Proceedings of Meetings on Acoustics

Volume 30 http://acousticalsociety.org/

Acoustics `17 Boston

173rd Meeting of Acoustical Society of America and 8th Forum Acusticum
Boston, Massachusetts
25-29 June 2017

Speech Communication: Paper 1aSC6

Toward real-time physically-based voice simulation: An Eigenmode-based approach

Zhaoyan Zhang
Head and Neck Surgery, UCLA School of Medicine, Los Angeles, CA, 90095; zyzhang@ucla.edu

While physically-based continuum models of voice production have potential applications in clinical intervention of voice disorders and personalized natural speech synthesis, their current use is limited due to the high computational cost associated with resolving the complex fluid-structure interaction during voice production process. The goal of this study is to summarize recent efforts in developing a physically-based, computationally-efficient continuum model of voice production toward near real-time applications. The model uses an eigenmode-based formulation of the governing equations, in which vocal fold eigen-modes are used as building blocks to reconstruct more complex vocal fold vibration patterns. Simulations show that a reasonable accuracy in the fundamental frequency, vocal intensity, and selected spectral measures can be reached with the use of the first 100 vocal fold eigenmodes, thus significantly reducing the degrees of freedom of the governing equations (as compared to tens of thousands in finite element models) and computational time. It is expected that for applications in which absolute values are not as essential, even a smaller number of eigenmodes would be acceptable. Examples are provided to demonstrate the capability of the model in modeling large range of voice qualities, natural voice quality change over time, and speech production in general.

Published by the Acoustical Society of America

https://doi.org/10.1121/2.0000572

1. INTRODUCTION

A physically-based continuum model of voice production is essential to applications in which realistic representation of the laryngeal physiology and material properties is required. For example, in the clinic, accurate representation of laryngeal physiology and material properties to the degree that matches actual clinical intervention is desirable. Although lumped-element models of phonation have been widely used [1-6], a major drawback of these models is that their control parameters cannot be directly measured or easily related to realistic properties of the vocal folds, which prevents direct translation of the findings from these models to clinical applications,

A continuum mechanics-based approach may also benefit synthesis of personalized, natural-sounding speech, in which the voice source may differ across speakers and vary significantly with the speaking style and emotional state of the speaker. In current speech synthesis models, such variations in the voice source are often simulated through careful, manual specification of the time contours of relevant acoustic features (e.g., the fundamental frequency, closed quotient, speed of glottal closure, etc.) in a parametric voice source model. However, because these acoustic features of the voice source are not independent from each other and often co-vary during phonation, it is difficult to specific physically-consistent time-varying voice source characteristics [7]. Additionally, although a source-filter model is often adopted in voice modeling, in reality the acoustics of the sub- and supra-glottal tracts may still interact with and affect the voice source. In contrast, a physically-based approach models the voice production process based on physical principles, thus avoiding the need to manually specify consistent time contours of voice source characteristics. A model based on continuum mechanics would further allow realistic representation of vocal fold physiology and thus voice control in a way similar to humans. In this sense, a physically-based voice production model holds the most promise in personalized natural speech synthesis.

One difficulty that limits the use of physically-based continuum models in practical applications is the excessively high computational costs, due to the computational complexity involved in modeling vocal fold posturing and glottal fluid-structure interaction. Often, simulation of one cycle of vocal fold oscillation may take days or weeks. In this paper, we describe our recent effort toward developing a reduced-order, physically-based voice production model that allows realistic representation of vocal physiology yet with much improved computational efficiency, for use in speech production research and potential practical applications such as simulation of clinical intervention of voice disorders or personalized speech synthesis.

2. MODEL DESCRIPTION

A sketch of our three-dimensional computational model is shown in Fig. 1. The model includes a respiratory system driven by a respiratory muscular pressure to provide the target subglottal pressure, a three-dimensional vocal fold model coupled with a one-dimensional glottal flow model, and a vocal tract model in which sound propagates. The fluid-structure-acoustic interaction within the glottis and sound propagation in the vocal tract are described in [8, 9], and the respiratory model is described in detail in [13]. The vocal folds and the respiratory system are coupled through the instantaneous glottal volume flow rate [10]. The input of the model includes the vocal fold properties (geometry, stiffness, and position), lung function parameters (total lung volume, vital capacity, functional residue capacity, and lung compliance), and the respiratory muscular pressure or the target subglottal pressure. Although a simplified vocal fold geometry is shown in Fig 1, because of the continuum nature of the model, the model can be adapted to realistic vocal fold geometry of a specific speaker, thus allowing personalized speech synthesis.

