ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 92 (2006) 731 – 740

Physiological Control of Low-Dimensional Glottal Models with Applications to Voice Source Parameter Matching

Federico Avanzini, Simone Maratea
Department of Information Engineering, University of Padova, Via G. Gradenigo 6/A, 35131 Padova, Italy

Carlo Drioli
Institute of Phonetics and Dialectology, ISTC-CNR, Via G. Anghinoni 10, 35121 Padova, Italy

Summary
This study explores the possibility for physical models of the glottis to accurately reproduce target glottal flow waveforms. A set of rules is proposed for controlling a two-mass physical model through activation levels of laryngeal muscles. The proposed rules convert muscle activities into physical quantities such as fold adduction, mass, thickness, depth, stiffness. Numerical simulations of the glottal model, controlled through this set of rules, are used in order to construct a codebook between muscular activations and a set of relevant acoustic parameters of the voice source, such as foundamental frequency, open quotient, speed quotient, and return quotient. The paper explores the potentialities of the derived codebook for the control of the glottal model, and its applications to voice source parameter matching. A fitting procedure is developed, and it is shown that the parameters of the physiologically-controlled two-mass model can be matched to fit target signals (glottal flow pulses) with good accuracy.

PACS no. 43.70.Bk, 43.72.Ja

1. Introduction

Among the challenges to be faced within speech synthesis technologies, audio quality of the synthesis, naturalness of speech, and speaker-specific synthesis, are topical themes. Features of the voice source signal (i.e., the glottal flow) are known to be relevant for characterizing voice quality and speaker identity (see e.g. [1, 2]), and accordingly research on voice source models is becoming increasingly important in speech synthesis.

Various approaches to voice source modeling have been proposed. In the analytical modeling approach the glottal flow waveform is described in terms of piecewise analytical functions: the model proposed by Liljencrants and Fant [3] is a well-known example of this approach, and characterizes one cycle of the flow derivative using four parameters (see section 5.1 and Figure 4). A second approach comprises physical models, which simulate the oscillatory characteristics of the vocal folds. Distributed-element models use systems of partial differential equations to describe the glottis, and can be numerically simulated using e.g. finite-element methods [4, 5]. Simpler, lumped-element models schematize the glottis by means of ideal masses coupled through springs and dampers [6, 7, 8, 9, 10]. The Ishizaka–Flanagan model [7] describes one vocal fold by means of two coupled mechanical oscillators, driven by the intraglottal pressure distribution.

Physical models of the glottis are easily integrated into articulatory models of the vocal apparatus for rule-based speech synthesis [11]. They can potentially provide realistic excitation signals, by reproducing such "natural" effects as transients in the dynamical behavior due to changes in the lung pressure, or occurrence of oscillatory ripples and skewing of the glottal flow waveform due to acoustic interaction with the vocal tract [12]. Physical models are potentially interesting for coding applications as well, as complex and fast variations in a speech signal are the result of slow variations in articulatory parameters that control the speech production process. The adoption of a dynamical model is in a sense a natural way of interpolating the slowly time-varying control parameters, and it seems natural to code these control parameters rather than the resulting speech signal [13].

The long-term goal of this research is to design physical models of the voice source oriented to synthesis applications and to model-based speech coding. Specifically, this paper explores the possibility for a two-mass model to accurately reproduce target glottal flow waveforms, derived e.g. from inverse filtering of real utterances. The model inversion problem (i.e., the process of deriving control parametr curves given a target waveform) has been studied in the context of analytical models. It has been shown that

Received 25 January 2006,
accepted 20 July 2006.

e.g. the Liljencrants–Fant model can be used for fitting target flow derivative waveforms [2, 14, 15]. On the contrary, using physical models for the same purpose is a non-trivial task, which involves inversion of non-linear dynamic systems with a large number of parameters involved. As an example, as many as 19 parameters have to be specified in the two-mass model proposed in [7], and later refinements such as the 3-mass model of Story and Titze [10] involve an even larger number of parameters.

The inversion problem was addressed by Flanagan *et al.* [16], who proposed a method for parametric control of a two-mass model by an adaptive procedure. The method was further improved by Schroeter and Sondhi [13]. In previous studies [17, 18] we have proposed a hybrid approach in which the vocal fold is treated as a linear oscillator, while a non-linear block accounts for interaction with glottal pressure. The non linear block is treated as a black-box element and modeled as a regressor-based mapping. As such, the model can be said to be *physically-informed*.

In this study we explore a different approach. A two-mass model with no black-box elements is employed, and a physical description of the aerodynamic forces is used. The two-mass model is presented in section 2. We then propose a set of physiologically-based control rules, developed after [19], that map activation levels of three laryngeal muscles to physical parameters of the model and consequently enable a drastic reduction of the dimensionality of the control space. The proposed set of control rules is discussed in section 3, and in section 4 the dynamic behavior of the two-mass model in this new control space is analyzed through numerical simulations. Having a physiologically-motivated, low-dimensional control space, we construct in section 5 a codebook between the muscle activation parameters and a set of relevant voice source parameters. The proposed codebook is applied to a voice source parameter matching procedure, described in section 6, by which the two-mass model is used to resinthesize target glottal flow waveforms.

2. A low-dimensional vocal fold model

The two-mass formulation, originally proposed in [7], describes one vocal fold with two mass-spring pairs in the coronal plane, plus an additional coupling spring. Despite its simplicity, this formulation is appealing for many reasons. First, a two-mass model captures both a shear mode (with the two masses in counterphase) and a compressional mode (with masses in phase), which are conceptually equivalent to the two eigenmodes determined by [4] from simulations of a finite-element vocal fold model: much of the success of the two-mass formulation is arguably due to its ability to capture these two modes. Second, a two-mass model requires limited computational resources with respect to more complex models, which makes this formulation still used for synthesis purposes [11]. Finally, the results obtained on the two-mass model can be extended to simpler low-dimensional physical models recently proposed by the authors [20, 21], which include several simplifications and present advantages in terms of controllability and computational load. Such an extension requires the adaptation of the rules for physiological control to the case where a different representation for both the folds and the flow is assumed: this will be the topic of a follow-up investigation.

The mechanical model consists of a pair of coupled differential equations for the mass displacements x_i:

$$\begin{aligned} m_1\ddot{x}_1 + r_1\dot{x}_1 + k_1[x_1 - x_{01}] + k_c[x_1 - x_2] &= LT_1 p_1, \\ m_2\ddot{x}_2 + r_2\dot{x}_2 + k_2[x_2 - x_{02}] - k_c[x_1 - x_2] &= LT_2 p_2, \end{aligned} \tag{1}$$

where x_{0i} are the equilibrium positions, and LT_i are the driving surfaces on which the pressure p_i act ($i = 1, 2$). See Figure 1 for the meaning of the parameters.

