Nonlinear aspects of analysis and synthesis of speech time series data

M. A. Trevisan, M. C. Eguia, and G. B. Mindlin
Departamento de Física, FCEN, UBA, Ciudad Universitaria, Pab. I (1428), Buenos Aires, Argentina

(Received 25 July 2000; published 25 January 2001)

In this work we study a simple model of voiced sound production. We analyze contributions that the qualitative theory of dynamical systems can make to the analysis and synthesis of human speech.

DOI: 10.1103/PhysRevE.63.026216 PACS number(s): 05.45.−a

I. INTRODUCTION

Speech data constitute an amazing test bench for nonlinear dynamics. From developed turbulence to relaxation oscillations, every classical problem studied in nonlinear dynamics plays a role in the production of human voice. Therefore, time series data from speech present important challenges for the time series analyst. For example, self-oscillations might be established in the vocal folds as parameters are changed, which dynamically enrich their spectral properties, with simultaneous changes in the fundamental frequency. Since our observations are mediated by filters, the time series can present features resembling bifurcations and even chaotic behavior [1]. Since realistic models of the folds can indeed present bifurcations of the self-oscillations, it is important to know what to expect from the simplest models of voice production.

The term voice refers only to the sounds produced by vocal fold oscillations. In fact, a first order classification of the sounds used in human speech can be made in terms of whether the vocal folds oscillate or not, i.e., voiced or unvoiced sounds. Vowels are a typical example of voiced sounds [2], and in this work we will analyze dynamical aspects of their production.

The spectral content of voiced sounds is fairly simple, displaying a discrete number of peaks (harmonics of the fundamental frequency of oscillation of the vocal folds), modulated by a smooth function. The classical theory accounting for this observation is known as the source filter theory [2]. In this framework, the time varying flow through the glottis is filtered by the vocal tract. Therefore, some of the harmonics of the fundamental frequency are enhanced and others are reduced, producing a rich variety of sounds. The frequencies enhanced by the vocal tract are known as formants, and in the case of the vowels the ratio between the first two determines them.

The study of voiced sounds includes the acoustic aspects of a very complex tract and the dynamics involved in the oscillations of the vocal folds. Moreover, the dynamics can become extremely rich as soon as we integrate into the study the effects of the coupling between the sound source and the filter [3]. In this work, we review, from the dynamical systems perspective, one of the simplest models of vocal fold oscillation [4]. We analyze the bifurcation structure of its solutions as parameters are changed. Finally, we show how a combined analysis of the bifurcating solutions, enriching their spectral content, and a proper scaling allows us to fit (and therefore synthesize) speech data.

The work is organized as follows. In the second section we review some important models accounting for vocal fold oscillations. In the third section we study the simplest of these models in detail, discussing its range of validity and the structure of its solutions. The fourth section is dedicated to a simple model for the vocal tract filter. The fifth section shows how to use these simple elements to synthesize human speech data, and a genetic algorithm is used to properly fit the model parameters. The last section discusses future work, applications, and conclusions.

II. MODELS

The construction of models of vocal folds has a long and rich history. Basically, the vocal folds are the source of voiced sounds. A flow induced instability of these opposed ligaments modulates the airflow, giving rise to a sequence of pulses that excites the vocal tract. A seminal work within this area was done by Ishizaka and Flanagan [5]. Interested in the problem of achieving a realistic synthesis of voiced sounds, they built a very successful model of two stiffness-coupled masses, which is used almost 30 years after its publication [6]. Although the authors of this model also mention the importance of understanding the critical parameters of the mechanism in order to address the diagnosis of voice disorders, it has been pointed out that the key parameters in the model have been difficult to relate to anatomical features [4]. Keeping the simplicity of the two mass model, Story and Titze introduced a three mass model that allows a better connection between physiological and model parameters [7]. This model builds upon the work of Hirano [8], who stressed the importance of understanding the vocal fold structure in order to properly explain the onset of vocal fold oscillations. The models in [5] and [7] represent an adequate compromise between very simplified one mass models (in which the vocal folds are modeled by one mass spring driven by airflow with an inertial coupling to the vocal tract [4]) and models that include several masses (as in [9,10]).

