EFFICIENT NUMERICAL MODELING OF VOCAL FOLD MECHANICS

PACS number: 43.70.Bk

Federico Avanzini
Università di Padova, Dip. di Elettronica ed Informatica.
Via Gradenigo 6/A. 35131 – Padova, Italy
Tel: +39.049.827.7665 Fax: +39.049.827.7699
<avanzini@dei.unipd.it> www.dei.unipd.it/~avanzini

ABSTRACT

The literature of glottal models is reviewed and the properties of one-mass models, as opposed to multi-mass models, are discussed. A one-mass model is then presented, in which the non-linear interaction between fold displacement and airflow is described through a modified equation that accounts for vertical phase differences between upper and lower fold edges. It is shown that the system behaves qualitatively as higher-dimensional models (e.g., the two-mass Ishizaka–Flanagan model). Numerical techniques for the simulation of the model are discussed, and a method is presented which allows accurate and efficient computation of the non-linear terms.

INTRODUCTION

Features of the glottal source signal (i.e. the glottal flow) carry most of the information that characterizes voice quality and speaker identity [1, 6], and accordingly research on source models is becoming increasingly important in speech synthesis research. Parametric models fit the glottal signal with piecewise analytical functions, and typically use a small number of parameters. As an example, the Liljencrants-Fant model [8] characterizes one cycle of the flow derivative using as few as four parameters. Physical models describe the glottal system in terms of physiological quantities. The Ishizaka-Flanagan (IF) model [9] is a known example of lumped model of the vocal folds. Physical models capture the basic non-linear mechanisms that initiate self-sustained oscillations in the glottal system, and can simulate features (e.g. interaction with the vocal tract) that are not taken into account by parametric models. However they typically involve a large number of control parameters, and are more computationally expensive than parametric models.

This paper presents results about efficient yet accurate numerical modeling of the glottal system. First, the lumped modeling approach and specifically the IF model is reviewed. Then a simplified "one-delayed-mass" model, originally presented in [2], is described. Finally, a numerical scheme is developed for efficient simulation of the model.

LUMPED MODELING

The IF model

In the lumped modeling approach, a complex mechanical system is described by means of basic elements such as springs, masses, and damping elements. The Ishizaka-Flanagan (IF) model addressed in this section describes each vocal fold using a two-mass approximation. It is assumed that the folds are bilaterally symmetric, so that only one needs to be modeled. As a consequence, the model is constructed using two masses m_1 and m_2, as in Fig. 1. The masses are permitted only lateral motion (see the coordinates x_1 and x_2 in Fig. 1(b)). Along this direction, the masses are assumed to behave as second-order mechanical oscillators, i.e. they are subject to elastic and dissipative forces. For the accurate simulation of the elastic properties of the fold, the springs

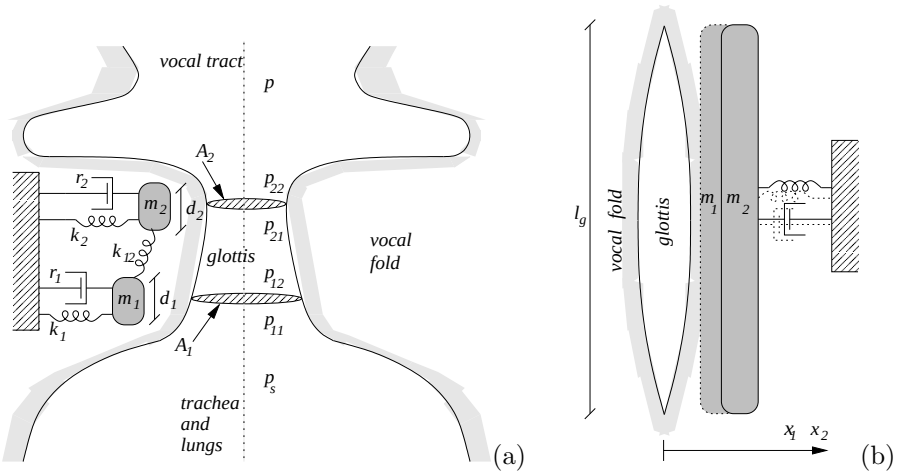

Figure 1: Schematic representation of the Ishizaka-Flanagan model: (a) coronal view and (b) superior view.

are non-linear and the coefficients $k_1(x_1)$ and $k_2(x_2)$ are modeled as quadratic functions of the corresponding displacements. The masses are coupled through a third spring k_{12}. The viscous forces are modeled as linear damping terms, with weights r_1 and r_2, respectively.