Modeling the glottal fluid-structural interaction in three-dimensions is computationally very challenging and may not be practical with the computational resources available in the near future. To reduce the computational costs toward practical applications, two major simplifications are made in our model, including the assumption of linear elasticity of the vocal folds and the use of a quasi-one-dimensional description of the glottal flow. Based on these two assumptions, the resulting governing equations of the vocal folds can be derived from Lagrange's equations as:

$$M\ddot{q} + C\dot{q} + Kq = Q$$

where M, C, and K are the mass, damping, stiffness matrices of the vocal folds, respectively, and Q is the generalized force due to external force applied to the vocal folds, including air pressure and contact force.

Although similar assumptions have been adopted in many previous models of phonation (e.g., [1, 11, 12, 13]), recent studies have provided evidence supporting these simplifications. For example, although the one-dimensional glottal flow model neglects many complex flow phenomena observed in phonation such as vortices, jet flapping, turbulence, our previous experimental studies suggest that these complex flow phenomena may only have minimal perceptual relevance [14, 15]. Phonation models using a one-dimensional flow description have been shown to be able to match reasonably with experimental measurements [16, 17, and 18]. Coupling a one-dimensional flow model with a linear elastic vocal fold model has been shown to qualitatively reproduce experimental observations regarding different vocal fold vibratory regimes and their transitions in pathological conditions of left-right stiffness asymmetry [19].

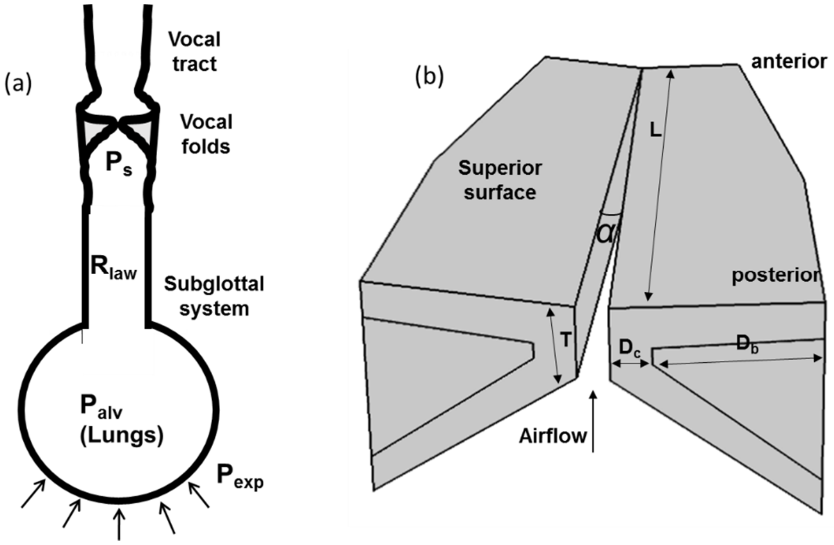

Figure 1. (a) Schematic diagram of the model and (b) the three-dimensional vocal folds. Although a simplified vocal fold geometry is shown here, the model can be adapted to a more realistic geometry.

3. EIGENMODE-BASED REDUCED-ORDER FORMULATION

Despite the above two major assumptions, the resulting system governing equations still have a large number of degrees of freedom. For a typical finite element model of sufficient spatial resolution, the degrees of freedom could easily in the order of tens of thousands (see e.g. [12]). To further improve the computational efficiency, an eigenmode-based formulation is used in our model, in which the in vacuo vocal fold eigenmodes are used as basis functions in formulating the governing equations of vocal fold dynamics, as described in detail in [8]. This is analogous to using the first few formants to approximate vocal tract acoustic response instead of calculating the vocal tract acoustic response in each time step of the simulation. Briefly, the vocal fold motion is approximated as linear superposition of the first *N in vacuo* eigenmodes of the vocal folds,

$$U(X_0,t) = \sum_{i=1}^{N} q_i(t)\varphi_i(X_0).$$

Substituting this approximation into the system governing equations would reduce the degree of freedom of the resulting governing equations to *N*. The advantage of this eigenmode-based formulation is that often only a small number of eigenmodes are required to reach reasonable accuracy. Fig. 2 shows the phonation frequency, sound pressure level, closed quotient of vocal fold vibration, and H1-H2 of the output acoustic spectrum as a function of the number of vocal fold eigenmodes included in the model.

