JASA EXPRESS LETTERS

https://doi.org/10.1121/10.0001276

Bayesian estimation of vocal function measures using laryngeal high-speed videoendoscopy and glottal airflow estimates: An *in vivo* case study

Gabriel A. Alzamendi,[1] Rodrigo Manríquez,[1] Paul J. Hadwin,[2] Jonathan J. Deng,[2] Sean D. Peterson,[2] Byron D. Erath,[3] Daryush D. Mehta,[4] Robert E. Hillman,[4] and Matías Zañartu[1,a)]

[1]*Department of Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso 2390123, Chile*

[2]*Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, Ontario N2L 3G1, Canada*

[3]*Department of Mechanical and Aeronautical Engineering, Clarkson University, Potsdam, New York 13699, USA*

[4]*Center for Laryngeal Surgery and Voice Rehabilitation, Massachusetts General Hospital, Boston, Massachusetts 02114, USA*

gabriel.alzamendi@usm.cl, rodrigo.manriquezp@usm.cl, pjhadwin@uwaterloo.ca, j8deng@uwaterloo.ca, peterson@uwaterloo.ca, berath@clarkson.edu, mehta.daryush@mgh.harvard.edu, hillman.robert@mgh.harvard.edu, matias.zanartu@usm.cl

Abstract: This study introduces the *in vivo* application of a Bayesian framework to estimate subglottal pressure, laryngeal muscle activation, and vocal fold contact pressure from calibrated transnasal high-speed videoendoscopy and oral airflow data. A subject-specific, lumped-element vocal fold model is estimated using an extended Kalman filter and two observation models involving glottal area and glottal airflow. Model-based inferences using data from a vocally healthy male individual are compared with empirical estimates of subglottal pressure and reference values for muscle activation and contact pressure in the literature, thus providing baseline error metrics for future clinical investigations.

[Editor: Brad H. Story] Pages: EL434–EL439

Received: 10 February 2020 **Accepted:** 3 May 2020 **Published Online:** 20 May 2020

1. Introduction

The clinical assessment of vocal function could be significantly enhanced by a better understanding of the underlying physical mechanisms of normal and disordered phonation. Recent studies using different inverse methods have produced subject-specific biomechanical models of the vocal folds (VFs) that provide access to relevant clinical features, such as subglottal pressure and VF contact pressure.[1–3] In this context, Bayesian estimation of subject-specific models of phonation is of particular interest, given that it takes into account the stochastic nature of the inverse problem, quantifies uncertainty in the form of confidence intervals, and naturally combines diverse signals.[4,5] However, this approach is relatively new and has not yet been investigated with multimodal *in vivo* clinical data.

In this study, we advance prior Bayesian efforts using an extended Kalman filter (EKF)[5] to now estimate model-based features using simultaneous *in vivo* recordings of laryngeal high-speed videoendoscopy (HSV) and oral volume velocity. An EKF is employed to estimate the activation levels of the cricothyroid (CT) and thyroarytenoid (TA) muscles, subglottal pressure, and contact pressure for a lumped-element body-cover model of the VFs. Two observation cases are considered to elucidate the effect of data aggregation in the Bayesian inference: case I uses the glottal area signal, whereas case II also includes the glottal airflow signal as part of the observation. The aim of this study is to illustrate the clinical application of the Bayesian method to provide access to relevant features that are difficult, if not impossible, to directly measure in some cases.

2. Methods

2.1 Data recording and calibration

The experimental setup allowed for the simultaneous recording of laryngeal high-speed videoendoscopy, radiated sound pressure, oral volume velocity (OVV), and intraoral pressure (IOP). A

[a)]Author to whom correspondence should be addressed.

https://doi.org/10.1121/10.0001276

transnasal fiberscope was used for flexible endoscopy, which allowed for simultaneous aerodynamic assessment and normal articulation. Videoendoscopy recordings were acquired at a frame rate of 4000 fps and a spatial resolution of 288 horizontal × 288 vertical pixels. The acoustic pressure was recorded using a head-mounted, high-quality condenser microphone situated approximately 4 cm from the lips. A circumferentially vented (Rothenberg) mask was modified in order to allow introducing the flexible endoscope, and to hold the OVV and IOP sensors. The IOP sensor was connected to a narrow tube inserted between the lips into the oral cavity. The analog signals were low-pass filtered (30 kHz cutoff frequency) and sampled at a 120 kHz sampling rate (16-bit quantization, and a +/−10 V dynamic range) that was synchronized with the video data using a common clock. All signals were subsequently resampled (including an antialiasing filter) at a 20 kHz sampling rate, and acoustic pressure and OVV signals were shifted backward in time to compensate for acoustic propagation. Additional details regarding this experimental setup are described in previous publications.[6,7]