Collisions between the folds are accounted for by adding an additional restoring contact force to the equations, which is represented by an equivalent non-linear spring. In other words, when one of the masses m_k collides (i.e., when the condition $x_k < 0$ holds for $k = 1 \dots 2$), its stiffness $k_k(x_k)$ is increased. Summarizing, the equations for the mechanical system are given by

$$\begin{cases} m_1\ddot{x}_1(t) + r_1\dot{x}_1(t) + k_1(x_1)[x_1(t) - x_{01}] + k_{12}[x_1(t) - x_2(t)] = l_g d_1 p_{m1}(t), \\ m_2\ddot{x}_2(t) + r_2\dot{x}_2(t) + k_2(x_2)[x_2(t) - x_{02}] - k_{12}[x_1(t) - x_2(t)] = l_g d_2 p_{m2}(t), \end{cases} \tag{1}$$

where p_{m1} and p_{m_2} are the mean pressures under m_1 and m_2, respectively, while $l_g d_1$ and $l_g d_2$ are the driving surfaces on which the two pressures act; x_{01} and x_{02} represent the rest positions.

The interaction with the glottal pressure distribution is derived under the assumption of quasi-steady glottal flow $u(t)$. The pressure distribution inside the glottis is approximated as successive discrete steps p_{ij} at each end j of each mass i (see Fig. 1(a)). The pressure drops along the glottis are given by the following equations:

$$\begin{cases} p_s - p_{11}(t) = 0.69\rho_{air}\frac{u(t)^2}{A_1(t)^2}, \\ p_{11}(t) - p_{12}(t) = 12\nu d_1 \frac{l_g^2 u(t)}{A_1(t)^3}, \\ p_{12}(t) - p_{21}(t) = \frac{1}{2}\rho_{air} u(t)^2 \left(\frac{1}{A_2(t)^2} - \frac{1}{A_1(t)^2}\right), \\ p_{21}(t) - p_{22}(t) = 12\nu d_2 \frac{l_g^2 u(t)}{A_2(t)^3}, \\ p_{22}(t) - p(t) = \frac{1}{2}\rho_{air}\frac{u(t)^2}{A_2(t)^2}\left[2\frac{A_2(t)}{S}\left(1 - \frac{A_2(t)}{S}\right)\right], \end{cases} \tag{2}$$

where ν and S are the air shear viscosity and the vocal tract input area, respectively. The authors also discuss the inclusion of air inertance in the equations, when time-varying conditions are considered. Given the pressure drops in Eq. (2), the driving pressures p_{m1} and p_{m2} acting on $m_{1,2}$ must be derived. In the IF model, these are defined as the mean pressures along each mass:

$$p_{m1}(t) = \frac{1}{2}[p_{11}(t) + p_{12}(t)], \qquad p_{m2}(t) = \frac{1}{2}[p_{21}(t) + p_{22}(t)]. \tag{3}$$

In conclusion, the IF model is completely described by Eqs. (1) and (3).

Properties of lumped models

The IF model can take into account features that are not reproduced by a parametric model; in particular, acoustic interaction with the vocal tract can be accounted for. Many refinements have been proposed to IF, in which a larger number of masses is used, or the description of the airflow though the glottis is modified. (e.g., the three-mass model by Story and Titze [12]). On the other

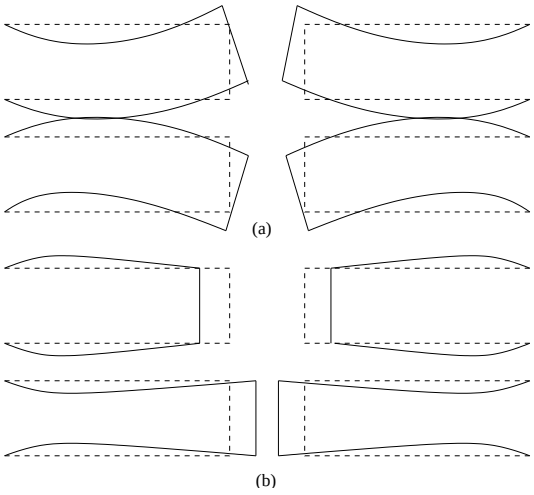

Figure 2: First two excited modes in a distributed model of the vocal folds.

hand, simpler one-mass models are used by many authors in articulatory speech synthesizers (see e.g. [10]), due to their reduced computational loads and better controllability. However, these models are not able to account for phase differences in the vocal fold motion, and consequently to generate a driving force which is asymmetrical within the cycle. According to some authors [11], this is the reason why one-mass models can not exhibit autonomous oscillations.

