A deep-learning based generalized reduced-order model of glottal flow during normal phonation

Yang Zhang,1 Weili Jiang,1 Luning Sun,² Jianxun Wang,² Simeon Smith,³ Ingo R. Titze,³ Xudong Zheng,1,a and Qian Xue1,b

1Department of Mechanical Engineering, University of Maine, Orono, ME, 04469, USA

2Department of Aerospace and Mechanical Engineering, University of Notre Dame, Notre Dame, IN, 46556, USA

3The National Center for Voice and Speech, University of Utah, Salt Lake City, UT, 84101, USA

(Dated: 26 May 2020)
This paper proposes a deep-learning based generalized reduced-order model (ROM) that can provide a fast and accurate prediction of the glottal flow during normal phonation. The approach is based on the assumption that the vibration of the vocal folds can be represented by a universal kinematics equation (UKE), which is used to generate a glottal shape library. For each shape in the library, the ground truth values of the flow rate and pressure distribution are obtained from the high-fidelity Navier-Stokes (N-S) solution. A fully-connected deep neural network (DNN) is then trained to build the empirical mapping between the shapes and the flow rate and pressure distributions. The obtained DNN based reduced-order flow solver is coupled with a finite-element method (FEM) based solid dynamics solver for FSI simulation of phonation. The reduced-order model is evaluated by comparing to the Navier-Stokes solutions in both statics glottal shapes and FSI simulations. The results demonstrate a good prediction performance in accuracy and efficiency.

a xudong.zheng@maine.edu
b qian.xue@maine.edu
I. **INTRODUCTION**

Voiced sound production in the human larynx is a complex fluid-structure interaction (FSI) process in which the forced air from the lungs interacts with vocal fold tissues to initiate sustained vibrations that modulate the glottal airflow (Titze, 1994). An accurate prediction of the vocal fold vibration and sound source relies on an accurate prediction of intraglottal pressure and glottal flow rate. In the past, the most commonly used glottal flow model for simulating FSI is the Bernoulli equation which simplifies the flow as a 1D inviscid flow (Ruty et al., 2007; Wurzbacher et al., 2006; Zanartu et al., 2007). By coupling with lumped-mass or continuum vocal fold models, the model has provided important understandings of the dynamics of FSI during voice production (Alipour et al., 2000; Erath et al., 2011; Ishizaka and Flanagan, 1972; Jiang and Zhang, 2002; Steinecke and Herzel, 1995; Story and Titze, 1995; Tao and Jiang, 2008; Titze, 1988; Zhang and Jiang, 2008). Yet, the inviscid assumption has made the model inaccurate in predicting the glottal flow rate and intraglottal pressures, especially during glottal closing when the glottis is typically in a divergent shape in which rich viscous effects occur such as flow separation, shear layer instability and intraglottal vortices (Deverge et al., 2003; Pelorson et al., 1994; Scherer et al., 1983). To improve the accuracy, research efforts have been made to incorporate various viscous loss terms into the Bernoulli equation (Deverge et al., 2003; Ishizaka and Flanagan, 1972; Van den Berg et al., 1957; Zhang and Yang, 2016). While the results showed improvement over the original Bernoulli equation, the modified model is largely based on assumptions of simple glottal shapes. On the other hand, the quick advancement of the continuum vocal fold model from
simple 2D configurations to complex 3D subject-specific configurations increasingly requires a more sophisticated glottal flow model that can represent glottal flow dynamics in complex glottal shapes. The Navier-Stokes (N-S) equation based model, i.e., the full-order model (FOM) can satisfy the requirement (Luo et al., 2008; Mittal et al., 2011; Xue et al., 2014; Zheng et al., 2010), but the very high computational cost limits its use in statistical studies. Therefore, there is a need and interest in developing a glottal flow model that can provide accurate and fast solution of glottal flow dynamics in complex glottal shapes.

