Comparison of Some Mechanical Models of Larynx in the Synthesis of Voiced Sounds

Edson Cataldo
Senior Member, ABCM
ecataldo@zipmail.com, ecataldo@vm.uff.br
Universidade Federal Fluminense - UFF
Department of Applied Mathematics
24020-140 Niterói, RJ. Brazil

Jorge Carlos Lucero
lucero@unb.br
Universidade de Brasília - UnB
Department of Mathematics
70910-900 Brasília, DF. Brazil

Rubens Sampaio
Emeritus Member, ABCM
rsampaio@mec.puc-rio.br
Pontifícia Universidade Católica do Rio de Janeiro
Mechanical Engineering Department
22453-900 Rio de Janeiro, RJ. Brazil

Lucas Nicolato
lucasnicolato@yahoo.com.br
Universidade Federal Fluminense - UFF
Telecommunications Engineering Department
24120-240 Niterói, RJ. Brazil

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 folds, which are set in motion in a frequency governed by the tension of their tissues. The vocal folds change the continuous 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 finally 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 simple mathematical models. The perception of a synthetic sound as natural is increased by choosing suitable waveforms for the time-varying subglottal pressure, rather than by augmenting the number of degrees of freedom of the mechanical model. This paper also shows possible ways to adapt the models to voices of men, women and children.
Keywords*: Voice production, articulatory synthesis, signal processing*

Introduction

The beginning of the voice production process occurs with a contraction-expansion of the lungs. The air pressure difference created between the lungs and the mouth exit causes airflow. The airflow passes through the larynx, where it is transformed into a series of pulses (glottal signal) of air that reaches the mouth and the nasal cavity. The pulses of air are modulated by the tongue, teeth, lips, and other articulators that shape the vocal tract (portion that goes from the larynx up to the mouth), and result in the final pressure wave that we perceive as voice. The glottal signal has important properties which are complex to reproduce and are intimately related to the anatomic and physiological characteristics of the larynx. Its generation mechanism has been explained by the so-called myoelastic-aerodynamic theory, proposed by van den Berg (1958) and Titze (1980).[1]

Looking at this system of voice production from a simpler perspective, we may consider four distinct groups: the first one, called *respiration group*, is related to the production of airflow, starts at the lungs and ends in the trachea. 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 vocal folds belong to this 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 may be characterized. This process occurs when the airflow passes through the vocal tract, which we call the *resonant group*. Finally, the pressure waves are radiated, when they reach the mouth, by the *radiation group*. Here, we will consider only the production of voiced sounds.

The present paper is a follow up from previous works (Cataldo *et al.*, 2004, 2006), in which we assessed and compared the capability of some models to simulate vocal productions, and introduced variations of the subglottal pressure to increase the perception of the synthesized sounds as natural sounds. Here, we will review our main results and consider adaptations of the models to the vocal physiology of children, women, and men, to reproduce their voices.

Nomenclature

A_g = *glottal area, m*
A_{g0} = *neutral glottal area, m*
B = *Constant of the damper, used in the one- mass- model, Ns/m*
B_1 = *Damping Matrix , used in the two-masses model,Ns/m*
B_2 = *Damping Matrix,, used in the two-masses model,Ns/m*
C_i = *Capacitances used in the acoustic model,* m^5/N
f_{sj} = *Force requested to produce the displacement* x_j *in the two-mass-model, N*
f_{hj} = *restauration force during the collision process in the two-mass-mode,Nl*
F_1 = *Forcing function in the one- mass- model,N*
F_2 = *Forcing function in the one- mass- model,N*
F = *Forcing function in the one- mass- model,N*
K = *Constant of the spring used in the one-mass-model,N/m*
k_c = *Constant of the linear spring used in the two-mass-model,N/m*
M = *Mass in the one-mass-model,kg*
P_s = *Subglottal pressure,*N/m^2

Presented at XI DINAME – International Symposium on Dynamic Problems of Mechanics, February 28th - March 4th, 2005, Ouro Preto. MG. Brazil.
Paper accepted: June, 2005. Technical Editors: J.R.F. Arruda and D.A. Rade.

