Available online at www.sciencedirect.com
SCIENCE DIRECT®
Mechanics Research Communications 33 (2006) 250–260
MECHANICS RESEARCH COMMUNICATIONS
www.elsevier.com/locate/mechrescom

Synthesis of voiced sounds using low-dimensional models of the vocal cords and time-varying subglottal pressure

Edson Cataldo [a,*], Fabiana R. Leta [b], Jorge Lucero [c], Lucas Nicolato [d]

[a] *Universidade Federal Fluminense, Department of Applied Mathematics, Mechanical Engineering Post Graduation Program, Telecommunications Engineering Post Graduation Program, Rua Mário Santos Braga, s/ No - 24020-140 Valonguinho, Centro, Niterói, 24020 140 Rio de Janiero, Brazil*
[b] *Universidade Federal Fluminense, Department of Mechanical Engineering, Mechanical Engineering Post Graduation Program, Niterói, RJ, Brazil*
[c] *Universidade de Brasília, Department of Mathematics, Campus Universitário Darcy Ribeiro, 70910-900 Brasília, DF, Brazil*
[d] *Universidade Federal Fluminense, Department of Telecommunications Engineering, Rua Passo da Pátria 156, 24120-240 São Domingos, Niterói, RJ, Brazil*

Available online 20 June 2005

Abstract

The vocal cords play an important role on voice production. Air coming from the lungs is forced through the narrow space between the two vocal cords that are set in motion in a frequency that is governed by the tension of the attached muscles. The motion of the vocal cords changes the type of flow, that comes from the lungs, to pulses of air, and as the flow passes through the oral and nasal cavities, it is amplified and further modified until it is radiated from the mouth. This complex process can be modeled by a system of integral-differential equations. This paper considers two mechanical models previously used for explaining the dynamics of the vocal cords. It shows that the level of naturalness of the sound generated by these models is rather poor, and it proposes temporal variations of the parameters of the models to increase such level. Examples of synthetic vowels and diphthongs are given to assess the models. In general, the results show that, although the system of voice production is complex, we can achieve satisfactory results with relatively simple low-dimensional models, by suitable temporal variations of the aerodynamic parameters.

Keywords: Voice synthesis; Mechanical models; Simulation; Signal processing

* Corresponding author. Tel.: +55 21 2629 5483; fax: +55 21 2629 5515.
E-mail addresses: ecataldo@zipmail.com (E. Cataldo), fabiana@ic.uff.br (F.R. Leta), lucero@mat.unb.br (J. Lucero), lucasnicolato@yahoo.com.br (L. Nicolato).

0093-6413/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.mechrescom.2005.05.007

1. Introduction

One of the main motivations to study the voice production mechanism is that the human voice is one of the main means of communications.

Voice production 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 the mouth, causing an airflow. This airflow passes through 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 be reproduced; 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-aerodynamic theory, proposed by Van den Berg (1958) and Titze (1980).

Studying 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 trachea. In the larynx, we find the organs of the second group, responsible for the production of the glottal signal, which is called the *vocalization group* (the vocal cords 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 can be characterized. This group is called *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 is called *radiation group.*

In this paper, we will discuss only the production of voiced sounds (vowels).

2. Mathematical modelling

In the production of voiced sounds, the airflow coming from the lungs is interrupted by a quasi-periodic vibration of the vocal cords, as illustrated in Fig. 1.

In the last two decades, the dynamics of vocal cords has been extensively studied, and a number of models of the vocal cords has 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 ideas used by Gardner et al. (2001) when they modelled the 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.

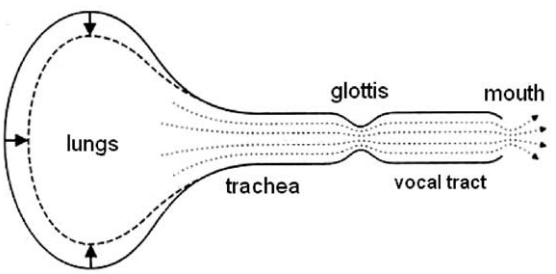

Fig. 1. A representation of the voice production system (adapted from Titze, 1994).