In this section we review in a qualitative way two simple models for vocal fold oscillations [4]. The motivation behind these models is the convenience of framing the basic mechanisms in simple mathematical terms and working out threshold conditions for the onset of oscillations in terms of parameters that could easily be compared with experimental ones. Both of them are based on the principle that vocal fold oscillation is induced by glottal airstream flow, and are conceived to account for the onset of the oscillations that build up from spread apart vocal folds, with no glottal closure. We

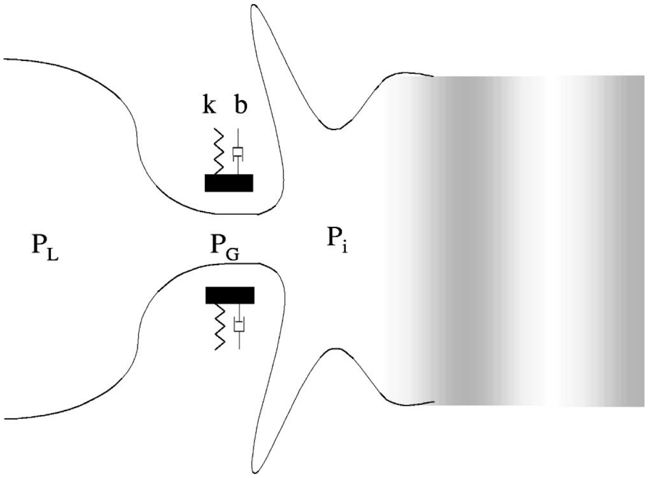

FIG. 1. The elements of the one mass model.

can think of the vocal folds as elastic masses that are pushed together by the (negative) pressure of the intraglottal airstream, but it is important to note that in order to create an oscillatory instability the driving force has to change in alternate quarter cycles. In this way, mechanical energy can be transferred to the vocal folds.

The simplest possible model reflecting this mechanism is known as the one mass model [4]. Each vocal fold is assumed to be equivalent to a mass, subjected to an elastic restitution force, a dissipative force, and the force due to the intraglottal pressure (see Fig. 1). In this model, it is possible to show that the intraglottal pressure is equal to the pressure at the entrance of the vocal tract. Therefore, in order to have oscillatory instabilities from the equilibrium position of the vocal folds, we need a positive pressure (larger than atmospheric) when the folds are spreading apart, and a negative pressure (smaller than atmospheric) when the folds are approaching each other. This model will be a good approximation as long as the air, for the range of frequencies involved, is mainly inert. As the vocal folds open, the flow rises and the air column gets accelerated. This implies a positive input pressure for the vocal tract, and therefore a positive intraglottal pressure, which further opens the folds. This model will be reviewed in detail in the following section, but one point has to be stressed: a one mass model is not capable of displaying self-oscillations without vocal tract loading.

The simplest way to achieve self-oscillations in the absence of coupling is by assuming a nonuniform tissue structure [4]. The idea is that the shape of the glottis can change over a cycle, giving rise to different pressure profiles. This can lead to the asymmetry that we need to transfer mechanical energy to the folds. In Fig. 2, we display such a scenario. If the glottis is convergent (diverging) when it is opening (closing), the average intraglottal pressure will be positive (negative), and therefore the motion will be enhanced. This model constitutes a simplified version of the two mass model [5], an interesting one displaying quite complex behavior for a wide region of its parameter space.

In this work, we will concentrate on the simplest model and show that its nonlinear self-oscillations (created in Hopf bifurcations) can lead, once properly filtered, to realistic outputs.

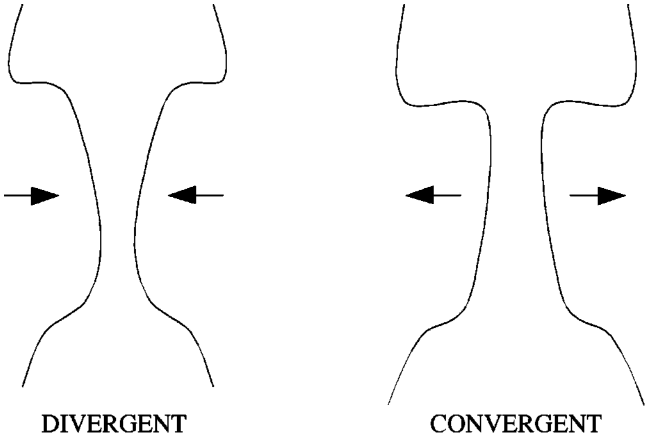

FIG. 2. The nonuniform tissue model: a convergent glottis (right) and a divergent glottis (left).

