ForumAcusticum
2 0 0 5

Physically oriented glottis models with inverse filtered waveform matching properties

Carlo Drioli
Institute of Phonetics and Dialectology, ISTC-CNR, Via Anghinoni 10, 35121 Padova, Italy, e-mail: **drioli@pd.istc.cnr.it**

A low-dimensional physically oriented model of the glottal source is discussed. The model relies on a lumped mechano-aerodynamic scheme based on the mass-spring paradigm. The vocal folds are represented by a mechanical resonator plus a delay line which takes into account the vertical phase differences. First, a simple flow model based on Bernoulli's law is assumed, and the properties of the system are discussed. The class of models under consideration is shown to be able to reproduce a broad range of phonation styles, and to provide interesting control properties. Secondly, an extended flow model is introduced with the aim of reproducing realistic glottal source waveforms obtained by inverse filtering. The new flow model is based on a general parametric nonlinear model. For this new scheme, the principal characteristics of the flow-induced oscillations are retained, and the overall model is suited for an identification approach where real inverse filtered glottal flow signals are to be reproduced. A data-driven identification procedure is outlined, where the parameters of the model are tuned in order to accurately match the target waveform. A set of inverse-filtered glottal flow wave forms with different characteristics are used to test the effectiveness of the approach. The results demonstrate that the model can reproduce a wide range of target waveforms.

1 Introduction

A wide number of speech analysis, coding, and synthesis techniques, are based today on the source-filter model of speech production [1], in which the laryngeal excitation can be reasonably considered independent of vocal tract. The glottal source has been recognized to play an important role with respect to the quality of the synthesis [2, 3], and speech synthesis schemes and systems relying on extended models of the glottal excitation are now available (e.g., [4]). In many cases, analytical models [5] are the preferred choice. The use of physical glottal models has been evaluated as well, although the parametric fitting to sampled waveforms, required in many applications, has turned out to be a extremely complex task [6].

Numerical models of the voice source production based on the physiology of the vocal folds have been proposed since 1968 [7]. The first attempts to simulate the flow-induced oscillations of the vocal folds were based on a lumped-element model made of a single spring-mass oscillator driven by airflow from the lungs. An essential improvement to the one-mass model was proposed by Ishizaka and Flanagan, with their two-mass model [8]. A wide range of variations and improvements have been proposed since the introduction of the original spring-mass models. However, the increase of accuracy in the modeling has disadvantages as well, as the growth of computational complexity and the difficulty to fit the model to observed data, due to the large amount of parameters involved. This prevented the physical modeling approach from being extensively adopted in practical applications in spite of its intrinsic potentialities, and at today the principal motivations for physical modeling of the voice source remain the understanding and learning about the phonatory process.

This article describes an investigation on a class of source models which combine physical knowledge and data-driven parametric fitting to the aim of reproducing inverse filtered glottal flow waveforms. We describe a waveform-matched mathematical model of the glottis loosely inspired to the myoelastic-aerodynamic theory and the lumped mass-spring paradigm. The principal difference with respect to the original one- and multi-mass models is the oversimplification of the mechanical resonator and the inclusion of a parametric nonlinear component in the mechano aerodynamic loop. The design of this component relies on an data-driven identification scheme which allows the model oscillation to be fitted to a target volume velocity waveform.

The properties of a low-dimensional self-oscillating glottal model are first reviewed, and the improvements introduced to support the data-driven modelling approach are described. The parametric identification of the model and issues concerning the synthesis of voiced sounds are then discussed. The experimental section is finally dedicated to the fitting of different inverse filtered volume-velocity waveforms to the model.

2 Description of the glottal model

The choice of the glottis model structure is inspired by the lumped multi-mass paradigm and by the body-cover wave model by Titze [9]. The body-cover model perspective looks at the vocal fold motion as a surface wave that propagates along the body-cover from the bottom of the

glottis to the top. We refer here to a simplified version of the body-cover model, introduced in [10], in which a single mass-spring system represents the entrance to the glottis, and a transmission line is responsible for the phase delay between the entry and exit to the glottis. One-delayed-mass models of this kind, also explored in [11], showed to retain the principal characteristics of flow induced oscillation, despite its simplicity and computational efficiency. The mechanical system is described by a mass-spring oscillator equation $m\ddot{x}_1+r\dot{x}_1+kx_1 = F_m$, where x_1 is the lateral displacement of the vocal fold at the entrance of the glottis, m, r, and k are respectively the mass, damping factor, and stiffness, and F_m is the force that drives the folds. The full details of the equations describing the model can be found in [10], and we focus here only on the design of the flow model.

