PHYSIOLOGICAL CONTROL OF LOW-DIMENSIONAL GLOTTAL MODELS WITH APPLICATIONS TO VOICE SOURCE PARAMETER MATCHING

Federico Avanzini[1], Simone Maratea[1], Carlo Drioli[2]
[1] Department of Information Engineering, University of Padova, Padova, Italy
[2] Institute of Phonetics and Dialectology, ISTC-CNR, Padova, Italy

A set of rules is proposed for controlling a 2-mass glottal model through activation levels of laryngeal muscles. The rules convert muscle activities into physical quantities such as fold adduction, mass, thickness, depth, stiffness. A codebook is constructed between muscular activations and a set of relevant voice source parameters, and its applications to voice source parameter matching are explored.

I. INTRODUCTION

Features of the voice source signal (i.e., the glottal flow) are known to be relevant for characterizing voice quality and speaker identity. Parametric models of the voice source fit the glottal signal with piecewise analytical functions, using a small number of parameters. As an example, the Liljencrants and Fant (LF) model [8] characterizes one cycle of the flow derivative using as few as four parameters (see section II and Fig. 1). Physical models of the glottal system describe the vocal fold with two [10] or more [14] coupled mechanical oscillators, driven by the intraglottal pressure. Physical models capture the basic non-linear mechanisms that initiate self-sustained oscillations, and can simulate subtle features (e.g. interaction with the vocal tract); however the large number of parameters typically involved makes it hard to employ these models for voice source matching purposes needed in many applications, ranging from rule-based speech synthesis [13] to analysis and assessment of voice quality, including the detection and classification of voice pathologies [6].

We have addressed the issue of identification of physically-based models in previous studies [3], [7] using a hybrid approach in which the vocal fold is treated as a linear oscillator, while a non-linear block that accounts for interaction with glottal pressure is modeled as a regressor-based mapping: given a target glottal flow signal, weights for the regressors can be estimated in order to fit the target.

In this study we explore a different approach, in which the dimension of the control space of a 2-mass model (see Fig. 2) is drastically reduced by applying a set of rules that map three muscular activation parameters to the low-level physical parameters of the model. The rules are developed after Titze and Story [18] and are described in section III.

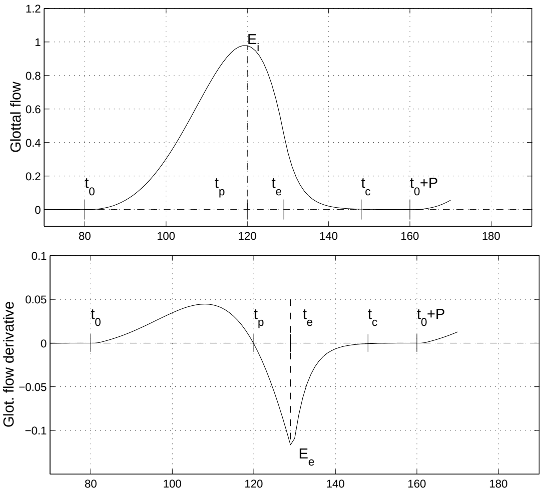

Fig. 1. Glottal flow and derivative: time of glottal opening t_o; time and value t_p, E_i of flow maximum; time and value t_e, E_e of flow derivative minimum; time of glottal closure t_c; glottal period P.

Having a physiologically-motivated, low-dimensional control space, we construct in section IV a codebook between the muscle activation parameters and a set of relevant voice source parameters, and we explore its potentials in fitting target flow waveforms.

II. VOICE SOURCE PARAMETERS

Some cues 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 Fig. 1, typical [8], [1] voice source quantification parameters extracted from the flow and the differentiated flow are: $T_o = t_p - t_o$ (opening phase duration), $T_{pp} = t_e - t_p$ (positive to negative peak interval duration), $T_{ret} = t_c - t_e$ (return phase duration), $T_c = t_o + P - t_c$ (closed phase duration), $T_{open} = T_o + T_{pp} + T_{ret}$ (open phase duration). Derived parameters are the *speed*

