A discrete-time model for the vocal folds

Gabriel E. Galindo, Matías Zañartu, Juan I. Yuz

***Abstract*— Numerical models of the vocal folds are broadly used in studies that aim to better understand the underline mechanism of normal and disordered speech. Lumped element models capture the most predominant modes of vibration with a low computational cost, with the body cover model being one of the most common. Model parameter extraction from vocal fold movement is possible using system identification techniques, an approach that has been used to study different pathologies and vocal behaviors. However, many estimation techniques require a discrete state space model, this force a discrete solution of the kinematic model. In this study, a discrete-time model was developed based on a continuous state space description of the body cover model. A comparison with the continuous solution of the original body cover model was performed in order to quantify the differences in the corresponding outputs. Additionally, a clinical comparison was performed to explore if the results obtained agree with clinical data. Finally, a computational cost comparison was performed. The results obtained validate the proposed time-discrete model, yielding a normalized correlation index of 0.9786 between both continuous and discrete models in steady state. The clinical data agreed the results obtained with the proposed model, and the computing time was decreased in 60%.**

I. INTRODUCTION

Vocal fold models have been used for many years to study the behavior of normal and pathological phonation. Different approaches have been taken to identify different characteristics of speech. Physical models and excised larynx emulate the human phonation process, allowing for a visualization that is hardly achievable on in-vivo recordings. Finite difference elements models are a good approximation to the kinematics of the vocal folds, but they are computationally expensive and have numerical problems when considering acoustic resonances, collisions, and fluid mechanics. Lumped element models, on the other hand, are simple representations of the vocal folds that efficiently capture the most predominant modes of vibration and reproduce many clinically relevant aspects of phonation with acceptable accuracy at a lower computational cost.

Lumped element vocal fold models consist of a collection of discrete coupled mass-spring-damper components that interact with an aerodynamic force and an acoustic load. Lumped element models have been capable of emulating the vocal fold kinematic and acoustic output in physiological ranges, and have been used to describe normal voice production, vocal fold paralysis, incomplete glottal closure, vocal hyperfunction, subject pathology classification, etc [1]. The

Department of Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso - Chile. Email: gabriel.e.galindo@gmail.com

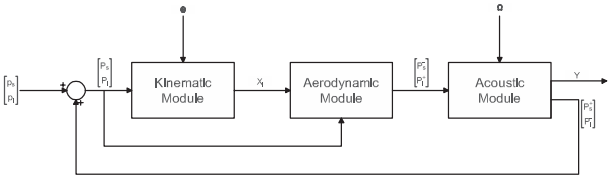

Fig. 1. Voice Production System Representation.

Lumped element model is described as a collection of non-linear time-variant differential system of equations, which is usually solved through ordinary differential equation (ODE) solvers like Runge-Kutta, Adams, Rosenbrock, etc. to obtain a "continuous" solution.

The continuous solution of the differential equations is not strictly needed due to the time discretization required in the vocal tract propagation models [2]. In addition, a discrete-time state space model allows for the application of Bayesian parameter identification techniques that can provide uncertainties for each estimated parameter[3]. Previous work in vocal fold parameter estimation [4] have not used Bayesian estimation to obtain a subject-specific parameters. Instead, an optimization scheme was used to classify the vocal behavior on a parameter based space distribution.

The goal of this study is to develop a new discrete-time solution method for the body cover model (BCM) [5], that provides a discrete state space model representation to enable subsequent studies on Bayesian parameter identification. The paper begins with a description of the voice production system for this study (II-A), followed by the analysis and modification of the lumped element model (II-B), and discrete state space model representation of the BCM (II-C). The simulation results are presented in Section III, along with a brief overall conclusion of the study in Section IV.

II. METHODS

A. Voice Production Structure

The voice production process could be separated as three interactive systems: An aerodynamic module, a tract propagation module, and an acoustic module. These three modules work as interconnected systems governed by a set of parameters controlled by the underlying physiological structure. In the present study, an open loop approach has been assumed due to the lack of information on the biological feedback that controls to the model parameters (figure 1).