III. THE ONE MASS MODEL

As stated in the previous section, this model assumes that the vocal folds move symmetrically, and each fold can be thought of as a mass subjected to a restitution force, a dissipative force, and the driving force due to the intraglottal pressure. Therefore, its dynamics will be given (in terms of its displacement x) by

$$Mx'' + Bx' + Kx = P_g, \tag{3.1}$$

where M, B, and K represent the mass, the dissipation constant, and the restitution constant, all per unit area. The variable P_g stands for the intraglottal pressure.

For the general case in which the entry glottal area (a_1) and exit glottal area (a_2) are different, it has been shown that the relationship between the transglottal pressure and the intraglottal pressure is given by

$$P_g = P_i + (P_s - P_i)(1 - a_2/a_1 - k_e)/k_t, \tag{3.2}$$

where P_i is the input pressure at the vocal tract [4] and P_s the subglottal pressure. The coefficients K_e and K_t are phenomenological quantities accounting for the differences between the relationships between pressure and velocity expected for steady flows, and are known as the pressure recovery coefficient, and transglottal pressure coefficient, respectively [11]. Since $a_1 = a_2$ in our case, and $K_e \approx 0$ whenever a_2 is much smaller than the vocal tract area S_t, we get $P_i = P_g$.

The last element that is needed in order to solve our dynamical equation for x is a relationship between the driving pressure and the intraglottal variables. Rothenberg [12] showed that, whenever the fundamental frequency is smaller than the first resonance (formant) of the vocal tract, its input impedance is mainly inertial [12], allowing us to write

$$P_i = R_2 U + I_2 U', \tag{3.3}$$

where the coefficients R_2 and I_2 are the resistance and the inertial constant and U stands for the flow. In order to proceed to form a closed dynamical system, we relate the lung pressure P_L and the input pressure by

$$P_L-(R_2U+I_2U')=k_t\rho v^2, \tag{3.4}$$

where v stands for the velocity of the air at the glottis, i.e., $v=U/a$. Notice that the glottal area a can be written in terms of the vocal fold length L and the equilibrium position of the fold x_0 as $a=2L(x+x_0)$.

Now it is possible to write the equation as a system of three equations of first order: defining F as

$$F=P_L-\rho/2v^2-2L[R_2(x_0+x)+I_2y]v/[2L(x_0+x)I_2], \tag{3.5}$$

it reads

$$x'=y, \tag{3.6}$$

$$y'=1/M[2LR_2x_0v+2LI_2x_0F$$
$$-(B-2LI_2v)y-(K-2LR_2v-2LI_2F)x], \tag{3.7}$$

$$v'=F. \tag{3.8}$$

Notice that this system of equations was derived under the assumption that the vocal folds do not close (as is implied, for example, in the relationship between P_L and P_s through a phenomenologically corrected version of Bernoulli's theorem). Therefore, it is expected to help us understand the system's behavior only for $x>-x_0$.

We are going to study the solutions of this system and their qualitative changes as parameters are varied [13]. The pressure at the lungs and the restitution constant K are a sensible set of parameters to explore, since they are typically controlled by a normal speaker.

The fixed points are easy to find. They will all be located at $y_f=0$, and will satisfy

$$P_L-\rho/2v_f^2-2L[R_2(x_0+x_f)]z_f=0, \tag{3.9}$$

$$2LR_2x_0vx_f-(K-2LR_2v_f)x_f=0. \tag{3.10}$$

In general there will be up to three solutions of this system, yet, only two in the domain of interest ($x>-x_0$), one of them at a positive $v_f=v_1$ value, and the other at a negative $v_f=v_2$ value. The regions in parameter space with qualitatively different fixed point local stability are displayed in Fig. 3. Although we are interested in the behavior of the fixed point at $v_1>0$, we will describe the whole flow.