It has been shown that self-sustained oscillation of vocal folds is dominated by a few modes of vibration, even when the motion is abnormal (Berry, 2001; Berry et al., 1994; Döllinger et al., 2005). This high predictability of the vibratory pattern of the vocal folds makes it feasible to model the glottal flow dynamics based on the glottal shapes using deep-learning approach. Nevertheless, related research focusing on this area is still rare. A deep-learning based reduced-order model (ROM) model for glottal flow was proposed in our previous study (Zhang et al., 2020). The model was based on the Bernoulli equation with a viscous loss term predicted by a deep neural network (DNN) model. With the trained DNN-Bernoulli model, the flow resistance coefficient as well as the flow rate and pressure distribution of a given glottal shape can be predicted. However, the DNN-Bernoulli model was developed under certain initial and geometry conditions and the generalization ability of the model may be limited. In this paper, a deep-learning based generalized ROM of the glottal flow during normal phonations is proposed. The underlying assumption of the approach is that the vocal fold kinematics can be approximated by a few vibration modes described by the surface-wave approach (Smith and Titze, 2018). A number of past studies
showed that the vocal fold vibration in normal phonation is dominated by two modes (Berry, 2001; Berry et al., 1994; Döllinger et al., 2005; Smith and Titze, 2018). Therefore, in the present work, we assume that the vibration of the vocal folds is approximated by a linear combination of the modal displacement of the two dominant modes, and then a universal kinematics equation (UKE) can be obtained. The UKE is examined by generating a large number of glottal shapes from FSI simulations with various vocal fold material properties and subglottal pressures and fitting the glottal shapes with the UKE using the genetic algorithm (GA) (Forrest, 1996; Goldberg, 2006; Mitchell, 1998). The probability distribution function (PDF) of each fitting parameter is then obtained and used to construct a generalized glottal shape library by appropriately resampling the PDF of the fitting parameters. For each shape in the library, the ground truth value of the flow rate and pressure distribution are obtained from high-fidelity N-S solutions. A fully-connected DNN (Goodfellow et al., 2016) is then used to build the empirical mapping between input parameters (fitting parameters in the UKE and subglottal pressure) and output parameters (flow rate and pressure distribution). K-fold cross validation is performed to fine tune the architecture and hyperparameters and evaluate the prediction performance of the DNN. The developed reduced order glottal flow model is therefore composed of two parts: (a) glottal shape parameterization using the UKE and GA, and (b) glottal flow rate and intraglottal pressure prediction using the trained DNN. The performance of the developed flow model (ROM) is evaluated by comparing to the N-S solutions (FOM) in both static glottal shapes and FSI simulations.

The outline of the paper is organized as follows: the three-dimensional shape of the vocal fold during vibration, including the prephonatory geometry and UKE, are introduced in
Section II; the process of building up the generalized glottal shape library is elaborated in Section III; details about the implementation and evaluation of the DNN model are discussed in Section IV; implementation and evaluation of the performance of the present ROM for FSI Simulation are discussed in Section V; finally, the conclusions are summarized in Section VI.

II. THREE-DIMENSIONAL SHAPE OF VOCAL FOLD DURING VIBRATION

A. Prephonatory Geometry

The prephonatory geometry of the vocal fold (right half) is shown in Figure 1. The length $L$ along the anterior-posterior direction ($z$), medial surface thickness $T$ along the inferior-superior direction ($y$) and depth $D$ along the lateral direction ($x$) are 1.5 cm, 0.3 cm and 0.75 cm, respectively. The subglottal angle $\alpha$ equals to arctan 0.5. An initial gap $\Delta x = 0.002 cm$ along the lateral direction ($x$) exists between the left and right counterpart. The vocal fold is divided into three layers including the cover, ligament, and body. The thickness of the cover ($T_C$) and ligament ($T_L$) layers are both 0.05 cm. Each layer is assumed to be invariant in the anterior-posterior direction.

