Proceedings of the XI DINAME, 28th February-4th March, 2005 - Ouro Preto - MG - Brazil
Edited by D.A. Rade and V. Steffen Jr.

COMPARISON OF SOME MODELS OF LARYNX IN THE SYNTHESIS OF VOICED SOUNDS

Edson Cataldo
Departamento de Matemática Aplicada, PGMEC- Programa de pós-graduação em Engenharia Mecânica, Universidade Federal Fluminense, Rua Mário Santos Braga s/n, Centro, Niterói RJ 24020-140
ecataldo@zipmail.com

Jorge Carlos Lucero
Departamento de Matemática, Universidade de Brasília, Brasília DF 70910-900
lucero@mat.unb.br

Rubens Sampaio
Departamento de Engenharia Mecânica, Pontifícia Universidade Católica do Rio de Janeiro, Rua Marquês de São Vicente, 225, Gávea, Rio de Janeiro, 22453-900
rsampaio@mec.puc-rio.br

Lucas Nicolato
Departamento de Engenharia de Telecomunicações, Universidade Federal Fluminense, Rua Passo da Pátria, 156, São Domingos, Niterói, RJ, 24120-240
lucasnicolato@yahoo.com.br

Abstract. *One of the main objectives in studying the production of voice is that it is the main means of communication. A great part of the framework in the modern systems of Telecommunications, for example, is dedicated to the transmission of voice signals. The process of voiced sounds production can be described as follows: air coming from the lungs is forced through the narrow space between the two vocal cords, which are set in motion in a frequency governed by the tension of the attached tissues. The vocal cords change the type of flow that comes from the lungs into a series of pulses. Then, as the flow passes through the oral and nasal cavities it is amplified and changed until it is radiated from the mouth. This complex process can be modeled by a system of integral-differential equations. In spite of such complexity, this paper shows that it is possible to obtain synthetic voice sounds of satisfactory realism using relatively simple mathematical models. It also shows that the degree of realism is better engnaced by choosing suitable waveforms for the time-varying subglottal pressure, rather than by increasing the degrees of freedom of the model.*

Keywords*: voice production, mechanical models, voice synthesis.*

1. Introduction

The production of voice starts with a contraction-expansion of the lungs. At this moment, an air pressure difference is created between the lungs and a point in front of our mouth, causing an airflow. This airflow passes by the larynx and, before homogeneous, it is transformed into a series of pulses (glottal signal) of air that reach the mouth and the nasal cavity. The pulses of air are modulated by the tongue, teeth and lips; that is, by the geometry of the vocal tract, to produce what we hear as voice. The glottal signal, however, has important properties which are complex to reproduce; they are intimately related to the anatomic and physiological characteristics of the larynx. The theory that has been more accepted to describe the glottal signal is the myoelastic-aerodinamic theory, proposed by van den Berg (1958) and Titze (1980).

To study the system of voice production in a simpler way, we consider four distinct groups: the first one, called *respiration group*, is related to the production of an airflow, that starts and ends in the ending of the traquea. In the larynx, we find the organs of the second group, responsible for the production of the glottal signal, which we call the *vocalization group*. The glottal signal is a signal of low intensity, which needs to be amplified and emphasized at determined harmonic components, so that the phonemes can be characterized. We call this the *resonant group.* This phenomenon occurs when the airflow passes through the vocal tract (portion that goes from the larynx up to the mouth). Finally, the pressure waves are radiated when they reach the mouth. This group we call *radiation group*.

This paper will discuss only the production of voiced sounds, more precisely, only the production of vowels. Its purpose is to compare the quality of the synthesized voice produced by low-dimensional models previously proposed in the literature.

2. Modeling

In vowel production, the airflow coming from the lungs is interrupted by a quasi-periodic vibration of the vocal folds, as ilustrated in the Fig.1.

