A novel source-filter stochastic model for voice production

Edson Cataldo, L. Monteiro, Christian Soize

A novel source-filter stochastic model for voice production

E. Cataldo[a], L. Monteiro[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

The novel stochastic model to produce voiced sounds proposed in this paper uses the source-filter Fant theory to generate voice signals and, consequently, it does not consider the coupling between the vocal tract and the vocal folds. Two novelties are proposed in the paper. The first one is the new model obtained from the unification of two other deterministic one mass-spring-damper models obtained from the literature and the second one is to build a stochastic model which can generate and control the level of jitter resulting even in hoarse voice signals or with pathological characteristics but using a simpler model than those ones discussed in the literature. An inverse stochastic problem is then solved for two cases, considering a normal voice and other obtained from a case of paralysis on the vocal folds. The parameters of the model are identified in the two cases allowing the validation of the model.

Keywords: Voice production, jitter, stochastic 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. Over the years, researchers have had interests in some aspects of the human voice due to its peculiarities and actions in society, considering that it is the main work tool for many agents, such as singers, speakers, teachers, and others. To study the production of the human voice, a brief introduction on how its generation takes place is needed.

It should be noted that the framework of this paper is not that of voice recognition and speech synthesis but is that of the development of a simple stochastic model that allows generating voice signals with jitter, even hoarse voice and/or with pathological characteristics.

The phonation occurs when the glottis closes (adduction) and a column of air, expelled from the lungs, forces passing through the vocal folds. Then, pulses of air are produced forming a (quasi-)periodic acoustic pressure signal called the glottal signal, which goes into the vocal tract (portion that goes from the glottis up to the mouth), where it is filtered, amplified, and finally, radiated by the mouth generating the sound we hear.

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 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). There are different objective ways to measure jitter as, for example, the absolute jitter and the relative jitter. 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 all, 1991).

Models of jitter could suggest or confirm the mathematical form of markers that would characterize perturbed cycle lengths statistically rather than heuristically. It has been verified empirically that vocal jitter increases for some dysphonic voices (Pinto and Titze, 1990; Schoentgen and De Guchteneere, 1995). In addition, models of jitter can be used to improve naturalness or simulate hoarse voices (Bangayan et al., 1997). Another interesting application of

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

models of jitter is the generation of voice signals to calibrate signal processing algorithms and to help in detecting glottal cycles (Muta et al., 1988).

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). Other authors have recently described a mechanical model for the vocal folds considering the 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 for carrying out its identification by solving a statistical inverse problem.

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 as a stochastic process following the ideas proposed by Cataldo and Soize (2018).

The fact that the model is relatively primitive is exactly one of the main ideas of this work. With this model that is experimentally validated, 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 addition, with this simplified model it is possible to obtain intelligible voiced sounds, characterizing normal voices but also hoarse voices or voices indicating pathologies due to the high level of jitter. It is important to note that Cataldo and Soize (2018) have used concatenated tubes to simulated the vocal tract and here only a frequency response function with bandwidths associated with the resonance frequencies is considered contributing to reduce the computational cost of the model.

2. 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.

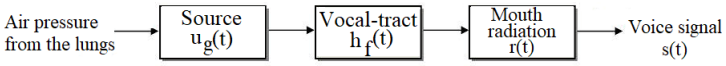

Figure 1: Sketch of source-filter.

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 $r(t)$. We then have,

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

or, in the frequency domain,

$$\widehat{s}(\omega) = \widehat{r}(\omega)\,\widehat{h_f}(\omega)\,\widehat{u_g}(\omega)\,, \tag{2}$$

where $\widehat{\cdot}$ 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. Both models generate the glottal signal and the source. The final idea is to consider 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 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

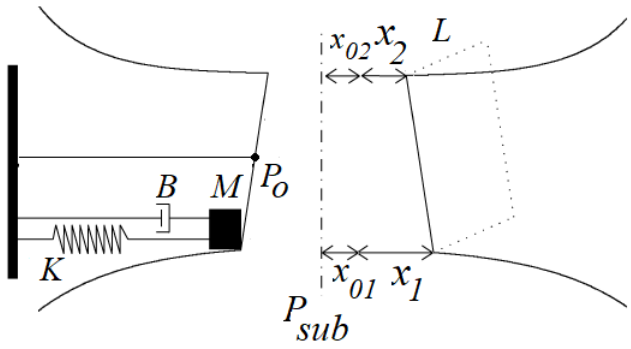

Figure 2: Sketch of the Qureshi model.