B. Universal Kinematics Equation (UKE)

Past studies have shown that vocal fold vibrations are dominated by a few vibratory modes (Berry, 2001; Berry et al., 1994; Döllinger et al., 2005). Following the surface-wave approach in (Titze, 1988), these modes can be described with a combination of $(m, n)$ modes,
where \( m \) and \( n \) correspond to the number of half-wavelengths in the anterior-posterior and inferior-superior directions, respectively. For normal phonation, the most dominant modes are the \((1,0)\) and \((1,1)\) modes, where \((1,0)\) represents the in-phase vibration and \((1,1)\) represents the out-of-phase vibration (Smith and Titze, 2018; Titze, 1988). Taking the right-half vocal fold model in Figure 1 as an example, the displacement of the medial surface over time can be represented by a linear combination of the modal displacement of these two modes,

\[
\xi(y, z, t) = \alpha \xi(y, z, t)_{(1,0)} + (1 - \alpha) \xi(y, z, t)_{(1,1)},
\]  

(1)
where the subscripts \((1, 0)\) and \((1, 1)\) respectively refer to modes \((1, 0)\) and \((1, 1)\), and \(\alpha\) is the weight coefficient of mode \((1, 0)\). An equivalent equation exists for the left-half vocal fold. An example of the modal shape of the right-half vocal fold is illustrated in Figure 2, where the subfigures (a) and (b) respectively represent the modal shape \((1, 0)\) and \((1, 1)\) of the vocal fold, and the subfigures (c), (d) and (e) represent the actual shape of the vocal fold obtained from Eq. (1) with \(\alpha = \frac{1}{4}, \frac{1}{2}, \frac{3}{4}\), respectively. Note that in our study, to simplify the model, only the lateral \((x)\) vibration is allowed and the vertical \((y)\) motion is fixed.

In (Smith and Titze, 2018), based on the surface-wave approach and small-angle approximation (Titze, 1988), the modal displacement of the medial surface of the vocal fold at any instant in time were defined as,

\[
\xi(y, z, t)_{(m,n)} = \xi_m \sin(m\pi z/L)[\sin \omega t - n(\omega/c)(y - y_m) \cos \omega t],
\]

(2)

where \(\xi_m\) is the modal displacement amplitude, \(y_m\) is the inflection point for the vertical half wavelength, \(\omega\) is angular frequency, and \(c\) is the speed of the mucosal wave.

The displacement of the medial surface of the vocal fold over time in Eq. (1) can then be expressed as,

\[
\xi(y, z, t) = \xi_m \sin(\pi z/L)[\sin \omega t - (1 - \alpha)(\omega/c)(y - y_m) \cos \omega t],
\]

(3)

where \(y_m\) is the inflection point for the vertical half wavelength (Smith and Titze, 2018).

Note that our later FSI simulation results reflected that the location of the inflection point changes along the anterior-posterior direction, therefore, the inflection location is modeled as

\[
y_m = T - \beta(\sin \frac{\pi z}{L} + 1),
\]

(4)
FIG. 2. Modal shape of the right-half vocal fold (Each 2D profile corresponds to mid-coronal plane).
where $0 \leq \beta \leq T/2$.

By superimposing the time-dependent displacement in Eq. (3) on the prephonatory geometry, the three-dimensional shape of the glottis at any time instant can be obtained. Eq. (3) is also termed as the universal kinematics equation (UKE) in this paper.