3. Vocal cords models

3.1. Flanagan and Landgraf model (1968)

The first model to be discussed is the one-mass model proposed by Flanagan and Landgraf (1968), whose acoustic circuit representation for the production of voiced sounds is shown in Fig. 2.

Assuming that the lungs appear as a low-impedance constant-pressure source, and that the pressure drop across the large-area bronchi and trachea is relatively small, then the subglottal pressure may be approximated by the variable battery P_s. Using the experimental results from Van den Berg (1958), 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 (L_g) due to the mass given in terms of the kinematic viscosity of air, the vocal-cord thickness, the vocal-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 and Landgraf (1968).

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 shown in Fig. 3. This force is distributed and its resultant is applied on mass M.

The equation that gives the dynamics of the system (the vocal cords) is given by

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

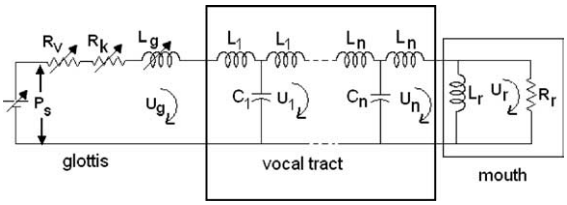

Fig. 2. Acoustic circuit representation for the production of voiced sounds (Flanagan and Landgraf, 1968).

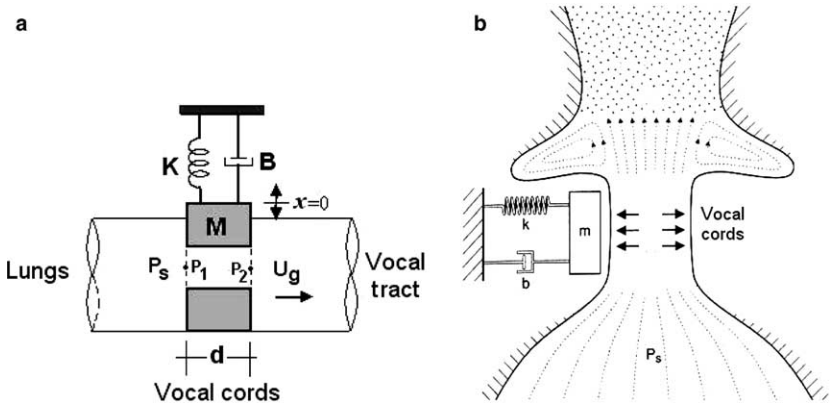

Fig. 3. (a) Mechanical model for the vocal cords and (b) vocal system used. (Flanagan and Landgraf model, 1968).

where $x(t)$ is the displacement of mass M. In the present study, the forcing function is taken as the mean inlet (P_1) and outlet (P_2) pressures; i.e.

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

Experimental measurements (Titze, 1980) show that these pressures can be approximated as

$$\begin{aligned} P_1 &= (P_s - 1.37P_B) \\ P_2 &= -0.50P_B \end{aligned} \tag{3}$$

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 and 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.

3.2. 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 details in the behaviour of the vocal cords. For example, the amount of acoustic interaction displayed between source and tract was greater than observed in human speech. Incorporating more physiological properties, multiple-mass representations of the cords were therefore considered. In Ishizaka and Flanagan's model (1972), each vocal cord is represented by two coupled mass–damper–spring oscillators, as shown in Fig. 4.