For this particular vocal fold condition, a reasonable accuracy of these four measures is reached with the use of the first 100 vocal fold eigenmode, which is significantly smaller than the degrees of freedom in conventional finite element formulations (often in the order of tens of thousands). In general, a half-second simulation can be finished in a few minutes with the use of 100 vocal fold eigenmodes in a computer with a 3.5 GHz CPU. Although this still does not allow real-time simulation, it is a much improvement from other continuum models currently used, and allows large scale parametric studies for research purposes (e.g., [8, 9]). One can argue that for applications in which prediction of relative changes is more important than absolute values such as speech synthesis, it may be sufficient to use even less eigenmodes. In the extreme case of using only two eigenmodes, the complexity of the model would reduce to that of the two-mass model.

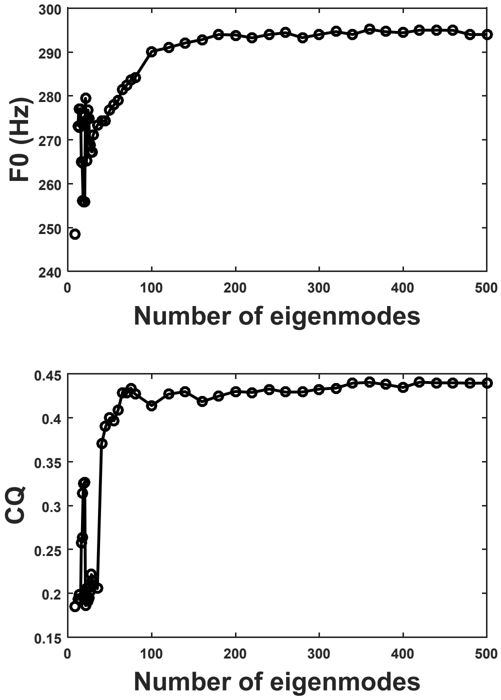

Figure 2. Dependence of the fundamental frequency of phonation (F0), sound pressure level SPL, the closed quotient, and H1-H2 of the source spectra on the number of eigenmodes included in the model.

4. MODELING THE VOICE SOURCE

With realistic representation of vocal fold geometry and stiffness, this model is capable of producing a large variety of voice source conditions. As an example, Fig. 3 shows the phonation frequency (F0), sound pressure level (SPL), closed quotient (CQ) of vocal fold vibration, and two spectral slope measures (H1-H2, H1-H2k) of the output acoustics as a function of the mean glottal flow rate (Qmean), for the 19008 vocal fold conditions investigated in [9]. These 19008 conditions include combinations of different values of the subglottal pressure, vocal fold stiffness along the anterior-posterior direction, resting glottal angle (controlling the degree of vocal fold approximation), and vertical thickness of the vocal fold medical surface. Such a large-scale parametric investigation is almost impossible using other continuum models with currently available computational resources.

Fig. 3 shows that the model is able to produce voice source parameters that cover the range observed in humans. In particular, cluster analysis reveals two regions of distinct characteristics in Fig. 3. One region is characterized by high F0, high H1-H2, high H1-H2k, low HNR, low CQ, and large flow rate, which corresponds to a falsetto-like phonation. The other region has low to moderate F0, low H1-H2, low H1-H2k, moderate to high CQ, and small to moderate flow rate, corresponding to a chest-like phonation. Examination of the physiological conditions producing these two major voice types indicates that the vertical thickness plays a critical role in differentiating these two voice types, with a thin vertical thickness often producing a falsetto-like phonation and a thick vocal fold more likely to produce a chest-like phonation [9]. Note that the vertical thickness of the vocal folds does not have a direct correspondence to any control parameters of the two-mass model, although similar effect can be achieved by varying the coupling stiffness connecting the two masses [1].