In this scheme, the kinematic module is the system that describes vocal folds displacements and velocities using as input the incident pressures from the sub-glottal tract and the supra-glottal tract. The aerodynamic module solves for the

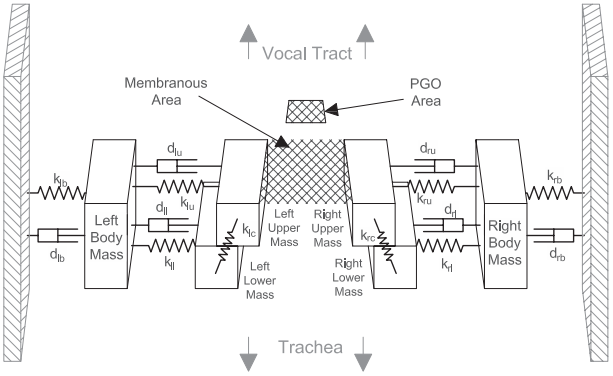

Fig. 2. Body Cover Model representation showing the vocal folds masses, the posterior glottal opening and the membranous area.

glottal airflow given the incident pressures in the glottis and the kinematic glottal configuration, which in turn produces an acoustic output. Finally, the acoustic module propagates the resulting sound from the glottis to the mouth, nose, or any other point in the sub/supra glottal tracts

B. Numerical Model Selection

Three-mass body-cover model [5] is an extension of the classical two-mass model [6] that is broadly accepted to study different glottal behaviors[1]. A schematic of the model representation is shown in Fig. 2. The BCM represents key physiological aspects of the vocal folds, and has been used to study source-filter interaction [7], [8], voice pathologies [9], [10], inverse filtering [11], and muscle activation [12], among others.

The model parameters were selected to produce a male modal voice using muscle activation principles [13], by selecting a 10 % cricothyroid and 20 % thyroarytenoid muscle activation. A wave-reflection analog scheme [2] was used to account for sound propagation and interaction, based on a sustained vowel /e/ [14]. A posterior glottal opening was included with an area of $5mm^2$ [15]. The subglottal area function was adapted from respiratory system measurements of human cadavers [16] and includes the trachea, bronchi, and a resistive termination impedance (zero$^{\text{th}}$ and first airway generations). The speed of sound was set at 350 m/s, simulation time 200 ms, sampling frequency 70 kHz, subglottal pressure 800 Pa, and Bernoulli is used as the flow solution.

The disposition of the forces has a small difference with the original publication of the BCM. In this study a separation of the static and the dynamic systems was made, obtaining the following force structure:

$$F_{u,t}(\theta,\varphi_t) = F_{uStat,t}(\theta) + F_{uDyn,t}(\theta,\varphi_t) \quad (1)$$
$$F_{l,t}(\theta,\varphi_t) = F_{lStat,t}(\theta) + F_{lDyn,t}(\theta,\varphi_t) \quad (2)$$
$$F_{b,t}(\theta,\varphi_t) = F_{bStat,t}(\theta) - F_{bDyn,t}(\theta,\varphi_t) \quad (3)$$

where the subindex "Stat" denotes the part of the system that does not change due to kinematic movement, and "Dyn" denotes the system components that change due to vocal fold dynamics (convergent or divergent vocal folds, collisions of the upper or lower masses, and complete closed glottis), generating a dynamic time varying system that swaps between 5 different configurations.

$$F_{uStat,t}(\theta) = F_{ku,t} + F_{du,t} - F_{kc,t} \quad (4)$$
$$F_{lStat,t}(\theta) = F_{kl,t} + F_{dl,t} + F_{kc,t} \quad (5)$$
$$F_{bStat,t}(\theta) = F_{kb,t} + F_{db,t} - F_{ku,t} - F_{du,t} - F_{kl,t} - F_{dl,t} \quad (6)$$
$$F_{uDyn,t}(\theta) = F_{eu,t} + F_{kuCol,t} + F_{duCol,t} \quad (7)$$
$$F_{lDyn,t}(\theta) = F_{el,t} + F_{klCol,t} + F_{dlCol,t} \quad (8)$$
$$F_{bDyn,t}(\theta) = F_{duCol,t} + F_{dlCol,t}. \quad (9)$$