The areas at entry and exit of the glottis can be respectively defined as

$$a_1(t) = 2L(x_{01} + x_1(t)) \quad (1a)$$

$$a_2(t) = 2L(x_{02} + x_1(t) - \tau\dot{x}_1(t)) \quad (1b)$$

where L is the length of the glottis, x_{01} and x_{02} are the rest positions of masses at entrance and exit to the glottis, and $\tau = T/c_f$, with T being the thickness of the folds and c_f being the wave velocity on the fold surface, is the time taken by the wave to propagate from the entrance to the upper end of the glottis. In the discrete-time implementation of the model, the wave propagation on the fold surface is represented by a simple delay line. Given the

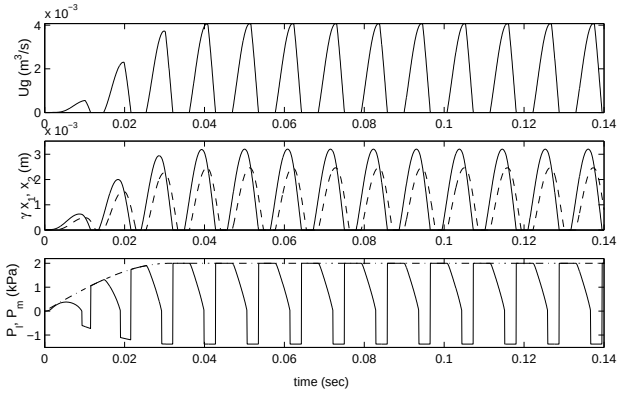

Figure 1: Simulation of the one-mass model. Upper panel: U_g, middle panel: γx_1 (solid) and x_2 (dashed), lower panel: lung pressure P_l (dot-dashed) and fold driving pressure $P_m = F_m/S_m$ (solid), with S_m being the fold surface.

description of the glottis area and a driving lung pressure P_l, there is a wide number of possible choices to write simplified formulas for the flow. The most simple one is derived from the stationary Bernoulli's law and assumes that the flow is proportional to the glottal area:

$$U_g = \sqrt{\frac{2P_l}{\rho}} \min\{\gamma a_1, a_2\}, \quad (2)$$

where $\gamma > 1$ is a term used to take into account that the flow separation point is located at the exit of the glottis in the case of a convergent glottis, or may move within the glottis, as the folds assume a divergent configuration. Figure 1 shows an example of the dynamical behavior of the system at the oscillation onset. This model has also shown to be able to simulate a wide range of voice qualities, such as pressed, breathy and bifurcated phonation [12].

2.1 Fitting inverse-filtered glottal flow waveforms with the one-mass model

The simple one-mass model described in the previous section can be improved with some extensions which permit to approach the modelling of the flow waveform with a data-driven waveform matching perspective. The structural extension consists in the definition of a more general flow model, i.e., in place of Eq. (2), the following parametric function is used:

$$U_g = w_0 \cdot \sqrt{P_l} \min\{\gamma x_1, x_2\} + \sum_{i=1}^{M} w_i \psi_i(x_1, x_2) \quad (3)$$

where w_i are weights to be identified, and $\psi_i(x_1, x_2)$ are nonlinear regressors of the input data. The regressors are functions that can be used to combine the inputs in a nonlinear fashion. The choice of the regressors can be made in several ways. Local models, such as gaussian functions or any other radial basis function, are often used. This approach leads to a model called *Radial Basis Function Network* (RBFN), and is adopted here.

Given a target flow waveform obtained by standard inverse-filtering procedures, the parameters of (part of the) model are first adapted in a way that the area function provided by the fold displacement is coherent with the target flow waveform. Then, the parametric model of the flow is designed to transform the area underlying the flow onto the actual flow waveform. The full details of the parametric identification procedure can be found in [10]. Here, only a brief description of the process is given:

1 Given the flow signal, a feasible lung pressure signal $P_l(n)$ is derived.

2 Considered the target flow and lung pressure as driving signals, the displacement of the fold edges are computed by running the mechanical part of the model, i.e., the fold driving pressure is computed and used to drive the mass-spring system. In this step, the

tuning of various parameters (i.e., the mass-spring parameters, the fold resting position, and the fold length) is required so to let the area function be coherent with the target flow.

3 Given the lung pressure and the displacement of the fold edges as input, and the desired flow as output, the set of parameters $\{w_0, w_1, \ldots, w_M\}$ of the flow model is identified by a LS algorithm performed on a steady-state portion of the flow signal (in general, a three-periods window was considered).