### III. GENERALIZED GLOTTAL SHAPE LIBRARY

The vocal fold shape during vibration can be described by Eqs. (3) and (4) with the following parameters: the vibration amplitude $\xi_m$, weight coefficient of mode $(1, 0) \alpha$, inflection point factor $\beta$, phase $\phi = 12\omega t/\pi$, and ratio between the angular frequency and mucosal wave speed $\omega/c$, which is related to the vibration frequency $f$. The estimated physiological range of these parameters for normal phonation (Smith and Titze, 2018) are listed in Table I.

| Parameters | Range          |
|------------|----------------|
| $\xi_m$    | $(0, 0.1cm]$   |
| $\alpha$   | $[0, 1]$       |
| $\beta$    | $[0, T/2]$     |
| $\phi$     | $[0, 24]$      |
| $f$        | $[100Hz, 250Hz]$ |
In this section, we aim to verify that the UKE can be used as a generalized equation to represent any glottal shape during normal phonation. To have a good estimation of the possible glottal shapes during FSI, FSI simulations of vocal fold vibration under various subglottal pressures and material properties are conducted. The simulations employ the finite-element vocal fold model coupled with the Bernoulli equations for fast solutions (Geng et al., 2016). A large number of glottal shapes are extracted from the simulation results and used to fit the UKE by using the genetic algorithm (GA) (Forrest, 1996; Goldberg, 2006; Mitchell, 1998). The fitting error is used to quantify the representative capability of the UKE. Finally, the probability density function (PDF) of each input parameter in the UKE is obtained and used to build the generalized glottal shape library through appropriate resampling.

A. Bernoulli-FEM FSI Simulation

The vocal fold tissue is modeled as the viscoelastic, transversely isotropic material. The baseline material properties of each layer of the vocal fold (Alipour et al., 2000; Xue et al., 2012) are listed in Table II.

Based on the baseline material properties listed in Table II, the ranges of the material properties for each layer can be obtained by simultaneously multiplying the corresponding $E_{pz}^0$ and $G_{pz}^0$ with a factor $k$, where the physiological range of $k$ is $[0.5, 5.0]$ with an increment size $\Delta k = 0.5$. Note that the value of $k$ for the cover layer and ligament layer are always the same. The various material property factors of the cover-ligament layers and body layer under selected subglottal pressure conditions at $P_0 = 0.5kPa, 0.75kPa, 1.0kPa$ can be
TABLE II. Baseline material properties of each layer of the vocal fold.

|        | ρ (g/cm$^3$) | $E_p$ (kPa) | ν$_p$ | $E_{pz}^0$ (kPa) | ν$_{pz}$ | $G_{pz}^0$ (kPa) |
|--------|--------------|-------------|-------|------------------|---------|-----------------|
| Cover  | 1.043        | 2.01        | 0.9   | 40               | 0.0     | 10              |
| Ligament | 1.043       | 3.31        | 0.9   | 66               | 0.0     | 40              |
| Body   | 1.043        | 3.99        | 0.9   | 80               | 0.0     | 20              |

ρ is the tissue density; $E_p$ and $E_{pz}^0$ are the transversal and longitudinal Youngs Modulus, respectively; ν$_p$ and ν$_{pz}$ are the in-plane transversal and longitudinal Poisson ratio, respectively; $G_{pz}^0$ is the longitudinal shear modulus (Alipour et al., 2000; Xue et al., 2012).

respectively expressed as

$$k_{CL} = m\Delta k, \quad m = 1, 2, ..., 10,$$

$$k_B = n\Delta k, \quad n = 1, 2, ..., 10,$$

where the subscript $CL$ and $B$ indicate the cover-ligament layers and body layer, respectively.

By systematically varying $k_{CL}$, $k_B$ and $P_0$, a total of 300 cases are generated for the FSI simulations. For each case, the density and kinematic viscosity of the air are $1.145 \times 10^{-3} g/cm^3$ and $\nu = 1.655 \times 10^{-1} cm^2/s$, respectively. The glottis are discretized with $N_S = 69$ uniformly spaced cross sections along the inferior-superior direction such that the spacing
is 0.01 cm. The contact surface is calculated as an average of the left and right surface coordinates. A uniform Rayleigh damping factor is used for each case. As an example, the vibration pattern of the vocal folds during one converged cycle at $P_0 = 1.0 \text{kPa}$, $k_{CL} = 1.0$, $k_B = 4.0$ is illustrated in Figure 3, where the left subfigure corresponds to the time history of the flow rate $Q$ during one converged cycle, and the right subfigure corresponds to the glottal shape at 5 representative phases probed from the left subfigure. The vibration shows a typical alternative convergent-divergent glottal shape change.