The sub index "t" denotes a continuous temporal variable, θ is the collection of all the model parameters, which in this simulations were assumed to be static (and thus omitted in most of the forces expressions), and φ_t is the input pressure from the vocal tract. The forces that influence each mass have the following description:

- $F_{ku,t}$, $F_{kl,t}$, $F_{kb,t}$, $F_{kc,t}$: Spring forces (respectively: body-upper mass, body-lower mass, body-Thyroid wall, upper-lower mass)
- $F_{du,t}$, $F_{dl,t}$, $F_{db,t}$: Damping forces (respectively: body-upper mass, body-lower mass, body-Thyroid wall)
- $F_{kuCol,t}$, $F_{klCol,t}$: Additional spring force during collision (respectively: left upper mass - right upper mass, left lower mass - right lower mass)
- $F_{duCol,t}$, $F_{dlCol,t}$: Additional damping force during collision (respectively: body-upper mass, body-lower mass)
- $F_{eu,t}$, $F_{el,t}$: force due to the incident pressure on the glottis (respectively: upper mass, lower mass)

The expression for the spring forces, the damping forces, and the force due to the incident pressure from the tract are as defined in [5], with the modification of the collision damping force that is now separated in an additional force such as:

$$F_{duCol}(t) = -d_{uCol}\left(\dot{x}_{u,t} - \dot{x}_{b,t}\right) \quad (10)$$
$$F_{dlCol}(t) = -d_{lCol}\left(\dot{x}_{l,t} - \dot{x}_{b,t}\right) \quad (11)$$
$$d_{uCol} = 2\zeta_{uCol}\left(m_u k_u\right)^{1/2} \quad (12)$$
$$d_{lCol} = 2\zeta_{lCol}\left(m_l k_l\right)^{1/2}, \quad (13)$$

which is the same as in [5], but expressed in a more convenient way for the purpose of this study.

C. Discrete State Space Model

The use of state space models (SSM) brings a set of tools for analysis and system identification that are hardly achievable with other methods, such as energy representation, stability and causality analysis, perturbations analysis, feedback control, etc. The SSM representation of the system is not unique. Thus, a physically meaningful mechanical approach has been used to maintain the structure of the BCM, thus obtaining a model of the following state vector:

$$X_t = \begin{bmatrix} v_{u,t} & v_{l,t} & v_{b,t} & x_{u,t} & x_{l,t} & x_{b,t} \end{bmatrix}^T, \quad (14)$$

where the state equation (considering a deterministic model), is a function of the state, the model parameters, and the incident pressure, thus being represented by $\dot{X}_t = \mathcal{F}(X_t, \theta, \varphi_t)$. The particular non-linear state function is:

$$\mathcal{F}(X_t, \theta, \varphi_t) = \begin{bmatrix} \frac{1}{m_u} F_u(X_t, \varphi_t, \theta) \\ \frac{1}{m_l} F_l(X_t, \varphi_t, \theta) \\ \frac{1}{m_b} F_b(X_t, \varphi_t, \theta) \\ v_u(t) \\ v_l(t) \\ v_b(t) \end{bmatrix} \tag{15}$$