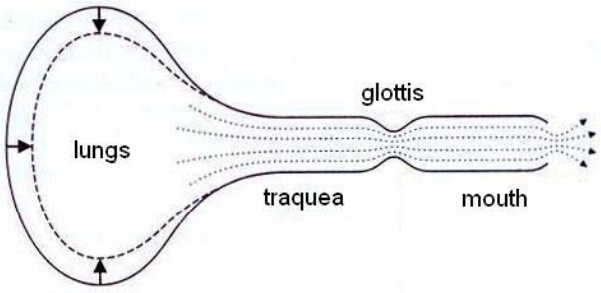


Figure1. Esquematic representation of the voice production system.
(adapted from Titze (1980)).

In the last two decades, the dynamics of vocal cords has been extensively studied, and a number of models of the vocal cords have been developed.

In this paper we will use a single mass model proposed by Flanagan and Landgraf (1968), a double mass (two-mass) model proposed by Ishizaka and Flanagan (1972) and some variations of these models including the ones proposed by Ishizaka and Isshiki (1976) and Gardner et al (2001), where the last one is a model of sound production in a songbird's vocal organ.

We will represent the vocal tract as a series of acoustic tubes concatenated, with section areas varying only with the position and not with time. We may adopt this representation because we are considering only the production of steady vowels.

3. Models Presentation

3.1. Introduction

In this section we will present the models that will be used to the voice production. We will discuss two basic models, one proposed by Flanagan and Landgraf (1968) and other proposed by Ishizaka and Flanagan (1972) and then we will discuss the variation of these models.

3.2. Flanagan-Landgraf model (1972)

The first model to be discussed is the one proposed by Flanagan and Landgraf (1972), whose acoustic circuit representation for the production of voiced sounds is schematized in Fig. 2.

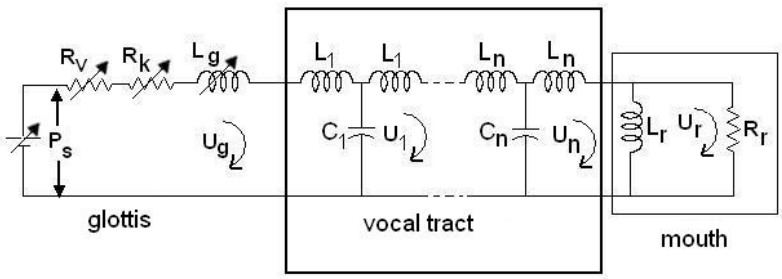


Figure 2. Acoustic circuit representation for the production of voiced sounds.

As the lungs appear as a low-impedance constant-pressure source, and because the pressure drop across the large-area bronchi and trachea is relatively small, the subglottal pressure is approximated by the variable battery P_s. Using the experimental results of van den Berg (1957), the time varying glottal impedance is represented by a viscous non-flow dependent resistance (R_v); a kinetic flow-dependent resistance (R_k) ; and an inertance owing to the mass of (L_g) given in terms of the kinematic viscosity of air, the vocal-cord thickness, the cord length, the area of the glottal orifice, the air density and the airflow through the glottal orifice. These values can be found in Flanagan e Landgraf (1972).

In this model, the vocal cords are considered as a mass-spring-damper system (M is the mass, K is the constant of the spring and B is the constant of the damper). The system is excited by a force $F(t)$, given by the product of the air pressure in the glottis by the area of the intraglottal surface. The force acts in the face of the vocal cord, as schematised in the Fig. 3. The force is distributed and its resultant, which does not appear in the figure, can be thought as applied on mass M .