We begin the description of the dynamical responses of the system with region I, which is the most relevant one in terms of voice production. In this region, the fixed point at $v_1>0$ is a saddle focus. It has a stable direction (approximately parallel to the v axis), and a two dimensional unstable manifold (associated with complex conjugate eigenvalues) that feeds an attractive limit cycle. The coexisting fixed point at $v_2<0$ is a saddle. Locally, it has a stable manifold approximately parallel to the x axis, and two unstable directions (approximately parallel to the y and v axes). A two dimensional manifold, tangent at the fixed point v_2 to its stable manifold and to the first unstable direction described above, separates the phase space between the basin of attraction of the limit cycle and the points that tend to infinity. This two dimensional manifold approaches the singular plane at $x=-E_0$ at the one dimensional curve in which the numerator of F is equal to zero. In Fig. 4 we show a two dimensional projection of the flow on the x,v axes. A piece of the boundary basin is displayed together with the trajectory of an initial condition right above it, approaching the attracting limit cycle. In Fig. 5 we display the time evolution of $P_i=P_L-1/2\rho v^2$. The line in parameter space separating regions I and II indicates the Hopf bifurcation in which the limit cycle is created from $v_1>0$. In other words, in region II, the fixed point at $v_1>0$ is an attractor.

For completeness, we describe the evolution of the flow as parameters are changed, even when they are referred to changes in the basin of infinity. Toward region III, the fixed

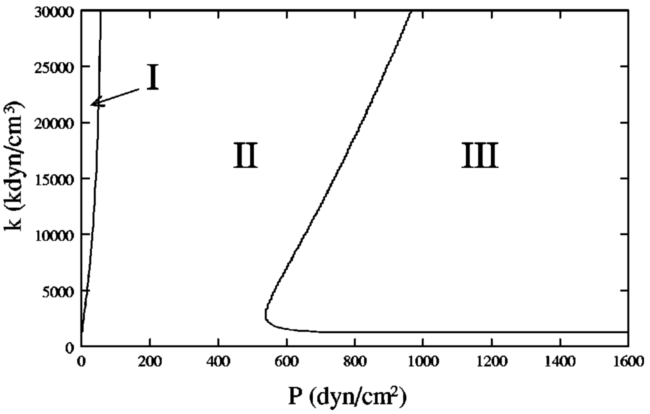

FIG. 3. Bifurcation diagram of the one mass model. Within region I, an attracting limit cycle represents the oscillation of the the vocal folds responsible for the production of voiced sounds.

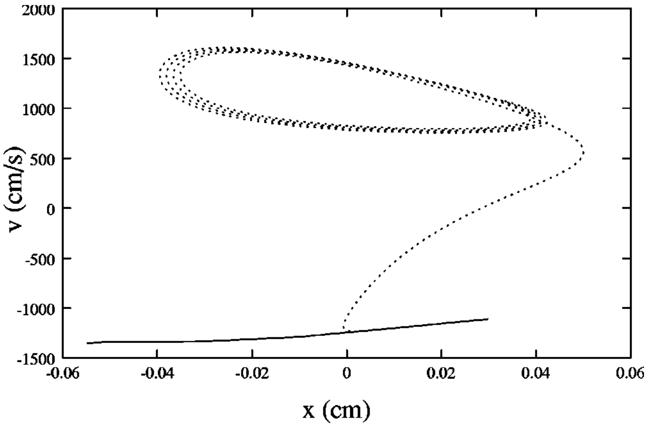

FIG. 4. A two dimensional projection of the flow on the x,v axes. A piece of the boundary basin is displayed together with the trajectory of an initial condition right above it, approaching the attracting limit cycle. The parameters used in this simulation are $L=0.14$, $M=0.04$ g/cm^2, $k=20\,000.0$ kdyn/cm^3, $x_0=0.06$ cm, $P=840.0$ dyn/cm^2, $r_0=0.001\,14$ g/cm^3, $R_2=3.1$ dyn s/cm^5, $I_2=0.1$ dyn s^2/cm^5, and $c=35\,000$ cm/s.

