One-delayed-mass model for efficient synthesis of glottal flow

Federico Avanzini[1,2]*, Paavo Alku*[2]*, Matti Karjalainen*[2]

[1]Dipartimento di Elettronica ed Informatica, Università di Padova, Italy
[2]Lab. of Acoustics and Audio Signal Processing, Helsinki University of Technology, Finland
avanzini@dei.unipd.it {paavo.alku;matti.karjalainen}@hut.fi
http://www.dei.unipd.it http://www.acoustics.hut.fi

Abstract

A lumped physical model of the glottal source is presented. Vocal folds are described as single masses, but vertical phase differences between upper and lower margins of the folds are taken into account by appropriately describing the non-linear interaction of the mechanical model with aerodynamics. This results in a modified one-mass model, or a "one-delayed-mass model". Analysis on numerical simulations shows that the system behaves qualitatively as higher-dimensional models (such as the two-mass model by Ishizaka and Flanagan); in particular, control over flow skewness is guaranteed, allowing for synthesis of realistic glottal flow waveforms. As only one degree of freedom (one mass) is needed in the model, structure and number of parameters are drastically reduced, thus making it suitable for real-time synthesis applications.

1. Introduction

Glottal source modeling is recognized to be a key feature for improving naturalness in speech synthesis, and for characterizing different voice qualities (e.g., modal, pressed and breathy phonation [1]). Among glottal models, both parametric and physical ones have been developed. One of the most widely used parametric models is the Liljencrants and Fant *(LF)* model: this characterizes one cycle of the derivative of the glottal flow by using as few as four parameters. It has been proved to be very flexible, and able to reproduce a variety of voice qualities [2]. Among physical models, the first and most widely known one was developed by Ishizaka and Flanagan (*IF*) in [3]; this describes one vocal fold as two lumped mechanical oscillators (two masses with springs and dampings plus a spring for coupling the two oscillators).

A physical model, such as IF, can take into account subtle features that are not reproduced by a parametric model; in particular, interaction with the vocal tract is considered, thus allowing to develop a full articulatory model [4, 5]. This interaction gives rise to several "natural" effects; among them are such phenomena as occurrence of oscillatory ripples on the glottal flow waveform, as well as a slight dependence of the pitch and the *open quotient* on the load characteristics. On the other hand, the IF model suffers from an over-parametrization: as many as 19 parameters have to be estimated in order to account for non-linear corrections in the elastic forces, for collisions between the two folds, and other features. This results in high computational loads and problems in tuning the parameters. Proposed refinements to the IF model involve an even larger number of parameters: an example is the three-mass model by Story and Titze [6]. Such models account for a very accurate description of the glottal system, but are hardly controllable and computationally expensive. On the other hand, simpler one-mass models suffer from insufficient description of the system; in particular they are not able to account for phase differences in the vocal fold motion, thus resulting in a wrong coupling with aerodynamics. Nonetheless many authors (see for instance [7]) prefer to use a one-mass model despite its poor accuracy, because of their reduced computational loads and better controllability.

Table 1: *Constants and parameters. The * indicates that the parameter is varied in simulations*

quantity	symbol	value	unit
Air density	ρ	1.14	[Kg/m^3]
Air shear viscosity	ν	$1.85 \cdot 10^{-5}$	[N$\cdot$ s/m^2]
Fold length	l_g	$1.3 \cdot 10^{-2}$	[m]
Fold thickness	$2d_1$	$3 \cdot 10^{-3}$	[m]
Fold mass	m_g	$4.4 \cdot 10^{-5}$	[Kg]
Fold spring constant	k_g	20	[N/m]
Fold viscous resist.	r_g	$0.1 \cdot \sqrt{m_g k_g}$	[N$\cdot$ s/m]
Fold equilib. area *	A_0	$5 \cdot 10^{-6}$	[m^2]
Voc. tract input area	A_1	$5 \cdot 10^{-4}$	[m^2]
Mass delay*	t_0	$8.5 \cdot 10^{-4}$	[s]
Sampling rate*	F_s	22.05	[kHz]

In this paper we develop an improved one-mass model, where interaction with aerodynamics is modified: the effect of a second mass on the aerodynamics equations is taken into account by introducing a delay t_0 in the mass position and describing the glottal airflow as a function of this "delayed mass". Results from simulations show that the model behaves qualitatively as IF, using only one degree of freedom (one mass) instead of two; as a consequence the structure is drastically simplified and computational costs are reduced. Moreover, less than half of the IF parameters are needed; among them, the delay t_0 gives control on the airflow skewness, which is known to be a perceptually relevant feature [2]. Having a small set of meaningful control parameters, the proposed physical model can be "competitive" with parametric ones, such as LF.