vocal fold is the same as its right one. The glottal entry displacement is represented by $x_1(t)$, while $x_2(t)$ corresponds to the glottal exit displacement and P_0 divides the glottis in two parts that are assumed here to have the same lengths. 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):

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

where M, B, and K are the mass, the damping, and the stiffness, *per area unit*, and where 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), is written as,

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

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. The input glottal area is written (Qureshi,2011) as,

$$A_1(t) = \begin{cases} 2L(x_{01} + x_1(t))\,, & x_1(t) > -x_{01} \\ 2Lx_{01}\,, & \text{otherwise}\,, \end{cases} \tag{5}$$

and the output glottal area is given by Eq. (6):

$$A_2(t) = \begin{cases} 2L(x_{02} + x_2(t))\,, & x_2(t) > -x_{02} \\ 2Lx_{02}\,, & \text{otherwise}\,, \end{cases} \tag{6}$$

in which x_{02} is the initial glottal displacement, pre-phonation, in relation to the mean point of oscillation, L is the width of the vocal folds, and $x_2(t) = -(x_1(t) - x_0)$, where the mean position of the oscillation is x_0, which is specified by the horizontal position of the fulcrum point. In this text, the same values are considered for x_{01} and x_{02}. Finally, according to Qureshi (2011) the glottal flow $u_g(t)$ is given by Eq. (7):

$$u_g(t) = \sqrt{\frac{2P_{sub}(t)}{k_t\,\rho}}A_2(t)\,, \tag{7}$$

where ρ is the air density. After generating $u_g(t)$, the convolution with the filter (vocal tract and the mouth) should be performed to produce the sound, which will be done further in this paper.

The next step is the construction of the stochastic model of the vocal folds based upon the unified deterministic model proposed here.

3. Proposed stochastic model

The objective is to vary the frequency of the voice signal. As the mass is fixed, the stiffness is considered as a stochastic process 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 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 will not be explained here, because they can be obtained in the reference (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 is chosen as described by

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

The random generator of stochastic process $\{Z(t), t \in \mathbb{R}\}$ can be constructed using the 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 (Krée and Soize, 1986; Soize, 1994). The power spectral density function of stochastic process Z considered here is given by Eq. (10):

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

The level of jitter will mainly be controlled by a and b. 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 stochastic differential equation,

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

All the other equation related to the unified deterministic model should be rewritten substituting $x_1(t)$ and $x_2(t)$ by random variables $X_1(t)$ and $X_2(t)$, respectively. The realizations of stochastic process K are obtained using Eqs. (8) and (9). Then the realizations of stochastic process X_1 are computed by solving Eq. (11) and the realizations of stochastic process X_2 are deduced. Finally, realizations of stochastic process U_g 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) opens and the second one

the instant when it closes 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, a 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.

(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

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

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

4. Simulations

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 (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, where $t \in [0, T_f]$. 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 given in Tab. 1. Figure 3 shows the graph of function $t \mapsto P_{sub}(t)$. In general, the transfer function corresponding to the digital filter (the vocal tract) is

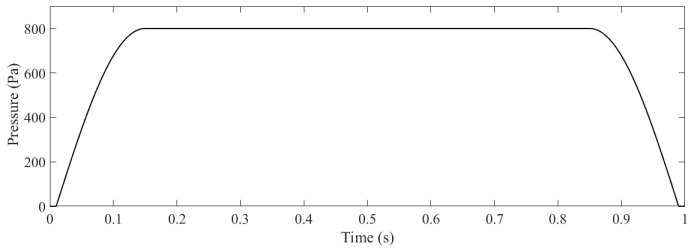

Figure 3: Graph of the subglottal-pressure function $t \mapsto P_{sub}(t)$.

characterized by having only poles associated with the resonance frequencies (the formants) related to voiced sounds. The transfer function 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 important to perform a convergence

Parameters	Values
M	$0.476\times10^{-3}\ kg/m^2$
B	$1\,000\,N\times s/m$
$\underline{K}$	$4.2\times10^{6}N/m^3$
K_0	$2\times10^{5}N/m^3$
η	$5\,000\ \times10^{4}/m^2$
P_m	$800\,Pa$
k_t	1.1
τ	$0.001\ s$
x_0	$4\times10^{-5}\ m$
x_{01}	$10^{-5}\ m$
x_{02}	$10^{-5}\ m$
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).