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 and 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, we consider that a contact force acts when the masses collide. 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

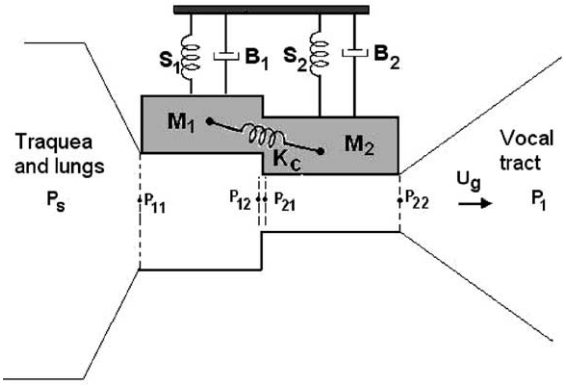

Fig. 4. Mechanical model for the vocal cords, proposed by Ishizaka and Flanagan (1972).

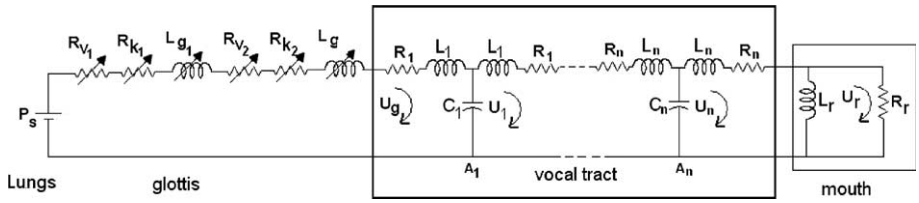

Fig. 5. The acoustic circuit representation proposed by Ishizaka and Flanagan (1972).

$$f_{h_j} = h_j\left(x_j + \frac{A_{g0j}}{2l_g}\right)\left\{1 + \eta_{h_j}\left(x_j + \frac{A_{g0j}}{2l_g}\right)^2\right\} \text{ for } x_j + \frac{A_{g0j}}{2l_g} \leqslant 0, \quad j = 1, 2 \tag{4}$$

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 cords in contact. A resultant force acting in M_j during the closure is given by $f_{S_j} + f_{h_j}$. A_{g0j} is the area of the region between the vocal cords when they are in rest.

The acoustic circuit representation is given in Fig. 5.

We consider the kinematic viscosity of air 1.84×10^{-5}, the thickness of the vocal cords 0.0032 m, the cord length 1.8×10^{-2}m, the neutral area 0.05 cm^2 and the air density 1.3 g/cm^3.

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

$$\begin{cases} M_1\ddot{x}_1 + S_1(x_1) + B_1(\dot{x}_1) + k_c(x_1 - x_2) = F_1 \\ M_2\ddot{x}_2 + S_2(x_2) + B_2(\dot{x}_2) + k_c(x_2 - x_1) = F_2 \end{cases} \tag{5}$$

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.

3.3. 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 canary's song 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 to the mechanics and can often be expressed through smooth and simple variations in the frequency and relative phase of two driving parameters.

This same idea will be applied here for voiced sound production. According to previous experimental results (Lieberman and Blumstein, 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 equation:

$P_s(t) = P_0 \sin\left(\frac{2\pi t}{T}\right)$, where T is the duration of the vowel and P_0 is a value to the subglottal pressure.

4. Simulation

All computer simulations were done using MATLAB software. The simulations were performed by numerical solution of the models' equation. Due to numerical instabilities when using classical numerical methods such as the Runge–Kutta algorithms, an Euler backward algorithm was applied. This algorithm consists in a map from the continuum frequency domain, represented by s, to the discrete frequency domain, represented by z. The relation between s and z is given by

$$s \approx \frac{1 - z^{-1}}{T} \tag{6}$$

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

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &\approx \frac{x[n] - x[n-1]}{T} \\ \frac{\mathrm{d}^2x}{\mathrm{d}t^2} &\approx \frac{1}{T}\left(\frac{x[n] - x[n-1]}{T} - \frac{x[n-1] - x[n-2]}{T}\right) \approx \frac{x[n] - 2x[n-1] + x[n-2]}{T^2} \end{aligned} \tag{7}$$

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

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

5. Simulation results and comparisons

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: P_s constant and P_s variable. The plots below show the variation of the glottal area (A_g), the glottal airflow (U_g) and the mouth sound pressure, in the production of vowel /a/. It is not clear how to assess either the naturalness of the synthesized sound or its approximation to actual voices, in a quantitatively way (Titze, 1994). Following previous studies on the subject (e.g., Koizumi et al., 1987), the naturalness of the synthesized sounds was then evaluated by perceptual comparisons.