The advantages of a SSM also apply to a discrete state space model (DSSM), which can use methods of system parameter estimation designed for discrete-time models such as Sequential Montecarlo methods and other Bayesian estimation methods[3]. To produce a discrete-time model, a Taylor Series approach has been used [17], with the following expression:

$$X_{k+1} = X_{k\Delta_T} + \Delta_T \dot{X}_{k\Delta_T} + \frac{\Delta_T^2}{2!} \ddot{X}_{k\Delta_T} + \ldots, \tag{16}$$

where Δ_T is the sampling time period used for discretization. The truncated series results in the following DSSM approximation:

$$\tilde{X}_{k+1} \approx \begin{bmatrix} v_{u,k} + \frac{\Delta_T}{m_u} F_{u,k}(\tilde{X}_k, \varphi_k, \theta) \\ v_{l,k} + \frac{\Delta_T}{m_l} F_{l,k}(\tilde{X}_k, \varphi_k, \theta) \\ v_{b,k} + \frac{\Delta_T}{m_b} F_{b,k}(\tilde{X}_k, \varphi_k, \theta) \\ x_{u,k} + \Delta_T v_{u,k} + \frac{\Delta_T^2}{2m_u} F_{u,k}(\tilde{X}_k, \varphi_k, \theta) \\ x_{l,k} + \Delta_T v_{l,k} + \frac{\Delta_T^2}{2m_l} F_{l,k}(\tilde{X}_k, \varphi_k, \theta) \\ x_{b,k} + \Delta_T v_{b,k} + \frac{\Delta_T^2}{2m_b} F_{b,k}(\tilde{X}_k, \varphi_k, \theta) \end{bmatrix} \tag{17}$$

$$= \mathcal{F}(\tilde{X}_k, \varphi_k, \theta, \Delta_T). \tag{18}$$

D. Selected measures of vocal function

A set of measures were selected to evaluate the behavior of the proposed model. Both functional and numerical behavior were analyzed to study the accuracy of the proposed model (17 - 18). The selected acoustic parameters were compared to the results obtained in [18] to validate the clinical relevancy of this model. Relevant clinical parameters were computed, including fundamental frequency, maximum flow declination rate (MFDR), radiated sound pressure level (SPL), and steady and unsteady glottal airflow components. SPL was obtained at the lips and was projected to a 15 cm distance by subtracting 30 dB, based upon empirical observations.

An algorithm to evaluate the simulation time was developed in order to characterize the complexity of the proposed model. The algorithm used a loop with increased time simulation on each iteration, simulating equal scenarios with both models and recording the time that each model used to compute each iteration. The result was normalized by the maximum simulation time achieved.

III. RESULTS

The simulations were performed, with model parameters selected from [13]. A simulation period of 200ms was used and a 5ms ramp (made of half Hanning window) was applied to the initial section of the sub glottal pressure in order to obtain a smooth transient period. The selected waveforms for the non-gap situation are presented in figure 3 (a) and (b). A transient portion and a steady state portion are shown. The transient portion shows a clear decoupling of the two systems, which is attributed to the interaction between the non-iterative solution of the discrete model and the dynamical structure of the model. While in the stationary period a delay is evident, no modification of the signal shape can be noticed, and a correlation analysis showed a maximum normalized value of 0.9786 on a window of 100ms.

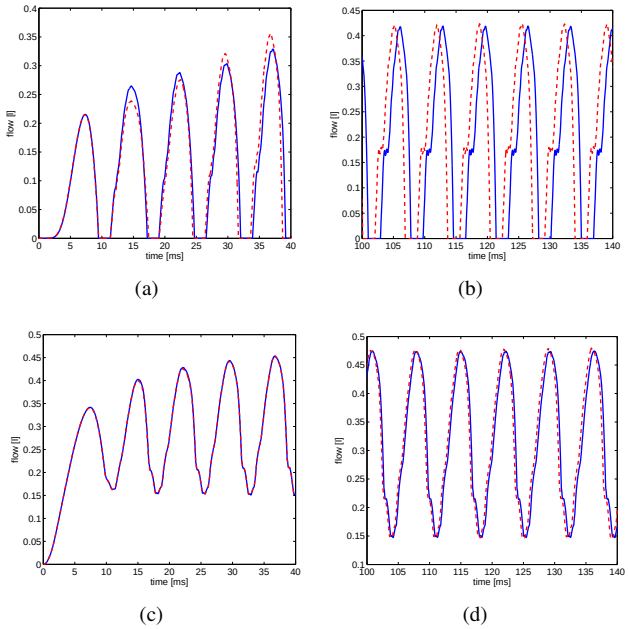

Fig. 3. Glottal flow with continuous and discrete models (– Continuous Model, -- Discrete Model). (a) Transient period without PGO, (b) Steady state without PGO, (c) Transient period with PGO of $5mm^2$, (d) Steady state with PGO of $5mm^2$