The IF model has two eigenmodes which are conceptually equivalent to those found by Berry and Titze [4] using a distributed three-dimensional numerical model of the vocal folds (see Fig. 2). The IF mode where the two masses move π-out of phase corresponds approximately to the one in Fig. 2(a), while the mode with the two masses in phase corresponds to that of Fig. 2(b). Berry and Titze suggest that the success of the IF model in describing the glottal behavior might be attributed to its ability to capture these two eigenmodes, and therefore facilitate self-oscillation.

A serious objection to all the models mentioned above has been raised by Villain et al. [13]. The authors remark that elementary mechanical constraints on the physiological problem are neglected in these models. Both the lumped and the distributed approach assume that the elastic structure is fixed to a rigid wall, which is a crude approximation since in reality a significant radiation of surface waves from the throat can be noticed when voiced sounds are produced. The effect of this radiation may be significant in terms of energy loss in the system. Villain et al. used an experimental setup where the folds are modeled by thin latex tubes filled with water, and showed that the behavior of the valve is strongly affected by the mechanical constraints. The authors therefore claim that modal analysis on glottal models such as the one developed in [4] is questionable.

One shortcoming of lumped models is that the glottal area (e.g. A_1 and A_2 in IF) is assumed to be rectangular. As a consequence, closure of the glottis occurs in an abrupt manner and the flow signals obtained from the model exhibit a sharp corner at the beginning of the closed phase. This affects the spectral tilt of the glottal source, introducing additional energy at high frequencies. In natural flow signals, a smoother glottal closure is usually observed. Stroboscopic measurements often show zipper-like movements of the glottal area during the closing phase. Lumped models do not take into account these phenomena.

When used for speech synthesis purposes, multi-mass models suffer from over-parametrization: as an example, as many as 19 parameters have to be estimated in the IF model. Proposed refinements to IF (see e.g. Story and Titze [12]) involve an even larger number of parameters and are hardly controllable and more computationally expensive.

One-delayed-mass model

Equation (2) shows that the positions x_1 and x_2 of both masses are needed in order to compute the pressure drops p_{ij} in the IF model. The "one-delayed-mass model", originally presented in [2], avoids the use of a second mass by exploiting additional information on the system.

- As already remarked, the two eigenmodes of IF correspond roughly to the first two excited modes of a distributed model [4] (see Fig. 2). Berry and Titze found that the two eigenfrequencies are very closely spaced. As a consequence, $1:1$ mode locking occurs during self-oscillation.

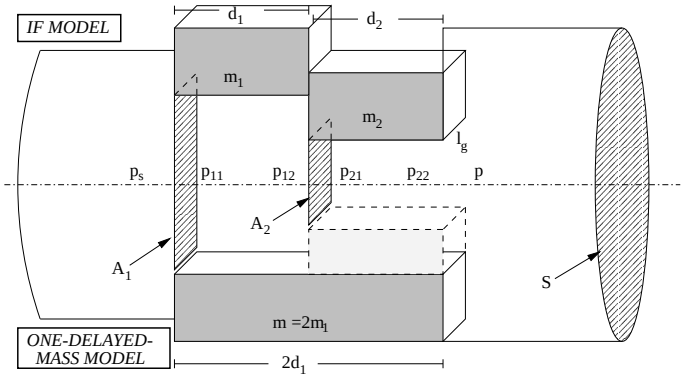

Figure 3: Scheme of the one-delayed-mass model (lower half) as opposed to the IF model (upper half).