	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.

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. The convergence for $K(t)$ is warranty from 3×10^5 time steps.

Figure 4 show three cases of glottal signals considering different levels of jitter, taking into account different values of a, showing that it is possible to generate the phenomenon with the model proposed.

Table 3 shows different values of jitter, and 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 voices with pathological characteristics (the last three values of a). The values presented here correspond to the vowel /a/ generated. Let $F_0 = 1/T_g$ be the random variable called the fundamental frequency. Considering

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.*

the values of the parameter a from Tab. 3, it is possible to construct the probability density function of the random

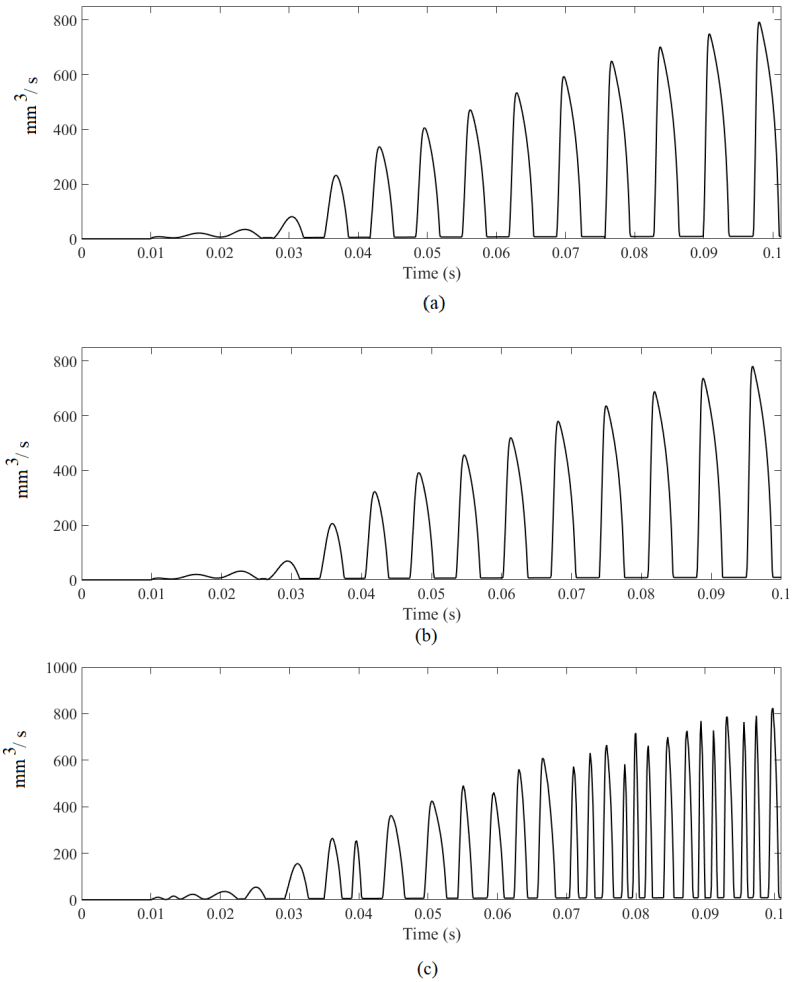

Figure 4: (a) Glottal signal without jitter $a = 0$, (b) Glottal signal with $Jit_{rel} = 0.5\%$ and (c) Glottal signal with $Jit_{rel} = 6.13\%$.

variable F_0. The probability density functions are 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.

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

5. Perspectives of the inverse problem

In order to validate the model proposed, parameters a, b, and $\underline{K}$ are identified using experimental voice signals. The role of $\underline{K}$ is to fit the fundamental frequency and a and b are to fit the values of jitter. This identification is carried out by introducing a cost

