A stochastic model of voice generation and the corresponding solution for the inverse problem using Artificial Neural Network for case with pathology in the vocal folds

Edson Cataldo, Christian Soize

A stochastic model of voice generation and the corresponding solution for the inverse problem using Artificial Neural Network for case with pathology in the vocal folds

E. Cataldo[a], C. Soize[b]

[a] *Universidade Federal Fluminense, Graduate program in Electrical and Telecommunications Engineering, Rua Mário Santos Braga, S/N, Centro, Niterói, RJ, CEP: 24020-140, Brazil*

[b] *Université Gustave Eiffel, Laboratoire Modélisation et Simulation Multi Echelle, MSME UMR 8208 CNRS, 5 Bd Descartes, 77454 Marne-La-Vallée, France*

Abstract

A novel stochastic model to produce voiced sounds is proposed and, mainly, the corresponding identification of some model parameters using an Artificial Neural Network (ANN). The procedure described in this paper is about an intermediate step, which has as final objective to identify pathologies in the vocal folds through the voice of patients, that is, through a non-invasive method. The proposed model presented here uses the source-filter Fant theory and three main novelties are presented: a new mathematical model to produce voice obtained from the unification of two other deterministic one mass-spring-damper models obtained from the literature; a stochastic model that can generate and control the level of jitter resulting even in hoarse voice signals and/or with pathological characteristics but using a simpler model than those usually discussed in the literature; and the most important novelty, the identification of parameters of the proposed model, from experimental voice signals, using an ANN, particularly in a pathological case. The proposed neural network-based identification method requires a construction of a database from which an ANN can be trained to learn the nonlinear relationship between the parameters of the stochastic model and some relevant quantities of interest. The corresponding inverse stochastic problem is then solved in two cases: for one utterance corresponding to a normal voice and for another utterance corresponding to a pathological case corresponding to a nodulus in the vocal folds, helping to validate the model.

Keywords: Voice production, jitter, stochastic biomechanical models, voice pathologies.

1. Introduction

Voice has a fundamental importance in the transmission of knowledge, feelings and emotions. It has also an important role in the culture of a people, such as singing and acting and also it is the main work tool for many agents, such as singers, speakers, teachers, and others. The phonation occurs when a column of air, expelled from the lungs, passes through the vocal folds. Then, pulses of air are generated forming a (quasi-)periodic acoustic pressure signal called the glottal signal, which is further filtered and amplified by the vocal tract and finally, radiated by the mouth generating the voice. The glottal signal is not exactly periodic due to small random deviations in relation to a mean value of the glottal time interval called jitter. This phenomenon has some practical applications as to help in the identification of pathologies from the vocal folds, identification of voice aging, voice recognition, speaker recognition, and other (Wilcox, 1980;Li et al., 2010; Mendonza et al., 2014; Fraile et al., 2012). In general, values of relative jitter between 0.1% and 1.04% indicate normal voice, that is, a voice that is not symptomatic of a pathology (Wong et al., 1991; Manfredi et al., 2009).

Some stochastic models of jitter have already been proposed taking into account only mathematical expressions of the glottal signal without considering a mechanical model for the vocal folds (Schoentgen et al., 1997, 2001; Vasilakis and Stylianou, 2009). Other authors have recently described a mechanical model for the vocal folds considering the

Email addresses: ecataldo@id.uff.br (E. Cataldo), christian.soize@univ-eiffel.fr (C. Soize)

generation of jitter (Cataldo et al., 2012; Cataldo and Soize, 2016, 2018). However these models consider the coupling between the vocal folds and the vocal tract causing a relative model complexity that induces a significant computational cost. In addition, for addressing the problem related to the statistical inverse identification of stochastic models the proposed methods require solving a statistical inverse problem classically formulated as a stochastic optimization problem, which may be computationally expensive, because it is necessary to run many times the computational model related to the direct problem.

It is important to say that the work presented here is a continuation and an extension of the paper previously published in Journal of Biomedical Signal Processing and Control (Cataldo et al., 2016) but with a new methodoly and also new results.

In this work, a simplified model is proposed for the generation of jitter, considering the unification of two deterministic models proposed by Qureshi (2011) and also by Titze (1984,1988) with posterior modifications (Lucero, 1999; Lucero et al., 2001), disregarding a coupling equation between the vocal tract and the vocal folds, and considering the stiffness associated to the vocal folds as a stochastic process following the ideas proposed by Cataldo and Soize (2016,2018). With simplified stochastic model it is possible to obtain very good intelligible synthesis of voiced sounds, including jitter, characterizing normal voices but also hoarse voices or voices indicating pathologies due to the high level of jitter. To validate the model, a stochastic inverse problem is solved through an Artificial Neural Network. The statistical inverse problem under consideration consists in finding the values of the parameters of the stochastic model to produce voice for some given observed quantities of interest of the voice signals. An initial database is first generated from forward numerical simulations of the computational model. A processed database can then be deduced by conditioning the initial database in order to improve the performance of the trained ANN and therefore the efficiency of the neural network-based identification method. A multilayer ANN can then be designed to learn the nonlinear relationship between the parameters (network outputs) and the quantities of interest (network inputs). As a consequence, the proposed ANN-based identification strategy is computationally cheap, easy to implement, and use.