P_0 = *Subglottal pressure used in the expression of the Ps variable,N/m²*
P_1 = *Inlet pressure used in the two-masses model,N/m²*
P_2 = *Outlet pressure used in the two-masses model,N/m²*
P_B = *Bernoulli pressure,N/m²*
L_g = *Inertance owing to the mass in the acoustic circuit,Ns/m⁵*
R_k = *Resistance used in the acoustic circuit,Ns/m⁵*
R_v = *Flow independent resistance used in the acoustic circuit in the one-mass model,Ns/m⁵*
U_g = *Flow volume velocity,m³/s*
S_1 = *spring force used in the two-masses model, non-linear,N/m*
S_2 = *spring force used in the two-masses model, non-linear,N/m*
T = *duration of the vowel, s*
x = *displacement of the mass* M, *in the one-mass-model,m*
x_1 = *displacement of the mass* M_1, *in the two-mass-model,m*
x_2 = *displacement of the mass* M_2, *in the two-mass-model,m*

Greek Symbols

ΔT = *sampling period,s*
η = coefficient that describes the nonlinearity of the spring in the one-mass model
η_{hj} = coefficient that describes the nonlinearity of the springs in the two-masses model

Subscripts

1 relative to the first mass in the two-masses model
2 relative to the second mass in the two-masses model

Modeling

In the last decades, the dynamics of vocal folds has been extensively studied and new models have been developed, but older models are still used as a basis to the development of new ones. We will base our considerations on the one-mass model proposed by Flanagan and Landgraf (1968), the two-masses model proposed by Ishizaka and Flanagan (1972), and also some variations of these two models including some ideas used by Gardner et al (2001) when modeling the sound production in a songbird's vocal organ.

The two base models assume perfectly symmetrical vocal folds, and motion in the direction perpendicular to air flow only. The airflow is assumed quasi-steady, and described by Bernoulli's energy equation. Other assumptions may be found in the references.

The vocal tract is represented as a series of concatenated uniform acoustic tubes with stepwise varying cross-sectional area.

Flanagan and Landgraf's Model (1968)

The acoustic circuit representation for voiced sounds production is shown in the Fig 1.

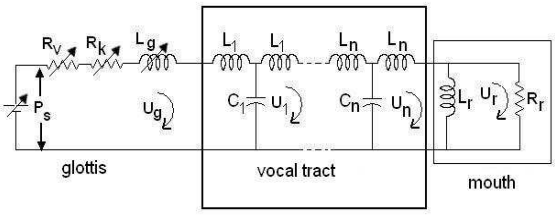


Figure 1. Acoustic circuit representation for production of voiced sounds.

The subglottal pressure is approximated by a voltage source (Ps) because the lungs appear as a low-impedance constant-pressure source and the pressure drop across the large-area bronchi and trachea is relatively small. Using the experimental results of van den Berg (1957), the time varying glottal impedance is represented by a flow-independent resistance (R_v); a kinetic flow-dependent resistance (R_k); and an inertance owing to the mass (L_g) given in terms of the kinematic viscosity of air, the vocal-fold thickness, the fold length, the area of the glottal orifice, the air density and the airflow through the glottal orifice. Representative values of these parameters can be found in Flanagan and Landgraf (1968).

Each vocal fold is represented as a mass-spring-damper system. The system is excited by a force F, given by the product of the air pressure in the glottis with the area of the intraglottal surface. The force acts on the medial surface of the vocal folds, as shown in the Fig. 2.

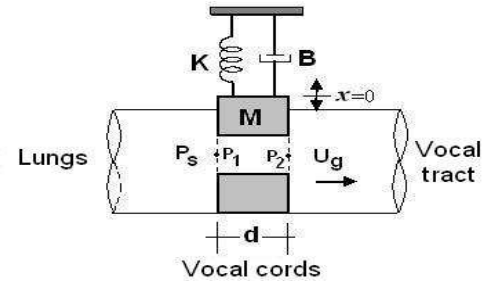


Figure 2. Mechanical model for the vocal folds.

The dynamics of the vocal folds is given by

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

where $x(t)$ is the displacement of the mass M and the force $F(x,t)$ drives the system.

In the present study, the forcing function is computed from the mean value of the inlet and outlet pressures; i.e., acting on the vocal fold surface.