FIG. 3. Vibration pattern during one converged cycle at $P_0 = 1.0 \text{kPa}$, $k_{CL} = 1.0$, $k_B = 4.0$. 
B. Glottal Shape Fitting with the GA

In this subsection, we aim to verify that those glottal shapes extracted from FSI simulations in Subsection III A can be represented by the UKE. The GA is employed to inversely determine the values of the fitting parameters from the range listed in Table I such that the difference between the optimized and target (FSI) values of the nodal displacement is minimal. In the optimization process, as the flow rate heavily relies on the minimum glottal area, an equal constraint between the optimized and target minimum cross-section area along the inferior-superior direction of the glottis is enforced. Therefore, the constrained minimization function for each glottal shape can be written as,

$$\xi_{m}, \alpha, \beta, \phi, f = \arg \min \frac{\sum_{i=1}^{n} [\xi_{\text{optimized}}^{i}(\xi_{m}, \alpha, \beta, \phi, f) - \xi_{\text{target}}^{i}]^2}{n}$$

subject to $\arg \min A_{j, \text{optimized}} = \arg \min A_{j, \text{target}}$, $(A_{j, \text{optimized}})_{\text{min}} = (A_{j, \text{target}})_{\text{min}}$

where the values of $\xi_{m}, \alpha, \beta, \phi, f$ are bounded by the corresponding ranges listed in Table I, $n$ is the number of nodal points of the glottis surface, and $A_{j, \text{optimized}}$ and $A_{j, \text{target}}$ are the optimized and target cross-section area function with $j$ the cross-section index, respectively.

The constraints imply that the location and value of the optimized minimum cross-section area are equal to the target one.

The population size and the number of generation for the GA are chosen as 160 and 100, respectively. The overall residual of the fitness function extracted from the FSI cases in Subsection III A is plotted in Figure 4. The residual for each phase is normalized by the corresponding maximum nodal displacement. The relative residuals for most of the phases are close to 0 and the maximum relative residual among all the phases is around 0.01, indicating that GA converges well for each glottal shape and therefore the UKE can be used
a generalized equation to represent the extracted glottal shapes. Furthermore, the kernel
density estimation (Freedman et al., 2007) is used as a non-parametric way to estimate the
probability density function (PDF) of the fitting parameters, and the corresponding PDF for
$P_0 = 0.75\text{kPa}$ is plotted in Figure 5. The PDF for $P_0 = 0.5\text{kPa}$ and $P_0 = 1.0\text{kPa}$ are highly
similar and thus not shown. Note that the PDF of the optimized frequency is not plotted
in those figures because the values for all cases are similar and the corresponding PDFs are
concentrated at $f = 210\text{Hz}$. Therefore, to reduce the number of redundant shapes, we fix
the value of the optimized frequency to be $f = 210\text{Hz}$. Based on the PDFs, the generalized

glottal shape library can be built by appropriately resampling the parameters. Concretely,
we first locate the parameter values with the local maximum probabilities from each PDF,
and then with this located value as the center value, conduct the uniform resampling from
each PDF such that the majority of the representative glottal shapes can be included in this
library. The re-sampled values of the input parameters under different subglottal pressure
conditions are listed in Table III. Note that for different subglottal pressure values, only the
amplitude $\xi_m$ is different, and the other parameters are all the same. A total of $N_L = 3960$
different shapes are generated by substituting the values in Table III into the UKE, and
these shapes constitute the generalized glottal shape library which are used as the raw data
for training the DNN in the next section.