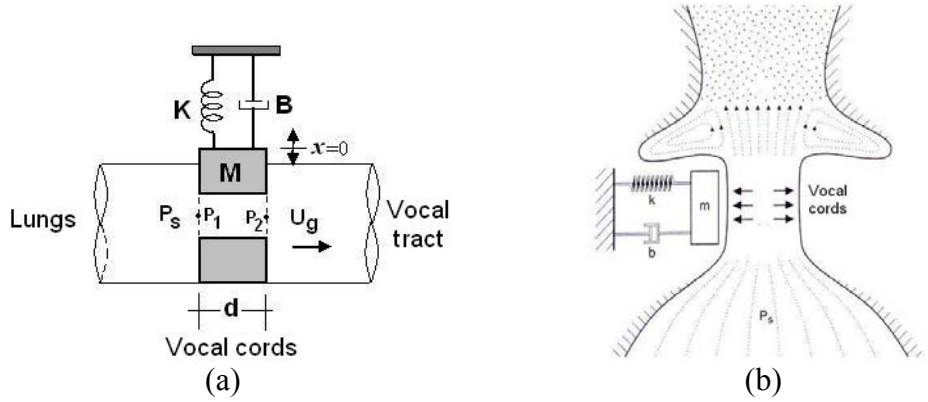


Figure 3. (a) Mechanical model for the vocal cords; (b) Vocal system used (adapted from Titze (1980)). (Flanagan and Landgraf model (1968)).

The equations that give the dynamic of the system (the vocal cords) are given by

$$M\ddot{x} + B\dot{x} + Kx = F(t) \tag{1}$$

where $x(t)$ is the displacement of mass M. The systems is driven by force $F(t)$. In the present study, the forcing function is taken as the mean inlet and outlet pressures; i.e.,

$$F(t) = \frac{1}{2}(P_1 + P_2)(\ell d) \tag{2}$$

acting on the vocal cord face. Experimental measurements show that these pressures can be approximated as

$$P_1 = (P_s - 1{,}37P_B) \qquad \text{and} \qquad P_2 = -0{,}50P_B \tag{3}$$

where $P_B = \frac{1}{2}\rho|U_g|^2 A_g^{-2}$, ρ is the air density, U_g is the acoustic volume velocity through the glottal orifice and A_g is the area of the glottal orifice. The constants ℓ e d are the cord length and the vocal-cord thickness (depth), respectively. The area A_g is variable and given by $A_g = A_{g0} + \ell x$, where A_{g0} is the neutral area.

3.3. Ishizaka and Flanagan model (1972)

Although the one-mass model could produce acceptable voiced-sound synthesis and simulate many of the properties of glottal flow, it was inadequate to produce other physiological detail in vocal cord behaviour. For example, the amount of acoustic interaction displayed between source and tract was greater than observed in human speech. To incorporate more physiological properties, multiple-mass representations of the cords were therefore considered. In this model, the vocal cords are assumed to be bilaterally symmetric. The properties of only one cord are therefore discussed, the same being implied for the opposing cord. A schematic diagram of the glottal system is shown in the Fig. 4.

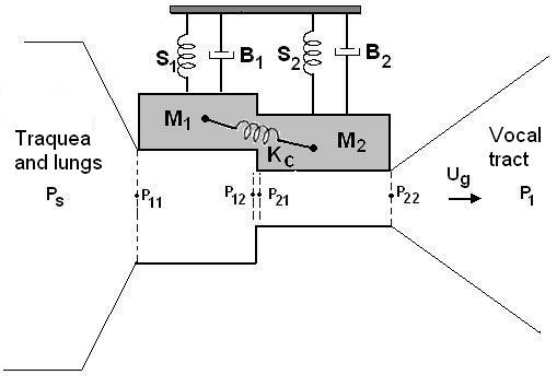


Figure 4. Mechanical model for the vocal cords, proposed by Flanagan and Ishizaka (1972).

The system considered is a system of two degrees of freedom. The springs S_1 and S_2 are non-linear, they represent the tension in the vocal cords, and the spring K_c is linear. The nonlinear relation between the deflexion from the position of equilibrium the the force requested to produce this deflexion is given by $f = Kx(1+\eta x^2)$, where f is the force requested to produce x, K is the non-linear stiffness and η is the coefficient that describes the nonlinearity of the spring S.

During the closing of the glottis, a contact force acting when the masses collide is considered. This force will cause deformation in the vocal cords. The restauration force during this collision process can be represented by a equivalent spring S_{h_j}; with a characteristic non-linear; that is,

$$f_{h_j} = h_j\left(x_j + \frac{A_{g0j}}{2l_g}\right)\left\{1+\eta_{h_j}\left(x_j+\frac{A_{g0j}}{2l_g}\right)^2\right\} \quad \text{for} \quad x_j + \frac{A_{g0j}}{2l_g} \le 0\,,\ j = 1,2 \tag{5}$$

where f_{h_j} is the force requested to produce the deformation in the M_j during the collision, h_j is the linear stiffness and η_{h_j} is a positive coefficient representing the non-linearity of the vocal folds in contact. A resultant force acting in M_j during the closure is given by the sum $f_{S_j} + f_{h_j}$. A_{g0j} is the area of the region between the vocal cords when they are in rest.