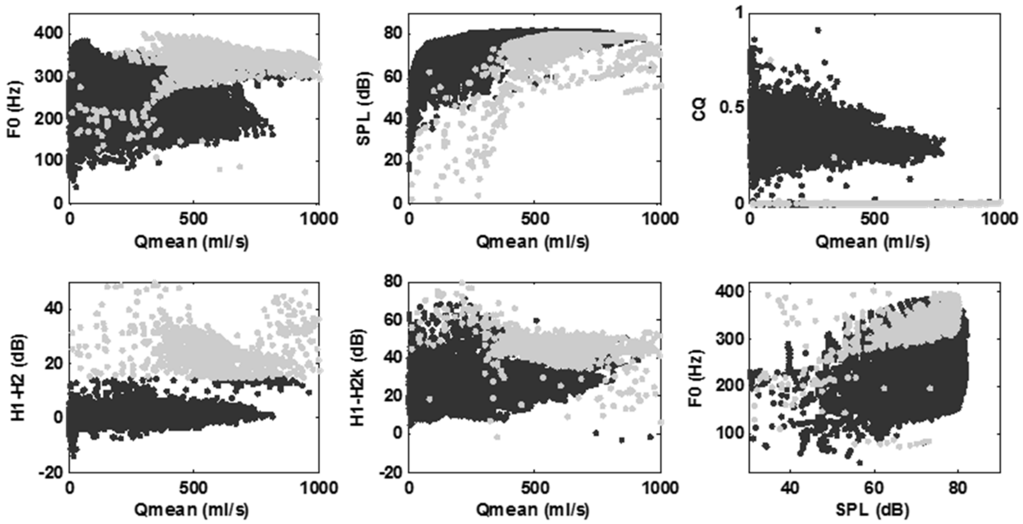

Figure 3. The model is able to produce a large variety of voice types. Cluster analysis reveals two distinct voice regimes, corresponding to a chest-like voice (black symbols) and a falsetto-like voice (gray symbols).

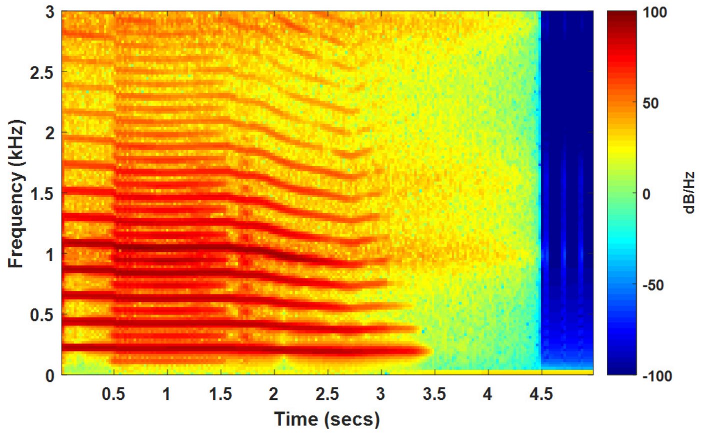

Figure 4. Spectrogram of a synthesized voice produced with a decreasing subglottal pressure. Vocal instabilities occur naturally in the continuum model. In this case, a short period of subharmonics can be observed between 0.5-1.5 seconds.

Voice instabilities (e.g., irregular phonation, sudden F0 jump or appearance of subharmonics, or register changes) often occur in human speech, leading to notable changes in voice quality. Such changes in voice quality are often due to a qualitative change in the vocal fold vibration pattern as different vocal fold eigenmodes are excited. With a sufficient number of eigenmodes included, vocal instabilities and their transitions occur naturally in our model. An example is shown in Fig. 4, which shows the spectrogram of an utterance synthesized with a decreasing subglottal pressure, as often occurring toward the end of an utterance. The continuous decrease in the subglottal pressure leads to a continuous decrease in the F0. A brief region of subharmonics can be observed between 0.5 and 1.5 seconds, probably due to an entrainment of vocal fold vibration to the vocal tract acoustics as the F0 approaches the first resonance of the vocal tract. The onset of this subharmonic region is marked by a notable increase in the sound pressure amplitude and a slight change in the vocal fold contact pattern. In this region the F0 remains relatively constant, in contrast to the regions before and after in which the F0 decreases continuously with time. Further decrease in the subglottal pressure eventually leads to a complete loss of vocal fold contact and a return to regular phonation, probably because the coupling of the vocal folds to the vocal tract is weakened as the F0 moves away from the vocal tract resonance, before the sound pressure amplitude gradually decreases toward phonation offset.

5. MODELING SPEECH PRODUCTION

Toward speech synthesis, the model currently includes a reflection-type line model of the vocal tract [20, 21], although other vocal tract models can be also implemented in the future [22, 23]. As an example, Fig. 5 show the acoustic waveform of a simulated /aha/ utterance, and the corresponding glottal opening area, spectrogram, and F0 as a function of time. For comparison, Fig. 5 also shows the waveform of a recorded human voice, although no effort has been made to reproduce all the features and details of this voice in the simulation.