Collisions between folds are modeled by including restoring contact forces in the equations. When one of the mass m_i collides (i.e., when the condition $x_i < 0$ holds), its stiffness k_i is increased. Additionally, dissipation during collision is accounted for by increasing the r_i value. This formulation has been proposed by other authors [22] because it has been recognized to provide a more realistic behavior during the closed phase. In summary during collision the ith mass is subject to a restoring force

$$f_i^{(rest)}(x_i, \dot{x}_i) = -k_i^{(rest)} x_i - r_i^{(rest)} \dot{x}_i. \tag{2}$$

The two-mass system (1) is coupled to the aerodynamic driving forces via the glottal areas $a_i(t) = 2Lx_i(t)$. The pressure equations originally proposed in [7] for the Ishizaka–Flanagan model are based on very crude approximations and have been recognized to provide inaccurate modeling of the intraglottal pressure distribution. Here we follow the equations proposed in [10], that are based on the assumptions of *(i)* Bernoulli-type flow in the region upstream of minimum glottal diameter, *(ii)* flow detachment at minimum glottal diameter, *(iii)* constant pressure in the region downstream of minimum glottal diameter, and *(iv)* pressure recovery after glottal exit. Next we briefly summarize the resulting pressure equations.

Let a be the cross-sectional area at a point upstream of minimum glottal diameter, a_s be the subglottal duct area, and ρ_{air} be the air density. Then, according to assumption *(i)*, the pressure $p(a)$ is computed as

$$p_s - p(a) = \frac{1}{2}\rho_{\text{air}}\, \text{sgn}[u] u^2 \left(\frac{1}{a^2} - \frac{1}{a_s^2}\right), \tag{3}$$

where u is the flow. Let $a_{\min}$ be the minimum glottal area, a be the cross-sectional area at any point downstream of minimum glottal diameter, and $p_{\min} = p(a_{\min})$ according to equation (3). Then, according to assumptions *(ii-iii)*, $p(a) = p_{\min}$. Note that for the two-mass model used in this work $a_{\min}$ is defined as $a_{\min} = \min(a_1, a_2)$.

Finally, let p be the pressure at vocal tract entrance and $k_e(a_{\min})$ be a pressure recovery coefficient. Then, according to assumption *(iv)*, p is computed as

$$p - p_{\min} = \frac{1}{2}\rho_{\text{air}} k_e(a_{\min})\, \text{sgn}\big[u\big] \frac{u^2}{a_{\min}^2}, \tag{4}$$

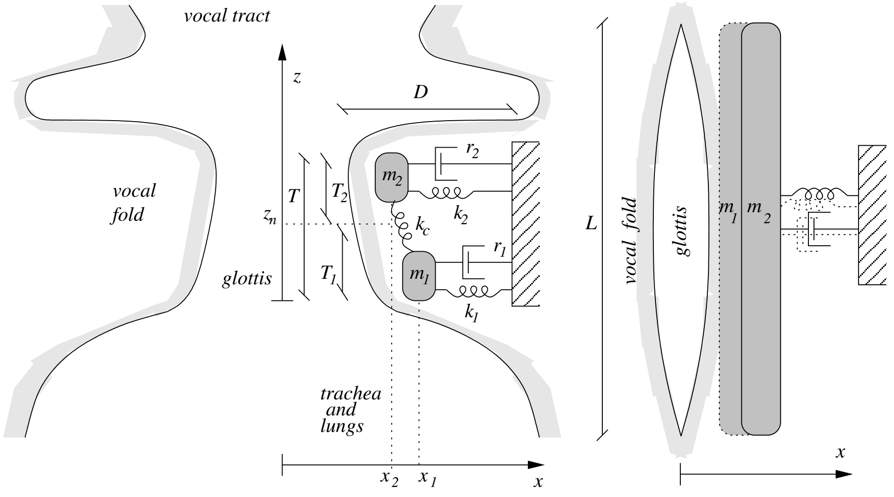

Figure 1. The two-mass model of the vocal folds used in this work: view in the coronal plane (left) and superior view in the horizontal plane (right).

The pressure recovery coefficient is defined as

$$k_e(a_{\min}) = 2\frac{a_{\min}}{a_e}\left(1 - \frac{a_{\min}}{a_e}\right), \tag{5}$$

where a_e is the epilarynx cross-sectional area.

Several works have proposed more refined pressure equations. In particular, Pelorson *et al.* [9] developed a theoretical model in which the flow detachment point is allowed to move continuously within the vocal fold profile. It is clear that the results presented in the next sections, and specifically those summarized in Table IV, are affected by a particular choice of pressure equations. Nonetheless the analysis developed in the remainder of the paper is independent on a specific model of intraglottal pressure.

3. Physiological control of the two-mass model

Low-level parameters (modal frequencies, effective mass in vibration, stiffness, fold thickness, fold length, rest position) are not independently controlled by a speaker: in order to understand the oscillatory characteristics of the vocal folds in a physiologically motivated control space, a set of rules has to be found that transforms muscle activations into geometrical and visco-elastic parameters of a lumped-element model of the glottis. In this section we propose a set of rules for the two-mass model described above, based on the analysis presented in [19].

3.1. From muscle activations to fold geometry

The empirical rules derived by Titze and Story in [19] link the geometry of the vocal folds to activation levels of three muscles: cricothyroid (a_{CT}), thyroarytenoid (a_{TA}) and lateral cricoarytenoid (a_{LC}). These levels are assumed to be normalized in the $[0, 1]$ range. The rules are summarized in Table I and briefly discussed in this section.

The fold length L is the sum of the rest length L_0 and an elongation term $L_0\epsilon$, where ϵ is the longitudinal vocal fold strain and is mainly controlled by the cricothyroid and thyroarytenoid muscles, which act with opposite effects. Additionally, the authors include an "adductory strain" term, Ha_{LC}, that occurs for prephonatory posturing. The factors G (gain of elongation), R (torque ratio), and H (adductory strain factor) are empirical constants. We let $G = 0.2$, $R = 3.0$, $H = 0.2$, in accordance with [19]).

Vocal fold thickness T increases as the vocal fold shortens. The rule given in Table I is derived by assuming that most of the length change is absorbed by thickness change (T_0 is the resting thickness). In the rule governing the fold depth D, the factor 0.2ϵ in the denominator is the complement to the 0.8ϵ factor for thickness. Note that this choice implies that the total fold volume LTD (and therefore the total fold mass) is approximately constant independently on the activation levels: $LTD \sim L_0T_0D_0 + \mathcal{O}(\epsilon^2)$. Note also that the depth rule proposed in [19] is further specialized into two separate rules for the depth of fold cover, and the depth of fold body.