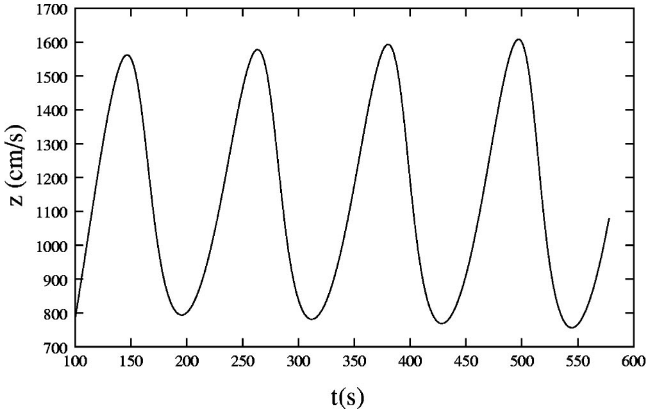

FIG. 5. The time evolution of the input pressure at the vocal tract, for the same parameters used to generate Fig. 4.

point at $v_2 < 0$ has two eigenvalues which become complex conjugate [with associated eigenvectors almost in a plane parallel to the plane (x',z)]. At the separatrix between regions II and III, an inverse Hopf bifurcation takes place: the saddle point at $v_2 < 0$ emits a saddle limit cycle, gaining stability.

In summary, the one mass model predicts the appearance of self-sustained oscillations as the flow is increased. Dynamically, the oscillation is created in a Hopf bifurcation, and therefore the solution, as the flow is increased, changes its frequency and its spectral content as a nonlinear oscillation is established.

IV. THE VOCAL TRACT

Measuring the pressure fluctuations generated as a voiced sound is produced, we see typically time series that look like the one displayed in Fig. 6 [14]. Notice the difference between this time series and the one displayed in Fig. 5. According to the source filter theory of voiced sounds, the difference is due to the filtering effects that occur in the vocal tract. A series of partial reflections and transmissions happen in different parts of it, enhancing some frequencies and suppressing others. The simplest model that one can conceive to reproduce these time series consists of three tubes, of different lengths (L_i) and areas (A_i) [2]. The input pressure P_i generates a wave, which is partially reflected and partially transmitted to the second tube that models the tract at the interface between the first two tubes. The coefficient of reflection at this interface is given by $r_{1,2} = (A_1 - A_2)/(A_1 + A_2)$, and the transmission coefficient $t_{1,2}$ by $t_{1,2} = 1 - r_{1,2}$. Clearly, the transmitted wave is partially reflected at the second interface, and partially reflected toward the third tube. At the interface between the last tube and the atmosphere, the wave is partially reflected and partially emitted toward the atmosphere. Calling $a(t)$ $[b_b(t)]$ the forward (backward) wave in the first tube, $b_f(t)$ $[c_b(t)]$ the forward (backward) wave in the second tube, and $c_f(t)$ $[d_b(t)]$ the forward (backward) wave in the third tube, the equations accounting for the boundary conditions are

$$a(t) = P_i(t) + b_b(t - \tau_1), \tag{4.1}$$

$$b_b(t) = r_{1,2} a(t - \tau_1) + t_{2,1} c_b(t - \tau_2), \tag{4.2}$$

$$b_f(t) = t_{1,2} a(t - \tau_1) + r_{2,1} c_b(t - \tau_2), \tag{4.3}$$

$$c_b(t) = r_{2,3} b_f(t - \tau_2) + t_{3,2} d_b(t - \tau_3), \tag{4.4}$$

$$c_f(t) = t_{2,3} b_f(t - \tau_2) + r_{3,2} d_b(t - \tau_3), \tag{4.5}$$

$$d_b(t) = \alpha c_f(t - \tau_3), \tag{4.6}$$

where α accounts for the reflection coefficient of the interface between the third tube and the atmosphere (with no losses, $\alpha = -1$), and τ_i is the time that it takes a sound wave to travel the length L_i.