At the start of the utterance, the vocal folds are abducted and gradually approximate toward each other. Phonation starts as the vocal folds are sufficiently approximated, starting the initial /a/ segment. The vocal folds then start to abduct and vocal fold vibration gradually decays as the utterance transitions to the /h/ segment. It is observed that a gradual abduction is essential to avoid perception of a glottal stop during this transition. Toward the end of the /h/, the vocal folds rapidly adduct and phonation starts again as the utterance transitions to the final /a/. The vocal folds then gradually abduct toward the end of the utterance, reducing both the F0 and sound pressure level. This gradual abduction at the end is important to achieve a natural-sounding quality at the end of an utterance [24].

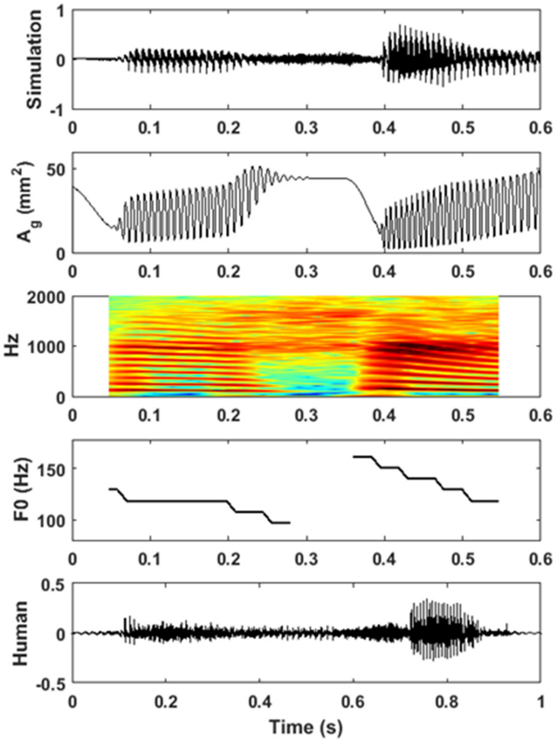

Figure 5. Simulation of an /aha/ utterance. The panels from the top down show the acoustic waveform, glottal opening area, spectrogram, and F0 as a function of time. For comparison, the bottom panel shows a typical waveform recorded from a human subject.

Note that the F0 at the beginning of the final /a/ is higher than that in the initial /a/. Such high F0 occurs because the vocal folds in the final /a/ are approximated more tightly than in the initial /a/ segment.

This F0 contour is similar to the general observation in humans that F0 at vowel onset is significantly higher following voiceless consonants, which is often ascribed to changes in vocal fold stiffness [25]. However, in this study the higher F0 following the /h/ is due to changes in vocal fold approximation alone, with the vocal fold stiffness remaining unchanged.

6. CONCLUSION

This paper introduces a reduced-order, physically-based continuum model of voice production. Based on continuum mechanics, this model allows voice production simulation using realistic geometry and mechanical properties of the vocal system, thus allowing adaptation toward personalized speech synthesis. The eigenmode-based formulation significantly reduces the degrees of freedom of the model and improves its computational efficiency, making it possible to use the model in large-scale parametric investigations of speech production and practical applications. Future work will aim to couple this model with a vocal fold posturing model [e.g, 26], toward modeling vocal control in a way similar to humans.

ACKNOWLEDGMENTS

This study was supported by research Grant Nos. R01DC009229 and R01 DC011299 from the National Institute on Deafness and Other Communication Disorders, the National Institutes of Health, USA.

REFERENCES

[1]K. Ishizaka, and J. Flanagan, “Synthesis of voiced sounds from a two-mass model of the vocal cords,” Bell Syst. Tech. J., vol. 51, pp. 1233–1267, 1972.

[2]K. Ishizaka, and N. Isshiki, “Computer simulation of pathological vocal-cord vibration,” J. Acoust. Soc. Am., vol. 60, pp. 1193–1198, 1976.

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

[4]R. McGowan, L. Koenig, and A. Lofqvist, 1995, “Vocal tract aerodynamics in /aCa/ utterances: Simulations,” Speech Communication, vol. 16, pp. 67-88, 1995.

[5]P. Birkholz, L. Martin, K. Willmes, B. Kroger, C. Neuschaefer-Rube, “The contribution of phonation type to the perception of vocal emotions in German: An articulatory synthesis study,” J. Acoust. Soc. Am., vol. 137, pp. 1503–1512, 2015.

[6]B. Elie, and Y. Laprie, “Extension of the single-matrix formulation of the vocal tract: Consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink,” Speech Communication, vol. 82, 85-96, 2016.