Sec. 2 describes details of the model; Sec. 3 illustrates its main properties and results from simulations; in Sec. 4 these are discussed and compared with other models. Table 1 lists symbols and values for constants and parameters used throughout the paper.

2. The model

The IF model is sketched in Fig. 1 (upper half). Ishizaka and Flanagan describe pressure drops p_{ij} along the vocal folds as

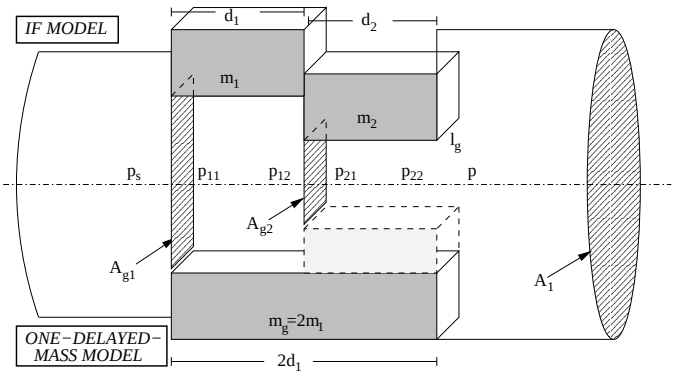

Figure 1: *Scheme of the IF model (upper half) as opposed to the one-delayed-mass model (lower half).* p_{ij} *are pressures at upper and lower margins of masses,* p *is pressure at the entrance of vocal tract.*

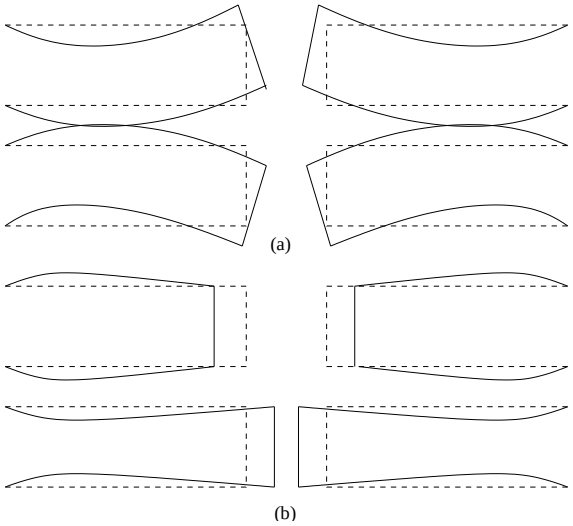

Figure 2: *First two excited modes in a distributed model of the vocal folds.*

follows [3]:

$$
\begin{aligned}
p_s - p_{11} &= 0.69\rho\frac{u_g^2}{A_{g1}^2},\\
p_{11} - p_{12} &= 12\nu d_1\frac{l_g^2 u_g}{A_{g1}^3},\\
p_{12} - p_{21} &= \frac{1}{2}\rho u_g^2\left(\frac{1}{A_{g2}^2} - \frac{1}{A_{g1}^2}\right),\\
p_{21} - p_{22} &= 12\nu d_2\frac{l_g^2 u_g}{A_{g2}^3},\\
p_{22} - p &= \frac{1}{2}\rho\frac{u_g^2}{A_{g2}^2}\left[2\frac{A_{g2}}{A_1}\left(1 - \frac{A_{g2}}{A_1}\right)\right]
\end{aligned}
\qquad (1)
$$

where pressures and areas are as depicted in Fig. 1; p_s in the lung (subglottal) pressure and u_g is the glottal airflow. Therefore positions of both masses $m_{1,2}$ are needed in order to compute pressure drops along the glottis and the resulting airflow.