For this case study, the data considered consisted of high-speed videoendoscopy and analog OVV and IOP signals corresponding to two sustained vowels (/a/ and /i/) and a repetitive /pa/ gesture from a male participant having no medical history of voice disorders. For both vowels, clinical data corresponding to segments of approximately 400 ms exhibiting stable VF oscillations were considered. OVV and IOP signals were calibrated in physical units using standard reference airflow and pressure levels, respectively.[7] The OVV signal was then inverse filtered to cancel out the vocal tract resonances, thus resulting in a calibrated glottal volume velocity (GVV) signal.[8] At the same time, video recordings were digitally processed to detect the glottal contour for every frame, and thereby the glottal area waveform (GAW) was computed using segmentation. Spatial calibration of the video-based GAW functions in physical units was achieved by identifying a reference laryngeal landmark (e.g., blood vessels patterns) near the glottis whose dimensions were measured independently using a calibrated endoscope system.[6] Finally, the GAW was resampled (interpolated) to a 20 kHz sampling rate, equivalent to the other analog signals.

Figure 1 illustrates a segment of a repetitive /pa/ gesture after calibration and inverse filtering procedures. In this common clinical exercise used to measure *in vivo* subglottal pressure, a steady sustained vowel is interrupted with a bilabial plosive that alters the pressure in the airways such that the IOP approximates the subglottal pressure. The IOP pressure peaks are extrapolated to compute the driving subglottal pressure during the intermediate voiced segments. In this study, we compare this standard lip occlusion method *in vivo* with the proposed Bayesian estimates that use the actual voiced segments information. This study is the first step toward the long-term aim of applying the proposed framework to obtain subglottal pressure and other measures of interest for various vocal gestures, where the direct measures cannot be obtained.

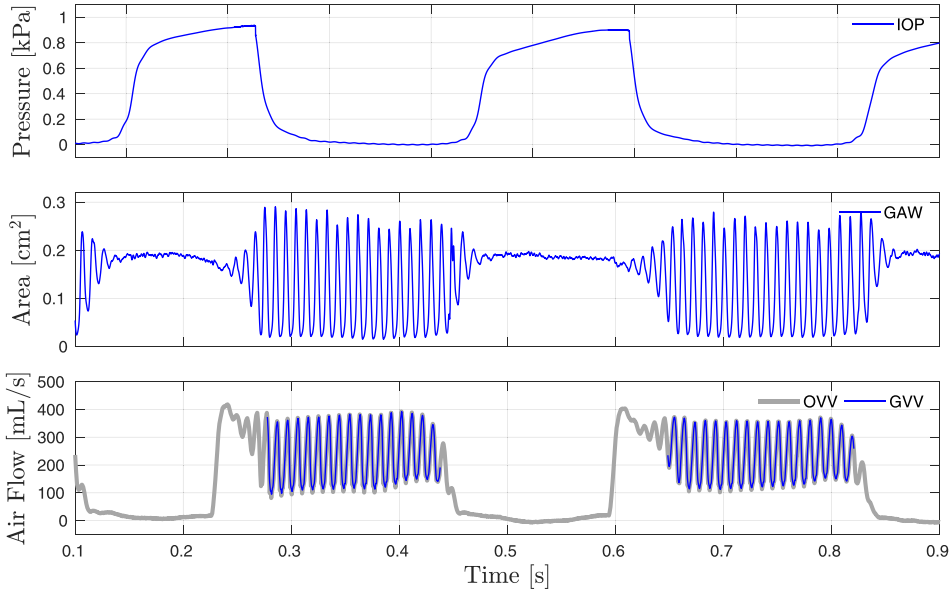

Fig. 1. (Color online) Clinical signals for a repetitive /pa/ gesture of the subject. Top: Intraoral pressure. Center: Glottal area waveform. Bottom: Oral (thick line) and inverse-filtered glottal (thin line) volume velocity. Differences in signal dynamics indicate voiced and plosive segments.