A simulation with a PGO of $5mm^2$ was also performed (fig 3 (c) and (d)). In this case scenario the transient period had a smaller decoupling behavior in comparison with the non gap scenario, which could be explained by the smaller amount of energy involved in the motion of the masses due to the constant leakage of sub glottal pressure. As a consequence, a smaller delay could be noticed in the stationary period, with a maximum normalized correlation value of 0.9940 on a window of 100ms.

Selected clinical measures for both simulations were obtained using the stationary part of the signal, as showing in tables I and II. It can be noticed that the selected measures are in the range of a normal male subject, thus validating the discrete-time model proposed here. Also, a reduction of the processing time was achieved for a simulation of 2 seconds of speech, by saving up to 60% of the computation time compared with the continuous solution (ODE4 solver from Matlab).

Parameter (Unit)	Continuous Model	Discrete Model	Perekell 1993 Loud (SD)
spl (dB)	86.0	86.0	85.6(4.7)
F_0 (Hz)	136.7	136.7	128.9(24.7)
mfdr ($^{l}/_{s^2}$)	807.9	828.4	650.2(251.2)
ac flow ($^{l}/_{s}$)	0.42	0.42	0.47(0.14)
min flow ($^{l}/_{s}$)	0.00	0.00	0.08(0.05)

TABLE I
COMPARISON BETWEEN MODELS AND CLINICAL DATA [18] FOR A NON PGO PRESENCE

Parameter (Unit)	Continuous Model	Discrete Model	Perekell 1993 Normal (SD)
spl (dB)	77.0	78.0	77.8(4.0)
F_0 (Hz)	136.7	136.7	112.4(11.8)
mfdr ($^{l}/_{s^2}$)	408.5	412.9	337.2(127.2)
ac flow ($^{l}/_{s}$)	0.33	0.33	0.33(0.07)
min flow ($^{l}/_{s}$)	0.15	0.15	0.08(0.05)

TABLE II
COMPARISON BETWEEN MODELS AND CLINICAL DATA [18] FOR A PGO OF $5mm^2$

IV. CONCLUSIONS

The proposed model offers a fast and sufficiently accurate solver solution to obtain the vocal folds kinematics, offering a discrete-time state space model of the vocal folds that enables the use of Bayesian techniques for parameter identification. These new tools for system identification can be used, for example, to estimate model parameters from high-speed video visualizations, aerodynamic and acoustic measurements. The proposed model provides good results for stationary and quasi stationary behavior. Such new clinical approach to identify the structural parameters of the vocal folds can assist not only in the diagnosis of vocal fold pathologies or vocal hyperfunction, but also in the prediction of results in patient specific voice treatment (therapy and/or surgery).

The specific expressions obtained here should not be directly extrapolated to other vocal fold models due to the assumptions made in the specific model that was selected. The use of a non-realistic one dimensional displacement and the lack of spatial resolution in the anterior-posterior glottal shape, support the need for a discrete-time model with more degrees of freedom (lateral, inferior-superior, anterior-posterior). Additionally, non-stationary scenarios must be evaluated to quantify the importance of the differences shown here. However, the lack of restriction in the discretization process, and the general mathematical approximation provided by the truncated Taylor series, provide a good perspective for the additional analysis required in more complex scenarios.

ACKNOWLEDGMENT

Gabriel Galindo acknowledges the support of CONICYT doctoral scholarship and UTFSM doctoral scholarship.