4 Once trained, the system is run with arbitrary control input (within the training range), and the stability of the oscillations is verified.

Items 1 to 3 are referred to as the analysis, or identification, step. Item 4 is referred to as the synthesis step. Figure 2, upper panel, shows how the tuning of the parameters representing the mechanical part of the model leads to the synchronization of the fold displacements with the flow waveform. In particular, the opening of the exit of the glottis (x_2 displacement) is synchronized with the beginning of the open phase of the flow waveform, and the closing of the entrance of the glottis (x_1 displacement) is synchronized with the beginning of the closed phase of the flow waveform. The tuning of the parameters at this stage is performed manually with an interactive procedure. Figure 2, mid and lower panels, shows the identification result of the flow model parameters by the LS algorithm, and demonstrates how the extended flow model is able to transform the rather crude representation of the fold displacements into the non-trivial waveform of the target flow.

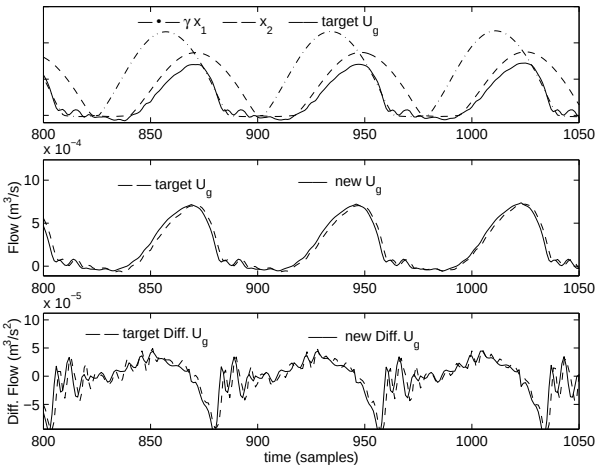

Figure 2: Open-loop identification result. From top to bottom: the fold edges displacement compared to the target flow; the target flow and the reproduced flow; the target flow rate and the reproduced flow rate.

2.2 Results and discussion

Figures 3 and 4 show the details of the self-oscillation properties of the model after training. The target flow period (upper-left panel), the reproduced flow (lower panel), and the comparison of the two in the frequency domain, are shown for two voice source examples. The comparison in the frequency domain was made using an auditory excitation pattern representation of the physical spectrum [13]; the auditory model transforms the physical spectrum into a pattern of specific loudness as a function of critical band rate, and has previously been found to efficiently predict perceived differences between vowel sounds [14]. These examples show how the system successfully adapts the parametric flow model so as to fit different flow shapes, while maintaining the original self oscillatory properties. A good matching of the excitation patterns is observed, at least in the lower part of the erb scale.

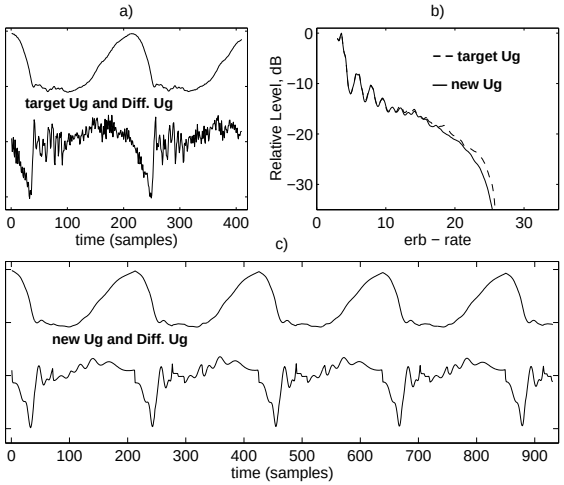

Figure 3: Target flow (panel a)) from a male speaker uttering a sustained vowel ($F_0 = 104$ Hz). Panel b): comparison of the excitation patterns of the target flow (dashed line) and reproduced flow (continuous line). Panel c): self-sustained oscillation produced by the model after training.