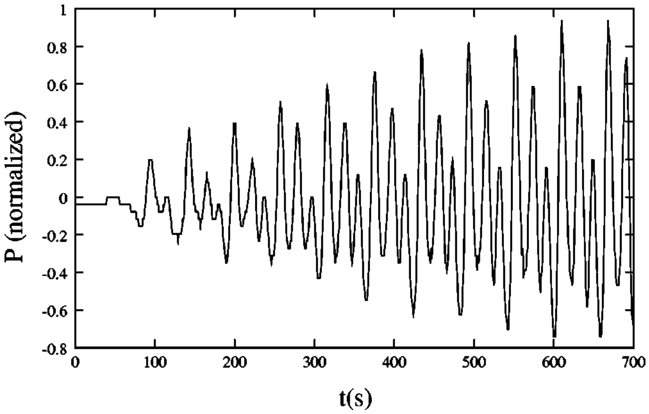

FIG. 6. The experimental record of the pressure fluctuations at the mouth as the vowel ‘‘u’’ is pronounced in Spanish.

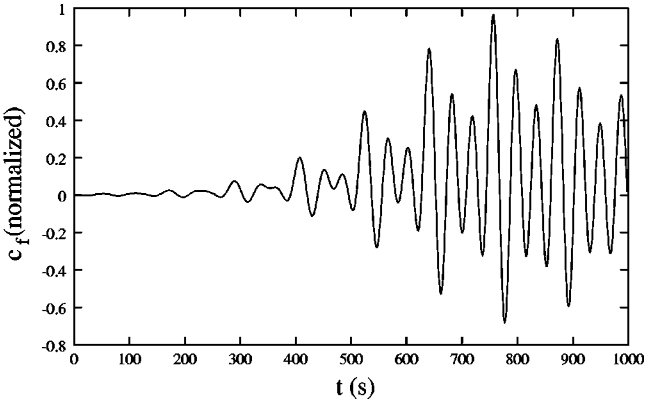

FIG. 7. A time series corresponding to $c_f(t)$, where the input pressure is the one displayed in Fig. 5.

In Fig. 7, we display a time series corresponding to $c_f(t)$, where the input pressure is the one displayed in Fig. 5. Notice that some frequencies have been enhanced from the set of harmonics of the original time series. As in the experimental record, the initial values present a simple oscillation, which becomes more complex as the time evolves. The reason is that the time self-fluctuation at the glottis arose as a fixed point lost its stability in a Hopf bifurcation. Therefore, its spectral content is dynamically enriched as the pressure increases.

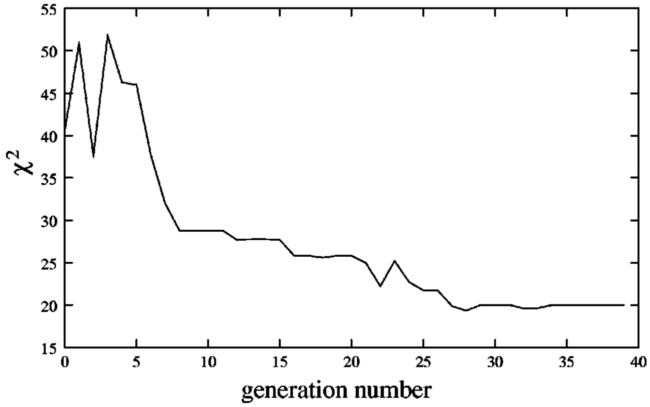

FIG. 8. The evolution of χ^2 for a segment of 150 points of the experimental record displayed in Fig. 4 for the best chromosome of each generation, as a function of the generation number.

V. FITTING OF THE PARAMETERS

At this point we have the basic building blocks needed to reproduce a realistic voiced sound. A mass subjected to restitution elastic forces, dissipation, and pressure begins to display self-oscillations as kinetic energy of the air is transferred to mechanical energy of the mass. The oscillation is created in a Hopf bifurcation, and as the control parameters are moved beyond the bifurcation the attracting limit cycle enriches its spectrum due to the influence of the coexisting invariant sets described above. The input pressure then displays several supraharmonics, which are filtered by the vocal tract as discussed in the previous section. In order to test the model, we have devised a genetic algorithm which allows us to find the appropriate parameters needed to reproduce a measured signal right at the mouth [15].