[7]Z. Zhang, “Mechanics of human voice production and control,” J. Acoust. Soc. Am., vol. 140, no. 4, pp. 2614–2635, 2016.

[8]Z. Zhang, “Regulation of glottal closure and airflow in a three-dimensional phonation model: Implications for vocal intensity control,” J. Acoust. Soc. Am., vol. 137, no. 2, pp. 898–910, 2015.

[9]Z. Zhang, “Cause-effect relationship between vocal fold physiology and voice production in a three-dimensional phonation model,” J. Acoust. Soc. Am., vol. 139, no. 4, pp. 1493–1507, 2016.

[10]Z. Zhang, “Respiratory laryngeal coordination in airflow conservation and reduction of respiratory effort of phonation,” Journal of Voice, vol. 30, no. 6, pp. 760.e7–760.e13, 2016.

[11]I. Titze, and D. Talkin, “A theoretical study of the effects of various laryngeal configurations on the acoustics of phonation,” J. Acoust. Soc. Am., vol. 66, pp. 60–74, 1979.

[12]F. Alipour, D. Berry, and I. Titze, “A finite-element model of vocal-fold vibration,” J. Acoust. Soc. Am., vol. 108, pp. 3003–3012, 2000.

[13]X. Pelorson, A. Hirschberg, R. van Hassel, A. Wijnands, and Y. Auregan, “Theoretical and experimental study of quasi-steady flow separation within the glottis during phonation: Application to a modified two-mass model,” J. Acoust. Soc. Am., vol. 96, pp. 3416–3431, 1994.

[14]Z. Zhang, and J. Neubauer, “On the acoustical relevance of supraglottal flow structures to low-frequency voice production,” J. Acoust. Soc. Am., vol. 128, no. 6, pp. EL378–EL383, 2010.

[15]M. Farahani, and Z. Zhang, “A computational study of the effect of intraglottal vortex-induced negative pressure on vocal fold vibration,” J. Acoust. Soc. Am., vol. 136, no. 5, pp. EL369–EL375, 2014.

[16]Z. Zhang, L. Mongeau, and S. Frankel, “Experimental verification of the quasi-steady approximation for aerodynamic sound generation by pulsating jets in tubes,” J. Acoust. Soc. Am., vol. 112, no. 4, pp. 1652–1663, 2002.

[17]N. Ruty, X. Pelorson, A. Van Hirtum, I. Lopez-Arteaga, and A. Hirschberg, “An in vitro setup to test the relevance and the accuracy of low-order vocal folds models,” J. Acoust. Soc. Am., vol. 121, pp. 479–490, 2007.

[18]M. Farahani, and Z. Zhang, “Experimental validation of a three-dimensional reduced-order continuum model of phonation,” J. Acoust. Soc. Am., vol. 140, no. 2, pp. EL172–EL177, 2016.

[19]Z. Zhang, and T. Luu, “Asymmetric vibration in a two-layer vocal fold model with left-right stiffness asymmetry: Experiment and simulation,” J. Acoust. Soc. Am., vol. 132, no. 3, pp. 1626–1635, 2012.

[20]J. Liljencrants, "Speech synthesis with a reflection-type line analogy," D.S. dissertation, Royal Inst. of Tech., Stockholm, Sweden, 1985.

[21]B. Story, “Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract,” Ph.D. dissertation, University of Iowa, 1995.

[22]S. Maeda, “A digital simulation method of the vocal-tract system,” Speech Communication, vol. 1, pp. 199–229, 1982.

[23]R. Blandin, M. Arnela, R. Laboissière, X. Pelorson, O. Guasch, A. Van Hirtum, and X. Laval, “Effects of higher order propagation modes in vocal tract like geometries,” J. Acoust. Soc. Am., vol. 137, pp. 832-843, 2015.

[24]J. Slifka, “Some physiological correlates to regular and irregular phonation at the end of an utterance,” J. Voice, vol. 20, no. 2, pp. 171-186, 2006.

[25]A. Lofqvist, T. Baer, N. McGarr, and R. Story, “The cricothyroid muscle in voicing control,” J. Acoust. Soc. Am., vol. 85, pp. 1314-1321, 1989.

[26]J. Yin, and Z. Zhang, “Laryngeal muscular control of vocal fold posturing: Numerical modeling and experimental validation,” J. Acoust. Soc. Am., vol. 140, no. 3, pp. EL280–EL284, 2016.