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

Experimental measurements show that these pressures can be approximated as

$$P_1 = P_s - 1{,}37P_B \tag{3}$$

and

$$P_2 = -0{,}50P_B \tag{4}$$

where

$$P_B = \frac{1}{2}\rho\left|U_g\right|^2 A_g^{-2} \tag{5}$$

is the Bernoulli pressure and

$$A_g = A_{g0} + \ell x \tag{6}$$

Ishizaka and Flanagan's Model (1972)

Although the one-mass model may produce acceptable voiced-sound synthesis and simulate many of the properties of the glottal flow, it is inadequate to produce other physiological details related to the vocal folds behaviour.

For example, the prediction of acoustic interaction displayed between source and tract is larger than observed in human speech. To incorporate such physiological properties, multiple-mass representations of the folds have been considered. In Ishizaka and Flanagan's two-mass model, vocal folds are represented by two coupled mass-damper-spring oscillators, as shown in the Fig.3.

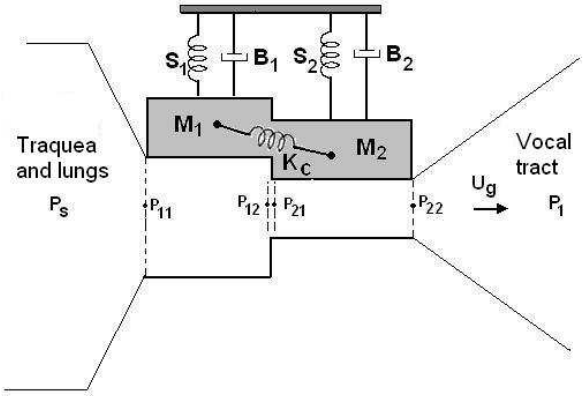


Figure3. Mechanical model for vocal folds proposed by Ishizaka and Flanagan. (1972).

Springs S_1 and S_2 are non-linear and represent the tension in the vocal folds. The nonlinear relation between the deflexion from the position of equilibrium and the force requested to produce this deflexion is given by

$$f_{sj} = K_j x_j (1 + \eta x_j^2) \tag{7}$$

When the opposite masses collide, during glottal closure, a reaction force acts on the masses, which causes deformation of the vocal folds. The reaction force f_{h_j} during collision may be represented by an equivalent spring with a non-linear characteristic of the form

$$f_{h_j} = h_j \left(x_j + \frac{A_{g0j}}{2l_g} \right) \left\{ 1 + \eta_{h_j} \left(x_j + \frac{A_{g0j}}{2l_g} \right)^2 \right\},$$
$$x_j + \frac{A_{g0j}}{2l_g} \le 0\,,\ j = 1,2 \tag{8}$$

where h_j is a linear stiffness coefficient and η_{h_j} is a positive coefficient representing the non-linearity introduced by the contact. The resultant force acting on M_j during the closure is given by the sum $f_{Sj} + f_{hj}$. The dynamics of the system is given by

$$\begin{cases} M_1 \ddot{x}_1 + S_1(x_1) + B_1(\dot{x}_1) + k_c(x_1 - x_2) = F_1 \\ M_2 \ddot{x}_2 + S_2(x_2) + B_2(\dot{x}_2) + k_c(x_2 - x_1) = F_2 \end{cases} \tag{9}$$

where F_1 and F_2 are the forces acting on M_1 and M_2, respectively, and they are computed in terms of mean pressures acting on the vocal fold surface.

The acoustic circuit representation used in the Ishizaka and Flanagan model (1972) is repeated in the Fig. 4.

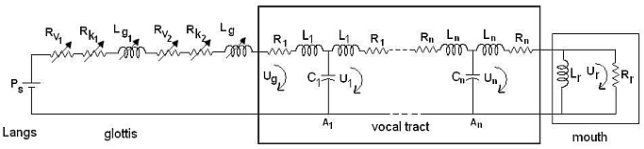


Figure 4. The acoustic circuit representation proposed by Ishizaka and Flanagan (1972).

Variations of the Subglottal Pressure

Gardner et al (2001) presented a model of sound production in a songbird's vocal organ and found that much of the complexity of the canary's singing can be produced introducing 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 used the same idea applied to the base models discussed above.

According to previous experimental results (Lieberman, 1991) the subglottal pressure follows smooth variations at the start and end of an utterance. For simplicity, we consider the variation of the subglottal pressure according to the Eq. (10):Variations of subglottal pressure

$$P_s = P_0 \sin\left(\frac{2\pi t}{T} \right) \tag{10}$$

where T is the duration of the vowel and P_0 is a value of the subglottal pressure.