The nodal point z_n of the shear mode of the vocal folds [4] is represented in the two-mass model by a mode where the two masses oscillate in counterphase. This point can move vertically redistributing mass in the vocal fold, and its position is mainly controlled by the thyroarytenoid muscle. When this muscle contracts, z_n moves up on the medial surface, suggesting that there is greater vibrational amplitude at the bottom relative to the top.

The adduction x_{top} (i.e., the resting glottal half-width) at the top of the vocal fold is mainly affected by the activation of the posterior and lateral cricoarytenoid muscles, which act with opposite effects. The analysis reported here includes only the lateral cricoarytenoid activation, however Titze and Story suggest that the effect of the posterior cricoarythenoid muscles may be taken into account by letting a_{LC} vary in the range $[-1, 0]$.

Finally, the convergence parameter x_c is defined as the difference $x_{bottom} - x_{\text{top}}$ between the the glottal half-widths at the bottom and at the top of the vocal fold, and is to a large extent governed by activation of the thyroarytenoid

Table I. Rules for physiological control of vocal fold geometry, after [19].

Fold elongation	$\epsilon = G(Ra_{CT} - a_{TA}) - Ha_{LC}$
Fold length	$L = L_0(1+\epsilon)$
Fold thickness	$T = T_0/(1+0.8\epsilon)$
Fold depth	$D = D_0/(1+0.2\epsilon)$
Nodal point position	$z_n = (1+a_{TA})T/3$
Adduction	$x_{\rm top} = 0.25L_0(1-2a_{LC})$
Convergence	$x_c = T(0.05 - 0.15a_{TA})$

muscle. The rule listed in Table I produce slightly convergent shapes for small a_{TA} values and moderately divergent shapes for large a_{TA} values.

Note the all the rules listed in Table I are independent on a specific model of the vocal folds. Next we derive a second set of rules that link $L, T, D, z_n, x_{\rm top}, x_c$ to lumped parameters of the two-mass model described in section 2.

3.2. From fold geometry to low-level parameters

In [19] the above discussed rules are used for controlling parameters of a 3-mass vocal fold model [10], in which two masses describe the cover tissue and a larger mass describes the vocal fold body. Their analysis can, with some additional care, be adapted to our two-mass model.

Let ρ be the density of the vocal fold tissue, then the total mass of tissue in vibration is $m = \rho LTD$. The fraction of this total mass assigned to the two masses depend on the nodal point z_n:

$$m_1 = m\frac{z_n}{T}, \qquad m_2 = m\left(1 - \frac{z_n}{T}\right). \tag{6}$$

The equilibrium position of the two masses are derived directly from the fold adduction and convergence:

$$x_{02} = x_{\rm top}, \qquad x_{01} = x_{\rm top} + x_c. \tag{7}$$

The derivation of the stiffness parameters k_1, k_2, k_c is less straightforward and is based on modal analysis of the two-mass model. We only report the final equations and refer to [19] for further details:

$$\begin{aligned} k_1 &= \frac{z_n}{T}\left(2\mu\frac{LT}{D} + \pi^2\sigma\frac{DT}{L}\right), \\ k_2 &= \left(1-\frac{z_n}{T}\right)\left(2\mu\frac{LT}{D} + \pi^2\sigma\frac{DT}{L}\right), \\ k_c &= \frac{z_n\mu L}{T}\left(1-\frac{z_n}{T}\right)\cdot \\ &\quad \left\{\frac{D}{2T}\left[\frac{1}{3} - \frac{z_n}{T}\left(1-\frac{z_n}{T}\right)\right]^{-1} - \frac{2T}{D}\right\}, \end{aligned} \tag{8}$$

where μ and σ are the shear modulus and the fiber stress of the fold tissue, respectively. Values for physical parameters are $\rho = 1030$ kg/m^3, $\mu = 700$ Pa, and $\sigma = 500$ Pa, in accordance with values reported in the literature [23, 24, 10, 19, 5].

The geometrical parameters L_0, D_0, T_0 contribute to the determination of the total fold mass in oscillation and of the driving surfaces on which the pressures p_i act. Values for these parameters must therefore be chosen with care, as they profoundly affect the dynamic behavior of the model. Estimates for the fold length at rest range from 1 to 1.6 cm, here we choose $L_0 = 1.3$ cm as in [5]. Concerning the resting thickness, the value $T_0 = 3$ mm is used in [19], which corresponds to the maximum vocal fold thickness. However, a smaller value seems more appropriate in order to account for medial surface bulging. A simple choice for the resting depth D_0 is the depth of the cover layer, since a two-mass model does not capture the body-cover vocal fold structure and is sometimes identified as a "cover" model [10]. However it seems more correct to state that the model accounts for the vibration of both the cover and the body layers, therefore a D_0 value in between the cover depth and the total fold depth would be more appropriate. Moreover, there is a certain variability in the literature for estimates of the cover depth, with values ranging from 0.75 to 3 mm [10, 19]. We then choose values for D_0, T_0 based on the following considerations.

First, we estimate a value for the product D_0T_0. In order to do that, we employ the results presented in [5], where parameter values of a two-mass model are determined by fitting its dynamic behavior to that of a finite-element model. From the mass values determined in [5], $m_1 + m_2 = 0.044$ g, the equivalent D_0T_0 value can be estimated as $D_0T_0 = (m_1 + m_2)/\rho L_0 \sim 3.29$ mm^2.

Second, we exploit a constraint in the ratio D_0/T_0. As discussed in [19], the condition $k_c > 0$ in equation (8) imposes a fairly stringent requirement on the ratio D/T, which is not allowed to become too small. It is easily verified that the condition $D > 2T/3$ guarantees $k_c > 0$.

Based on the above considerations, we finally define $T_0 = 2$ mm, $D_0 = 1.65$ mm. These values satisfy the discussed constraints, and are compatible with values reported in the literature.

4. Numerical simulations

The two-mass model described by equations (1,3,4), together with the control model summarized in Table I and equations (6,7,8), was implemented in Matlab/Octave[1]. The numerical realization of the model employs a discretization technique that ensures accurate simulation of the system (details about this technique are reported in [25]). The realization was used to run a set of simulations for the exploration of the 3-D control space (a_{TA}, a_{LC}, a_{CT}). All the simulations used a sampling rate $F_s = 22.05$ kHz. The subglottal pressure p_s was held fixed at the value 0.8 kPa.