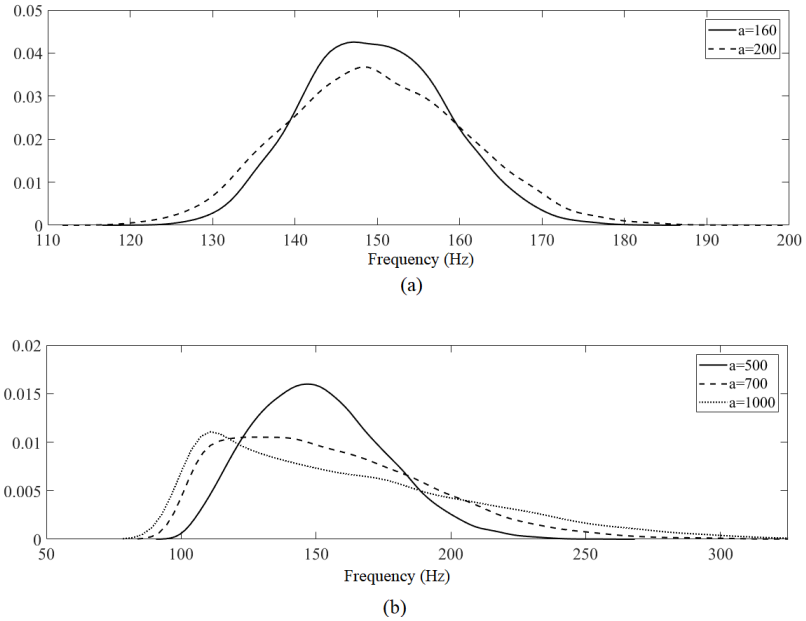

Figure 5: pdf of the fundamental frequency considering (a) $a = 160$ and $a = 200$; (b) $a = 500$, $a = 700$ and $a = 1000$.

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.

function denoted by $V_{cost}(a, b, \underline{K})$ defined by

$$\begin{aligned} V_{cost}(a,b,\underline{K}) = & \left|E\{F_0(a,b,\underline{K})\} - E\{F_0^{exp}\}\right| \\ & + \left|\frac{Jit_{abs}(a,b,\underline{K}) - Jit_{abs}^{exp}}{Jit_{abs}^{exp}}\right| \\ & + \left|\frac{Jit_{rel}(a,b,\underline{K}) - Jit_{rel}^{exp}}{Jit_{rel}^{exp}}\right| . \end{aligned} \tag{15}$$

For each given value of vector $(a, b, \underline{K})$, an utterance of the voice signal is obtained, which allows for computing the expected value of the random variable F_{fund} and the values of jitter (Jit_{abs} and Jit_{rel}) given by $E\{F_0(a, b, \underline{K}\}$, $Jit_{abs}(a, b, \underline{K})$, and $Jit_{rel}(a, b, \underline{K})$, respectively. For each experimental signal, it is possible to calculate the expected value of the fundamental frequency $E\{F_0^{exp}\}$ and the values of jitter Jit_{abs}^{exp} and Jit_{rel}^{exp}. The optimal values a^{opt}, b^{opt} , and $\underline{K}^{opt}$ are then computed by solving the following optimization problem, in which C is the admissible set.

$$(a^{opt}, b^{opt}, \underline{K}^{opt}) = \arg \min_{(a,b,\underline{K})\in C} V_{cost}(a, b, \underline{K}) . \tag{16}$$

5.1. Methodology used

- Step 1: From the experimental voice signal obtained with the vowel produced all the values corresponding to the random variable $F_0 = 1/T_g$ are obtained. The mean value of random variable F_0(denoted by F_0^{exp}) is calculated and it is used in the other steps. From this signal, the two measures of jitter are obtained: Jit_{abs} and Jit_{rel}, denoted by Jit_{abs}^{exp} and Jit_{rel}^{exp}, respectively.
- Step 2: Values for a, b, and $\underline{K}$ have been calculated so that values of the expected value of the fundamental frequency and values of jitter are not so far from those calculated for the experimental signal. This step takes time due to the essays. Values

are obtained and they will serve as start for the grid variation of the values of the parameters a, b, and $\underline{K}$ and, consequently, three loops are constructed.

- Step 3: For each value of the triplet $(a, b, \underline{K})$, the Monte Carlo Method is used for the computation of cost function $V_{cost}(a, b, \underline{K})$.
- Step 4: The minimum value of the cost function estimated in Step 3 is the objective that has to be reached.