Simulations

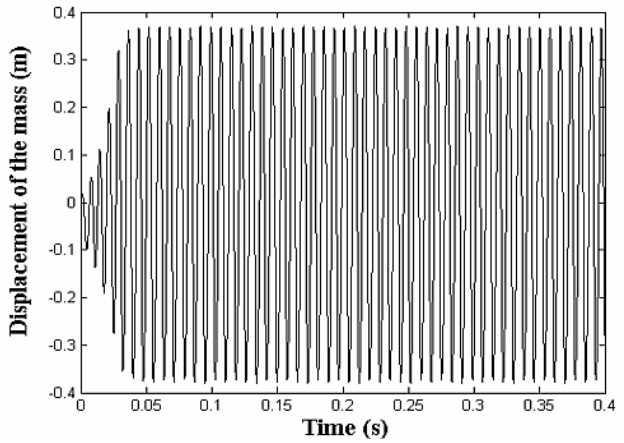
Simulations of voiced sounds were obtained by numerical resolution of the above equations, using Euler's backward method and Matlab software. Using standard values of the parameters, as described by Flanagan and Landgraf (1968) and by Ishizaka and Flanagan(1972), we simulated, or synthetized, the sounds that are available at http://www.professores.uff.br/ecataldo/resultados.htm.

The Fig. 5 shows results from Flanagan and Landgraf's model (1968), for constant (a, b) and varying (c) subglottal pressure. The Fig. 6 shows the same signals, when computed from Ishizaka and Flanagan's model (1972).

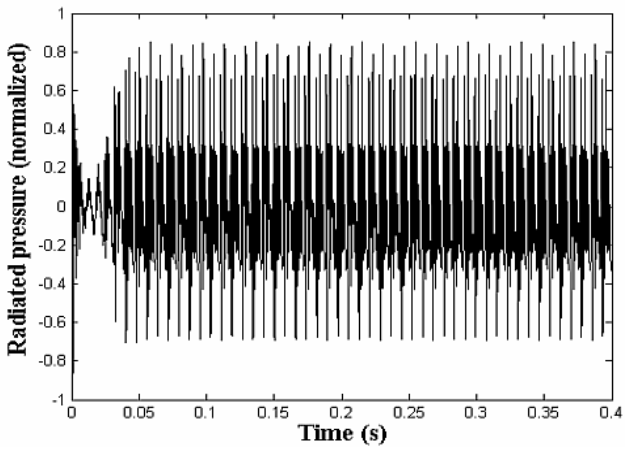
From these plots we may observe an amplitude modulation caused by the varying pressure, and a higher maximum flow. The perception of synthesized sounds as natural is better for varying pressure than for constant. This result seems to imply that introducing simple variations of the subglottal pressure is a more efficient technique to improve the perception of a sound as natural than increasing the complexity of the mechanical representation of the vocal folds by adding additional masses.

Diphtongs Generation and a Plosive Consonant Generation

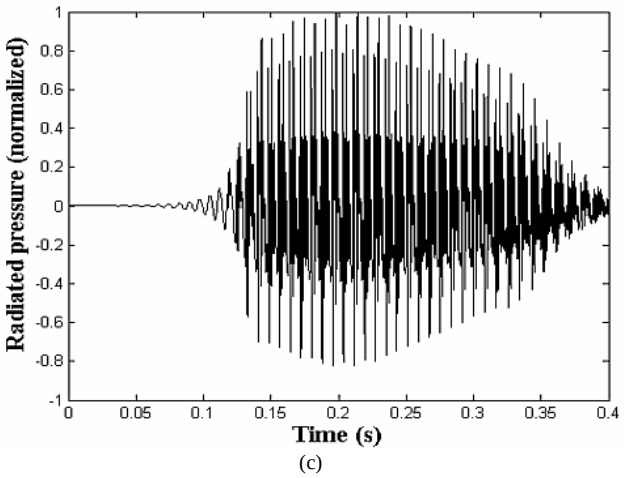
We may also use Flanagan and Landgraf`s model to generate diphthongs and the plosive /p/. Diphthongs were simulated by varying linearly the vocal tract areas between the configurations corresponding to the two vowels of the diphthong. Mouth closure was simulated reducing by 1000 times the last section of the vocal tract (it is not possible to reduce it to zero, lest the occurrence of divisions by zero during the numerical solution of the equations). The example in the Fig. 7 shows a simulation for the Portuguese word /papai/ ("daddy").