With this model experimentally validated, through Artificial Neural Networks, it is possible to generate the random phenomenon that is present in all voice signals. By numerical simulation, a big dataset can be generated for different voice signals, with different levels of jitter, and for different kinds of pathologies. Such a big dataset can be used for training an Artificial Neural Network. Finally, it should be noted that simple models have been used for the vocal folds, even nowadays, to better understand their movement as, for example, in the recently published paper by Lucero et al. (2020).

In this moment and in this work, it is very important to say that the discussion is an intermediate step to reach a final objective that is to help the pre-diagnosis of pathologies related to the vocal tract and the vocal folds in a noninvasive method. The idea is to find a set of parameters from the model, which can generate voices with pathological characteristics, even if only a specific group of pathologies is considered. Then, given a voice signal, the inverse problem is solved and if the parameters obtained belong to the range of the initial group of parameters, it will indicate that a possible pathology is present to the speaker of that voice signal.

In this paper, two experimental cases are considered, one of them corresponding to a normal voice and another from an experimental signal with pathological characteristics, the case of a nodulus in the vocal folds.

2. A proposed unified deterministic model for the vocal folds

The complete model presented here is based on the source-filter Fant theory (Fant, 1981), illustrated in Fig. 1.

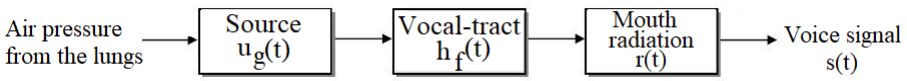

Figure 1: Sketch of the source-filter theory.

The voice signal generated, $s(t)$, is given by the convolution of the glottal signal $u_g(t)$, the corresponding impulse response function $h_f(t)$ of the filter that models the vocal tract, and the radiation by the mouth for which the impulse response function is $r(t)$. Then, Eq. (1) can be written as,

$$s(t) = (r * (h_f * u_g))(t)\,, \tag{1}$$

where $*$ is the convolution product. In the frequency domain, Eq. (2) can be written as,

$$\widehat{s}(\omega) = \widehat{r}(\omega)\,\widehat{h}_f(\omega)\,\widehat{u}_g(\omega)\,, \tag{2}$$

where $\widehat{.}$ means the Fourier Transform. This equation is well discussed in the literature about voice synthesis (Rabiner and Schafer, 1978; Prasad, 2017). With this formulation, there is no coupling between the vocal folds and the vocal tract, simplifying the model. In this paper, $u_g(t)$ is constructed using the proposed model, $h_f(t)$ is one coming from the literature and detailed later, and $r(t)$ is a first order high-pass FIR (finite impulse response) filter, such as suggested in (Rabiner and Schafer, 1978). Before discussing the stochastic model, the also original deterministic model is constructed based on two other models from the literature, which generate the glottal signal. The final idea is to join the best characteristics of each model to construct the proposed unified deterministic model.

The sketch considered is the one proposed by Qureshi (2011) and reproduced in Fig. 2. The movement of each

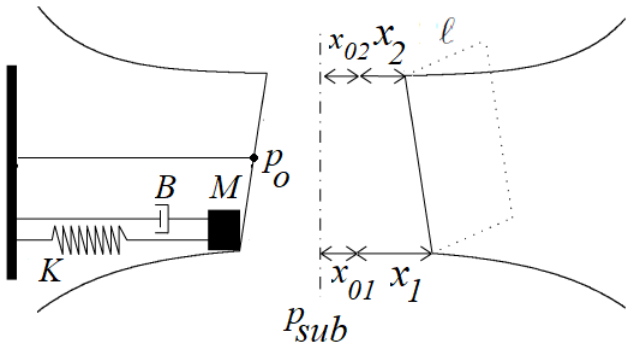

Figure 2: Sketch of the Qureshi model.

vocal fold is given by a rotary motion about its support point p_0. A single mass-spring-damper system is attached to the glottis at its entrance. The model is assumed to be symmetric about its central line so that the left side of the vocal fold is the same as its right one. The glottal entry displacement is represented by $x_1(t)$. The characteristic of a nonlinear damping is introduced in the model, according to (Laje et al.,2001), and the equation of the motion for a single vocal fold is described by (Lucero et al., 2011):

$$M\ddot{x}_1(t) + B\,(1 + \eta\, x_1^2(t))\,\dot{x}_1(t) + K x_1(t) = p_g(t; x_1(t), \dot{x}_1(t))\,. \tag{3}$$

where M, B, and K are the mass, the damping, and the stiffness, *per area unit*, and η is the nonlinear coefficient damping. The mean glottal pressure p_g is obtained using the Bernoulli law, and after some simplifications given by Eq. (4) (Titze, 1988;Lucero, 1999):

$$p_g(t; x_1(t), \dot{x}_1(t)) = \frac{2\,\tau\, p_{sub}(t)\,\dot{x}_1(t)}{k_t}\,(x_{01} + x_1(t))\,, \tag{4}$$

where x_{01} is the initial glottal displacement, pre-phonation, in relation to the mean point of oscillation, in the entry, τ is a short glottal delay time, k_t is the coefficient of transglottal pressure, and $p_{sub}(t)$ is the subglottal pressure. According to Qureshi (2011), the glottal flow $u_g(t)$ is given by Eq. (5):