Simulations were run using two different conditions of vocal tract loading: in the first "setup" we assume that no vocal tract load is present, so that the vocal tract input pressure is atmospheric ($p = 0$): this roughly corresponds to the configuration of excised larynges typically used for experimental measures. In the second set-up we

[1] Open source software, a high-level language for numerical computations mostly compatible with MATLAB (www.octave.org).

take into account the effect of a vocal tract load and couple the two-mass glottal model with an idealized vocal tract, modeled as an inertive load: this amounts to saying that the vocal tract input pressure is $p(t) = Ru(t) + I\dot{u}(t)$, where R and I are the vocal tract input resistance and inertance, respectively. Similar low-loss, low-frequency approximations of the tract load have been employed by many authors for analysis and simulation purposes, see e.g. [26, 27]. Realistic values for R, I can be chosen from the analysis in [28], according to which in the limit of low losses, low frequency, and narrow epilarynx the vocal tract impedance reduces to the epilarynx impedance. If the values $a_e = 5$ cm^2 and $l_e = 3.174$ cm are chosen for the epilarynx cross-section and length, then the values $R \sim 2.53 \cdot 10^5$ Ns/m^5 and $I \sim 724$ Ns^2/m^5 are found.

A set of simulations was performed in order to determine the phonation region in the 3-D control space, i.e. the control values (a_{TA}, a_{LC}, a_{CT}) for which self-sustained oscillation is achieved. The phonation region was searched for each of the two vocal tract loading conditions. At each point (a_{TA}, a_{LC}, a_{CT}) the existence of self-sustained stable phonation was determined by applying a zero-crossing multiple-detector to the last 50 ms of the simulated glottal area signal. With this choice we arbitrarily do not consider "always-open glottis" phonation.

The results in the case of no vocal tract loading are shown in Figure 2(top). Phonation is restricted to a small portion of the control space, and occurs in a left neighborhood of $a_{LC} = 0.5$. This behavior is mainly determined by the adduction rule (see Table I): phonation is achieved for sufficiently small adduction values, obtained with a_{LC} values close to 0.5. For values $a_{LC} > 0.5$ both x_{01} and x_{02} become negative. Figure 2(bottom) depicts a hizontal section of the phonation region, for a constant a_{CT} value, and serves as a comparison with the results obtained by Titze and Story [19] on their 3-mass model. Although the results are qualitatively similar, our simulations did not provide evidence of the second, downward sloping phonation path discussed in [19]. This discrepancy may be explained by the lower dimensionality of the two-mass model, although we have not investigated this issue any further.

Results in the case of inertive vocal tract loading are shown in Figure 3, and confirm the expectation that a vocal tract load coupled to the glottis profoundly affects the behavior of the system and facilitates self-sustained oscillation. Again, phonation is obtained within a neighborhood of $a_{LC} = 0.5$. However in this case the phonation region is considerably wider, meaning that self-sustained oscillation is achieved for larger adduction values. No comparison with results in [19] is possible in this case, since they do not include vocal tract simulation.

Table II summarizes ranges, means, and standard deviations for the low-level parameters of the model, computed within the phonation regions. Note that in the case of no-vocal tract loading m_2 and k_2 are on average greater than m_1 and k_1. This finding is consistent with the plot of Figure 2(top), which shows that phonation mainly occurs for values $a_{TA} < 0.5$. This implies that $z_n < T/2$ (and consequently $m_2 > m_1$ and $k_2 > k_1$) in most of the phonation region. On the contrary, in the case of vocal tract loading the distributions are approximately symmetrical.

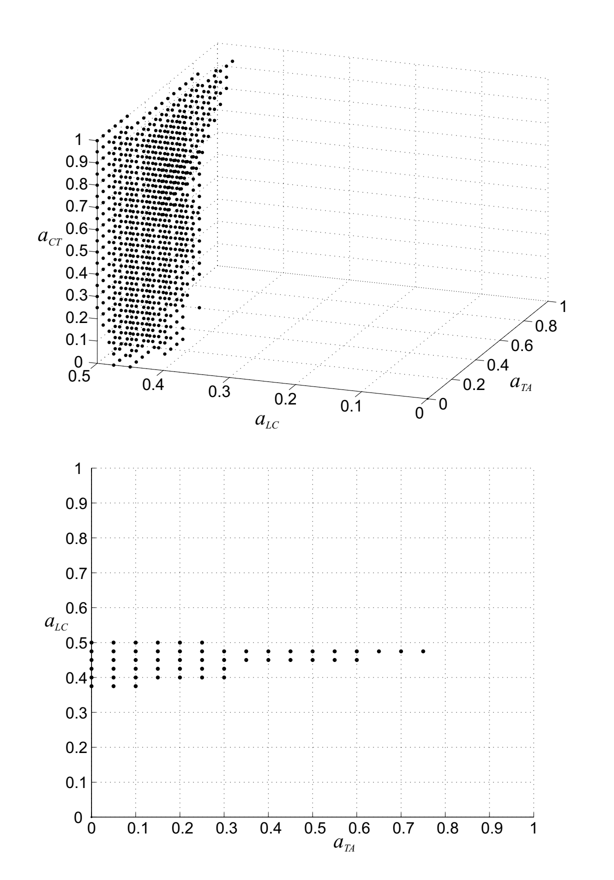

Figure 2. Phonation region without vocal tract loading (top) in the 3-D muscle activation space, and (bottom) in the $a_{LC} - a_{TA}$ plane (with $a_{CT} = 0.7$). Data points show the region of self-sustained oscillation. The subglottal pressure p_s is fixed at 0.8 kPa.

5. Acoustic properties of voice source signals

5.1. Voice source parameters

A wide range of glottal configurations allows a speaker to choose over different phonation modalities: geometric and mechanical fold properties determine the frequency and mode of vibration; the degree of vocal fold adduction has an important role in determining the closed phase duration and the abruptness of closure, and affects the perceived phonation quality. As opposed to "normal" voice quality, *breathy*, *pressed*, *creaky*, are terms commonly found in the literature to denote special phonation types. In breathy voice the vocal folds are vibrating qualitatively as in normal voicing, but glottal closure is incomplete, and consequently the voicing is inefficient and air leaks between folds throughout the vibration cycle. A distinctive characteristic of breathy voice is hence an audible friction noise. On the opposite side, pressed voice occurs when vocal

Table II. Ranges, means, and standard deviations of the low-level parameters of the two-mass model within the phonation regions. Values obtained by de Vries *et al.* [5] are listed as a term of comparison.

	m_1 (g)	m_2 (g)	k_1 (N/m)	k_2 (N/m)	k_c (N/m)	x_{01} (mm)	x_{02} (mm)
Without vocal tract loading							
Range	[0.014, 0.027]	[0.017, 0.029]	[7.5, 16.1]	[9.7, 26.2]	[1.7, 13.3]	[0.01, 0.90]	[0, 0.81]
Mean	0.018	0.026	10.4	15.1	7.2	0.33	0.30
St. dev.	0.003	0.003	1.8	2.4	2.8	0.22	0.21
With vocal tract loading							
Range	[0.014, 0.029]	[0.014, 0.029]	[7.5, 17.6]	[6.7, 29.6]	[0.1, 14.2]	[0, 2.61]	[0, 2.60]
Mean	0.022	0.022	12.2	12.3	6.3	1.13	1.18
St. dev.	0.004	0.004	2.3	3.2	3.0	0.67	0.68
de Vries *et al.* [5]							
	0.024	0.020	22	14	10	N.A.	N.A.