5.2. Cases

5.2.1. First case: sustained vowel /a/ - voice signal from a woman without pathological characteristics.

After solving the inverse stochastic problem, the obtained optimal values are $a^{opt} = 100$, $b^{opt} = 1\,100\,000$, and $\underline{K}^{opt} = 5\,700\,000\,N/m^3$. Table 6 shows the values of the expected value of the fundamental frequency and the values of jitter calculated for the experimental voice and for the simulated voice, after solving the inverse problem. As a way to verify what happens when a

Parameters	Experimental	Simulated
$E\{F_0\}$	$272.14\,Hz$	$273.76\,Hz$
Jit_{abs}	$8.57\times10^{-6}\,s$	$7.72\times10^{-6}\,s$
Jit_{rel}	0.23%	0.21%

Table 5: Parameters for a female production of a vowel /a/, without pathological characteristics.

sound is synthesized considering these optimal values of the parameters, a voice signal has been simulated with the optimal values of the parameters. The sounds of the experimental signal (*firstexperimentalcase.wav*) and the corresponding optimal simulated one (*firstsimulatedcase.wav*) can be heard following

`www.dropbox.com/sh/fw0qmqfn3amo5yv/AABFHfe6Dt6-MpEuw7Ha9wcPa?dl=0` .

5.2.2. Second case: sustained vowel /e/ - voice signal from a woman with unilateral paralysis.

Similarly, after solving the inverse stochastic problem, the obtained optimal values are $a^{opt} = 200$, $b^{opt} = 1\,200\,000$, and $\underline{K} = 900\,000\,N/m^3$. Table 6 shows the values of the expected value of the fundamental frequency and the values of jitter calculated for the experimental voice and for the simulated voice, after solving the inverse problem. As a way to verify what happens when a sound is synthesized considering these optimal values of the parameters,

Parameters	Experimental	Simulated
$E\{F_0\}$	$213.05\,Hz$	$215.51\,Hz$
Jit_{abs}	$1.85\times10^{-4}\,s$	$2.04\times10^{-4}\,s$
Jit_{rel}	3.90%	4.35%

Table 6: Parameters for a female production of a vowel /a/, without pathological characteristics

a voice signal has been simulated with the optimal values of the parameters. The sounds of the experimental signal (*secondexperimentalcase.wav*) and the corresponding optimal simulated one (*secondsimulatedcase.wav*) can be heard following

`www.dropbox.com/sh/fw0qmqfn3amo5yv/AABFHfe6Dt6-MpEuw7Ha9wcPa?dl=0`

6. Conclusions

A stochastic model to generate voiced sounds with jitter has been proposed based on the unification of two deterministic models from the literature and considering the parameter corresponding to the stiffness as a stochastic process. Two control parameters have been considered and it is shown that a model based on the source-filter theory, without taking into account the coupling between the vocal tract and the vocal folds, can generate synthesized sounds near real voices, even voiced sounds with pathological characteristics. The level of jitter is measured for each sound synthesized and also the probability density functions related to random variable associated with the fundamental frequency are constructed showing that jitter is created. Two cases are considered for the corresponding inverse problem

helping to validate the model. In addition, the sounds produced with the stochastic model proposed are intelligible, which shows that the model created is very reasonable. Consequently, such a stochastic model is useful for generating big datasets involving different voice signals, with different levels of jitter, and for different kinds of pathologies. Such big datasets can be used for training artificial neural networks.

7. Acknowledgments

This work was supported by CNPq.

8. Conflict of interest statement

The authors disclose any financial and personal relationships with other people or organizations that could inappropriately influence their work.

9. References

Bangayan, P., Long, C., Alwan, A., Kreiman, J. and Gerrat, B., 1997. Analysis by synthesis of pathological voices using the Klatt synthesizer. Speech Communication, 22, 343–368.

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., Sampaio R., 2012. Using Bayesian method for updating the probability density function related to the tension parameter in a voice production model. Journal of Biomechanics, 45 (1), S481.

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.

Krée, P., Soize C., 1986. Mathematics of Random Phenomena. Reidel, Dordrecht.

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

Li, L., Saigusa, H., Hakazawa, Y., et al, 2010. A pathological study of bamboo nodule of the vocal fold. Journal of Voice, 24(6), 738-–741.

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., 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, J. Acous. Soc. Am. 129, 1568-1579.

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

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.

Muta, H., Baer, T., Wagatsuma, K., Muraloka, T., Fukuda, H., 1988. A pitch-synchronous analysis of hoarseness in running speech. The Journal of the Acoustical Society of America, 84, 1292–1301.

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. Backend 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.

Soize, C., 1994. The Fokker-Planck Equation for Stochastic Dynamical Systems and its Explicit Steady State Solutions. World Scientific, Singapore.

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.

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.