The main values used in the simulation were

Subglottal pressure constant (P_s) = 783 Pa.
Subglottal pressure variable (P_s) = $P_{s0}(\frac{\pi}{2})\sin(2\pi f_P t)$, where f_P = 1,25 Hz and P_{s0} = 783 Pa.

For the Flanagan and Landgraf model (one mass model)

mass (M) = 0.12×10^{-3} kg;
stifness of each vocal cord (K) = $2\pi M f_0^2$ N/m;
natural frequency of the vocal cords (f_0) = 25 Hz;
neutral area (A_{gO}) = 5×10^{-6} m^2.

For the Ishizaka and Flanagan model (double mass model):

$M_1 = 0.1563 \times 10^{-3}$ kg, $M_2 = 0.0313 \times 10^{-3}$ 3 kg, K_1 = 100 N/m, K_2 = 10 N/m, K_c = 31.25 N/m;
$A_{g01} = 5 \times 10^{-6}$ m^2, $A_{g02} = 5 \times 10^{-6}$ m^2.

5.1. Flanagan and Landgraf model (1968)

We first show the results obtained with the simulation of the Flanagan and Landgraf model when considering Ps constant (Fig. 6) and considering Ps variable (Fig. 7).

The synthesized sounds may be heard on www.professores.uff.br/ecataldo. It may be noticed the naturalness of the sound in case of variable P_s.

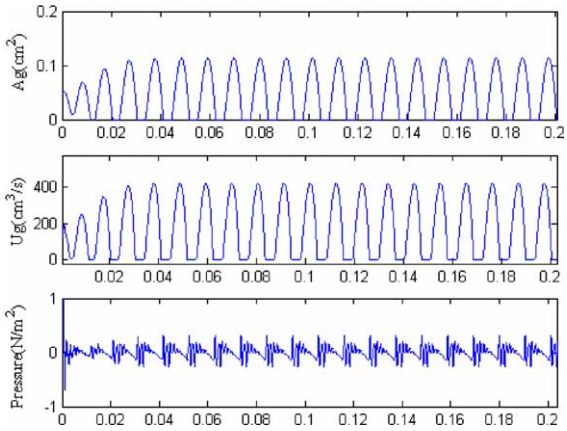

Fig. 6. P_s constant.

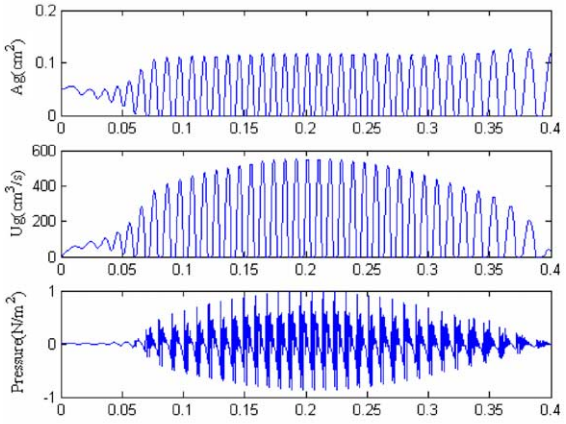

Fig. 7. P_s variable.

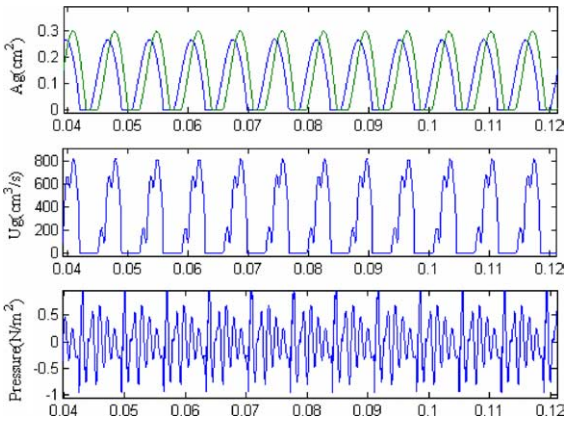

Fig. 8. P_s constant (windowed).

