-1}\displaystyle\sum_{i-1}^{N-1} \mid T_i - T_{i+1} \mid}{\dfrac{1}{N}\displaystyle\sum_{i=1}^{N} T_i} . \tag{9}$$

In general, the value 1.040% is considered as a threshold for the occurrence of a pathology.

(iii) *Jitter RAP*. It is the relative average perturbation, the average absolute difference between a period and the average of it and its two neighbors, divided by the average period. In general, 0.680% is considered as a threshold for the occurrence of a pathology.

(iv) *Jitter PPQ5*. It is the five-point period perturbation quotient, computed as the average absolute difference between a period and the average of it and its four closest neighbors, divided by the average period. In general, 0.840% is considered as a threshold for pathology; as this number was based on jitter measurements influenced by noise, the correct threshold is probably lower.

4.1 Frequency domain parameters

To estimate frequency domain parameters, the frequency or the power spectrum of the glottal pulse is considered. Here, two frequency domains parameters are

taken into account. First is the $dH12$ which is the difference of the first and second harmonics of the glottal frequency spectrum waveform in decibel. Another similar parameter is harmonic richness factor (HRF), which is defined as the ratio between the sums of the amplitudes of harmonics above the fundamental frequency and the magnitude of the fundamental frequency of the first harmonic in decibels. Both are defined in [17].

4.2 Simulations related to the direct problem

As an example, a glottal signal was generated; that is, a realization of the stochastic process associated, taken into account some values for the random variables and the parameters considered. The values considered were: $\overline{T_{fund}} = 1/200$, $\overline{OT} = 1/0.25$, $\overline{TC} = 0.35$, $\delta = 0.3$. In this case, the jitter can be clearly noted. Figure 2 shows one realization considering the data.

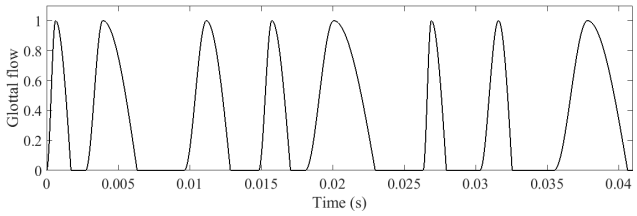

Fig. 2. Simulation of a glottal signal with jitter.

For each realization, features were extracted, as discussed before: MFCCs, measures of jitter, DH12 and HRF.

The details about the solution of the inverse problem will be described in the next section.

5 Solving the inverse problem using Artificial Neural Networks

The methodology applied is described in the following:

Step 1: Using the probability density functions constructed for T_{Fund}, OT and TC, glottal signals are generated in the following manner: for each mean value of T_{Fund}, OT and TC, and considering an unique value for the dispersion parameter δ (the same value was considered for the three random variables), glottal pulses are constructed and, consequently, glottal signals. Each realization of a glottal signal is composed by several glottal pulses.

Step 2: For each realization obtained in Step 1, features were extracted: 15 MFCCs, 4 measures of jitter, $dH12$ and HRF.

Step 3: Steps 1 e 2 were varied considering a grid for mean values of T_{Fund}, OT and TC and also a grid for values of δ. That is, let $\overline{T_{Fund_i}}$, $i = 1, \ldots, n$, $\overline{OT_j}$, $j = 1, \ldots, m$, $\overline{TC_k}$, $k = 1, \ldots, p$ be the mean values of the random variables and let $\delta_\ell \, \ell = 1, \ldots, q$ be the values of the dispersion coefficient.

For each vector $\mathbf{u} = (\overline{T_{Fund_i}}, \overline{OT_j}, \overline{TC_k}, \delta_\ell)$, a vector denoted by $\mathbf{w}$ containing 15 MFCCs, 4 measures of jitter, $dH12$ and HRF is associated.
This is the direct problem.

Step 4: An ANN (Artificial Neural Network) is designed for solving the corresponding stochastic inverse problem considering as input the vector $\mathbf{w}$ and as output the vector $\mathbf{u}$ from Step 3.

6 The ANN designed

As described on section 4, the extracted parameters from the glottal signal are used to train the Artificial Neural Network.

The ANN chosen in this paper is based on a multilayer perceptron (MLP) with the back-propagation algorithm [18] using Levenberg - Marquardt [19, 20] as the resolution method of optimization, both created and simulated in MATLAB.

A 27x20x4 architecture was proposed for solving the input-output fitting problem, which the 27 chosen parameters ($\mathbf{w}$ input vector) are non linear mapped on output inverse problem parameters of interest ($\mathbf{u}$) through a supervised training paradigm.

The labeled database consists on 4455 samples. About 70% of these samples are used for training, 15% are used for validating and the 15% remaining samples are used for testing.

Simulations results were performed with a second set of data, where the input data was presented to the Artificial Neural Network (after the training phase) and recording the outputs.

6.1 Some results obtained

Two main cases were discussed. The first one, considering the same set of data that had been already simulated and the other one, from experimental data.

In the first case, the direct problem was constructed considering the following variations: $\overline{T_{fund}}$ from 110 up to 310, with step 10; $\overline{OT}$ from 0.1 up to 0.9, with step 0.1 and $\overline{CT}$ from 0.1 up to 0.9, with step 0.1. In addition, δ from 0.001 up to 0.009, with step 0.001. And, with these considerations for the random variables, glottal signals were generated and features extracted. Then, the ANN corresponding to the inverse problem designed.