REFERENCES

[1] B. D. Erath, M. Zañartu, K. C. Stewart, M. W. Plesniak, D. E. Sommer, and S. D. Peterson, "A review of lumped-element numerical models of voiced speech," *Speech Comm.*, 2012, submitted for review on September 12, 2012.

[2] B. H. Story, "Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract," Ph.D. dissertation, the University of Iowa, Iowa City, IA, 1995.

[3] J. Candy, *Bayesian Signal Processing: Classical, Modern and Particle Filtering Methods*, ser. Adaptive and Cognitive Dynamic Systems: Signal Processing, Learning, Communications and Control. Wiley, 2009.

[4] T. Wurzbacher, R. Schwarz, M. Döllinger, U. Hoppe, U. Eysholdt, and J. Lohscheller, "Model-based classification of nonstationary vocal fold vibrations," *J. Acoust. Soc. Am.*, vol. 120, pp. 1012–1027, 2006.

[5] B. H. Story and I. R. Titze, "Voice simulation with a body-cover model of the vocal folds," *J. Acoust. Soc. Am.*, vol. 97, pp. 1249–1260, 1995.

[6] K. Ishizaka and M. Matsudaira, "Fluid mechanical considerations of vocal fold vibration," in *Speech Communication Research Laboratory, Mongraph No. 8*, Santa Barbara, CA, 1972.

[7] I. R. Titze and B. H. Story, "Acoustic interactions of the voice source with the lower vocal tract," *J. Acoust. Soc. Am.*, vol. 101, no. 4, pp. 2234–2243, 1997.

[8] I. R. Titze and A. S. Worley, "Modeling source-filter interaction in belting and high-pitched operatic male singing," *J. Acoust. Soc. Am.*, vol. 126, no. 3, pp. 1530–1540, 2009.

[9] J. Kuo, "Voice source modeling and analysis of speakers with vocal-fold nodules." Ph.D. dissertation, Harvard-MIT Division of Health Sciences and Technology, 1998.

[10] D. D. Mehta, M. Zañartu, T. F. Quatieri, D. D. Deliyski, and R. E. Hillman, "Investigating acoustic correlates of human vocal fold phase asymmetry through mathematical modeling and laryngeal high-speed videoendoscopy," *J. Acoust. Soc. Am.*, vol. 130, pp. 3999–4009, 2011.

[11] P. Alku, C. Magi, S. Yrttiaho, T. Bäckström, and B. Story, "Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering," *J. Acoust. Soc. Am.*, vol. 125, no. 5, pp. 3289–3305, 2009.

[12] I. R. Titze, "Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model," *J. Acoust. Soc. Am.*, vol. 111, pp. 367–376, 2002.

[13] I. R. Titze and B. H. Story, "Rules for controlling low-dimensional vocal fold models with muscle activation," *J. Acoust. Soc. Am.*, vol. 112, pp. 1064–1076, 2002.

[14] H. Takemoto, K. Honda, S. Masaki, Y. Shimada, and I. Fujimoto, "Measurement of temporal changes in vocal tract area function from 3D cine-MRI data," *J. Acoust. Soc. Am.*, vol. 119, pp. 1037–1049, 2006.

[15] M. Zañartu, J. C. Ho, D. D. Mehta, R. E. Hillman, and G. R. Wodicka, "Acoustic coupling during incomplete glottal closure and its effect on the inverse filtering of oral airflow," *J. Acoust. Soc. Am.*, vol. POMA 19, pp. 1–7, 2013.

[16] E. R. Weibel, *Morphometry of the Human Lung*. New York: Springer, 1963.

[17] J. Yuz and G. Goodwin, *Sampled-Data Models for Linear and Nonlinear Systems*, ser. Communications and Control Engineering. Springer London, 2013.

[18] J. S. Perkell, R. E. Hillman, and E. B. Holmberg, "Group differences in measures of voice production and revised values of maximum airflow declination rate," *J. Acoust. Soc. Am.*, vol. 96, no. 2, pp. 695–698, 1994.