(a)

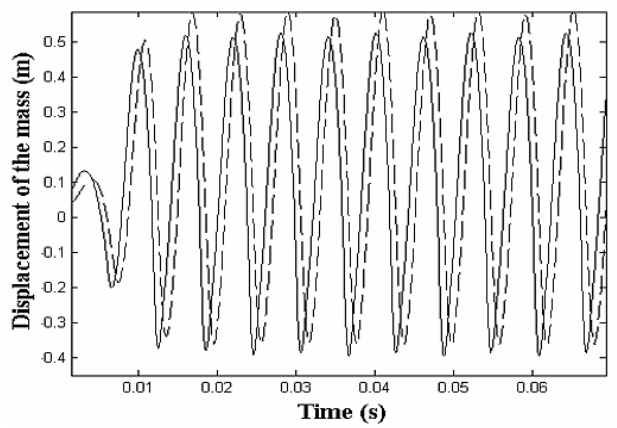


(b)

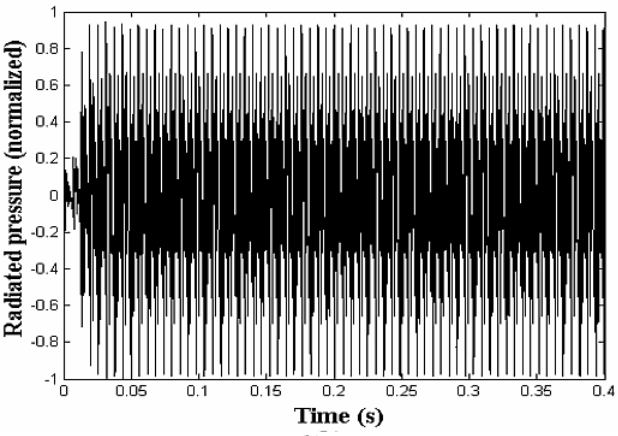


(c)

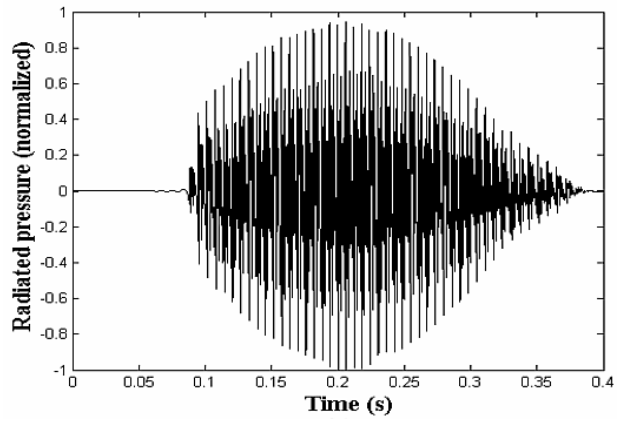
Figure 5. One-mass model. (a) Displacement of the mass (Ps constant). (b) Radiated pressure (Ps constant). (c) Radiated pressure (Ps variable).



(a)



(b)



(c)

Figure 6. Two-masses model. (a) Displacement of the masses. (b) Radiated pressure (Ps constant). (c) Radiated pressure (Ps variable).

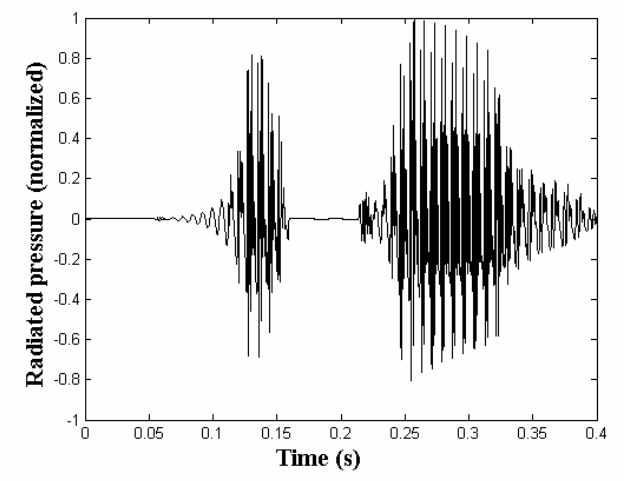


Figure 7. Mouth acoustic pressure in the simulation of the Portuguese word /papai/.