2.1. Discussing the system

The "one-delayed-mass model" presented here avoids the use of a second mass by exploiting additional information on the system.

• The IF model has two eigenmodes: the one with two masses in phase and the one with two masses π-out of phase. As pointed out by Berry, Titze *et al.* [8, 9], these modes correspond roughly to the first two excited modes found in a distributed model of the vocal folds (see Fig. 2). Berry and Titze remark that the two eigenfrequencies are very closely spaced; as a consequence, $1:1$ mode locking occurs during self-oscillation.

• In a recent paper de Vries *et al.* [10] make use of a similar distributed model for estimating "correct" values for the IF parameters: these are found by requiring the behavior of the IF model to resemble as close as possible that of the distributed model. Results show significant differences with the values originally stated by Ishizaka and Flanagan; in particular the parameter values for the two masses are found in [10] to be much more symmetrical: the ratio between m_1 and m_2 is close to one (while it is close to five in typical IF parameters), and the same holds for the spring constants, dampings and geometrical parameters (d_1 and d_2 in Fig. 1 are found to be the same).

Using this additional information we can consistently simplify the model.

2.2. Simplifying the model

From the discussion in the previous section, we derive the main assumptions of our model: first, $m_{1,2}$ are taken to be equal, together with their thicknesses $d_{1,2}$ and their spring constants and dampings. Moreover, the two masses are taken to move with constant phase difference, because of mode locking; this means that the area $A_{g2}(t)$ under the second mass follows the first on $A_{g1}(t)$ with a constant phase difference:

$$A_{g2}(t) = A_{g1}(t - t_0), \qquad (2)$$

t_0 being a given delay. Substituting expression (2) in Eq. (1) results in a set of pressure drops that are nonlinear functions of area A_{g1} and the same area delayed by t_0. In this way only one degree of freedom is needed in the model; in other words, we can treat the vocal fold as a single mass $m = 2m_1$, and describe phase differences between the upper and lower margins of the folds by means of delay t_0. If we assume the vocal fold to be driven by the mean pressure p_m at glottis ($p_m = 1/4\sum_{i,j=1}^{2} p_{ij}$), we can describe the fold motion as

$$m_g\ddot{A}_{g1} + r_g\dot{A}_{g1} + k_g(A_{g1} - A_0) = 2l_g^2 d_1 \cdot p_m \qquad (3)$$

The driving pressure p_m and pressure at vocal tract entrance p are derived from Eq. (1):

$$
\begin{cases}
p_m(t) = p_m\left(A_g(t), A_g(t - t_0), u_g(t)\right)\\
p(t) = p\left(A_g(t), A_g(t - t_0), u_g(t)\right).
\end{cases}
\qquad (4)
$$

One last equation relates the glottal flow to pressure p:

$$u_g(t) = z_{load}(t) * p(t), \qquad (5)$$

where the load impedance z_{load} can be, for instance, the input impedance of the vocal tract. Eqs. (3),(4),(5) form our one-delayed-mass model; this is outlined in Fig. 1 (lower half). From Eq. (3) it can be seen to be a one-mass model, but the dependence on $A_{g1}(t - t_0)$ in Eq. (4) results in a modified interaction with the aerodynamics. As shown in the following section, this allows to preserve the main features of a two-mass model using a single degree of freedom.