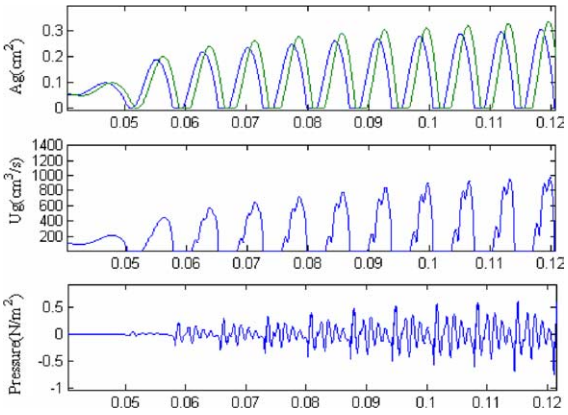

Fig. 9. P_s variable (windowed).

5.2. Ishizaka and Flanagan model (1972)

The next plots in Figs. 8 and 9 show the results obtained using Ishizaka and Flanagan's model, in cases of P_s constant and P_s variable, respectively.

The synthesized sounds may be heard on www.professores.uff.br/ecataldo.

One of the most interesting results is that naturalness of the synthetic sound does not seem to improve when increasing the dimensionality of the model by adding one mass. Rather, it becomes much better when the subglottal pressure is varied.

6. Diphthongs generation

We also used Flanagan and Landgraf model (1968) to generate diphthongs. This was done by varying linearly the vocal tract areas between the configurations corresponding to the two vowels of the diphthong. Fig. 10 shows an example of synthesized mouth acoustic pressure for diphthong /ai/.

The synthesized sounds may be heard on www.professores.uff.br/ecataldo.

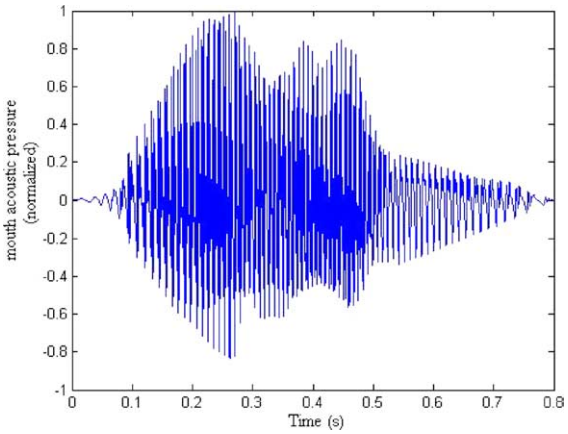

Fig. 10. Mouth acoustic pressure in the simulation of the diphthong /ai/.