Simulations of Man, Woman and Child Voices

As in previous works (Lucero and Koenig, 2003, 2005), the standard parameters of the above models are adopted as reference for a male configuration. Woman and child configurations are then achieved by adjusting the dimensions of the models to their laryngeal anatomy.

A single scaling factor β is used for all dimensions. This is a simplification of the actual size variations of the larynx and vocal tract. According to Titze (1989), two main scaling factors for the size relation between male and female larynges may be identified, depending on the specific dimension. The relative lengths between the pharynx and oral cavity also differs to men, women and children. Here, the single factor β was adopted as a convenient and simple way to control the overall size of the model. Thus, an adult female configuration would correspond to an approximate factor of $\beta = 0.72$ (according to data by Titze, 1989), and a 5-year-old configuration to $\beta = 0.64$ (according to data by Goldstein, 1980).

All linear dimensions are then scaled multiplying by β. Masses are accordingly computed multiplying by β^3. For the tissue stiffness we assume a constant elasticity modulus for all sizes. This assumption is again a simplification since Titze (1989) reported slighter stiffer tissue for females than for males, probably as a result of differences in tissue composition. Similar differences between child and adult tissues have also been reported (Kurita, Hirano and

Nakashima, 1983). For a constant elasticity modulus, the stiffness coefficient is directly proportional to the cross-sectional area of the tissues, and inversely proportional to their length. Hence, the scaling of all dimensions by a factor β implies that stiffness is also scaled by this same factor. For the tissue damping we assume a constant damping ratio for all sizes.

Using Flanagan and Landgraf's model, with a vocal tract set in a configuration for vowel /a/, we synthesized voices for adult male, female, and a 5-year-old child. We also produced simulations of the Portuguese word /papai/ ("daddy") for the three cases, varying appropriately the vocal tract geometry.

The Fig. 8 shows the plots of the radiate pressure, in the Flanagan and Landgraf's model, for an adult male (a), an adult female (b), and a child voice (c). The three cases correspond to the configuration for the vowel /a/.

The simulations show an increase in the fundamental frequency as the laryngeal size is reduced. In the case of synthesis of vowel /a/, the fundamental frequencies are 126 Hz in the male, 173 Hz in the female, and 192 Hz for the child case.

This increase is approximately proportional to β^{-1}, which is consistent with reported experimental data (e.g., Titze, 1989,1994; Lucero and Koenig, 2003,2005).