Considering some particular representative results, listed on Table 1, target is the real value of interest for a certain input, and estimated is the ANN output for this same input. For the sake of clarity, the $\mathbf{w}$ vector input was omitted from the table.

	δ_ℓ	$\overline{OT_j}$	$\overline{TC_k}$	$\overline{T_{Fund_i}}$
target	0.100	0.700	0.001	210
Estimated	0.093	0.712	0.001	207
target	0.200	0.400	0.007	230
estimated	0.203	0.408	0.007	231
target	0.300	0.400	0.001	250
estimated	0.296	0.388	0.001	254
target	0.400	0.300	0.003	290
estimated	0.403	0.285	0.0029	301

Table 1. Artificial Neural Network estimates for **u** output vector

The maximum relative error on Table 1 is about 7%, relative to δ_ℓ, with 0.100 target and 0.093 estimated value. This is considered a small error for this type of application.

7 Conclusions

A methodology to estimate the probability density functions of three random variables associated to control parameters in a non-linear model for producing voice was developed, through their mean values and the dispersion parameter. The methodology consists in solving an inverse stochastic problem using an artificial neural network, in the place of considering the model itself.

Although the system used was non-linear and stochastic, this paper showed that it is possible to identify some parameters using an Artificial Neural Network. Mainly, a simply model, like the Rosenberg model for the glottal pulse could be used and validated.

The future idea is then to use more parameters extracted from the voice signals in order to improve the quality of the parameters estimation.

Acknowledgements

The authors acknowledge CNPq (Brazilian Agency: Conselho Nacional de Desenvolvimento Científico e Tecnológico) for the financial support they gave to this research.

References

1. Ishizaka, K., Flanagan, J.: Synthesis of voiced sounds from a two-mass model of the vocal folds. Bell Syst. Tech. J., 51, 1233–1268, 1972.
2. Fant, G.: The acoustic theory of speech production. Mouton, The Hague, 1960.
3. Rosenberg, P.: Effect of glottal pulse shape on the quality of natural vowels. Journal of the Acoustical Society of America, 49, 183–190, 1981.
4. Fant, G., Liljencrants, J., Lin, Q.: A four-parameter model of glottal flow. Journal STL-QPSR, 26, 4, 1–13, 1985.

5. Fujisaki,H., Ljungqvist,M.: Proposal and evaluation of models for the glottal source waveform. Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP),3, 1605–1608, 1986.
6. Veldhuis,R.: A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation. Journal of the Acoustical Society of America, 103 (1), 566–571, 1998.
7. Cataldo, E., Soize, C.: Voice Signals Produced With Jitter Through a Stochastic One-mass Mechanical Model. Journal of Voice, 31, 1, 111e9–111e18, 2017.
8. Wong, D., Ito, M. R., Cox, N. B., Titze, I. R.: Observation of perturbations in a lumped-element model of the vocal folds with application to some pathological cases. The Journal of the Acoustical Society of America 89, 1, 383–394, 1991.
9. Hirose, H.: Clinical Aspects of Voice Disorders. Interuna Publishers, Tokyo, 1998.
10. Lieberman P.: Some acoustic measures of the fundamental periodicity of normal and pathologic larynges. J. Acoust. Soc. Am. 35, 344–353, 1963.
11. Cataldo, E., Soize, C., Sampaio, R., Desceliers, C.: Probabilistic modeling of a nonlinear dynamical system used for producing voice. Computational Mechanics, 43, 265–275, 2009.
12. Schoengten, J.: Stochastic models of Jitter. Journal of the Acoust. Soc. Am., 109, 1631-1650, 2001.
13. Jaynes, E. : Information theory and statistical mechanics. Phys. Rev., 106, 4, 620–630, 1957.
14. Jaynes, E.: Information theory and statistical mechanics II. Phys. Rev., 108, 171–190, 1957.
15. Shannon, C.E.: A mathematical theory of communication. Bell System Tech. J, 27, 379–423, 623-659, 1948.
16. Rabiner, L. R., Schafer, R. W.: Digital Processing of Speech Signals. Prentice-Hall, Englewood Cliffs, NJ, 1978.
17. Mongia, P. K., Sharma, R. K.: Estimation and Statistical Analysis of Human Voice Parameters to Investigate the Influence of Psychological Stress and to Determine the Vocal Tract Transfer Function of an Individual. Journal of Computer Networks and Communications, ID 290147. doi:10.1155/2014/290147.
18. Rumelhart, D. E., Hinton, G. E. and Williams, R. J.: Learning representations by back-propagating errors. Nature, vol. 323, pp. 533â536, Oct. 1986.
19. Levenberg, K.: A method for the solution of certain non-linear problems in least squares. Quart. J. Appl. Math., vol. 2, no. 2, pp. 164â168, Jul. 1944.
20. Hagan, M. T. and Menhaj, M. B.: Training feedforward networks with the Marquardt algorithm. IEEE Trans. Neural Netw., vol. 5, no. 6, pp. 989â993, Nov. 1994.