$$u_g(t) = \sqrt{\frac{2p_{sub}(t)}{k_t\,\rho}}\,area(t)\,, \tag{5}$$

where ρ is the air density and $area(t)$ is the area at the exit of the glottis and given by

$$area(t) = \begin{cases} 2\,\ell\,(x_{02} + x_2(t))\,, & x_2(t) > -x_{02} \\ 2\,\ell\,x_{02}\,, & \text{otherwise} \end{cases} \tag{6}$$

in which x_{02} is the initial glottal displacement, pre-phonation, in relation to the mean point of oscillation, in the exit, ℓ is the width of the vocal folds, and

$$x_2(t) = -(x_1(t) - x_0)\,, \tag{7}$$

where the mean position of the oscillation is x_0, which is specified by the horizontal position of the fulcrum point p_0. In this text, without loosing generality, the same values are considered for x_0, x_{01} and x_{02}. After generating $u_g(t)$, the convolution with the filter (vocal tract and the mouth) should be performed to produce the sound.

The next step is the construction of the corresponding stochastic model to generate the glottal pulse, which will be then a stochastic process, based upon the unified deterministic model proposed here.

3. Proposed stochastic model

The objective is to vary instantaneously the frequency of the voice signal to generate jitter. As the mass is fixed, the stiffness K is considered as a stochastic process $\{K(t), t \in \mathbb{R}\}$ following some ideas proposed by (Cataldo and Soize, 2018) with the corresponding changes, because the model created here does not consider the coupling between the vocal tract and the vocal folds. The consideration is that jitter is generated due to the variation of the stiffness of the vocal folds, giving possible additional information to the biomechanics of the vocal folds. In (Cataldo and Soize, 2018) a more complex model was considered, taking into account a coupling equation between the vocal folds and the vocal tract. The idea is to reproduce similar results using the same consideration of the stochastic model associated with the stiffness and to show that it is possible to generate jitter and very good intelligible voice sounds. The details about the construction of the stochastic model associated with the stiffness will not be explained here, because they can be obtained in (Cataldo and Soize, 2018). Only the most important characteristics will be listed.

Let E be the mathematical expectation. The stochastic process $\{K(t), t \in \mathbb{R}\}$ is constructed according to the properties defined as follows.

- (i) For all t in $\mathbb{R}$, it is assumed that $0 < K_0 \leq K(t)$ almost surely, where K_0 is a positive constant independent of t.
- (ii) $\{K(t), t \in \mathbb{R}\}$ is a non-Gaussian stationary stochastic process such that $E\{K(t)^2\} < +\infty$ for all t (second-order stochastic process), for which its mean function (that is independent of t) is written as $E\{K(t)\} = \underline{K} > k_0 > 0$, and which is assumed to be mean-square continuous in order to guaranty the existence of a power spectral measure.

A representation of the stochastic process $K(t)$ is chosen as,

$$K(t) = K_0 + (\underline{K} - K_0)(\bar{z} + Z(t))^2\,. \tag{8}$$

in which $\bar{z}$ is a positive real number and where $\{Z(t), t \in \mathbb{R}\}$ is a Gaussian second-order real-valued stochastic process, centered, mean-square continuous, stationary and ergodic, physically realizable, such that $Z = h * N_\infty$, where N_∞ is the centered Gaussian white noise (generalized stochastic process) whose power spectral density function is written, for all real ω, as $S_N(\omega) = 1/(2\pi)$, and $h = \mathcal{F}^{-1}\{H\}$ is the inverse Fourier transform of the complex-valued frequency response function $\omega \mapsto H(\omega)$ chosen, for all real ω, as $H(\omega) = a/(b + i\omega)$, in which a and b are positive constants, so that stochastic process Z can be constructed as the asymptotic stationary random solution of the following linear Itô stochastic differential equation,

$$dZ(t) = -b\,Z(t)\,dt + a\,dW(t)\ \ ,\ \ t > 0\,, \tag{9}$$

with the initial condition $Z(0) = 0$, where $\{W(t), t \geq 0\}$ is the real-valued normalized Wiener stochastic process. The power spectral density function of stochastic process Z is then given by Eq. (10):

$$S_Z(\omega) = |\,H(\omega)\,|^2 = \frac{a^2}{2\pi(\omega^2 + b^2)}\,. \tag{10}$$

The level of jitter will mainly be controlled by a (and b), which should satisfy the condition $a^2 < 4b$ (Cataldo and Soize, 2016). It could be intoduced a filter with a high order and then with more parameters to be identified. However,

it has been chosen a filter of first order with two parameters. The simulations, discussed later, show that this filter chosen is enough to generate jitter. Following Eq. (3), the displacement $x_1(t)$ of the vocal folds becomes the stochastic process $X_1(t)$. The dynamics of the vocal folds is then given by the following nonlinear stochastic differential equation,

$$M\ddot{X}_1(t) + B\,(1 + \eta\, X_1^2(t))\,\dot{X}_1(t) + K(t)\,X_1(t) = p_g(t; X_1(t), \dot{X}_1(t))\ . \tag{11}$$

All the other equations related to the unified deterministic model should be rewritten substituting $x_1(t)$ and $x_2(t)$ by $X_1(t)$ and $X_2(t)$, and also p_g by P_g. The realizations of the stochastic process $\{K(t), t \in \mathbb{R}\}$ are obtained using Eqs. (8) and (9). Consequently, the realizations of the stochastic process $X_1(t)$ are computed by solving Eq. (11) and the realizations of the stochastic process $X_2(t)$ are deduced. Finally, realizations of the stochastic process $U_g(t)$, corresponding to the glottal signal, are computed. The voice signal is generated through the convolution given by Eq. (1). The voice signals generated will indicate the presence of jitter and to quantify its level some measures can be used.