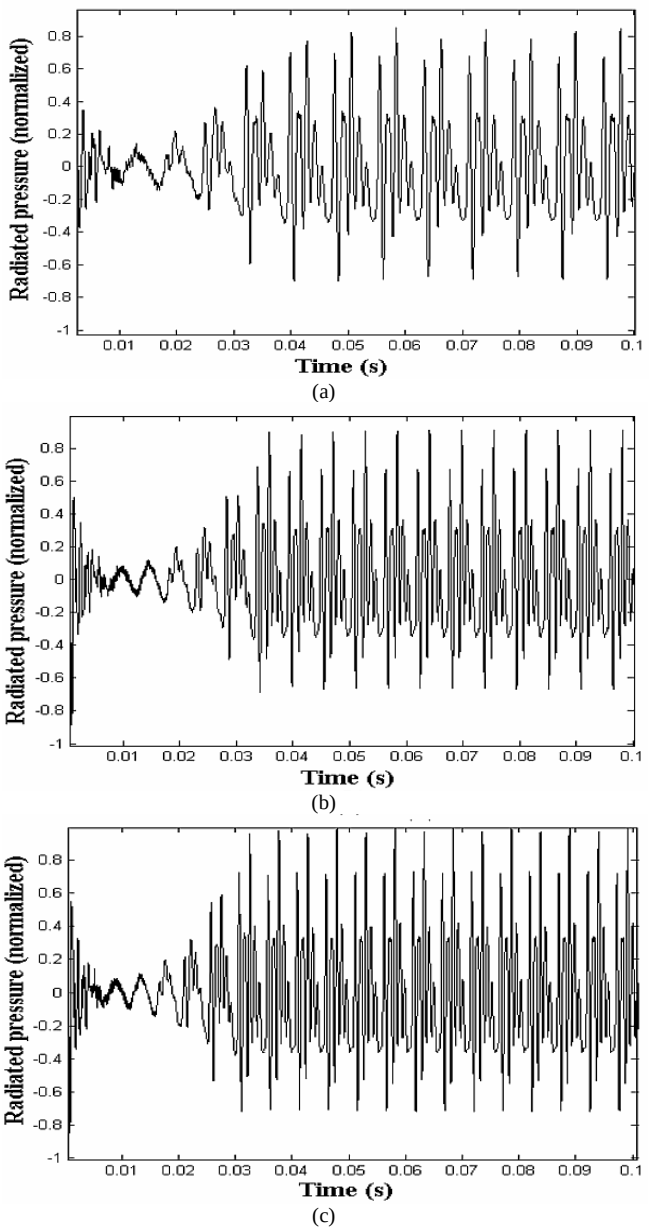


Figure 8. Radiated pressure in the Flanagan and Landgraf's model, for parameter β . (a) β =1 ; (b) β =0.71; (c) β =0.64.

Conclusions

Although the system of voice production is complex, we have shown that it may be modelled with good approximation using low-dimensional systems for adult and child configurations of the vocal organs. This result is in agreement with past studies on vocal fold vibrations, which have relied on simple models to characterize details of the dynamics (e.g. Flanagan et al. 1975; Flanagan and Landgraf, 1968; Ishizaka and Flanagan, 1972; Lucero, 1999; Lucero

and Koenig, 2003,2005). We have also shown that variation of the subglottal pressure is a relevant factor that, when properly chosen, improves the perception of the synthesized sound as natural. This result might be expected, since the actual subglottal pressure never jumps from zero to a non-zero constant value, even in sustained vowels. Due to the physiology of the respiration process, the subglottal pressure must always follow a smooth variation (Lieberman, 1991). For simplicity, we have assumed that such variation has a sinus shape, following Gardner et al. (2001).

On the other hand, an increase of the number of degrees of freedom (masses) of the model did not lead to any noticeable improvement on the generated sound. These results suggest that the perception of a synthesized sound as natural depends more on dynamic variations of the model's parameters than on the complexity of the model itself. Further studies on this issue are deemed necessary to confirm or reject this conclusion, which might have important implications for voice and speech computer synthesis.

References

Cataldo, E., Sampaio, R., Nicolato, L., 2004, "Uma discussão sobre modelos mecânicos de laringe para síntese de vogais" , Engevista, Vol. 6, No. 1, pp. 47-57.

Cataldo, E., Lucero, J.C., Leta, F., Nicolato, L., 2006 (available on line June 20, 2005), "Synthesis of voiced sounds using low-dimensional models of the vocal cords and time-varying subglottal pressure", *Mechanics Research Communications*, Elsevier, v. 33, n. 2, p. 250-260, 2006.

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.

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

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.

Goldstein, U., 1980, "An articulatory model for the vocal tracts of growing children", Ph.D. Thesis, Massachussetts Institute of Technology, Cambridge, MA, USA.

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

Ishizaka, K., Isshiki, N., 1976, "Computer simulation of pathological vocal-cord vibration", *Journal of the Acoustical Society of America*, Vol. 60, pp.1193-1198.

Kurita, S., Hirano, M.,and Nakashima, T., 1983, "Growth, development, and aging of human vocal folds", Vocal fold physiology: Contemporary research and clinical issues, College Hill Press, San Diego, USA, pp. 22-43.

Lieberman, S.E. Blumstein, "Speech physiology, Speech perception and Acoustic Phonetics", Cambridge Studies in Speech Science and Communication, Cambridge University Press, 1991.

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.

Lucero, J. C.; Koenig, L.L., 2003, "Simulations of VhV sequences in children", Proceedings of the 15th International Conference on Phonetic Science, pp. 2905-2908.

Lucero, J. C.; Koenig, L.L., 2005, "Simulations of temporal patterns of oral airflow in men and women using two-masses model of the vocal folds under dynamic control", *Journal of the Acoustical Society of America*, Vol. 117, pp. 1362-1372.

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., 1989, "Physiologic and acoustic differences between male and female voices", *Journal of the Acoustical Society of America*, Vol. 85, pp. 1699-1707.

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

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