3 Conclusions

A vocal fold model based on a one-mass scheme and enhanced with a data-driven identification component, was described. First, the properties of an oversimplified one-mass model were explored, and it was shown how flow-induced oscillations can be produced even without vocal tract load, provided that a vertical phase delay is reproduced with a propagation line. The resultant discrete scheme has the advantage of simplicity and of being computationally efficient. Then, the simple one-mass model

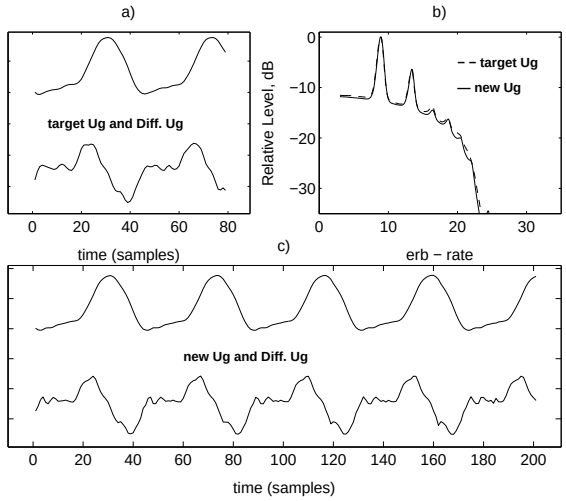

Figure 4: The target flow (panel a)) is from a female speaker uttering a sustained vowel ($F_0 = 372$ Hz). Panel b): comparison of the excitation patterns of the flow (dashed line: target, continuous line: reproduced). Panel c): self-sustained oscillation produced by the model after training.

was improved with a parametric model of the flow, and a data-driven analysis/synthesis procedure was described, that allows the model to fit to an arbitrary target flow. The training with respect to different flow waveforms demonstrates the versatility of the model, and its potential to represent a wide class of source flow signals. The advantage of using a physically-based description of the vocal folds, as we do here, is that the fitting accuracy provided by the parametric component is coupled with physically consistent dynamical behaviors and control properties.

For this approach to be useful for practical applications, e.g., for speech synthesis or speech coding, some points still require to be investigated: an algorithmic procedure to identify the parameters with physical interpretation is necessary and has to deal with the problem of non-uniqueness of the solution. Moreover, analytical conditions for closed-loop stability after the flow model fudge factor parametric identification remain an open issue, and the definition of constraints to account for during training, should be addressed.

References

[1] G. Fant, *Acoustic theory of speech production*, The Hague: Mouton, 1960.

[2] J. N. Holmes, "The influence of the glottal waveform on the naturalness of speech from a parallel formant synthesizer," *IEEE Trans. Audio Electroacoust.*, vol. AU-21, pp. 298–305, 1973.

[3] N. B. Pinto, D. G. Childers, and A. L. Lalwani, "Formant speech synthesis: improving production quality," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-37, no. 12, pp. 1970–1887, December 1989.

[4] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoust. Soc. Am.*, vol. 87, no. 2, pp. 820–857, February 1990.

[5] G. Fant, J. Liljencrants, and Q. Lin, "A four-parameter model of glottal flow," *STL-QPSR*, pp. 1–13, 1985.

[6] J. Schroeter and M. Sondhi, *Advances in Speech Signal Processing*, chapter Speech coding based on physiological models of speech production, pp. 231–263, Dekker, New York, 1992.

[7] J. L. Flanagan and L. L. Landgraf, "Self-oscillating source for vocal-tract synthesizers," *IEEE Trans. Audio and Electroacoustics*, vol. AU-16, pp. 57–64, 1968.

[8] K. Ishizaka and J. L. Flanagan, "Synthesis of voiced sounds from a two-mass model of the vocal cords," *The Bell Syst. Tech. J.*, vol. 51, no. 6, pp. 1233–1268, July-August 1972.

[9] I. R. Titze, "The physics of small-amplitude oscillations of the vocal folds," *J. Acoust. Soc. Am.*, vol. 83, no. 4, pp. 1536–1552, April 1988.

[10] C. Drioli, "A flow waveform-matched low-dimensional glottal model based on physical knowledge," *J. Acoust. Soc. Am.*, vol. 117, no. 5, pp. 3184–3195, May 2005.

[11] F. Avanzini, P. Alku, and M. Karjalainen, "One-delayed-mass model for efficient synthesis of glottal flow," *Proc. of Eurospeech Conf*, pp. 51–54, September 2001.

[12] C. Drioli and F. Avanzini, "Non-modal voice synthesis by low-dimensional physical models," *in Proc. 3rd Int. Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA)*, 2003.

[13] B.C.J. Moore and B.R. Glasberg, "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," *J. Acoust. Soc. Am.*, vol. 74, no. 3, pp. 750–753, September 1983.

[14] P. Rao, R. van Dinther, R. Veldhuis, and A. Kohlraush, "A measure for predicting audibility discrimination thresholds for spectral envelope distortions in vowel sounds," *J. Acoust. Soc. Am.*, vol. 109, no. 5, pp. 2085–2097, May 2001.