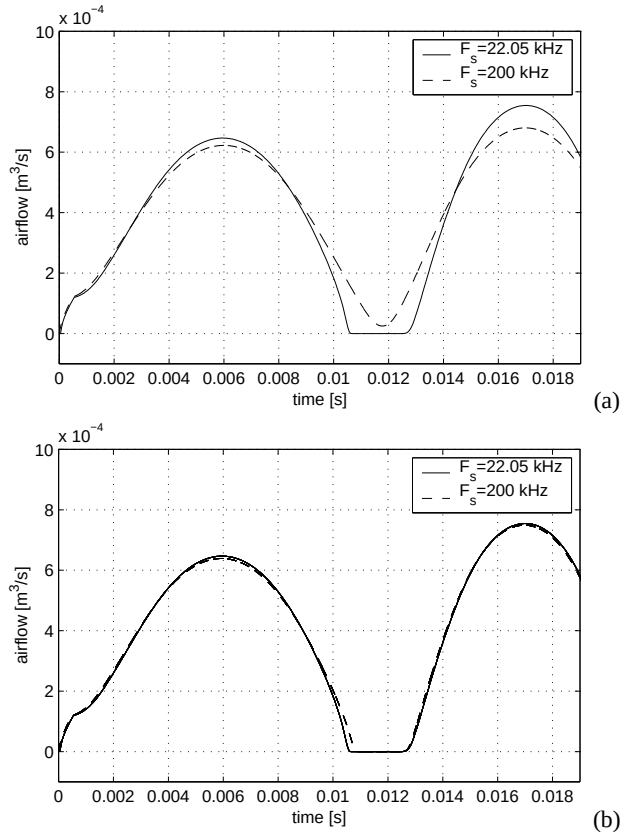

Figure 3: *Attack transients in the one-delayed-mass model. (a)* $t_0 = 1 \cdot 10^{-4}$ *[s],system at* $F_s = 22.05$ *kHz is unstable. (b)* $t_0 = 2 \cdot 10^{-4}$ *[s], system at* $F_s = 22.05$ *kHz is stable.*

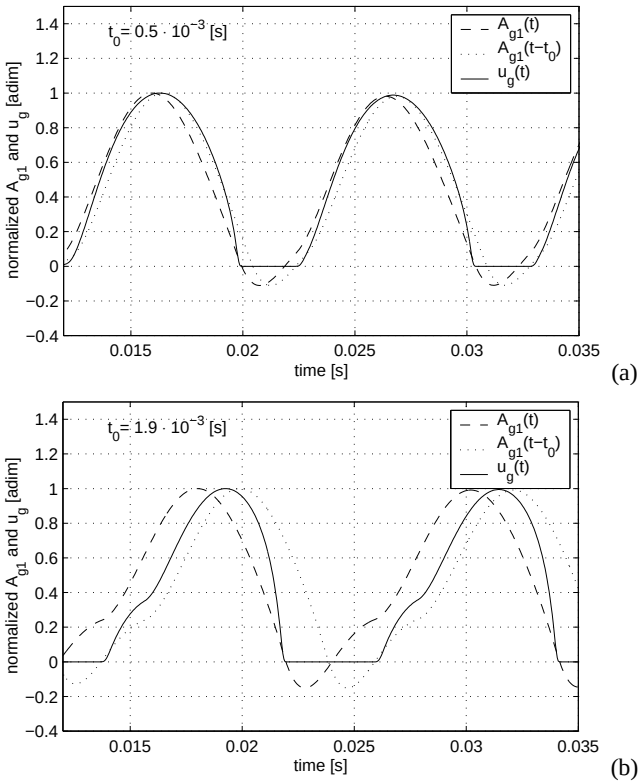

Figure 4: *Dependence of flow skewness on* t_0*. Simulations were run at* $F_s = 22.05$ *kHz.*

3. Results

A numerical implementation of the model described in Sec. 2.2 was developed by discretizing Eq. (3) with the bilinear transform; computational problems concerned with the non-linear block given in Eq. (4) were solved using the K method [11]; such a method has been found effective in dealing with non-linear aero-mechanical systems [12].

3.1. Stability and accuracy

The dependence on $A_{g1}(t-t_0)$ in Eq. (4) results in a delay loop in the system; this is a potential source of instability [13]. Due to the non-linear nature of the system, analytical conditions for stability are not easily found; stability properties of the system were therefore investigated experimentally, in two steps: first, simulations were run at a very high F_s ($=$ 200 kHz); these were taken as a reference for the behavior of the continuous-time system. In a second step, simulations were run at standard F_s ($= 11.025, 22.05$ kHz) and compared with the reference.