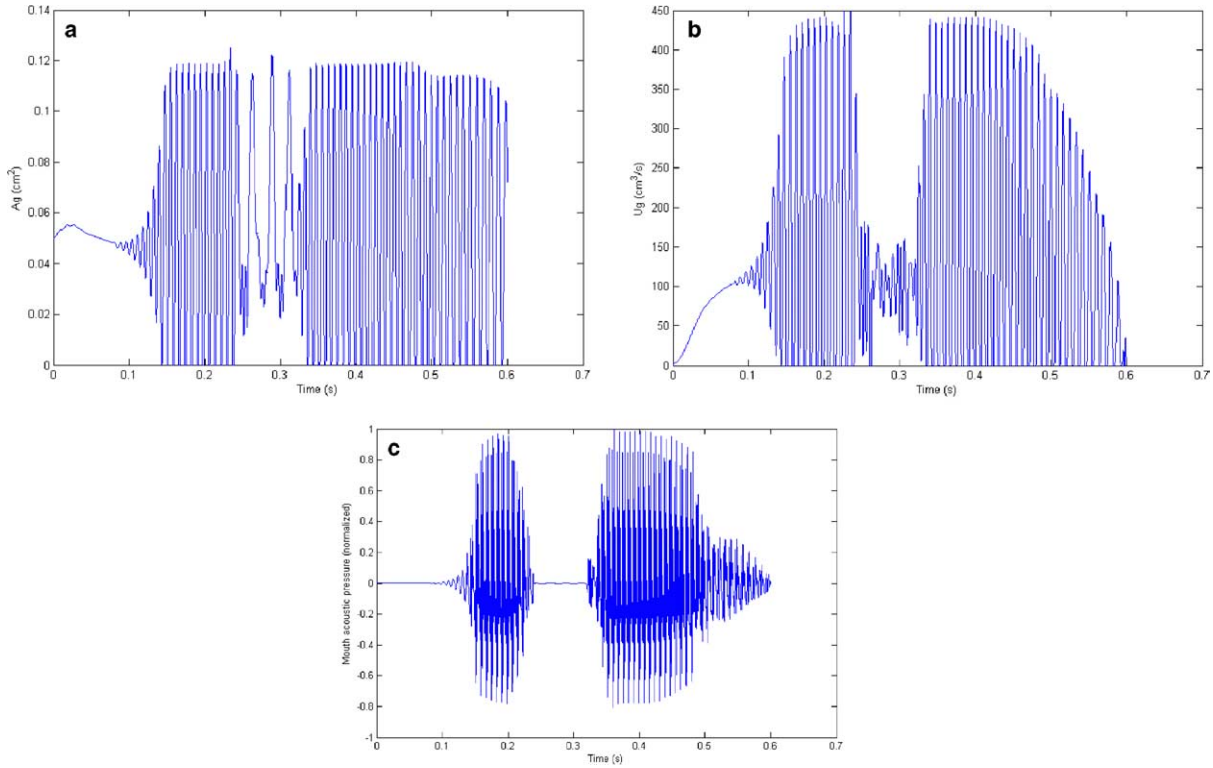

Fig. 11. (a) Glottal area (A_g); (b) glottal airflow (U_g) and (c) pressure (normalized)—synthesis of the Portuguese word /papai/.

7. Generation of a plosive consonant (/p/)

In the following, we show an example using the synthesis of the plosive /p/. We varied the vocal tract shape so that only its last section was reduced 1000 times, to simulate the mouth closure. It is not possible to reduce it to zero, because it could occur divisions by zero in the numerical solution of the system.

In this example, we simulate the Portuguese word /papai/ (“daddy”) using the Flanagan and Landgraf model (1968), varying the subglottal pressure.

The respective plots are shown in Fig. 11. The synthesized sounds may be heard on www.professores.uff.br/ecataldo.

8. Listening test

A listening test was performed to see if the synthetic vowels obtained using time-varying subglottal pressure sounded more natural than those using constant subglottal pressure.

We follow perceptual test setups used in previous studies (Allen and Strong, 1985; Trouvain et al., 1998). Particularly, Allen and Strong (1985) applied such tests to assess the naturalness of synthetic sounds generated by models of the vocal tract.

We tested two types of sounds: a sequence of vowels (/a/, /e/, /i/, /o/, /u/) and the Portuguese word (/papai/). For each type of sound, two models were used (Flanagan and Landgraf; Ishizaka and Flanagan).

For each model, we considered time-varying subglottal pressure and constant subglottal pressure. Each listener was asked to rank the sounds for each type (sequence /a/, /e/, /i/, /o/, /u/, and the word /papai/) in a scale from 1 to 4.

The results of the listening test appear in Table 1 (sequence of vowels) and in Table 2 (\papai \).