The acoustic circuit representation is given in the Fig. 5.

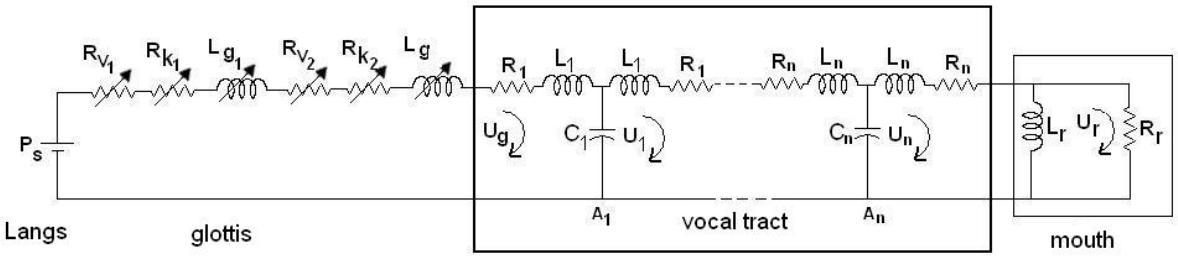


Figure 5. The acoustic circuit representation proposed by Flanagan and Ishizaka (1972).

We consider the kinematic viscosity of air $1.84\text{x}10^{-5}$, the thickness of the vocal cords equal to 0,0032 m, the cord length $1.8\text{x}10^{-2}$ m, the neutral area of the glottal 0,05 cm^2 and the air density 1,3 g/cm^3.

The equations that describe the dynamic of the system (the vocal cords) are given by

$$\begin{cases} M_1\ddot{x}_1 + S_1(x_1) + B_1(\dot{x}_1) + k_c(x_1 - x_2) = F_1 \\ M_2\ddot{x}_2 + S_2(x_2) + B_2(\dot{x}_2) + k_c(x_2 - x_1) = F_2 \end{cases} \tag{6}$$

where F_1 and F_2 are forces acting on M_1 and M_2 over their displacements x_1 and x_2, given in terms of mean pressures acting on the vocal cords exposed faces and the subglottal pressure P_s.

3.4. Gardner et al model (2001): birdsongs

Gardner et al (2001) presented a model of sound production in a songbird's vocal organ and find that much of the complexity of the song of the canary can be produced from simple time variations in forcing functions. The starts, stops and pauses between syllables, as well as variation in pitch and timbre are inherent in the mechanics and can often be expressed through smooth and simple variations in the frequency and relative phase of two driving parameters. We use the same idea to the voiced sound production.

Gardner et al. (2001) considered the following model for the vocal cords dynamics in birdsongs:

$$M\ddot{x} + D\dot{x} + D_2(\dot{x})^3 + Kx = F(t) \tag{7}$$

where the parameters M, K, D and D_2 describe the mass, restitution constant and coefficients of a nonlinear disspation – all per unit area. The nonlinear dissipation term $(\dot{x})^3$ is introduced *ad hoc* so that the variable x can take values only between precise boundaries, mimicking collisions. The force $F(t)$ acts on the vocal cord face and is given

in terms of the intraglottal surface area and the bronchial pressure. They assume that the bird controls vocalizations through the bronchial pressure (P_b) and the labial elasticity (K) given by

$$P_b = P_0 + A\cos[\phi(t) + \phi_i] \quad \text{and} \quad K = K_0 + B\cos[\phi(t) + \phi_j] \tag{8}$$

where $\dot{\phi}(t) = c$ or $\dot{\phi}(t) = 1 - e^{[\phi(t)-\phi_0]^2/\sigma^2}$ for constants ϕ_0 and σ.