In Fig. 3 results for $F_s = 22.05$ kHz are plotted: these show that the numerical delay $n_0 = t_0 F_s$ affects stability. Indeed, with very small delays ($t_0 < 2 \cdot 10^{-4}$ [s], i.e. $n_0 < 4$ at $F_s = 22.05$ kHz) the system is unstable, as seen from Fig. 3(a): the first few cycles in the oscillation show the increasing error with respect to the reference; in the following cycles this trend continues until a steady state far from the reference is reached. Above the "stability threshold" $n_0 = 4$ the system appears to be stable, as seen from Fig. 3(b). Similar results are found for $F_s = 11.025$ kHz. However, we remark that realistic values for t_0 are sensibly higher than those used in Fig. 3. Therefore n_0 is typically well above the threshold, and stability is guaranteed.

3.2. Airflow features

The effect of the delay t_0 in shaping the glottal waveform was investigated by simulations. In Fig. 4 the areas $A_{g1}(t), A_{g1}(t-t_0)$ and the airflow $u_g(t)$ are plotted for two different t_0's (for the sake of clarity the signals are normalized). From this, it can be clearly seen that the airflow skewness is controlled by the delay t_0. A quantitative measure of flow skewness is given by the *speed quotient (SQ)*, defined as the ratio between the opening phase ($\dot{u}_g(t) \gtrsim 0$) and the closing phase ($\dot{u}_g(t) \lesssim 0$). This is recognized to have perceptual relevance in characterizing different voice qualities: for instance, analysis on real signals by Childers and Ahn [2] show that the *SQ* ranges from about 1.6 to 3 when voice quality changes from breathy voice to vocal fry and finally to modal voice.

Simulations with the one-delayed-mass system were run in order to investigate the dependence of *SQ* on t_0. Each simulation was 0.3 [s] long, and automatic analysis was developed for extracting significant parameters (such as pitch, open quotient, speed quotient, max. amplitude) from the flow signal. Fig. 5(a) shows the results for *SQ*: it turns out to be an almost linear function of t_0. By appropriately choosing t_0, one can range from very low up to extremely high -even unrealistic- *SQ* values. Fig. 5(b) shows another interesting feature of the system: the max. amplitude for u_g exhibits a peak around $t_0 = 8 \cdot 10^{-4}$ [s], thus suggesting the existence of an optimum delay t_0 for maximal aerodynamic input power (defined as mean subglottal pressure times mean glottal flow).

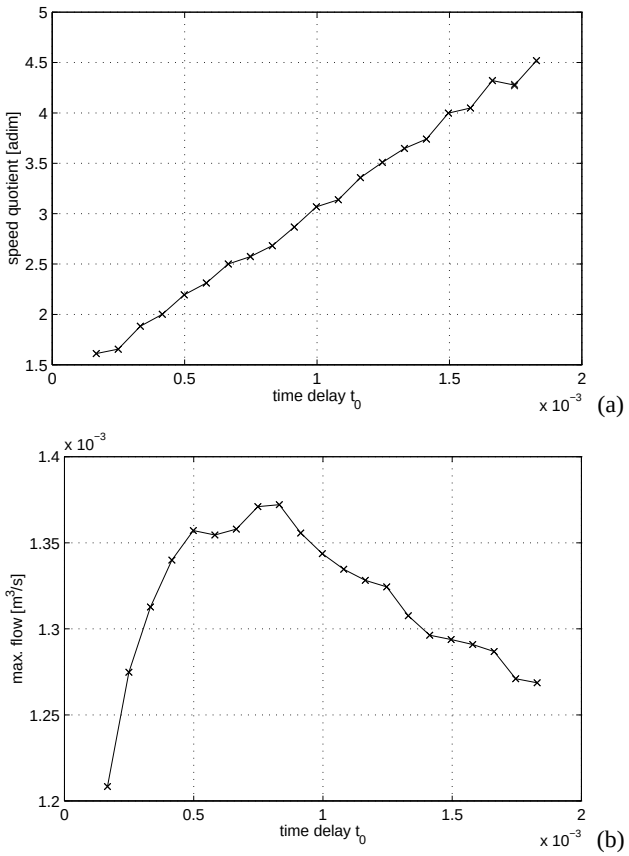

Figure 5: *Dependence of (a) speed quotient and (b) max. amplitude on t_0 for the airflow ($F_s = 22.05$ kHz, $p_s = 1000$ Pa).*

4. Discussion

The main advantages of the proposed model are its simple structure and its low number of control parameters. On the one hand, only one degree of freedom is needed, instead of two [3] or more [6] usually assumed in higher-dimensional lumped models of the vocal folds. On the other hand, the dependence on t_0 in Eq. (4) results in realistic glottal flow waveforms, that are not obtained with usual one-mass models [7]; in particular, from results given in Sec. 3 t_0 is shown to give control on the airflow skewness. The model is therefore a reasonable trade-off between accuracy of the description and simplicity of the structure: thanks to its low computational costs it can be suitable for real-time tasks.