IV. IMPLEMENTATION OF THE DNN MODEL

For each shape in the generalized glottal shape library, the subglottal pressure $P_0$ and
the parameters $\xi_m$, $\alpha$, $\beta$ and $\phi$ are the input features, and the corresponding output targets
are the flow rate $Q$ and the pressure distribution $P_i$, where $i$ is the index of the discretized cross sections in the inferior-superior direction of the vocal folds. The ground truth values of the flow rate $Q$ and pressure distribution $P_i$ are obtained by solving the N-S equations. Then, the mapping relationship between the input features and the corresponding output targets can be established by a fully-connected DNN as follows:

$$Q, P_i = f(P_0, \xi_m, \alpha, \beta, \phi; \theta)$$ (8)
TABLE III. Resampled values of input parameters.

| $P_0$ (kPa) | $\xi_m$   | $\alpha$ | $\beta$ | $\phi$ |
|-------------|-----------|-----------|---------|--------|
| 0.02, 0.03, | 0.04, 0.1 | 1, 2, 3, 4, |
| 0.5         |           |           |         |        |
| 0.025, 0.04,| 0.0, 0.2, 0.4, 0.0, 0.015, 0.03, |
| 0.75        |           | 5, 6, 7, 8, |
| 0.055, 0.1  | 0.6, 0.8, 1.0 | 0.135, 0.15 |
| 0.035, 0.055,|         | 9, 10, 11 |
| 1.0         |           |           |         |        |
| 0.075, 0.1  |           |           |         |        |

where $f$ is the function representing the overall DNN, and $\theta$ denotes all learnable parameters of the DNN. With this trained DNN, the flow rate and pressure distribution along any glottal shape generated by the UKE can be well predicted.

A. N-S Solution of the Output Targets

The fluid flow is governed by the incompressible N-S equations as follows,

$$ \frac{\partial u_i}{\partial x_i} = 0 $$  
$$ \frac{\partial u_i}{\partial t} + \frac{\partial u_i u_j}{\partial x_j} = -\frac{1}{\rho_f} \frac{\partial p}{\partial x_i} + \nu_f \frac{\partial^2 u_i}{\partial x_j \partial x_j}, $$

where $u_i$, $\rho$, $p$, $\nu$ are the incompressible flow velocity, density, pressure, and kinematic viscosity, respectively. An in-house sharp-interface immersed-boundary N-S flow solver (Zheng et al., 2010) is used to obtain the ground truth solution of the output targets. The setup of
FIG. 5. PDF of optimized input parameters for $P_0 = 0.75kPa$. 

(a) $\xi_m$  
(b) $\alpha$  
(c) $\beta$  
(d) $\phi$
the computational domain is depicted in Figure 6. The size of the computational domain is $1.5\text{cm} \times 21.0\text{cm} \times 1.5\text{cm}$ in the $x$ (lateral), $y$ (inferior-superior) and $z$ (anterior-posterior) direction. The vocal folds are placed $3.2\text{cm}$ and $17.0\text{cm}$ away from the inlet and outlet of the computational domain, respectively. The grid independence study is performed by comparing the flow rate and average pressure distribution on coarse, medium and fine meshes with fixed CFL number. The mesh number $N_x \times N_y \times N_z$ on the coarse, medium and fine meshes are $64 \times 64 \times 24$, $128 \times 128 \times 48$, and $256 \times 256 \times 96$ in the $x$, $y$ and $z$ direction, respectively, where $N_x$, $N_y$ and $N_z$ are the number of mesh nodes in the $x$, $y$ and $z$ direction, respectively. The mesh is stretched to the far field in the $x$ and $y$ direction, while uniformly distributed in the $z$ direction. The grid independence results of the flow rate and average pressure distribution are illustrated in Figure 7. From this figure, we can see that the medium mesh is adequate to obtain the ground truth solution of the output targets from the shape library. The relative error of the flow rate obtained on the coarse and medium mesh with respect to that obtained on the fine mesh are $12.1\%$ and $1.0\%$, respectively. The minimum interval of the medium mesh is $0.003\text{cm}$ and $0.01\text{cm}$ in the $x$ and $y$ direction, respectively. Moreover, the total CPU time required for convergence on the coarse, medium and fine meshes are respectively $0.2$, $2.3$ and $35$ hours on a parallel computer with 32 CPUs.