3.1. Jitter measures

Let T_g be the random variable associated with the duration of the glottal cycle, which is defined as the duration between two successive times, the first one corresponding to the instant the vocal folds (glottis) open and the second one the instant when they close completely. To calculate T_g from $U_g(t)$, it is used an algorithm based on an implementation of the RAPT pitch tracker (Talkin, 1995).

For each glottal cycle j, and each realization θ_j of the random variable T_g, the duration denoted by $T_g(\theta_j)$ can be associated with. Considering that the set $\{T_g(\theta_j), j = 1, \ldots, N\}$ constitutes N realizations of random variable T_g (corresponding to all the glottal cycles of the voice signal), jitter can be measured by the following equations (Mongia and Sharma, 2014).

(i) The absolute jitter, denoted by Jit_{abs}, is defined by

$$Jit_{abs} = \frac{1}{N-1}\sum_{j=1}^{N-1} |T_g(\theta_{j+1}) - T_g(\theta_j)|\,. \tag{12}$$

(ii) The relative jitter, denoted by Jit_{rel}, is defined by

$$Jit_{rel} = \frac{\dfrac{1}{N-1}\displaystyle\sum_{j=1}^{N-1} |T_g(\theta_j) - T_g(\theta_{j+1})|}{\dfrac{1}{N-1}\displaystyle\sum_{j=1}^{N-1} T_g(\theta_j)}\,. \tag{13}$$

In general, values of Jit_{rel} from 0.1% up to 1.04% are associated to non pathological characteristics for the voice.

4. Simulation

In this section, simulations are performed using the proposed stochastic model. The variation of the parameter a (see Eq. (10)) is considered. The subglottal pressure pattern, more realistic (Lucero et al., 2011) is given by

$$p_{sub}(t) = \begin{cases} 0, & if \;\; 0 \le t \le t_0 \;\; or \;\; T_f - t_0 \le t \le T_f \\ p_m sin(\frac{\pi(t-\tau)}{0.280}), & if \;\; t_0 \le t \le t_0 + 0.140 \\ p_m, & if \;\; t_0 + 0.140 \le t \le T_f - (t_0 + 0.140) \\ p_m sin(\frac{-\pi(t+t_0-T_f)}{0.280}), & if \;\; T_f - 0.140 - t_0 \le t \le T_f - t_0\,, \end{cases} \tag{14}$$

in which p_m is the maximum glottal pressure and $t \in [0, t_f]$. In this work, the values considered are $p_m = 800\,Pa$, $t_0 = 0.01\;s$, and $t_f = 1\;s$. All the other parameters are fixed and their values are summarized in Tab. 1.

Parameters	Values
M	$4.76\,kg/m^2$
B	$100\,(N\times s/m)/m^2$
$\underline{K}$	$4.2\times 10^6 (N/m)/m^2$
K_0	$2\times 10^5 (N/m)/m^2$
η	$5\times 10^3/m^2$
p_m	$800\,Pa$
k_t	1.1
τ	$0.001\;s$
x_0	$10^{-3}\;m$
x_{01}	$10^{-3}\;m$
x_{02}	$10^{-3}\;m$
ρ	$1.15\,kg/m^3$
ℓ	$0.014\,m$
b	10^6

Table 1: List of parameters and their values (M , B, K and K_0 are given per area unit).

Figure 3 shows the graph of the function $t \mapsto p_{sub}(t)$. In general, the frequency response function $\widehat{h}_f(\omega)$ corre-

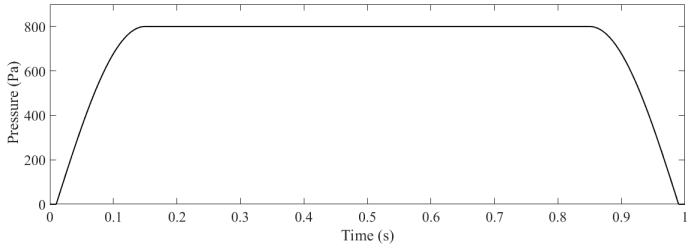

Figure 3: Graph of the subglottal-pressure function $t \mapsto p_{sub}(t)$.

sponding to the digital filter (the vocal tract) is characterized by having only poles associated with the resonance frequencies (the formants) related to voiced sounds. Function $\widehat{h}_f$ considered here takes into account the speech-forming frequencies and the effects of soft-wall, friction, thermal conduction losses and radiation on lips, which will be given by the bandwidths associated with the formant frequencies. Table 2 shows the values (in Hertz) for the first four resonance frequencies of the vocal tract, the formants, for five vowels, and also losses in the vocal tract represented by the bandwidths (Bw) of the formant frequencies based on the findings of Titze et al. (2014). For the simulations, it is