• De Vries et al. [7] used a distributed model for estimating "correct" values for the IF parameters (i.e. the values such that IF behaves as closely as possible like the distributed model). The parameter values were found in [7] to be very much symmetrical: the ratio m_1/m_2 is close to one (while it was chosen to be clese to five in [9]), and the same holds for the other parameters.
From these results, the IF model can be consistently simplified using the following assumptions.
a1. The masses $m_{1,2}$ are taken to be equal, together with their parameters.
a2. The masses move with constant phase difference, because of mode locking:

$$A_2(t) = A_1(t - t_0), \tag{4}$$

where t_0 represents the delay (in seconds) between the motion of upper and lower fold edges. From Eq. (4) it follows that the pressure drops of Eqs. (2) depend only on $A_1(t)$ and $A_1(t - t_0)$: this suggests that only one degree of freedom is needed in the model. Therefore, the fold is described as a single mass m subject to elastic and viscous forces, with stiffness k and damping weight r_1. Similarly to IF, the driving pressure acting on the fold is defined as the mean pressure p_m at the glottis: $p_m = 1/4 \sum_{i,j=1}^{2} p_{ij}$. The fold displacement $x(t)$ is thus given by

$$m\ddot{x}(t) + r\dot{x}(t) + k(x(t) - x_0) = l_g d p_m(t), \tag{5}$$

where $l_g d$ is the driving surface and x_0 is the rest position. From assumption **a1**, the mass m is given by $m = 2m_1$. Explicit expressions for the driving pressure p_m and the pressure at vocal tract entrance p are derived from Eq. (2), and depend only on $(x(t), xt - t_0), u(t))$:

$$p_m(t) = p_m(x(t), x(t - t_0), u(t)), \qquad p(t) = p(x(t), x(t - t_0), u(t)). \tag{6}$$

Equations (5) and (6) describe the one-delayed-mass model. From Eq. (5) it is seen to be a one-mass model, but the dependence on the delayed displacement $x(t - t_0)$ in Eq. (6) results in a modified non-linear interaction, and the effects due to phase differences between the upper and lower margins of the folds are controlled by the delay t_0. A graphic representation of the model, as opposed to IF, is depicted in Fig. 3.

NUMERICAL SIMULATIONS

Discrete-time equations

The IF model was implemented numerically in [9] using the backward Euler method. Coupling between the non-linear equations was avoided by inserting fictitious delays in the equations. This section develops a more accurate numerical scheme. The linear differential Eq. (5) is discretized using the bilinear transform, and the resulting numerical system can be schematically written as

$$\left\{ \begin{array}{rcl} \boldsymbol{w}(n) & = & \tilde{\boldsymbol{w}}(n) + \bar{\boldsymbol{C}}\boldsymbol{p}(n), \\ \boldsymbol{x}(n) & = & \tilde{\boldsymbol{x}}(n) + \boldsymbol{K}\boldsymbol{p}(n), \\ \boldsymbol{p}(n) & = & \boldsymbol{f}_{n_0}(\boldsymbol{x}(n)) = \boldsymbol{f}_{n_0}\left(\tilde{\boldsymbol{x}}(n) + \boldsymbol{K}\boldsymbol{p}(n)\right), \end{array} \right. \tag{7}$$

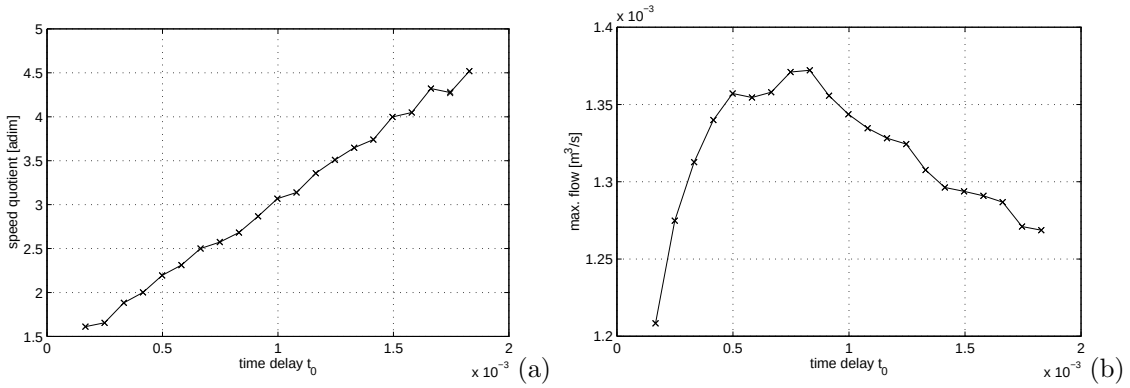

Figure 4: Dependence of (a) the speed quotient SQ and (b) the maximum amplitude on t_0 for the airflow ($F_s = 22.05$ [kHz], $p_s = 1000$ [Pa]).