We use the same idea for modeling the production of voiced sound. However, we considered $\dot{\phi}(t) = c$ and the bronchial pressure (P_b) in our case is the subglottal pressure (P_s). We observed the influence of P_s variations only, while keeping K constant. In future works, we will treat both variations (P_s and K).

4. Simulation

To develop the necessary simulations, we used the environment MATLAB. The physical models used in this paper are described by systems of equations that involve integrals and differentials, in the time domain. To simulate these systems we need to use numerical methods. Some trials were done using classical numerical methods as Runge-Kutta, for example. However, some numerical instabilities appeared. So, we decided to use the Euler backward method, that consists in a mapping from the continuum frequency domain, represented by S, in the discrete frequency domain, represented by z. The relation between s and z is given by:

$$s \approx \frac{1 - z^{-1}}{T} \tag{9}$$

where T is the sampling period used. In the time domain, we can write the equivalent relations:

$$\frac{dx}{dt} \approx \frac{x[n] - x[n-1]}{T} \tag{10}$$

and

$$\frac{d^2x}{dt^2} \approx \frac{1}{T}\left(\frac{x[n] - x[n-1]}{T} - \frac{x[n-1] - x[n-2]}{T}\right) \approx \frac{x[n] - 2x[n-1] + x[n-2]}{T^2} \tag{11}$$

We approximate the integrals by $\int_0^T x\,dt \approx T\sum_{i=0}^{n-1} x[i]$.

$x[n]$ represents the samples of the signal $x(t)$;i.e, $x[n] = x(nT)$.

5. Comparing the variation of the glottal section area (A_g), the glottal flow (U_g) and the mouth acoustic pressure (P_s)

5.1. Introduction

In this section, we will compare plots obtained from the simulation of the models. We consider for each model (Flanagan and Landgraf model and Ishizaka and Flangan model) the following situations: Ps constant and Ps variable. We show plots that represent the variation of the glottal area (Ag), the glottal airflow (Ug) and the mouth sound pressure, in the production of an /a/ vowel. The time considered for the simulations was 400 ms.

The main values used in the simulation were:

Subglottal pressure constant (P_{s0}) = 783 Pa and subglottal pressure variable (P_s) = $P_{s0}\left(\frac{\pi}{2}\right)sin(2\pi f_P t)$, where $f_P = 1.25\,Hz$ and $P_{s0} = 783\,Pa$.

For the Flanagan and Landgraf model (one mass model): Mass (M) = 0.12 x 10^{-3} Kg, stifness of each vocal cord (K) = $2\pi Mf_0^2$ N / m , natural frequency of the vocal cords (f_0) = 25 Hz, neutral area (A_{gO}) = 5x10^{-6} m^2.

For the Ishizaka and Flanagan model (double mass model): M_1 = 0.1563x10^{-3} Kg , M_2 =0.0313x10^{-3} Kg, K_1 =100N/m, K_2 =10 N/m , K_c =31.25 N/m , Ag_{01} =5x$10^{-6}$$m^2$ and Ag_{02} =5x$10^{-6}$$m^2$.

5.2. Flanagan and Landgraf model (1968)

We compare the results obtained with the simulation of the Flanagan and Landgraf model considering P_s constant (plots shown in the Fig.6) and considering P_s variable (plots shown in the Fig. 7).

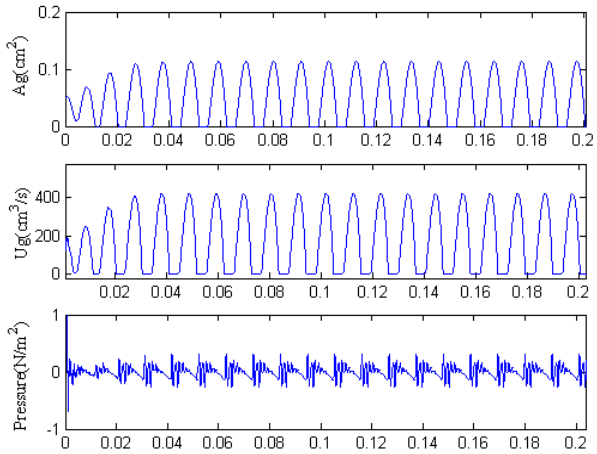


Figure 6. P_s constant.