	F1	F2	F3	F4
/a/	860	1513	2489	3600
/e/	423	1899	2017	3546
/i/	283	2113	2800	3566
/o/	504	905	2624	3439
/u/	352	809	2394	3450
Bw	20	25	200	50

Table 2: Formants, in Hertz, of the vocal tract, the digital filter, for the case of vowels production and the corresponding bandwidths.

important to perform a convergence analysis, mainly for the solution of the Itô stochastic differential equation used to generate realizations of $K(t)$. Let $\underline{K} = E\{K(t)\}$ be the mean value and $\overline{K}^2 = E\{K(t)^2\}$ be the second-order moment of $K(t)$, which are classically estimated using the asymptotically stationary and ergodic solution. After performing tests, it was concluded that the convergence for $K(t)$ is warranty from 3×10^5 time steps.

Figure 4 shows three cases of glottal signals considering different levels of jitter, taking into account different values of a, illustrating the generation of the jitter phenomenon with the model proposed. Table 3 shows different

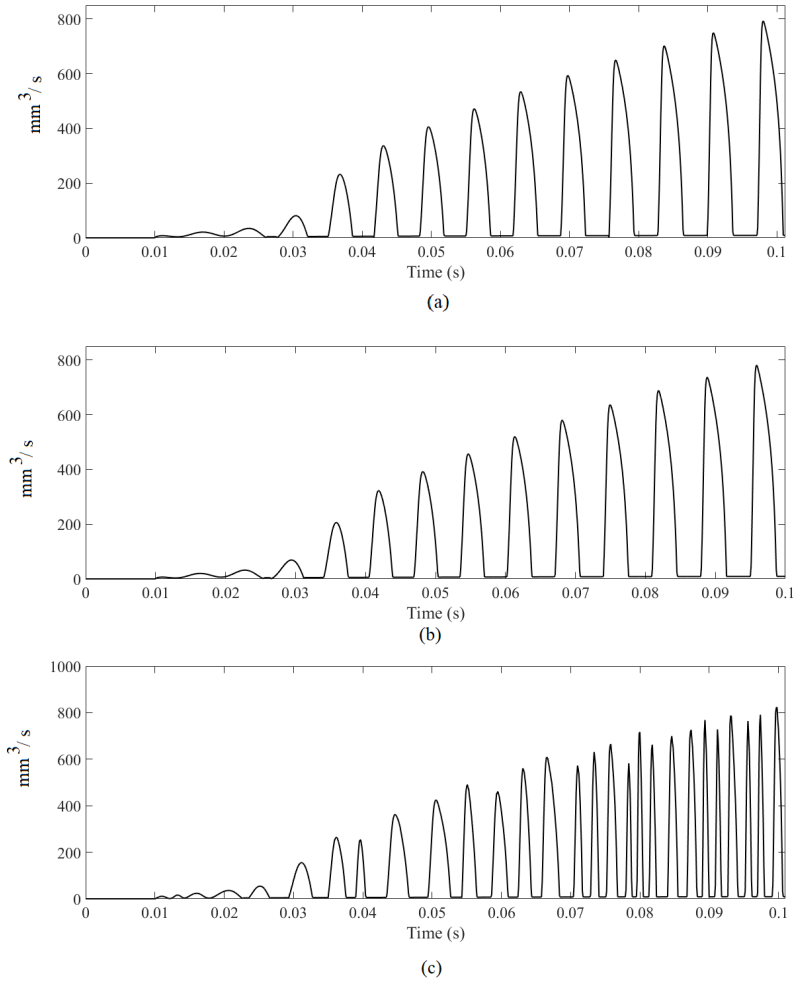

Figure 4: (a) Glottal signal without jitter $a = 0$, (b) Glottal signal with $Jit_{rel} = 0.5\%$ and (c) Glottal signal with $Jit_{rel} = 6.13\%$.

values of jitter, for two corresponding measures, considering different values for the parameter a, in the case of normal voices, without pathological characteristics (the first two values of a) and also three cases of hoarse voices and/or with pathological characteristics (the last three values of a). The values presented here correspond to the voice signal for the producing of a vowel /a/.

a	Jit_{abs}	Jit_{rel}
160	2.27×10^{-5}	0.34%
200	3.25×10^{-5}	0.50%
500	1.72×10^{-4}	2.70%
700	1.88×10^{-4}	3.00%
1000	3.78×10^{-4}	6.13%

Table 3: Absolute and local jitter for voices without pathological characteristics and also with pathological characteristics (hoarse voices).

Let $F_0 = 1/T_g$ be the random variable called the fundamental frequency. Considering the values of the parameter a from Tab. 3, it is possible to construct the probability density function of the random variable F_0, estimated by using the Gaussian kernel estimation method from the nonparametric statistics (Bowman and Azzalini, 1997), and shown in Fig. 5, for all cases of the Tab. 3.