2.2 Bayesian estimation of vocal function measures

In comparison with other Bayesian estimators, the EKF performs very well for inferring vocal fold biomechanical variables at a dramatically lower computational cost.[5] Herein, a previous effort[5] is extended by incorporating the inference of activation level of the TA muscle, subglottal pressure, and VF contact pressure, as well as the observation of multimodal *in vivo* signals (GAW and GVV). Our EKF estimator uses a state-space model representation of a symmetric three-mass body-cover model with posterior glottal opening.[9] Cover masses delimited the membranous glottal area, A_m, whereas the posterior glottal opening, A_{PGO}, was assumed an unknown, constant parameter. Thus the total glottal area was defined as $A_g = A_m + A_{PGO}$.[9] Physiologically inspired rules[10] were applied for controlling the body-cover model parameters, where normalized activation levels for the CT and TA muscles (a_{CT} and a_{TA}, respectively) simulate muscle contraction effects. Activation of the lateral cricoarytenoid muscle was held constant at 0.5 to anchor the "just touching" vocal fold configuration.[10] The subglottal pressure, P_s, during phonation was inferred. The three-way interaction among tissue, flow, and sound was considered. Glottal volume velocity, U_g, was estimated as a function of A_g, P_s, and the acoustic waves impinging on the glottis.[11,12] Acoustic wave propagation through the vocal tract was modeled with the wave reflection analog method. To this end, known vocal tract area functions[13] were initially used, and subsequently tuned to match observed formant frequencies in the acoustic pressure signal using sensitivity functions.[14] The VF contact pressure, P_{CP}, refers to the non-linear collision term introduced in the body-cover model to account for the overlapping of the left and right cover masses during closure.[12]

The phonation process can be described for time index $n = 1, 2, \ldots, N$ through the state vector

$$\mathbf{x}[n] = (x_u[n],\, v_u[n],\, x_l[n],\, v_l[n],\, x_b[n],\, v_b[n],\, A_{PGO}[n],\, P_s[n],\, a_{CT}[n],\, a_{TA}[n],\, P_{CP}[n])^{\mathrm{T}}, \quad (1)$$

where x and v represent the positions and velocities of the upper (u), lower (l), and body masses (b), respectively. This state vector gathers the minimum set of variables describing or controlling the most important features of the selected voice production model. This formulation brings forth the Bayesian estimation of all the variables in Eq. (1), including four additional states (A_{PGO}, P_s, a_{TA}, and P_{CP}) not present in the prior Bayesian framework.[5] Two different observation models are investigated herein. Case I considers only the GAW as observable, whereas case II incorporates both GAW and GVV in the observation vector. The aim is to investigate the effects on the estimated phonatory parameters when combining observable data. State transition and measurement covariance matrices as well initial state information were set to promote stability and convergence in the estimation procedure. This tuning was performed only once (for vowel /a/) and maintained unchanged for all other scenarios.

3. Results and discussion

3.1 Inference from sustained vowels

Results obtained for the two observation cases are presented in Fig. 2 for vowels /a/ (left column) and /i/ (right column). The first and second rows show the measured GAW and computed GVV signals (solid lines), and the corresponding model approximations A_g and U_g extracted from the EKF (case I: dashed lines, case II: dash-dotted lines). Similarly, the estimated states P_s and P_{CP} are plotted in the third and fourth rows, respectively. The 95% confidence bounds are shown as shaded regions.

The approximated area waveforms illustrate that the proposed method captures the overall glottal area behavior for both observation cases reasonably well, with better tracking during the closing phase of the cycle. Root-mean-square errors for both cases were 0.03 cm^2 for vowel /a/, and 0.06 cm^2 for vowel /i/. Estimated A_{PGO} for all cases was less than 0.01 cm^2 for both vowels, in accordance with the negligible minimum area exhibited in the GAW signals in Fig. 2. Glottal flow pulses obtained in case II portrayed right-skewed shapes and superimposed fluctuations due to the non-linear source-filter coupling, in addition to a flat response during the closed phase. Root-mean-square errors for flow approximations were 40 mL/s for vowel /a/ and 55 mL/s for vowel /i/.