The first column for each sound refers to the time-varying subglottal pressure and the second column refers to the constant subglottal pressure case. The results allow us to extract two conclusions: first, the synthetic sounds using time-varying subglottal pressure are perceived as more natural than those using a constant pressure. Second, the sounds generated with Ishizaka and Flanagan's model (IF72), in spite of having an aditional degree of freedom, were not perceived as more natural than those generated by Flanagan and Landgraf's model (FL68).

Table 1
Listener responses from comparison between models, with constant subglottal pressure—CSP (first column) and time-varying subglottal pressure—TVSP (second column), in the case of the sequence of vowels (/a/, /e/, /i/, /o/, /u/)

Listeners	FL68		IF72	
	Constant pressure	Time-varying subglottal pressure	Constant subglottal pressure	Time-varying subglottal pressure
L1	2	4	1	3
L2	1	3	2	4
L3	2	4	1	3
L4	2	4	1	3
L5	1	4	1	3
L6	2	4	1	3
L7	2	4	1	3
L8	1	3	2	4
Total	13	30	10	26
Mean	1.63	3.75	1.25	3.25

Table 2
Listener responses from comparison between models, with constant subglottal pressure (first column) and time-varying subglottal pressure (second column), in the case of the Portuguese word (/papai/)

Listeners	FL68		IF72	
	Constant subglottal pressure	Time-varying subglottal pressure	Constant subglottal pressure	Time-varying subglottal pressure
L1	3	4	1	2
L2	3	4	1	2
L3	3	4	2	2
L4	2	3	1	1
L5	3	4	2	2
L6	3	4	1	2
L7	1	3	1	3
L8	4	2	2	3
Total	22	28	11	17
Mean	2.75	3.50	1.38	2.13

9. 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 cord vibration, which have relied on such simple models to characterize details of its dynamics (e.g., Lucero, 1999). We have also 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, when properly chosen improves the naturalness of the synthesized sound. This fact was verified through listening tests on the synthetic sounds. This result might be expected, since the actual subglottal pressure never jumps from zero to a constant value, even in sustained vowels. Due to the physiology of the respiration process, the subglottal pressure must always follow a smooth variation (Lieberman and Blumstein, 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 degrees of freedom (masses) of the model did not lead to any noticeable improvement on the generated sound. These results suggest that the naturalness of the synthesis depends more heavily 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

Allen, D.R., Strong, W.J., 1985. A model for synthesis of natural sounding vowels. J. Acoust. Soc. Am. 78 (1), 58–69.

Flanagan, J., Landgraf, L., 1968. Self-oscillating source for vocal-tract synthesizers. IEEE Trans. Audio Eletroacoust. 16, 57–64.

Gardner, T., Cecchi, G., Laje, R., Mindlin, G.B. 2201. Simple motor gestures for birdsongs. Phys. Rev. Lett. 87(20) 208101-1–208101-4.

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

Koizumi, T., Taniguchi, S., Hiromitsu, S., 1987. Two-mass models of the vocal cords for natural sounding voice synthesis. J. Acoust. Soc. Am. 82 (4), 1179–1192.

Lieberman, P., Blumstein, S.E., 1991. Speech Physiology, Speech Perception and Acoustic Phonetics. In: Cambridge Studies in Speech Science and Communication. Cambridge University Press.

Lucero, J.C., 1999. A theoretical study of the hysteresis phenomenon at vocal fold oscillation onset–offset. J. Acoust. Soc. Am. 105, 423–431.

Titze, I.R., 1980. Comments on the myoelastic-aerodynamic theory of phonation. J. Acoust. Soc. of Am. 23, 495–510.

Titze, I.R., 1994. Principles of Voice Production. Prentice-Hall, NJ, Englewood Cliffs, NJ.

Trouvain, J., Barry, W.J., Nielsen, C., Andersen, O., 1998. Implications of energy declination for speech synthesis. In: Proceedings 3rd ESCA/CCOSDA Workshop on Speech Synthesis. Jenolan Caves, pp. 47–52.

Van den Berg, J., 1958. Myoelastic-aerodynamic theory of voice production. J. Speech Hear. Res. 1, 227–244.