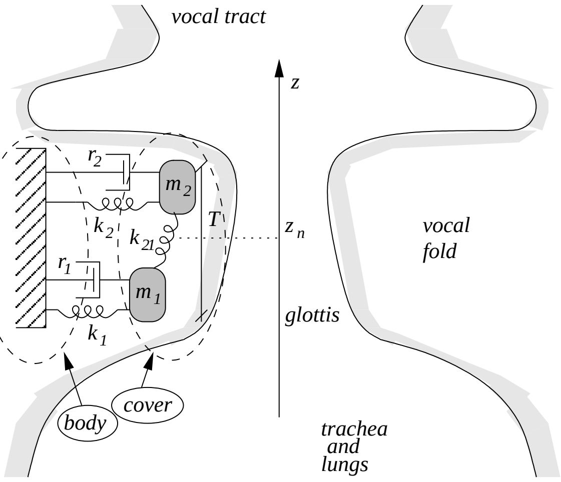

Fig. 2. The 2-mass model used in this work.

quotient $SQ = T_o/(T_{pp} + T_{ret})$, the *open quotient* $OQ = T_{open}/T$, the *opening quotient* $OingQ = T_o/T$, the *closing quotient* $CingQ = (T_{pp} + T_{ret})/T$, the *return quotient* $RQ = T_{ret}/T$, the *peak-to-peak quotient* $PPQ = T_{pp}/T$, and the *amplitude quotient* $AQ = E_i/E_e$. 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.

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; vocal fold adduction (i.e., relative distance) 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 glottal closure is incomplete, 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 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, additionally the tight compression of the folds may occasionally produce irregular vibrations, perceived as a crackling quality.

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 [11], [4], [5], [12], [15], [1].

TABLE I

RULES FOR PHYSIOLOGICAL CONTROL OF THE 2-MASS MODEL.

Fold elongation	$\epsilon = G(Ra_{CT} - a_{TA}) - Ha_{LC}$
Fold length	$L = L_0(1+\epsilon)$
Cover depth	$D_c = \frac{D_{muc}+0.5D_{lig}}{1+0.2\epsilon}$
Fold thickness	$T = \frac{T_0}{1+0.8\epsilon}$
Nodal point position	$z_n = (1+a_{TA})T/3$
Adduction	$\xi_0 = 0.25L_0(1-2a_{LC})$

III. A PHYSIOLOGICALLY CONTROLLED 2-MASS MODEL

In this section we search a link between laryngeal muscle activation and mechanical properties of the low-dimensional 2-mass model depicted in Fig. 2 and based on the Ishizaka-Flanagan model [10]. Low-level parameters (modal frequencies, effective mass in vibration, stiffness, fold thickness, fold length, rest position) are not independently controlled by the vocalist: in order to understand the oscillatory characteristics in a physiologically motivated control space, a set of rules has to be found that transforms muscle activations to geometrical and viscoelastic parameters of the model.

We follow the analysis by Titze and Story [18] who, based on experimentations and cadaveric examinations, developed a set of rules for controlling parameters of their 3-mass vocal fold model [14]. Specifically, the model is controlled by the (normalized) activation levels of three muscles: cricothyroid (a_{CT}), thyroarytenoid (a_{TA}) and lateral cricoarytenoid (a_{LC}).

The 3-mass model developed in [14] uses two masses to describe the cover tissue and a third, larger mass to describe the body. In this work we adapt Titze and Story’s rules set to the 2-mass model by ignoring any references to this third mass. Therefore we select the rule subset given in table I. Here D_{mus}, D_{muc}, and D_{lig} are the anatomical resting depths for thyroarytenoid muscle, mucosa, vocal ligament, respectively; T_0 and L_0 are the resting thickness and length, respectively. The factors G (gain of elongation), R (torque ratio), and H (adductory strain factor) are empirical constants (for this study we let $G = 0.2$, $R = 3.0$, $H = 0.2$ in accordance with [18]). Values for D_{mus}, D_{muc}, D_{lig} are chosen after [14]. The low-level parameters $k_1, k_2, k_{12}, m_1, m_2$ of the 2-mass model are then derived from the geometrical parameters D_c, T, L, z_n, together with the tissue density ρ, the cover shear modulus μ_c, and the cover fiber stress σ_c [18].

We have developed a MATLAB/Octave[1] implementation of the 2-mass model, completed by the physi-

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

ological link between laryngeal muscle activation and mechanical properties of the model, with the activations a_{TA}, a_{LC} and a_{CT} varying in the range $[0, 1]$.

IV. Numerical simulations

The 2-mass model with physiological control was used to run a set of simulations for the exploration of the 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. The anatomical resting depths of layers of vocal folds tissue were chosen in accordance with Titze and Story [14].