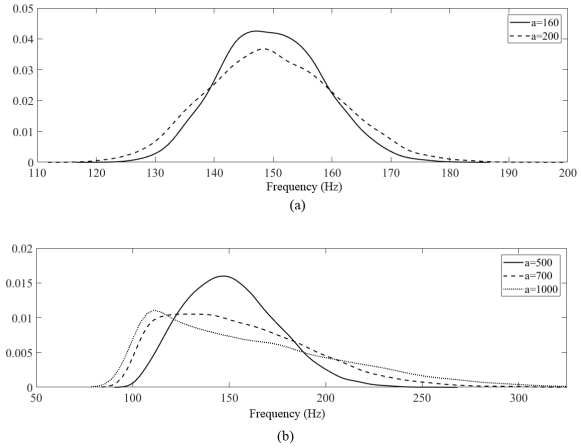

Figure 5: pdf of the fundamental frequency considering (a) $a = 160$ and $a = 200$; (b) $a = 500$, $a = 700$ and $a = 1000$.

Some synthesized sounds corresponding to the five vowels can be heard following the link:
`www.dropbox.com/sh/fw0qmqfn3amo5yv/AABFHfe6Dt6-MpEuw7Ha9wcPa?dl=0`
and are described in Tab. 4: Although the levels of jitter in Tab. 4 correspond to the vowel/a/, all of the other voice

a	*Local jitter*	file
100	0.16%	UnifiedA100Jit00016
500	2.42%	UnifiedA500Jit00242
1000	5.89%	UnifiedA1000Jit00589

Table 4: Synthesized sounds with values corresponding to the vowel /a/ generated.

signals corresponding to the other synthesized vowels have similar levels of jitter.

5. Solution of the inverse problem using an artificial neural network (ANN)

In order to validate the model proposed, parameters a, b, and $\underline{K}$ are identified using experimental voice signals. The main role of $\underline{K}$ is to fit the fundamental frequency while the parameters a and b fit the values of jitter. The objective is to use an ANN to learn the nonlinear relationship between the values of the hyperparameters a, b, and $\underline{K}$ and the measures extracted from the voice signal: the jitter measures and the mean value of the fundamental frequency.

5.1. Constructing the database

The database is constructed using the stochastic model. Values of a, b and $\underline{K}$ are varied and for each given triplet $(a, b, \underline{K})$ in the model, a voice signal is generated, and Jit_{Rel}, Jit_{Abs}, and the value of the fundamental frequency are extracted form it. All the other parameters of the model are fixed.

The fundamental frequency varies from low values up to higher values to cover a reasonable range of frequencies. The main parameter related to the variation of the fundamental frequency is $\underline{K}$, because the mass is fixed. The parameter a is the main responsible for the level of jitter. Then, it is important to vary this parameter from $a = 0$ (the deterministic case) up to a value corresponding to a pathology not much severe or the case corresponding to a hoarse voice. The third parameter, b, is directly related to the convergence of the solution of the Itô Differential Equation, although it also has some contribution to the level of jitter. For this parameter b, a convenient value is chosen so that the solution converges and variations around this value are considered. It is important to say that for each triplet $(a, b, \underline{K})$, different voice signals (realizations) can be generated because they are related to stochastic process $\{K(t)\}_t$. The strategy adopted is then to generate a group of signals related to each triplet $(a, b, \underline{K})$. For each voice signal of this group the parameters Jit_{Rel}, Jit_{Abs}, and the mean value of the fundamental frequency, denoted by $\underline{F}_0$ are obtained. Then, mean values of these parameters are evaluated for all the group and these mean values are associated with each triplet $(a, b, \underline{K})$.

In this work, the variations of the parameters that are considered are as follows: a varies from 0 up to 240, with a step of 30; b varies from 800 000 up to 1 200 000, with a step of 50 000; and $\underline{K}$ varies from 2 000 000 up to 6 000 000, with a step of 200 000 . With these values, the measures obtained from the voice signal have the following variations: the mean value of the random variable associated with the fundamental frequency varies from $103.34\,Hz$ up to $180.89\,Hz$, the absolute jitter varies from 0 up to 3.4389×10^{-4}, and the relative jitter varies from 0 up to $4.1370\,\%$.

5.2. *Methodology and configuration of the ANN used*

A sensitivity analysis of the network target data with respect to the network input data has been performed for the databases. Figure 6 shows a classical estimate of the matrix of the correlation coefficients between each of the parameters a, b and $\underline{K}$ and each of the observed Jit_{rel}, Jit_{abs} and $\underline{F}_0$. generated by the computational model.

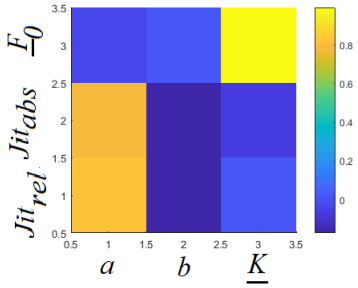

Figure 6: Matrix of correlation coefficients between input and output parameters.

It can be observed that parameter a is highly correlated with the values of jitter and the mean value of the stiffness is highly correlated with the values of the mean of the fundamental frequency. The parameter b has very small correlation with the values of jitter and also with the mean value of the fundamental frequency. All these observations are expected.