**B. Implementation Details of the DNN**

As mentioned above, the input features and corresponding output targets extracted from the shape library can be organized as a vector $\mathbf{x}$ and $\mathbf{y}$, respectively,
FIG. 6. Setup of the computational domain.

\[
\begin{align*}
\mathbf{x} &= \begin{bmatrix} P_0 \\ \xi_m \\ \alpha \\ \beta \\ \phi \end{bmatrix}, & \mathbf{y} &= \begin{bmatrix} Q \\ P_1 \\ P_2 \\ \vdots \\ P_{N_P} \end{bmatrix},
\end{align*}
\]

(11)
where $N_P = 68$ is the dimension of the output pressure distribution.

The mapping relationship between the input features $x$ and corresponding output targets $y$ can be established by a fully-connected DNN (Goodfellow et al., 2016; LeCun et al., 2015). In the fully-connected DNN, the input and output layers are denoted as $z_0$ and $z_L$, respectively. The layers between the input and output layers are called the hidden layers $z_l$, where $l = 1, ..., L - 1$. Neurons in the hidden layer $z_l$ have connections to all neurons of the previous layer $z_{l-1}$,

$$z_l = \sigma_l(W_l^T z_{l-1} + b_l)$$

(12)

where $W_l$ is the learnable weights, $b_l$ is the additive bias, and $\sigma_l$ is the nonlinear activation function.

The loss function $J$ of the DNN is

$$J = \frac{1}{N} \sum \|z_L - y\|^2 + \lambda \|W\|^2$$

(13)

where $z_L$ is the predicted value and $\lambda$ is the regularization coefficient to prevent the overfitting of the DNN model.
Note that the range of values of $Q$ and $P_i$ are different, i.e., $Q \geq 0$ while $P_i/P_0 \leq 1$; therefore for the ease of training the DNN, the input features $x$ are respectively mapped to the subsets of the output targets $y$ (i.e., $Q$ and $P_i$) with different architectures of the DNN.

The whole data set from the shape library is randomly split into the training and test sets. To avoid the overfitting of the model, we use 5-fold cross validation (Goodfellow et al., 2016) to fine tune the architecture and hyperparameters of the DNN, such as the number of hidden layers, the number of neurons on each hidden layer, the initialization of the weights, the activation function, the optimization method, the mini-batch size, and the number of epochs (Goodfellow et al., 2016). The final architecture and hyperparameters of the DNN are chosen from those that have the lowest errors on the validation set. The final DNN model is then trained on the full training set, and the prediction performance of the trained model is evaluated on the test set.

The final architectures of the DNN for $Q$ and $P_i$ are illustrated in Figure 8 and denoted as DNN-Q and DNN-P, respectively. The input layer for both DNNs has 5 neurons which correspond to the dimension of the input vector. The output layer of DNN-Q has a single neuron which corresponds to the ground truth value of the flow rate $Q$, while that of DNN-P has 68 neurons which correspond to the ground truth value of the pressure distribution on the discretized cross sections along the inferior-superior direction of the vocal folds. Since $Q$ and $P_i$ are bounded by different ranges ($Q \geq 0$ and $P_i/P_0 \leq 1$), the softplus and tanh activation function (Goodfellow et al., 2016) are used on the output layer of DNN-Q and DNN-P, respectively. Besides the input layer and output layer, there are two hidden layers for both DNNs. The number of neurons on the hidden layers of DNN-Q are 64, and the
softplus activation function is used on each hidden layer, whereas the number of neurons on
the hidden layers of DNN-P are 256, and the relu activation function (Goodfellow et al., 2016)
is used on each hidden layer. All of the weights on each layer are initialized with a random
normal distribution. Both of the DNN models are optimized using a mean-squared loss
function with an adaptive version of the stochastic gradient descent algorithm called Nadam
(Nesterov Adam) (Ruder, 2016). Both of the DNN models are trained with 10000 epochs,
where one epoch consists of one full training cycle on the training set, and the mini-batch
size is 128 for each epoch. The DNN models are implemented on the open-source machine
learning platform Keras (Chollet et al., 2015) using TensorFlow (Abadi et al., 2015) as the
backend.