where the variables are given by

$$\boldsymbol{w}(n) = \begin{bmatrix} x(n) \\ \dot{x}(n) \end{bmatrix}, \quad \boldsymbol{x}(n) = \begin{bmatrix} u(n) \\ x(n) \end{bmatrix}, \quad \boldsymbol{p}(n) = \begin{bmatrix} p_m(n) \\ p(n) \end{bmatrix},$$

and where non-linear mapping $\boldsymbol{f}_{n_0} : \mathbb{R}^2 \to \mathbb{R}^2$ is obtained from Eq. (6). The mapping is parametrized with the numerical delay n_0. Given a sampling rate F_s, n_0 is defined as $n_0 = t_0 F_s$. The vectors $\tilde{\boldsymbol{w}}(n)$ and $\tilde{\boldsymbol{x}}(n)$ have no instantaneous dependence on the pressures $\boldsymbol{p}(n)$, and are therefore computable at each step from known quantities. System (7) shows that a delay-free path is generated, such that the vocal fold state vector $\boldsymbol{w}(n)$ and the pressure vector $\boldsymbol{p}(n)$ have mutual instantaneous dependence. Due to the non-linear term $\boldsymbol{f}_{n_0}$, the difference equations can not be solved analytically. Instead, the K method [5] is used: this method operates a geometric transformation on the non-linearity, in such a way that the delay-free path can be computed without introducing fictitious delays in the discrete-time equations. By exploiting the implicit function theorem, the mapping $\boldsymbol{f}_{n_0}$ in system (7) is turned into a new-one:

$$\boldsymbol{p}(n) = \boldsymbol{f}_{n_0}\left(\tilde{\boldsymbol{x}}(n) + \boldsymbol{K}\boldsymbol{p}(n)\right) \quad \overset{\text{K method}}{\longmapsto} \quad \boldsymbol{p}(n) = \boldsymbol{h}_{n_0}\left(\tilde{\boldsymbol{x}}(n)\right). \tag{8}$$

At each time step, the pressure vector $\boldsymbol{p}$ is found as a function of the known vector $\tilde{\boldsymbol{x}}$. The mapping $\boldsymbol{h}_{n_0}(\tilde{\boldsymbol{x}})$ can be computed off-line and stored in a look-up table: in this case the implementation requires only linear operations and one look-up at each time step. However, when the control parameters vary over time $\boldsymbol{h}_{n_0}(\tilde{\boldsymbol{x}})$ needs to be recomputed on-line. In these case, a more efficient implementation strategy amounts to computing $\boldsymbol{h}(\tilde{\boldsymbol{x}})$ iteratively at the sampling rate, in two steps: **(1)** the current value of $\tilde{\boldsymbol{x}}$ is computed, and **(2)** the current value of $\boldsymbol{p}$ is found by imposing $\boldsymbol{f}_{n_0}(\tilde{\boldsymbol{x}} + \boldsymbol{K}\boldsymbol{p}) - \boldsymbol{p} = 0$. The Newton-Raphson method can be used to iteratively solve this implicit non-linear equation. Using the past value $\boldsymbol{y}(n-1)$ as the starting point in the Newton-Raphson iterations provides a fast convergence of the algorithm to the new value. This approach has been recently adopted for developing numerical simulations of single reed wind instruments [3].

Properties of the model

The numerical implementation described above was used to study the properties of the one-delayed-mass model. Specifically, the influence of the delay t_0 on the signal parameters (such as pitch, open quotient, speed quotient, maximum amplitude), was investigated through automatic analysis on the numerical simulations. The results given in the following were obtained by analyzing 0.3 [s] long flow signals, where the values of t_0 range from 0.1 to 1.9 [ms].

The speed quotient provides a quantitative measure of the flow skewness: it is defined as the ratio between the opening phase ($\dot{u}(t) > 0$) and the closing phase ($\dot{u}(t) < 0$). The speed quotient is known to have perceptual relevance in characterizing different voice qualities: for instance, analysis on real signals by Childers and Ahn [6] show that the speed quotient ranges from about 1.6 to 3 when the voice quality changes from breathy voice to vocal fry and finally to modal voice. Figure 4(a) shows the dependence of the speed quotient on t_0: it is seen that, in the range

under consideration, the speed quotient is approximately a linear function of t_0. By appropriately choosing t_0, one can range from very low up to extremely high values of the speed quotient.