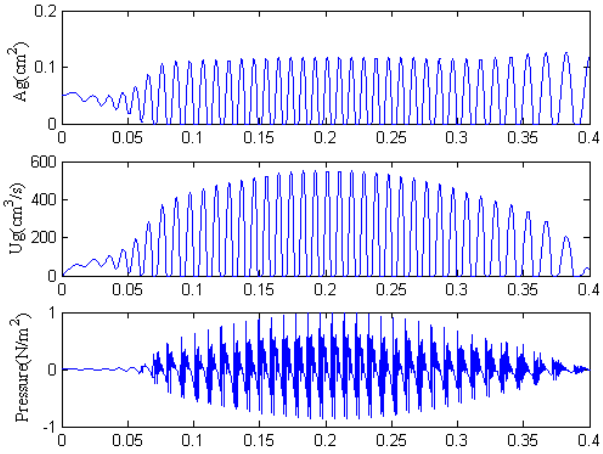


Figure 7. P_s variable.

First, we compare the signal obtained from a real voice, when a steady vowel /a/ is produced, with the two situations shown above. Our intention here is only to show the similarities between the periods of the signals.

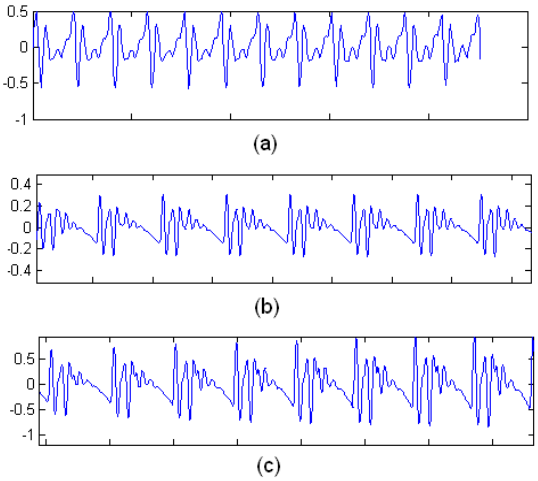


Figure 8. Comparison of signals: (a) real voice signal, vowel /a/ sustained;
(b) and (c) synthetic voice signal

Using the signals generated in Fig. 6 and Fig. 7, we evaluate the maximum airflow (U_g) and the mean airflow (U_g) in both cases and we obtained:

Considering P_s and K constants: Maximum airflow = 423 cm^3 / s and mean airflow = 172,3 cm^3 / s .

Considering P_s variable and K constant - Maximum airflow = 554 cm^3 / s and mean airflow = 163,2 cm^3 / s

From these plots we can observe that when we vary the pressure an amplitude modulation occurs and the maximum airflow increase.

Although the periods of the signals in the Fig. 6 and Fig. 7 are similar, when Ps is constant the synthesized sound is not so natural as when Ps is variable. You can heard the sounds in http://geocities.yahoo.com.br/lucasnicolato.

5.4. Ishizaka and Flanagan model (1972)

We compare the results obtained with the simulation of the Ishizaka and Flanagan model considering P_s constant (graphics shown in the Fig.9) and considering P_s variable (graphics shown in the Fig. 10).

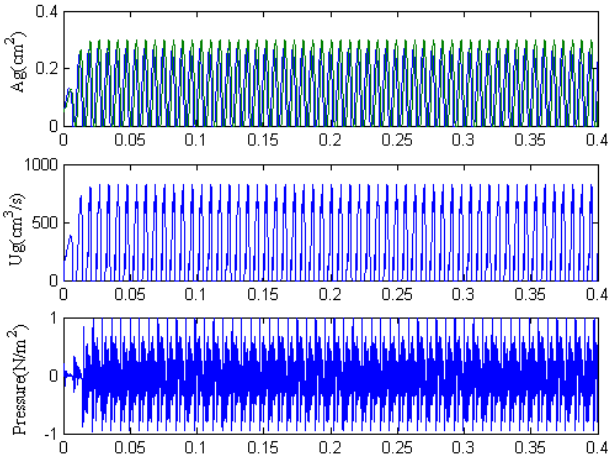


Figure 9. P_s constant.