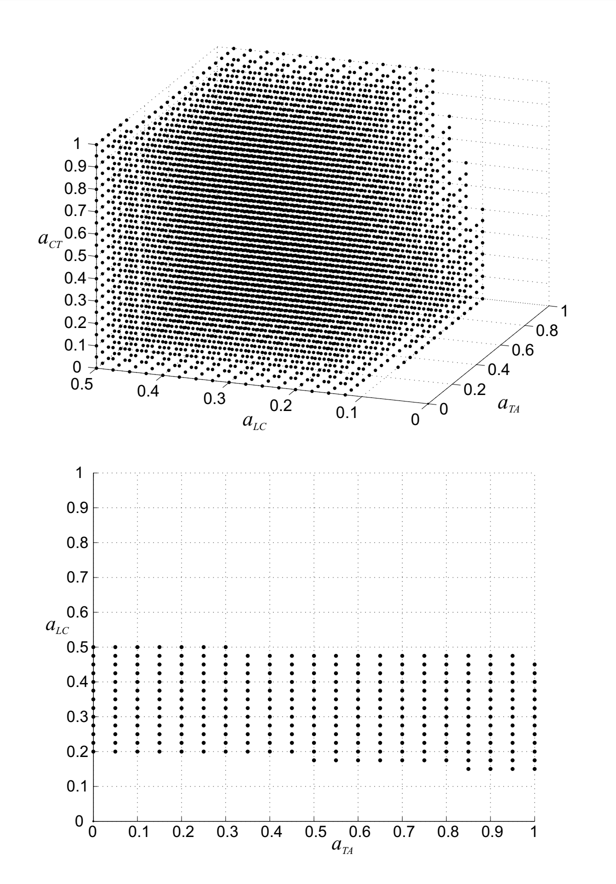

Figure 3. Phonation region with vocal tract loading (top) in the 3-D muscle activation space, and (bottom) in the $a_{LC} - a_{TA}$ plane (with $a_{CT} = 0.7$). Data points show the region of self-sustained oscillation. The subglottal pressure p_s is fixed at 0.8 kPa.

folds are pressed together and the glottal cycle is characterized by an abrupt closure, a reduced open phase duration, and a small vibration amplitude. Creaky voice is characterized in a somewhat similar way, except for the fact that the tight compression of the folds may occasionally produce irregular vibrations, perceived as a crackling quality.

Table III. Time-domain acoustic parameters of the glottal flow waveform.

Speed quotient	$SQ = (t_p - t_o)/(t_c - t_p)$
Open quotient	$OQ = (t_c - t_o)/P$
Opening quotient	$OingQ = (t_p - t_o)/P$
Closing quotient	$CingQ = (t_c - t_p)/P$
Return quotient	$RQ = (t_c - t_e)/P$
Peak-to-peak quotient	$PPQ = (t_e - t_p)/P$
Amplitude quotient	$AQ = E_i/E_e$

Some features of the glottal waveform have been recognized to be particularly relevant for the study of the perceptual influence of the voice source characteristics, and for comparing different voice qualities.

Referring to Figure 4, typical [1, 3] voice source quantification parameters extracted from the flow and the differentiated flow are: P (the glottal cycle period), $F_0 = 1/P$ (the fundamental frequency of oscillation), E_i (the maximum flow amplitude), E_e (the amplitude of the differentiated flow negative peak), t_o (the opening instant), t_p (the maximum flow amplitude instant), t_e (the negative peak instant), t_c (the closing instant).

Derived parameters are the *speed quotient* SQ, the open quotient OQ, the opening quotient $OingQ$, the closing quotient $CingQ$, the return quotient RQ, the peak-to-peak quotient PPQ, and the amplitude quotient AQ. Definitions for these parameters are provided in Table III. The spectral tilt of the voice source can be quantified by parameters such as the *harmonic richness factor* HRF= $(\sum_{i=2}^{N} H_i)/H_1$, where H_i denotes the amplitude of the ith harmonic partial.

The analysis and matching of inverse filtered voice samples from subjects with varying voice quality, age, and sex, permitted to gain understanding of the relations between the voice source characteristics and the perceived voice quality [29, 30, 2, 14, 15, 1]. For example, it has been observed that OQ decreases monotonically when phonation changes from breathy to pressed whereas SQ increases monotonically, or that the speed quotient SQ is closely re-

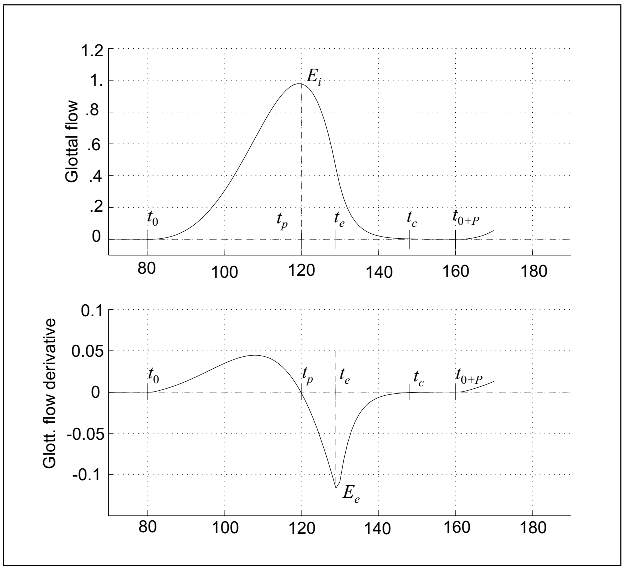

Figure 4. Glottal flow and derivative: time of glottal opening t_o; time and value t_p, E_i of flow maximum; time and value t_e, E_e of flow derivative minimum; time of glottal closure t_c; glottal period P.

lated to the perceptual sensation of vocal effort, and a high SQ results in general in a tense or hyperfunctional voice quality whereas a small SQ results in a lax or hypofunctional voice quality.