A genetic algorithm is a fitting procedure vaguely inspired by natural selection. Given a set of parameters for a model, its success is measured according to how similar the simulation with those parameters is to the experimental record. The algorithm consists of a number of iterations (*generations*). In each one, the simulation of the model is performed a large number of times (*population number*), each one for a set (*chromosome*) of parameters (*genes*). The chromosomes are ordered according to their success (the most successful chromosome being the set of parameters that better fits the time series). For the next generation, the better chromosomes are more likely to be chosen again, some chromosomes are discarded, and new chromosomes are generated by a set of operations over the better chromosomes of the previous generation. These operations include ‘‘crossing over’’ of some parameters between successful chromosomes and random mutation of a given parameter from a successful chromosome.

In Fig. 8, we display the evolution of the best chromosome of each generation, measured by the χ^2 of a segment of 150 points of the experimental record displayed in Fig. 4 and the simulation, as a function of the generation number. The simulation of the model with the parameters of the best chromosome at the tenth generation is displayed in Fig. 9. In Fig. 10 we display the results of applying this fitting procedure to a different record (this time corresponding to an ‘‘e’’ vowel, pronounced in Spanish).

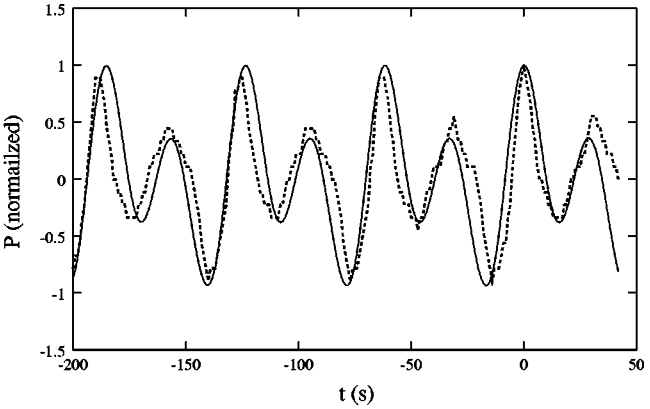

FIG. 9. The simulation of the model with the parameters of the best chromosome at the tenth generation. These are $A_1=1.5$ arb. units, $A_2=1.5$ arb. units, $A_3=4.5$ arb. units, $L_1=7.0$ cm, $L_2=6.4$ cm, $L_3=4.1$ cm.

In both cases the equations have been rewritten rescaling the time. Notice that if $t\rightarrow\alpha t$ the equations ruling the dynamics of (x,x',z) are unchanged, provided that $m\rightarrow\alpha^2 m$, $b\rightarrow\alpha b$, and $I\rightarrow\alpha I$. In this way, we can easily generate different time series for the pressure P_i with the same spectral content, but different fundamental frequency. In our simulations, this parameter was part of the chromosome, as well as the lengths and areas of the vocal tract. We have fitted normalized pressures, but the scaling of the equations allows us to fit the amplitude of P_i (and therefore of the pressure at the mouth). This can be done by scaling z in such a way that the equations remain unchanged: $z\rightarrow\gamma z$, together with $L\rightarrow L/\gamma$, $P_L\rightarrow P_L/\gamma^2$, $b\rightarrow b/\gamma^2$, $m\rightarrow m\gamma^2$, and $k\rightarrow k/\gamma^2$.

VI. CONCLUSIONS

In this work, we have analyzed in detail a simple model for vocal fold oscillation. The dynamical identification of the bifurcations taking place allows us to understand the time

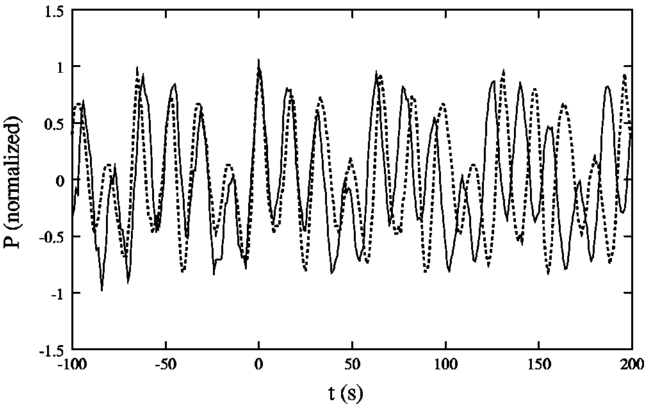

FIG. 10. The results of applying this fitting procedure to a different record (this time corresponding to an ‘‘e’’ vowel, pronounced in Spanish.