Interaction of the model with vocal tract loads has not yet been investigated in detail. Preliminary results with a uniform tube model show the occurrence of ripples in the airflow signal, mainly due to interaction with the first formant. Moreover, automatic analysis reveals a slight dependence of pitch on vocal tract characteristics. Further efforts will be devoted to this issue, in order to discuss applications of the proposed glottal model in articulatory speech synthesis.

One drawback of the model (which is present also in IF and in general in lumped vocal fold models) concerns closure: as the glottal area A_{g1} is assumed to be rectangular, closure of the glottis occurs abruptly and results in a sharp corner in the airflow (or equivalently in a narrow negative peak in the airflow derivative). This affects the spectral tilt of the glottal source, introducing additional energy at high frequencies. In natural flow signals, closure usually occurs in a smoother manner due to, for example, a zipper-like movement of the glottal area during closing phase. Further studies will therefore concentrate on how to integrate such features into the model.

5. Acknowledgments

This research has been partially supported by the Academy of Finland ("Sound Source Modeling" project).

6. 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*, vol. 48, pp. 240–254, 1996.

[2] D. G. Childers and C. Ahn, "Modeling the Glottal Volume-Velocity Waveform for Three Voice Types," *J. Acoust. Soc. Am.*, vol. 97, no. 1, pp. 505–519, Jan. 1995.

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

[4] M. M. Sondhi and J. Schroeter, "A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer," *IEEE Trans. ASSP*, vol. 35, no. 7, pp. 955–967, July 1987.

[5] V. Välimäki and M. Karjalainen, "Improving the Kelly-Lochbaum Vocal Tract Model Using Conical Tube Sections and Fractional Delay Filtering Techniques," in *Proc. Int. Conf. Spoken Language Processing*, Sept. 1994, pp. 615–618.

[6] B. H. Story and I. R. Titze, "Voice Simulation with a Body-Cover Model of the Vocal Fods," *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1249–1260, Feb. 1995.

[7] P. Meyer, R. Wilhelms, and H. W. Strube, "A Quasiarticulatory Speech Synthesizer for German Language Running in Real Time," *J. Acoust. Soc. Am.*, vol. 86, no. 2, pp. 523–539, Aug. 1989.

[8] D. A. Berry, H. Herzel, I. R. Titze, and K. Krischer, "Interpretation of Biomechanical Simulations of Normal and Chaotic Vocal Fold Oscillations with Empirical Eigenfunctions," *J. Acoust. Soc. Am.*, vol. 95, no. 6, pp. 3595–3604, June 1994.

[9] D. A. Berry and I. R. Titze, "Normal modes in a continuum model of vocal fold tissues," *J. Acoust. Soc. Am.*, vol. 100, no. 5, pp. 3345–3354, Nov. 1996.

[10] 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.*, vol. 106, no. 6, pp. 3620–3628, Dec. 1999.

[11] G. Borin, G. De Poli, and D. Rocchesso, "Elimination of Delay-free Loops in Discrete-Time Models of Nonlinear Acoustic Systems," *IEEE Trans. on Speech and Audio Process.*, vol. 8, pp. 597–606, 2000.

[12] F. Avanzini and D. Rocchesso, "Efficiency, Accuracy, and Stability Issues in Discrete Time Simulations of Single Reed Wind Instruments," *submitted for publication*, 2000.

[13] R. J. Anderson and M. W. Spong, "Bilateral Control of Teleoperators with Time Delay," *IEEE Trans. on Automatic Control*, vol. 34, no. 5, pp. 494–501, May 1989.