5.2. A glottal articulatory codebook

So called *articulatory codebooks* are used in the literature to map the acoustical properties of a speech signal to corresponding geometries of the vocal tract. An articulatory codebook consists of vectors of geometric parameters describing vocal tract shapes, that are linked to corresponding vectors of acoustic parameters representing vocal tract filters [13]. In this section we describe an analogous codebook for the two-mass vocal fold model, in which the articulatory parameters involved in the codebook are glottal articulators rather than vocal tract articulators.

Simulations of the two-mass model, with and without vocal tract loading, were run on a grid of muscle activation values $\boldsymbol{y} \triangleq (a_{TA}, a_{LC}, a_{CT})$ in the phonation region. For every point $\boldsymbol{y}$, a vector $\boldsymbol{x}$ of acoustic parameters was extracted via automatic analysis on the generated glottal flow waveforms. We have chosen to work on a minimal set of relevant acoustic parameters, namely $\boldsymbol{x} = (F_0, SQ, OQ, RQ)$ (see section 5.1 for definitions). The result is a glottal articulatory codebook of M pairs of entries $(\boldsymbol{x}_j, \boldsymbol{y}_j)$ $(j = 1 \dots M)$.

A grid-step of 0.05 was used along the a_{TA} and a_{CT} directions. A smaller step, namely 0.025 was used along the a_{LC} direction, since as already discussed in section 4 phonation is only obtained within a neighborhood of $a_{LC} = 0.5$. With these grid-steps, the generated codebook has $M \sim 10^3$ entries in the case of no vocal tract loading, and $M \sim 6.1 \cdot 10^3$ entries in the case of inertive vocal tract loading.

Table IV. Glottal articulatory codebook: ranges for the relevant voice source parameters.

	F_0 (Hz)	SQ	OQ	RQ
Without vocal tract loading				
Range	[78, 171]	[0.61, 5.53]	[0.36, 0.97]	[0, 0.307]
Mean	128	1.86	0.70	0.091
St. dev.	13	0.51	0.06	0.059
With vocal tract loading				
Range	[79, 152]	[1.43, 5.73]	[0.34, 0.88]	[0, 0.020]
Mean	114	2.49	0.79	0.004
St. dev.	6	0.40	0.04	0.001

Table IV provides indications about the ranges of the voice source parameters within the codebook. From this, a few remarks can be made. First, the range for F_0 is within realistic values, confirming the validity of the analysis reported in section 3.2 for the choice of physical parameter values. Although the average F_0 value is lower when vocal tract loading is present, comparison of flow signals within the common subregion of phonation shows that the inertive load has minimal effects on pitch. Values for the speed quotient SQ are markedly higher in the presence of vocal tract loading. This finding is consistent with expectation, as the vocal tract is known to have a major influence on the flow skeweness. Finally, values for the return quotient RQ are extremely low, especially in the presence of vocal tract loading. This reflects a general limitation of low-dimensional physical models of the glottis, in which glottal closure always occurs abruptly and results in poor modeling of the closing phase.

6. Voice source parameter matching

The codebook described in the last section has been tested in order to verify its potentials in fitting target glottal flow waveforms through an analysis-by-synthesis procedure, in which values for the physiological control parameters are chosen in order to minimize an acoustic distance between the target signal and the resynthesized one. More precisely, given a codebook of the form $(\boldsymbol{x}_j, \boldsymbol{y}_j)$ with M entries, and given a target vector $\boldsymbol{x} = (F_0, SQ, OQ, RQ)$ of acoustic parameters, we wish to find the vector of muscular activations $\boldsymbol{y}_j = (a_{TA}, a_{LC}, a_{CT})_j$ that minimizes the acoustic distance between $\boldsymbol{x}_j$ and $\boldsymbol{x}$ (where $j \in [1, M]$ is the index of the codebook entry).

In order to approach this problem one must first define a proper distance in the acoustic space. In the case of vocal tract shapes estimation, where the acoustic parameters are frequency-domain parameters, well-founded acoustic distances can be defined (e.g., a symmetrized likelihood ratio distance between LPC vectors has been used in [31]). In the case considered here, however, all the acoustic parameters (F_0, SQ, OQ, RQ) are time-domain parameters and it is less clear how to define a distance on the acoustic vec-

Table V. Results from the fitting procedure. Acoustic parameters of the target signals, acoustic parameters and relative errors of the fitting signals for F_0 = 100 Hz and F_0 = 120 Hz.

	Target signals								
	1	2	3	4	5	6	7	8	9
OQ	0.356	0.503	0.780	0.680	0.705	0.719	0.594	0.562	0.506
SQ	2.925	2.964	3.145	1.542	2.987	3.877	2.157	3.863	3.373
RQ	0.014	0.014	0.009	0.152	0.041	0.002	0.091	0.005	0.002
	Fitting signals – F_0 = 100 Hz								
F_0	106.010	103.644	103.037	106.010	103.644	107.561	90.370	107.561	104.009
Rel. error	0.060	0.036	0.030	0.060	0.036	0.076	0.096	0.076	0.040
OQ	0.670	0.674	0.668	0.750	0.674	0.654	0.539	0.654	0.660
Rel. error	0.882	0.340	0.143	0.103	0.044	0.091	0.093	0.162	0.306
SQ	3.059	2.960	3.146	1.557	2.960	3.786	1.990	3.786	3.446
Rel. error	0.046	0.001	0.000	0.010	0.009	0.024	0.077	0.020	0.022
RQ	0.005	0.005	0.005	0.000	0.005	0.000	0.000	0.000	0.005
Rel. error	0.647	0.655	0.485	1.000	0.885	1.000	1.000	1.000	1.080
	Fitting signals – F_0 = 120 Hz								
F_0	106.010	115.445	120.361	121.725	120.492	118.548	114.844	120.823	111.364
Rel. error	0.117	0.038	0.003	0.014	0.004	0.012	0.043	0.007	0.072
OQ	0.670	0.686	0.824	0.727	0.699	0.661	0.704	0.656	0.657
Rel. error	0.888	0.365	0.059	0.071	0.007	0.079	0.185	0.164	0.299
SQ	3.059	2.970	3.081	1.723	3.000	3.920	2.201	3.790	3.483
Rel. error	0.040	0.002	0.018	0.110	0.003	0.009	0.019	0.010	0.024
RQ	0.005	0.005	0.005	0.004	0.000	0.005	0.000	0.001	0.005
Rel. error	0.602	0.588	0.355	0.973	1.000	0.779	1.000	0.677	3.178

tors $\boldsymbol{x}$. Therefore in the following analysis we resorted to the use of a modified euclidean distance:

$$d(\boldsymbol{x}_1, \boldsymbol{x}_2) = \sqrt{\sum_{i=1}^{4} \frac{(x_{1i} - x_{2i})^2}{\bar{x}_i^2}}, \qquad (9)$$

where $\bar{x}_i$ is the mean of the ith acoustic parameter within the phonation region, as reported in Table IV.