A. Phonation regions

A first set of simulations was performed in order to determine the phonation region in the control space. Simulations were run using two configurations. First an ideally open glottis (i.e., with zero supraglottal pressure) was considered. Second, a vocal tract load was taken into account by coupling the 2-mass glottis model with a cylindrical vocal tract model.

The phonation region was searched for each of the two configurations. Following Titze *et al.* [18], 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. In this way we arbitrarily not consider “always-open glottis” phonation.

For both configurations, phonation regions are comparable with results by Titze and Story [18] on the 3-mass model. In particular, a_{CT} has little influence on the shape of the self-sustained phonatory region. For the open-glottis configuration, it simply acts as a switch that restricts phonation in the range $a_{CT} \in [0, 0.7]$, while for the cylindrical vocal tract configuration phonation occurs in the entire range $a_{CT} \in [0, 1]$.

The 2-D phonation region in the a_{LC}-a_{TA} plane (with a_{CT} fixed) is wedge-shaped. For the open-glottis configuration, the region is contained in the rectangle $a_{TA} \in [0, 0.9]$ and $a_{LC} \in [0.35, 0.5]$, while for the cylindrical vocal tract configuration the bounding rectangle is given by $a_{TA} \in [0, 1]$ and $a_{LC} \in [0.2, 0.5]$. Thus, following expectations the phonation region is larger when a vocal tract load is coupled to the glottis. Given the similarity between these results and those reported in [18], we consider our selected rules a valid link between laryngeal muscle activation and mechanical properties of the 2-mass model.

TABLE II

Physiological-to-acoustic codebook: ranges for the relevant voice source parameters.

	F_0	SQ	OQ	$OingQ$	$CingQ$	RQ
Open-glottis configuration						
Mean value	251	1.36	0.63	0.36	0.26	0.02
Min. value	217	0.90	0.51	0.29	0.19	0
Max. value	367	2.01	0.94	0.52	0.43	0.13
Cyl. vocal tract configuration						
Mean value	253	1.66	0.80	0.49	0.30	0.02
Min. value	179	1.13	0.35	0.23	0.12	0
Max. value	816	2.79	0.90	0.59	0.41	

B. A physiological-to-acoustic codebook

Having determined the phonation regions in the control space, we analyze the properties of the voice source signal in such regions. We chose a set of relevant acoustic parameters, namely $F_0, SQ, OQ, OingQ, CingQ, RQ$ (see section II for definitions) and developed a MATLAB/Octave script for automatic analysis and extraction of these parameters from the glottal flow signal. Using this tool, the signals produced by every triple a_{TA}, a_{LC}, a_{CT} in the phonation region were analyzed, resulting in a physiological-to-acoustic codebook of the form

$$(a_{TA}, a_{LC}, a_{CT}) \mapsto (F_0, SQ, OQ, OingQ, CingQ, RQ).$$

Table II provides indications about the ranges of the voice source parameters within the codebook. From this, a few remarks can be made.

First, F_0 values appear to be high, considering that a set of parameters typical for males has been used. This suggests that the choice of physical parameters made in section III (specifically, keeping the same values used in [18] for the vocal fold cover tissue, while discarding any description of the vocal fold body) is not optimal.

Second, values for the return quotient RQ are extremely low. 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.

The codebook has been tested in order to verify its potentials in fitting target flow waveforms. Target signals were constructed by superimposing a noisy component to synthetic glottal flow waveforms obtained from the LF model [8]. The fitting procedure works as follows:

1. The set $F_0, SQ, OQ, OingQ, CingQ, RQ$ of voice source parameters is extracted from the target signal.
2. A triple a_{TA}, a_{LC}, a_{CT} is determined in such a way that it minimizes the distance between its image in the codebook and the target voice source parameter vector.
3. A fitting signal is resynthesized with the 2-mass model controlled by the selected triple a_{TA}, a_{LC}, a_{CT}.