To solve the inverse problem, a multilayer feed forward static neural network is used with three neurons in the input layer, corresponding to the values of a, b, and $\underline{K}$, and three neurons in the output layer, corresponding to the values of absolute jitter (Jit_{abs}), relative jitter (Jit_{rel}), and the mean value of the fundamental frequency ($\underline{F}_0$). Several ANN configurations have been tested, in relation to the number of hidden layers and the number of the neurons in each one of them. Sigmoid hidden neurons are used in the hidden layers, while linear output neurons are used in the output layer. The input vectors and target vectors have been randomly divided into three distinct sets for training, validation, and testing, with 70% of the complete dataset assigned to the training set, 15% to the validation set, and 15% to the test set. The values of the input and target vectors were preprocessed and mapped into the normalized range $[-1, 1]$ before presenting to the neural network for training. After training, the network output vectors are then transformed back to the original scale (units) of the network target vectors. The performances of the trained neural networks have been also evaluated by computing the normalized mean-square error between the network outputs and the corresponding targets. The values of the relative jitter, the absolute jitter, and the mean value of the fundamental frequency are used as inputs, and the values of parameters a, b, and $\underline{K}$ are used as outputs. Several hidden layers have been considered, by starting with one hidden layer. However, the differences for the mean-square error (MSE) have not been so significant when more hidden layers or more neurons in each hidden layer have been used. Then, the ANN considered has only one hidden layer containing four neurons. The performance is illustrated in Fig. perfinverseproblem: As expected, the best identification corresponds to parameters a and $\underline{K}$. For parameter b, the identification is not very good due to the fact that its correlation with the inputs are not so high. However, its value does not need to be very precise because this parameter is directly related to the convergence of the solution for the Itô equation and some differences in its identified values will not modify significantly the corresponding voice signals.

5.3. *First example: solving the inverse stochastic problem with a real voice signal without pathological characteristics*

Now, an experimental signal is considered. It is an utterance of a vowel /e/ produced by a man with the following extracted parameters: absolute jitter = 2.5131×10^{-5}, relative jitter = 0.3091%, and the mean value of the fundamental frequency = 123.1086 Hz. The experimental voice used in this case is called *expnormalused* and can be accessed in the link:

`www.dropbox.com/sh/eu7xvwd84dvadbv/AACqHVQ6rdkmap6pG8Uqcy17a?dl=0` .

Solving the inverse problem, through the trained ANN, the values obtained are: $a = 119.37$, $b = 1.01 \times 10^6$, and $\underline{K} = 2.83 \times 10^6\, N/m$. Using these values in the computational model yields: absolute jitter= 2.0750×10^{-5}, relative

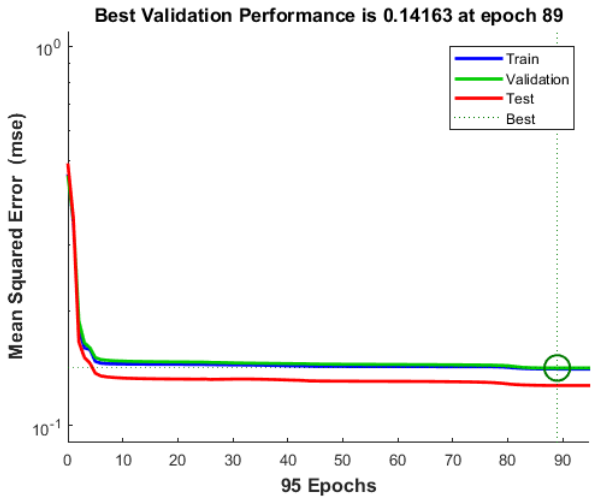

Figure 7: Performance for the ANN considered in the inverse problem.

jitter = 0.2932%, and the mean value of the fundamental frequency = 123.7539 Hz. It can be noted that these values are very near from the obtained with the ANN.

5.4. *Second example: solving the inverse stochastic problem with a real voice signal with a pathology in the vocal folds*

In this case, the utterance of a vowel /e/ is produced by a man with nodulus in the vocal folds and the following parameters were extracted: absolute jitter = 1.2323×10^{-4}, relative jitter = 2.1897%, and the mean value of the fundamental frequency = 178.62 Hz. The experimental voice used in this case is called *exppathologyused* and can be accessed in the link:

`www.dropbox.com/sh/eu7xvwd84dvadbv/AACqHVQ6rdkmap6pG8Uqcy17a?dl=0` .

Solving the inverse problem, through the trained ANN, the values obtained were: $a = 238.47$, $b = 8.8928 \times 10^5$, and $\underline{K} = 5.9591 \times 10^6\, N/m$. If these values are the inputs of the mathematical model used for the direct problem, the values obtained for the outputs are: absolute jitter= 1.5890×10^{-4}, relative jitter = 2.9208%, and the mean value of the fundamental frequency 171.2046 Hz. The values obtained with the experimental signal are near from those obtained with the computational model.

6. Conclusions

A neural network-based identification method has been presented for solving the statistical inverse problem related to the identification of parameters of a stochastic model for producing voice. A database has been generated using the simulations of the proposed stochastic model. A sensitivity analysis of the target data with respect to the input data has been performed. Different configurations of multilayer feedforward neural networks have been tested and the one that has given the minimum value of the mean square error (MSE) has been selected. Two experimental examples have been used to identify the parameters of the model through the defined ANN. One of them corresponding to a healthy voice and other one corresponding to a voice from a patient with nodulus in the vocal folds.

The results obtained showed that it is possible to use an artificial neural network to solve the problem with a small computational cost instead of using the stochastic optimization problem that we had previously been used. In this first essay to solve the inverse problem, it was considered a simpler mathematical model, which does not consider the coupling between the vocal tract and the vocal folds. Only three input parameters of the model have been considered for three measures from the voice signals.