Figure 4(b) shows the dependence of the maximum flow amplitude on t_0. This exhibits a peak around $t_0 = 8 \cdot 10^{-4}$ [s]. This suggests the existence of an optimum delay t_0 that maximizes the aerodynamic input power (defined as mean subglottal pressure times mean glottal flow), which is in turn related to the glottal efficiency, usually defined as the ratio of radiated acoustic power to aerodynamic power (i.e., the power delivered to the vocal system by the lungs). Further analysis is needed in order to assess the precise influence of t_0 on the glottal efficiency.

CONCLUSIONS

This paper has focused on efficiency issues in numerical simulations of vocal fold models. First, the one-delayed-mass model has been described. On the one hand, only one degree of freedom is needed, instead of two [9] or more [12] usually assumed in higher-dimensional lumped models of the vocal folds. On the other hand, the dependence on t_0 in Eq. (6) results in realistic glottal flow waveforms, that are not obtained with usual one-mass models [10].

Second, a discretization scheme has been described, which allows efficient and accurate numerical simulations of the one-delayed-mass model. Results from numerical experiments show that t_0 provides control on the airflow skewness. The numerical model is therefore a reasonable trade-off between accuracy of the description and simplicity of the structure.

Further studies will investigate the interaction of the glottal model with vocal tract loads, in order to discuss applications of the proposed glottal model in articulatory speech synthesis. Preliminary results, obtained by coupling the one-delayed-mass model with a digital waveguide bore model, show the occurrence of ripples in the airflow signal, mainly due to interaction with the first resonance of the tract. Moreover, automatic analysis reveals a slight dependence of pitch on the vocal tract characteristics.

REFERENCES

[1] P. Alku and E. Vilkman. A Comparison of Glottal Voice Quantification Parameters in Breathy, Normal and Pressed Phonation of Female and Male Speakers. Folia Phoniatrica et Logopaedica, 48:240–254, 1996.

[2] F. Avanzini, P. Alku, and M. Karjalainen. One-delayed-mass Model for Efficient Synthesis of Glottal Flow. In Proc. Eurospeech Conf., pages 51–54, Aalborg, 2001.

[3] F. Avanzini and D. Rocchesso. Efficiency, Accuracy, and Stability Issues in Discrete Time Simulations of Single Reed Wind Instruments. J. Acoust. Soc. Am., 111(5), 2002.

[4] D. A. Berry and I. R. Titze. Normal Modes in a Continuum Model of Vocal Fold Tissues. J. Acoust. Soc. Am., 100(5):3345–3354, 1996.

[5] G. Borin, G. De Poli, and D. Rocchesso. Elimination of Delay-free Loops in Discrete-Time Models of Nonlinear Acoustic Systems. IEEE Trans. Speech Audio Process., 8(5):597–606, 2000.

[6] D. G. Childers and C. Ahn. Modeling the Glottal Volume-Velocity Waveform for Three Voice Types. J. Acoust. Soc. Am., 97(1):505–519, 1995.

[7] M. P. de Vries, H. K. Schutte, and G. J. Verkerke. Determination of Parameters for Lumped Parameter Model of the Vocal Fold Using a Finite-Element Method Approach. J. Acoust. Soc. Am., 106(6):3620–3628, 1999.

[8] G. Fant, J. Liljencrants, and Q. Lin. A Four-Parameter Model of Glottal Flow. In Speech Transmiss. Lab. Q. Prog. Stat. Rep., pages 1–13, 1985.

[9] K. Ishizaka and J. L. Flanagan. Synthesis of Voiced Sounds from a Two-Mass Model of the Vocal Cords. Bell Syst. Tech. J., 51:1233–1268, 1972.

[10] P. Meyer, R. Wilhelms, and H. W. Strube. A Quasiarticulatory Speech Synthesizer for German Language Running in Real Time. J. Acoust. Soc. Am., 86(2):523–539, 1989.

[11] X. Rodet. One and two mass model oscillations for voice and instruments. In Proc. Int. Computer Music Conf. (ICMC'95), Banff, 1995.

[12] B. H. Story and I. R. Titze. Voice Simulation with a Body-Cover Model of the Vocal Folds. J. Acoust. Soc. Am., 97(2):1249–1260, 1995.

[13] C. Villain, L. Le Marrec, W. Op't Root, J. Willems, X. Pelorson, and A. Hirschberg. Towards a New Brass Player's Lip Model. In Proc. Int. Symp. Mus. Acoust. (ISMA'01), pages 107–110, Perugia, 2001.