Note that the distance function (9) is well defined since the components of the acoustic parameter vectors $\boldsymbol{x}$ are always positive by definition. The normalization of the contribution of each acoustic parameter with respect to its mean is necessary because the parameters take values in very different ranges: without the normalization, the distance would be dominated by the first component, i.e. the pitch component.

In order to validate the proposed fitting procedure, experiments were carried out on a set of synthetic test signals. Each test signal was obtained from the Liljencrats–Fant model, with a noisy componenent superimposed. For consistency, the acoustic parameters of the test flow waveform were extracted using the same automatic analysis tool used in section 5 to analyze flow signals of the two-mass model.

The pulses in the set were selected so that their acoustic parameters span as widely as possible the parameter ranges in the phonation regions of the physical model (see Table IV), except for a few cases where the values are out of range. Two F_0 values, 100 Hz and 120 Hz, were selected, and for each pitch nine pulses were generated, giving a total number of 18 test signals. The first three rows in Table V show the OQ, SQ, RQ values of the nine test pulses for pitch F_0 = 100 Hz: for pulses 1 to 3, OQ increases from 0.36 to 0.78, while SQ and RQ are approximately constant; For pulses 4 to 6, SQ increases from 1.54 to 3.88, while OQ is approximately constant and RQ present sligthly decreasing values. For pulses 7 to 9, RQ decreases from about 0.1 to 0.002 while OQ is approximately constant and SQ is limited to the range 2.16–3.86. Values of the pulses for pitch F_0 = 120 Hz are almost identical, with small discrepancies.

The remainder of Table V shows the results of our fitting procedure on the 18 test signals, and reports values and relative errors for the acoustic parameters of the fitting signals resynthesized using the two-mass model. The configuration with inertive vocal tract loading was used, since the Liljencrants–Fant flow model takes into account skewing effects of the vocal tract.

The pitch F_0 is in general accurately matched. The speed quotient SQ is also matched with good accuracy, with only two target signals (signal 7 with F_0 = 100 Hz and signal 4 with F_0 = 120 Hz) resulting in a relative error above 7%. Low OQ values (see especially signal 1) are poorly matched by the fitting signals. Finally, the closing phase is in general grossly mismatched in the resynthesis.

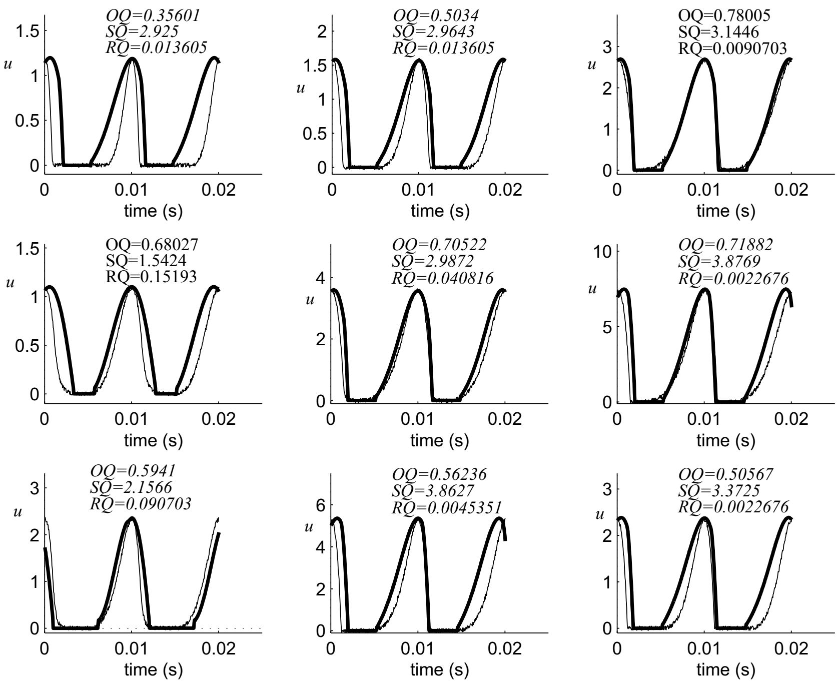

Figure 5. Results from the fitting procedure (F_0 = 100 Hz): target signals (thin lines) and fitting signals (thick lines).

As already mentioned, this is mainly a consequence of the intrinsic limitations of the two-mass model in describing the closing phase. Figure 5 shows the nine target waveforms and the fitting signals obtained from the two-mass model, in the case of $F_0 = 100$ Hz.

7. Conclusions

The results discussed in the last section suggest that the proposed approach can be successfully used for voice source parameter matching applications. First, the muscle activation control space allows exploration of a wide region of the acoustic parameter space, as shown in Table IV. Second, when used as a synthesis-by-analysis tool the approach proposed in this paper leads to robust resynthesis since the stability of the physical model is guaranteed “by construction”: this is an advantage with respect to our previous works [17, 18], where a regressor-based black-box element is used, and the stability of the system consequently depend upon the regressor coefficients and cannot be guaranteed “a priori”. Finally, similar sets of physiologically-based control rules can be developed for other models. The authors recently proposed a class of low-dimensional physical models that include several simplifications over the two-mass model and present some advantages in terms of controllability and computational load [20, 21]. The use of such models require the adaptation of the rules for physiological control to the case where a different representation for both the folds and the flow is assumed. The analysis for the derivation of the control rules presents substantial differences and is presently under study.

A number of weaknesses of the proposed approach have also been evidenced by the results of section 6. First, it has already been noted that the two-mass model provides a poor description of the glottal flow near closure. While accurate finite-element models are able to provide qualitative behaviors in agreement with observations of glottal closure during normal voice production [32], such behaviors are not easily simulated with a low-dimensional model. This is a major limitation, since the closing phase is known to carry very relevant perceptual features of the speech signal. Second, the codebook proposed in this work does not include the subglottal pressure p_s in the set of articulatory parameters. However p_s is known to have a major influence on relevant voice source parameters, in particular the fundamental frequency of phonation F_0 is known to increase almost linearly with p_s [33]. For this reason the physiological control space should be expanded to a 4-D space that includes p_s.

In order to adapt the fitting procedure to non-stationary voice source signals, the procedure needs to be refined to allow dynamic access to the codebook. In particular, techniques that have already been used for the estimation of vocal tract shapes in articulatory models can be adapted to the estimation of glottal articulators. One such technique makes use of dynamic programming methods in order to estimate parameters over a sequence of analysis frames rather than a single frame. In this way large “articulatory efforts” (i.e., fast changes in the articulatory parameters) are penalized in the estimation procedure, and smoothly evolving articulatory trajectories are identified [13]. Finally, the proposed fitting procedure has so far been tested using synthetic target signals. For a more thorough validation it needs to be tested on experimental data, i.e. glottal flow signals obtained from inverse filtering of real utterances.