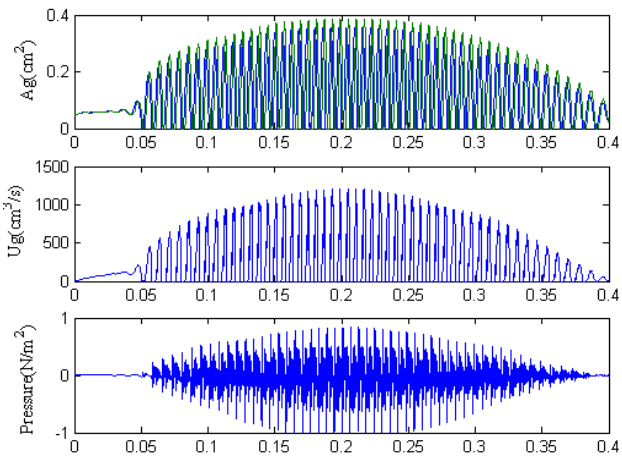


Figure 19. P_s variable.

We show in Fig. 11 and Fig. 12 parts of the signal with Ps constant and Ps variable. The objective is to compare the shape of some periods of the signals.

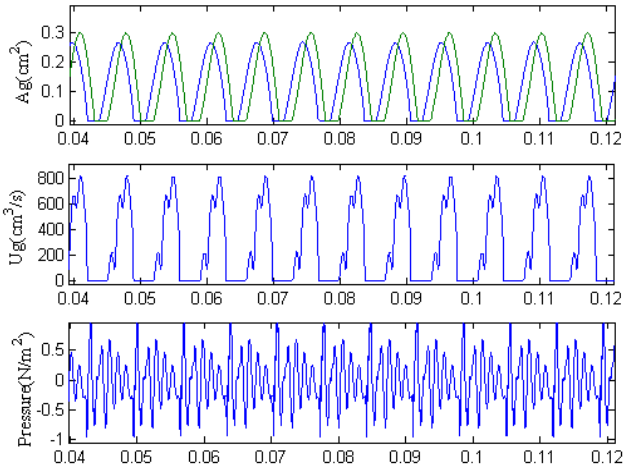


Fig. 11. P_s constant (windowed)

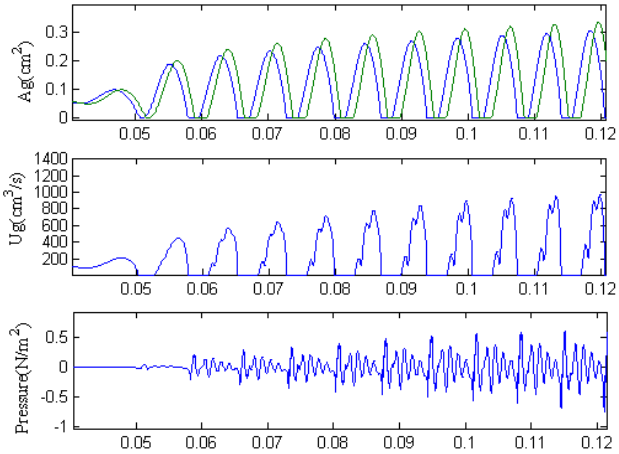


Fig. 12. P_s variable (windowed).

Using the signals generated in Fig. 9 and Fig. 10, we evaluate the maximum airflow (U_g) and the mean airflow (U_g) in both cases and we obtained:

For P_s and K constants: Maximum airflow = $826\,\mathrm{cm}^3/\mathrm{s}$ and mean airflow = $258{,}4\,\mathrm{cm}^3/\mathrm{s}$

For P_s variable and K constant: Maximum airflow = $1208\,\mathrm{cm}^3/\mathrm{s}$ and mean airflow = $249{,}9\,\mathrm{cm}^3/\mathrm{s}$

From these plots we can observe that when we vary the pressure an amplitude modulation occurs and the maximum airflow increase, as in the Flanagan and Landgraf model.