Estimates of subglottal pressure shown in Fig. 2 are nearly constant, with some observed fluctuations. This is more noticeable for case II, and could result from a better representation of the three-way interaction in the larynx with the added flow observation. The P_s contours are consistent with the fact that we are estimating the unsteady subglottal pressure. The rather constant value observed during the closed phase for this estimate can be associated with the driving lung pressure. The statistical information for the estimates of P_s is reported in Table 1. The estimated mean subglottal pressure significantly differs across observation cases,

JASA EXPRESS LETTERS
https://doi.org/10.1121/10.0001276

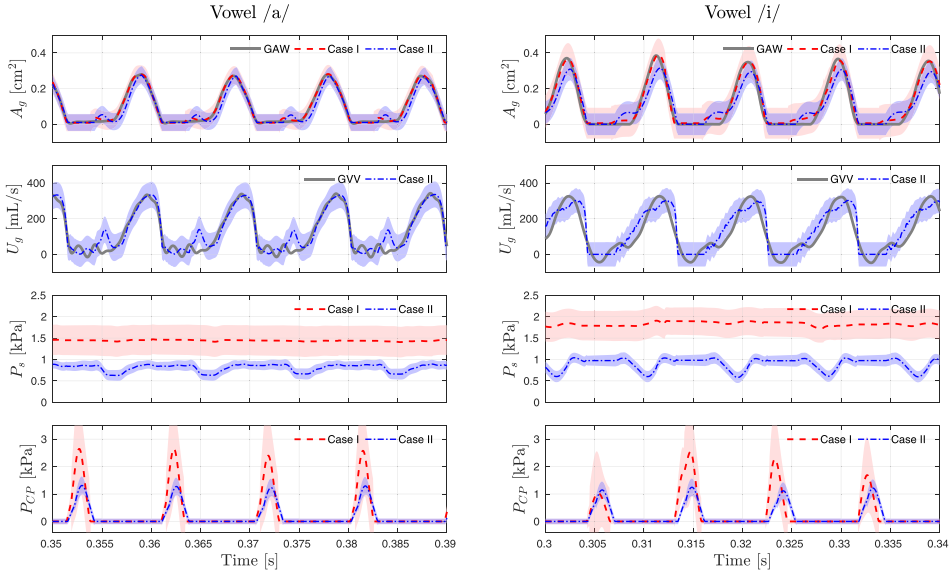

Fig. 2. (Color online) Observed data (solid line), model approximations, and estimates extracted from EKF and sustained vowel data for case I (dashed line) and case II (dash-dotted line). Shaded regions represent 95% confidence bounds.

with case I producing significantly higher pressure levels than case II and with higher uncertainty. Incorporating the additional volume velocity observation, however, reduces the estimates to the range of subglottal pressure levels reported for comfortable loudness during human phonation,[8] highlighting the importance of the additional observation for the given vocal function model.

Estimates of VF contact pressure in Fig. 2 for vowels /a/ and /i/ exhibit a peak value during the contact phase, with differences in magnitude and shape between the observation cases. The maximum peak contact pressure, P_{CP}^{+}, for vowels /a/ and /i/ are also reported in Table 1. Case I shows a high contact pressure peak for both vowels with high uncertainty, consistent with the higher subglottal pressure estimate. The additional flow observation in case II allows for a more robust and reliable behavior in the contact pressure estimate, and it better resembles other recent results.[1,3] Although it is difficult to assert that contact pressure estimates are physiologically accurate, the results illustrate that the proposed Bayesian processor is able to suitably aggregate clinical data and biomechanical modeling to give insights into VF collision.

Estimates of a_{CT} and a_{TA} for both observation cases are reported in Table 1. The estimates converged to constant values, as previously demonstrated for sustained phonation.[5] All estimates are in accordance with low muscle activation levels representative of conversational speech. For each case, no significant differences are observed between a_{CT} and a_{TA} for the two vowels. This can be partly explained by the fairly similar fundamental frequency (102.6 Hz for vowel /a/ and 107.0 Hz for vowel /i/). However, the differences are more significant between observation cases, with case II exhibiting lower a_{CT} and higher a_{TA} in comparison with case I. Direct *in vivo* validation for these estimates is difficult.

3.2 Analysis of a repetitive /pa/ gesture

The proposed Bayesian estimation method was also assessed in contrast with the standard clinical evaluation of subglottal pressure from repeated /pa/ gestures. Measured GAW and computed

Table 1. Mean (standard deviation) of subglottal pressure and muscle activation extracted from sustained vowel data for both observation cases. Maximum peak contact pressure pressure P_{CP}^{+} is also reported.