evolution of the spectral content of voiced sounds as they are pronounced. The description of the invariant manifolds coexisting with the bifurcating limit cycle describing the self-oscillation of the folds allows us to understand its spectral evolution, as the parameters are changed. Even if this model is an oversimplification of the rich dynamics that the vocal folds can display, we have shown that an appropriate set of tubes modeling the vocal tract allowed us to adequately fit experimental observations.

As the parameters are changed in the simple model studied here, self-oscillations are established which dynamically enrich their spectral properties. These changes are simultaneous with variations in their fundamental frequencies. We saw that the effect of the filters is that the observed time series can present complex features. Since realistic models of the folds can indeed present complex features at the level of the vocal fold oscillations, and even chaos [16,6], it is important to know what to expect from the simplest models.

Finally, we claim that a fitting procedure of physical parameters (of eventually richer models) can constitute an additional approach to the problem of speaker verification. This is an important issue, since the current paradigm for both speaker identification and verification is based on the analysis of the statistical properties of the recorded utterance, through LPC analysis, the computation of Cepstrum coefficients, pattern recognition applied to the Gabor transformation of the signal, or other techniques of spectral nature [17]. On the other hand, an analysis based on the reconstruction within a model of the parameters that are necessary to reproduce an utterance is able to distinguish the ergonomic features of a speaker (such as the typical lengths of his/her vocal tract) from circumstantial parameters (such as the lung pressure used).

This work was partially funded by UBA, CONICET, and Fundacion Antorchas.

[1] D. Sciamarella and G. B. Mindlin, Phys. Rev. Lett. **82**, 1450 (1999).

[2] I. R. Titze, *Principles of Voice Production* (Prentice-Hall, Englewood Cliffs, NJ, 1993).

[3] I. R. Titze and B. H. Story, J. Acoust. Soc. Am. **101**, 2234 (1997).

[4] I. R. Titze, J. Acoust. Soc. Am. **83**, 1536 (1988).

[5] K. Ishizaka and J. L. Flanagan, Bell Syst. Tech. J. **51**, 1233 (1972).

[6] M. S. Fee, Boris Shraiman, B. Peseran, and P. P. Mitra, Nature (London) **395**, 67 (1998).

[7] B. Story and I. Titze, J. Acoust. Soc. Am. **97**, 1249 (1995).

[8] M. Hirano, Folia Phoniatr. **26**, 89 (1974).

[9] I. Titze, Phonetica **29**, 1 (1974).

[10] D. Wong, M. R. Ito, N. B. Cox, and I. Titze, J. Acoust. Soc. Am. **89**, 383 (1991).

[11] K. Ishizaka and M Matsudaira, Speech Communication Research Laboratory, Santa Barbara, CA, Monograph 8, 1972.

[12] M. Rothenberg, in *Vocal Fold Physiology*, edited by K. Stevens and M. Hirano (University of Tokyo Press, Tokyo, 1981), pp. 304–323.

[13] H. Solari, M. Natiello, and G. B. Mindlin, *Nonlinear Dynamics: A Two Way Trip From Physics to Math* (IOP, London, 1996).

[14] The data were recorded with a PC running the software ‘‘SOUND STUDIO’’ for Linux, written by Paul Sharp, University of Leeds, 1998.

[15] M. Mitchell, *An Introduction to Genetic Algorithms* (MIT Press, Cambridge, 1993).

[16] *Vocal Fold Physiology: Controlling Complexity and Chaos*, edited by P. J. Davis and N. H. Fletcher (Singular Publishing Group, San Diego, 1996).

[17] L. Rabiner, B. Juang, and B. H. Juang, *Fundamentals of Speech Recognition* (Prentice-Hall, Englewood Cliffs, NJ 1993).