Acknowledgement

The authors would like to thank two anonymous reviewers, whose constructive criticism and helpful suggestions led to a significantly improved version of the manuscript.

References

[1] P. Alku, E. Vilkman: A comparison of glottal voice quantification parameters in breathy, normal and pressed phonation of female and male speakers. Folia Phoniatr. Logop. **48** (Sep. 1996) 240–254.

[2] D. G. Childers, C. Ahn: Modeling the glottal volume-velocity waveform for three voice types. J. Acoust. Soc. Am. **97** (Jan. 1995) 505–519.

[3] G. Fant, J. Liljencrants, Q.-G. Lin: A four-parameter model of glottal flow. STL-QPSR **26** (1985) 1–13.

[4] D. A. Berry, I. R. Titze: Normal modes in a continuum model of vocal fold tissues. J. Acoust. Soc. Am. **100** (Nov. 1996) 3345–3354.

[5] M. P. de Vries, H. K. Schutte, 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** (Dec. 1999) 3620–3628.

[6] J. L. Flanagan, L. L. Landgraf: Self-oscillating source for vocal-tract synthesizers. IEEE Trans. Audio and Electroacoust. **16** (1968) 57–64.

[7] K. Ishizaka, J. L. Flanagan: Synthesis of voiced sounds from a two-mass model of the vocal cords. Bell Syst. Tech. J. **51** (1972) 1233–1268.

[8] J. Liljencrants: A translating and rotating mass model of the vocal folds. STL-QPSR **32** (1991) 1–18.

[9] X. Pelorson, A. Hirschberg, R. R. van Hassel, A. P. J. Wijnands: Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. Application to a modified two-mass model. J. Acoust. Soc. Am. **96** (Dec. 1994) 3416–3431.

[10] B. H. Story, I. R. Titze: Voice simulation with a body-cover model of the vocal folds. J. Acoust. Soc. Am. **97** (Feb. 1995) 1249–1260.

[11] M. Sondhi: Articulatory modeling: a possible role in concatenative text-to-speech synthesis. Proc. 2002 IEEE Workshop on Speech Synthesis, S. Monica (CA), Sep. 2002, 73–78.

[12] D. G. Childers, C.-F. Wong: Measuring and modeling vocal source-tract interaction. IEEE Trans. Biomedical Engineering **41** (Jan. 1994) 663–671.

[13] J. Schroeter, M. Sondhi: Speech coding based on physiological models of speech production. – In: Advances in Speech Signal Processing. S. Furui, M. Sondhi (eds.). Dekker, New York, 1992, 231–263.

[14] E. L. Riegelsberger, A. K. Krishnamurthy: Glottal source estimation: Methods of applying the LF-model to inverse filtering. Proc. IEEE Int. Conf. Acoust. Speech and Signal Process. (ICASSP'93), Minneapolis, 1993, 542–545.

[15] H. Strik: Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses. J. Acoust. Soc. Am. **103** (May 1998) 2659–2669.

[16] J. L. Flanagan, K. Ishizaka, K. L. Shipley: Signal models for low bit-rate coding of speech. J. Acoust. Soc. Am. **68** (Sep. 1980) 780–791.

[17] F. Avanzini, C. Drioli, P. Alku: Synthesis of the Voice Source Using a physically informed model of the glottis. Proc. Int. Symp. Mus. Acoust. (ISMA'01), Perugia, Sep. 2001, 31–34.

[18] C. Drioli, F. Avanzini: Hybrid parametric physiological glottal modelling with application to voice quality assessment. Medical Engineering & Physics **24** (Sep. 2002) 453–460.

[19] I. R. Titze, B. H. Story: Rules for controlling low-dimensional vocal fold models with muscle activation. J. Acoust. Soc. Am. **112** (Sep. 2002) 1064–1027.

[20] F. Avanzini, P. Alku, M. Karjalainen: One-delayed-mass model for efficient synthesis of glottal flow. Proc. Eurospeech Conf., Aalborg, Sep. 2001, 51–54.

[21] C. Drioli: A flow waveform-matched low-dimensional glottal model based on physical knowledge. J. Acoust. Soc. Am. **117** (May 2005) 3184–3195.

[22] M. Sondhi, J. Schroeter: A hybrid time-frequency domain articulatory speech synthesizer. IEEE Trans. Acoust., Speech, and Sig. Process. **35** (July 1987) 955–967.

[23] Y. Kakita, M. Hirano, K. Ohmaru: Physical properties of vocal tissue: measurements on excised larynges. – In: Vocal Fold Physiology. M. Hirano, K. Stevens (eds.). University of Tokyo Press, Tokyo, 1981, 377–398.

[24] M. Hirano, S. Kurita, T. Nakashima: The structure of vocal folds. – In: Vocal Fold Physiology. M. Hirano, K. Stevens (eds.). University of Tokyo Press, Tokyo, 1981, 33–41.

[25] F. Avanzini, D. Rocchesso: Efficiency, accuracy, and stability issues in discrete time simulations of single reed wind instruments. J. Acoust. Soc. Am. **111** (May 2002) 2293–2301.

[26] G. Fant: Preliminaries to analysis of the human voice. STL-QPSR **23** (1982) 1–27.

[27] I. R. Titze: The physics of small-amplitude oscillation of the vocal folds. J. Acoust. Soc. Am. **83** (Apr. 1988) 1536–1552.

[28] I. R. Titze, B. H. Story: Acoustic interactions of the voice source with the lower vocal tract. J. Acoust. Soc. Am. **101** (Apr. 1996) 2234–2243.

[29] P. J. Price: Male and female voice source characteristics: inverse filtering results. Speech Commun. **8** (Feb. 1989) 261–277.

[30] D. Childers, C. Lee: Vocal quality factors: analysis, synthesis, and perception. J. Acoust. Soc. Am. **90** (Nov. 1991) 2394–2410.

[31] J. Schroeter, M. Sondhi: Techniques for estimating vocal-tract shapes from the speech signal. IEEE Trans. Speech Audio Process. **2** (Jan. 1994) 133–150.

[32] H. E. Gunter: A mechanical model of vocal-fold collision with high spatial and temporal resolution. J. Acoust. Soc. Am. **113** (Feb. 2003) 994–1000.

[33] I. R. Titze: On the relation between subglottal pressure and fundamental frequency in phonation. J. Acoust. Soc. Am. **85** (Feb. 1989) 901–906.