	Vowel /a/				Vowel /i/			
Case	P_s [kPa]	a_{CT} [—]	a_{TA} [—]	P_{CP}^{+} [kPa]	P_s [kPa]	a_{CT} [—]	a_{TA} [—]	P_{CP}^{+} [kPa]
I	1.22 (±0.16)	0.24 (±0.03)	0.19 (±0.03)	2.63 (±0.87)	1.64 (±0.16)	0.20 (±0.03)	0.27 (±0.03)	2.37 (±0.79)
II	0.81 (±0.06)	0.12 (±0.02)	0.37 (±0.02)	1.31 (±0.16)	0.91 (±0.06)	0.15 (±0.02)	0.41 (±0.02)	1.24 (±0.16)

JASA EXPRESS LETTERS

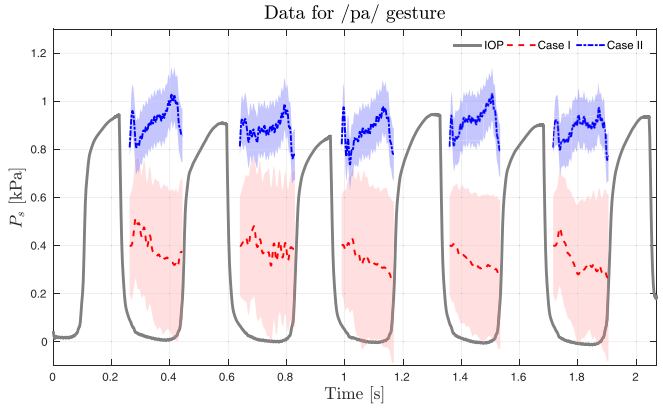

Fig. 3. (Color online) Comparison of IOP signal (solid line) for a repetitive /pa/ phonation, and estimated P_s obtained through EKF (case I: dashed line, case II: dash-dotted line). Shaded regions represent 95% confidence bounds.

GVV signals for the /pa/ gesture were segmented and every voiced segment was extracted. Bayesian inference was applied for each segmented section for the two observation cases, and the subglottal pressure was estimated. The voiced segments in this vocal gesture clearly exhibited incomplete glottal closure in the GAW signal (see Fig. 1), with an average A_{PGO} of 0.03 and 0.04 cm^2 for cases I and II, respectively. The IOP signal for five consecutive /pa/ gestures, and the corresponding P_s estimates for the two observation cases are shown in Fig. 3. All data were low-pass filtered at 80 Hz with a fifth-order Butterworth filter to enhance data visualization.[8]

Detailed information for the IOP pulses and the statistical information for P_s estimates are reported in Table 2. Similar to the simulations involving vowel data, the two observation cases produce significantly different subglottal pressure estimates. Case I seems to dramatically underestimate the subglottal pressure and to produce overly wide confidence intervals. On the other hand, case II takes advantage of the additional observation and improves the estimates and confidence interval bounds. The absolute relative error across the five segments is larger than 50% for case I, whereas for case II it is below 5.0%. Thus, the results suggest that the Bayesian processor requires multimodal information (observation) in order to return viable sugblottal pressure estimates for the model used in this study. Furthermore, this case study illustrates the potential of the Bayesian approach with *in vivo* clinical data, as the error in estimating subglottal pressure for our best-case scenario is smaller than recent studies with an excised larynx,[2] and comparable with those from silicone model experiments.[3] Future works will investigate the applicability of subglottal pressure estimation in more complex vocal gestures and with more comprehensive clinical data.

4. Conclusion

This single-subject study is, to the best of the authors' knowledge, the first attempt to apply the Bayesian inverse analysis framework to *in vivo* vocal fold data. This proof-of-concept illustrates that the proposed Bayesian framework with a lumped-element model of phonation can successfully fuse data from HSV and glottal airflow signals to produce meaningful estimates of clinically relevant variables that are difficult, if not impossible, to directly measure. This is the first step toward the long-term goal of applying the proposed framework to obtain clinical measures of interest, such as subglottal pressure, muscle activation, and vocal fold contact pressure, in running speech. The study highlights the sensitivity to the observation data, where both glottal area and glottal airflow were required to obtain robust and reliable estimates. Further studies

Table 2. IOP peak measures for five consecutive /pa/ emissions and mean (standard deviation) subglottal pressure P_s estimated through EKF for both observation cases. All the pressure values are reported in kPa.

	Segment 1	Segment 2	Segment 3	Segment 4	Segment 5	Average
IOP	0.94	0.91	0.85	0.95	0.90	0.91
Case I	0.39	0.39	0.34	0.34	0.34	0.36
	(±0.14)	(±0.15)	(±0.16)	(±0.14)	(±0.14)	(±0.15)
Case II	0.92	0.88	0.87	0.92	0.88	0.90
	(±0.05)	(±0.05)	(±0.05)	(±0.05)	(±0.05)	(±0.05)

https://doi.org/10.1121/10.0001276

involving data from a greater number of participants with normal and disordered voices are required to corroborate and extend the observations in this study.