Figure 3 shows an example of the results. The opening, closing, and flow maximum points (t_o, t_c, t_e) are accurately matched. On the other hand the opening and

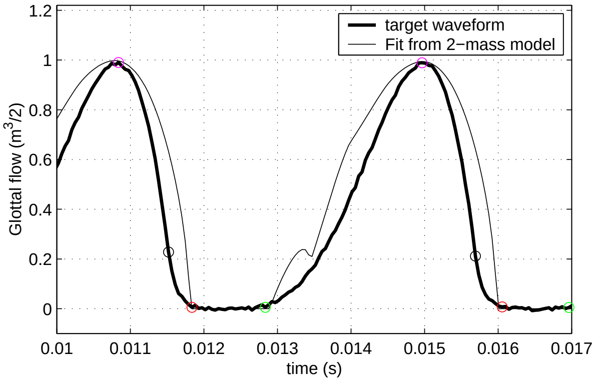

Fig. 3. Results from the fitting procedure. The target waveform is constructed by superimposing a noisy component to synthetic glottal flow waveforms obtained from the LF model.

especially the closing phases are poorly matched. As already mentioned, this is an intrinsic limitation of the 2-mass model. As a consequence the time and value of the negative peak of the flow derivative are mismatched.

V. Discussion

The results presented in this work are still very preliminary. Among the points that need further discussion and refinements, the following can be mentioned.

The codebook described in section IV does not include the subglottal pressure p_s among the varying physiological parameters. The parameter p_s is known to have a major influence on relevant voice source parameters, in particular the phonation fundamental frequency is known to increase almost linearly with p_s [16]. For this reason the physiological control space should be expanded to include p_s.

A second limitation of the results comes from the characteristics of the vocal tract load: neither the open glottis nor the cylindrical vocal tract configurations provide a realistic simulation of the load, while it is known that the load characteristics also influence relevant voice source parameters (e.g., the glottal flow skeweness). Better simulations of voice source/vocal tract interaction can be realized, see e.g. [17].

Finally, as already mentioned, the 2-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 [9], such behaviors are not easily simulated with a low-dimensional model.

Nonetheless, the preliminary results suggests that the proposed approach can be successfully used for voice source parameters matching applications. The following points can be mentioned.

First,the nuscle activation control space allow exploration of a wide region of the voice source parameter space. Second, with respect to our previous works [3], [7], this approach leads to more robust resynthesis, since no regressor-based black-box element is used and consequently stability is guaranteed by construction. Finally, the same approach can be extended to lower dimensional glottal models (e.g., [2]), in order to construct an efficient analysis/synthesis tool.

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 Phoniatr. Logop.*, 48(5):240–254, Sep. 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, Sep. 2001.

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

[4] D. Childers and C. Lee. Vocal quality factors: analysis, synthesis, and perception. *J. Acoust. Soc. Am.*, 90(5):2394–2410, November 1991.

[5] D. G. Childers and C. Ahn. Modeling the glottal volume-velocity waveform for three voice types. *J. Acoust. Soc. Am.*, 97(1):505–519, Jan. 1995.

[6] M. Döllinger, U. Hoppe, F. Hettlich, J. Lohscheller, S. Schuberth, and U. Eysholdt. Vibration parameter extraction from endoscopic image series of the vocal folds. *IEEE Trans. Biomedical Engineering*, 49(8):773–781, Aug. 2002.

[7] C. Drioli and F. Avanzini. Hybrid parametric physiological glottal modelling with application to voice quality assessment. *Medical Engineering & Physics*, 24(7–8):453–460, Sep. 2002.

[8] G. Fant, J. Liljencrants, and Q. guang Lin. A four-parameter model of glottal flow. *STL-QPSR*, 26(4):1–13, 1985.

[9] H. E. Gunter. A mechanical model of vocal-fold collision with high spatial and temporal resolution. *J. Acoust. Soc. Am.*, 113(2):994–1000, Feb. 2003.

[10] 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.

[11] P. J. Price. Male and female voice source characteristics: inverse filtering results. *Speech Commun.*, 8(2):261–277, February 1989.

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

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

[14] B. Story and I. Titze. Voice simulation with a body cover model of vocal folds. *J. Acoust. Soc. Am.*, 97:1249–1260, 1995.

[15] H. Strik. Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses. *J. Acoust. Soc. Am.*, 103(5):2659–2669, May 1998.

[16] I. R. Titze. On the relation between subglottal pressure and fundamental frequency in phonation. *J. Acoust. Soc. Am.*, 85(2):901–906, Feb. 1989.

[17] I. R. Titze and B. H. Story. Acoustic interactions of the voice source with the lower vocal tract. *J. Acoust. Soc. Am.*, 101(4):2234–2243, Apr. 1996.

[18] I. R. Titze and B. H. Story. Rules for controlling low-dimensional vocal fold models with muscle activation. *J. Acoust. Soc. Am.*, 112(3):1064–1027, Sep. 2002.