It should be highlighted that only two voice signals were used to validate the model and the only parameter considered was jitter. The initial idea was to show that the consideration about the modelling for the stochastic

process associated to the stiffness of the vocal folds was coherent and could generate jitter. In addition, parameters were identified, showing that it is possible to generate stochastic process corresponding to experimental voices. Not only normal voices but also with pathological characteristics.

It is important to note that the main objective of this research is to identify pathologies through the voice and then it is important to have a low computational cost because the application should be used in real time.

7. Acknowledgments

This work was supported by CNPq.

8. Conflict of interest statement

The authors do not disclose any financial and personal relationships with other people or organizations that could inappropriately influence their work.

9. References

Bowman, A. W., Azzalini, A., 1997. Applied smoothing techniques for data analysis: The kernel approach with S-Plus illustrations. Oxford University Press.

Cataldo, E., Soize, C., 2016. Jitter generation in voice signals produced by a two-mass stochastic mechanical model. Biomedical Signal Processing and Control, 27, 87–95.

Cataldo, E., Soize, C., 2018. Stochastic mechanical model of vocal folds for producing jitter and for identifying pathologies through real voices. Journal of Biomechanics, 74, 126–133, 2018.

Fant, G., 1963. The acoustic theory of speech production. Mouton, The Hague.

Fraile, R., Kobb, M., Godino-LlorentecNicolás, J.I., Sáenz-Lechónc, N., Osma-Ruiz, V. O, Gutiérrez-Arriolac, M. M., 2012. Physical simulation of laryngeal disorders using a multiple-mass vocal fold model. Biomedical Signal Processing and Control, 7 (1), 65–78.

Laje, R., Gardner, T., and Mindlin, G. B., 2001. Continuous model for vocal fold oscillations to study the effect of feedback, Physics Review, 64.

Lucero, J. C., 1999. A theoretical study of the hysteresis phenomenon at vocal fold oscillation onset-offset. The Journal of the Acoustical Society of America, 105, 423–431.

Lucero, J. C, Perlorson, X., Hirtun, A. V., 2020, Phonation threshold pressure at large asymmetries of the vocal folds, Biomedical Signal Processing and Control, vol. 62. https://doi.org/10.1016/j.bspc.2020.102105".

Lucero, J. C., Koenig, L. L., Lourenço, K. G., Ruty, N., and Pelorson, X., 2011. A lumped mucosal wave model of the vocal folds revisited: recent extensions and oscillation hysteresis. The Journal of the Acoustical Society of America, 129, 1568–1579.

Manfredi, C., Bocchia, L., Cantarella, G., 2009. A multipurpose user-friendly tool for voice analysis: Application to pathological adult voices. Biomedical Signal Processing and Control, 4(3), 212–220.

Mendonza, L., Vellasco, M., Cataldo, E., Silva M. B., Apolinario, A. A., 2014. Classification of Vocal Aging Using Parameters Extracted From the Glottal Signal. Journal of Voice, 21(2), 157–68.

Mongia, P. K., Sharma, R. K., 2014. 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, Article ID 290147.

Pinto, N. R., Titze, I. R., 1990. Unification of perturbation measures in speech signals, The Journal of the Acoustical Society of America, 87, 1278–1289.

Prasad, K. S., Ramaiah, G. K., Manjunatha, M. B., 2017. Back end Tools for Speech Synthesis in Speech Processing. Indian Journal of Science and Technology, vol. 10, n. 1, 1–8.

Qureshi, T. M., 2011. A one-mass physical model of the vocal folds with seesaw-like oscillations. Archives of acoustics, 36(1), 15–27.

Rabiner, L. R., Schafer, R. W., 1978. Theory and Applications of Digital Speech Processing, Prentice Hall.

Schoentgen, J., De Guchteneere, R., 1995. Time series analysis of jitter. Journal of Phonetics, 23, 189-201.

Schoengten J., De Guchteneere R., 1997. Predictable and random components of jitter. Speech Communication, 21, 255–272.

Schoengten J., 2001. Stochastic models of Jitter. The Journal of the Acoustical Society of America, 109, 1631–1650.

Talkin, D., 1995. A robust algorithm for pitch tracking (rapt). Speech coding and synthesis, 495–518.

Titze, I. R., Palaparthi, A., Smith, S., 2014. Benchmarks for time-domain simulation of sound propagation in soft-walled airways: steady configurations. The Journal of the Acoustical Society of America, 136(6), 3249–3261.

Titze, I. R., 1984. Parametrization of the glottal area, glottal flow, and vocal fold contact area, The Journal of the Acoustical Society of America, 75, 570–580.

Titze, I. R., 1988. The physics of small-amplitude oscillation of the vocal folds. The Journal of the Acoustial Society of America, 83, 1536-1552.

Vasilakis, M., Stylianou, Y., 2009. Spectral jitter modeling and estimation. Biomedical Signal Processing and Control, 4(3), 183–193.

Wilcox, K. A., Horii, Y., 1980. Age and changes in vocal jitter. Journal of Gerontology, 35(2), 194-–198.

Wong, D., Ito M. R., Cox N. B., Titze, I. R., 1991. 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.