Acknowledgments

This research was supported by the National Institutes of Health (NIH) National Institute on Deafness and Other Communication Disorders through Grants Nos. R01 DC007640 and P50DC015446, and by CONICYT/ANID through Grants Nos. FONDECYT 1191369 and BASAL FB0008. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References and links

[1]M. E. Díaz-Cádiz, S. D. Peterson, G. E. Galindo, V. M. Espinoza, M. Motie-Shirazi, B. D. Erath, and M. Zañartu, "Estimating vocal fold contact pressure from raw laryngeal high-speed videoendoscopy using a Hertz contact model," Appl. Sci. **9**(11), 2384–2405 (2019).

[2]P. Gómez, A. Schützenberger, S. Kniesburges, C. Bohr, and M. Döllinger, "Physical parameter estimation from porcine *ex vivo* vocal fold dynamics in an inverse problem framework," Biomech. Model. Mechanobiol. **17**(3), 777–792 (2018).

[3]P. J. Hadwin, M. Motie-Shirazi, B. D. Erath, S. D. Peterson, P. J. Hadwin, M. Motie-Shirazi, B. D. Erath, and S. D. Peterson, "Bayesian Inference of vocal fold material properties from glottal area waveforms using a 2D finite element model," Appl. Sci. **9**(13), 2735–2754 (2019).

[4]P. J. Hadwin, G. E. Galindo, K. J. Daun, M. Zañartu, B. D. Erath, E. Cataldo, and S. D. Peterson, "Non-stationary Bayesian estimation of parameters from a body cover model of the vocal folds," J. Acoust. Soc. Am. **139**(5), 2683–2696 (2016).

[5]P. J. Hadwin and S. D. Peterson, "An extended Kalman filter approach to non-stationary Bayesian estimation of reduced-order vocal fold model parameters," J. Acoust. Soc. Am. **141**(4), 2909–2920 (2017).

[6]D. D. Mehta, D. D. Deliyski, S. M. Zeitels, M. Zañartu, and R. E. Hillman, "Integration of transnasal fiberoptic high-speed videoendoscopy with time-synchronized recordings of vocal function," in *Technology, Vol. 1 of Normal and Abnormal Vocal Folds Kinematics: High Speed Digital Phonoscopy (HSDP), Optical Coherence Tomography (OCT) & Narrow Band Imaging (NBI®)*, 1st ed. (CreateSpace, Scotts Valley, CA, 2015), pp. 105–114.

[7]M. Zañartu, D. D. Mehta, J. C. Ho, G. R. Wodicka, and R. E. Hillman, "Observation and analysis of in vivo vocal fold tissue instabilities produced by nonlinear source-filter coupling: A case study," J. Acoust. Soc. Am. **129**(1), 326–339 (2011).

[8]V. M. Espinoza, M. Zañartu, J. H. Van Stan, D. D. Mehta, and R. E. Hillman, "Glottal aerodynamic measures in women with phonotraumatic and nonphonotraumatic vocal hyperfunction," J. Speech Lang. Hear. Res. **60**(8), 2159–2169 (2017).

[9]M. Zañartu, G. E. Galindo, B. D. Erath, S. D. Peterson, G. R. Wodicka, and R. E. Hillman, "Modeling the effects of a posterior glottal opening on vocal fold dynamics with implications for vocal hyperfunction," J. Acoust. Soc. Am. **136**(6), 3262–3271 (2014).

[10]I. R. Titze and B. H. Story, "Rules for controlling low-dimensional vocal fold models with muscle activation," J. Acoust. Soc. Am. **112**(3), 1064–1076 (2002).

[11]J. C. Lucero and J. Schoentgen, "Smoothness of an equation for the glottal flow rate versus the glottal area (L)," J. Acoust. Soc. Am. **137**, 2970–2973 (2015).

[12]B. H. Story and I. R. Titze, "Voice simulation with a body-cover model of the vocal folds," J. Acoust. Soc. Am. **97**(2), 1249–1260 (1995).

[13]B. H. Story, "Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002," J. Acoust. Soc. Am. **123**(1), 327–335 (2008).

[14]B. H. Story, "Technique for 'tuning' vocal tract area functions based on acoustic sensitivity functions," J. Acoust. Soc. Am. **119**(2), 715–718 (2006).