Although the periods of the signals in the Fig. 9 and Fig. 10 are similar, when P_s is constant the synthesized sound is not so natural as when P_s is variable. You can also heard these sounds in http://geocities.yahoo.com.br/lucasnicolato.

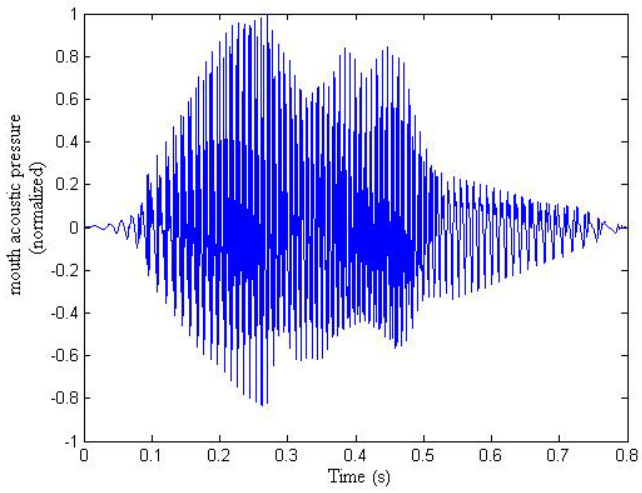
We can also observe that the two-mass model produce a greater airflow.

However, one of the more interesting results is that the naturalness of the sound seems not to be better when we add one mass. The naturalness is better when we vary the subglottal pressure.

6. Diphthongs generation

We use the Flanagan and Landgraf model and we consider Ps variable to generate diphthongs. We do it in the following way: the parameters of the equivalent circuit are evaluated in each sample, considering the vocal tract areas variable. During the first third part of the simulation, the areas are constant and are the same to the corresponding areas of the first vowel. During the middle third part, the areas vary linearly, finishing with the same area of the second vowel. Finally, during the last third part, the vocal tract areas are the same of the corresponding areas of the second vowel.

As an example, we can generate the "ditongo" /ai/. Figure 13 shows the plot of simulated mouth acoustic pressure.



7. Conclusions

Although the system of voice production is complex, we have shown that we can model it with good approximation, using low dimensional systems. This result is in agreement with past studies on vocal fold vibration, which have relied on such simple models to characterize details of its dynamics (e.g.,Lucero (1999)). Further, we have seen that even a simple system mass-spring-damper can be used to model the dynamics of the vocal cords. In this case, the variation of the subglottal pressure is a relevant factor, which properly chosen results in a good realism of the synthesized voice.

7. References

Van den Berg, J.,1958, "Myoelastic-aerodynamic theory of voice production", Journal of Speech and Hearing Research, Vol.1, pp. 227-244.

Titze, I. R., 1980, "Comments on the myoelastic-aerodynamic theory of phonation", The Journal of the Acoustical Society of America, Vol. 23, pp. 495-510.

Titze, I. R., 1994, "Principles of voice production". Prentice-hall, NJ: Englewood Cliffs, NJ.

Flanagan, J. and Landgraf, L., 1968, "Self-oscillating source for vocal-tract synthesizers", IEEE Trans. On Audio and Eletroacoustics, Vol. 16, pp. 57-64.

Ishizaka, K. and Flanagan, J., 1972, "Synthesis of voiced sounds from two-mass model of the vocal cords", Bell Syst. Tech. Journal, Vol. 51, pp. 1233-1268.

Ishizaka, K. and Isshiki, N., 1976, "Computer simulation of pathological vocal-cord vibration", Vol. 60, pp.1193-1198.

Flanagan, J.L.; Ishizaka, K. ; Shipley, K.L., 1975, "Synthesis of speech from a dynamic model of the vocal cords and vocal tract", Bell Syst. Tech. J., Vol. 54 (3), pp. 485-506.

Gardner, T.; Cecchi, G.; Laje, R.; Mindlin, G.B., 2001, "Simple motor gestures for birdsongs", Physical review letters, Vol. 87, No. 20, pp 208101-1 – 208101-4.

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

8. Responsibility notice

The authors are the only responsible for the printed material included in this paper.