C. Evaluation of the Trained DNN Models

The relative percent difference (RPD) between the true and predicted outcomes is used
to evaluate the trained DNN models. The expression of the RPD for \( Q \) and \( P_i \) for each
glottal shape in the training data are as follows:

\[
E_Q = \frac{|Q - \hat{Q}|}{\max(|Q|, |\hat{Q}|)}
\]  

(14)

\[
E_P = \frac{\sum_{i=1}^{N_P} \frac{|P_i - \hat{P}_i|}{\max(|P_i|, |\hat{P}_i|)}}{N_P}
\]  

(15)

where \( Q, P_i \) and \( \hat{Q}, \hat{P}_i \) are respectively the true and predicted outcomes.

The history of the 5-fold cross validation results for DNN-Q and DNN-P is plotted in
Figure 9. The horizontal axis corresponds to the number of epochs, and the vertical axis
FIG. 8. Architecture of the DNN.

(a) DNN-Q

(b) DNN-P
corresponds to the mean RPD between the true and predicted outcomes. The comparison is between the training and validation sets. It took 10000 epochs for the mean RPD on the training and validation sets to converge for DNN-Q and DNN-P. The converged mean RPD on the training and validation sets are 1.71% and 1.89% for DNN-Q, and 1.97% and 4.12% for DNN-P, respectively. The performance of the trained DNN-Q and DNN-P on the test set is plotted in Figure 10 and 11, respectively. The first subfigure for each figure shows the history of the model accuracy where the horizontal and vertical axes correspond to the number of epochs and mean RPD, respectively. The comparison is between the full training and test sets. After running 10000 epochs, the mean RPD on the test set converges at 1.74% and 3.52% for DNN-Q and DNN-P, respectively. The second subfigure illustrates the scatter plot of the true and predicted outcomes on the test set, and the good prediction performance on the test set for both DNN-Q and DNN-P can be observed. The final mean RPD on the training, validation and test sets for DNN-Q and DNN-P are summarized in Table IV.

TABLE IV. Mean RPD on the training, validation and test sets.

|       | train | validation | test |
|-------|-------|------------|------|
| Q     | 1.71% | 1.89%      | 1.74%
| P     | 1.97% | 4.12%      | 3.52%

Furthermore, 6 shapes under different subglottal pressures are randomly selected from the test set, and the comparison of the true and predicted pressure distribution of these shapes
FIG. 9. 5-fold cross validation results.
FIG. 10. Performance of the trained DNN-Q model on the test set.
(a) History of the model accuracy

(b) Scatter plot

FIG. 11. Performance of the trained DNN-P model on the test set.
are shown in Figure 12. From these figures, we can observe that the pressure distribution can be well predicted by the trained DNN-P model.

To summarize, the diagram of the implementation of the present reduced-order flow solver is illustrated in Figure 13. Concretely, it is divided into the following steps: firstly, various glottal shapes are extracted from 300 converged Bernoulli-FEM FSI results under different subglottal pressure and material properties. Secondly, these extracted shapes are fitted with the UKE using the GA and the PDF of the fitted input parameters of the UKE are determined. Thirdly, 3960 different glottal shapes are generated by appropriate resampling from the PDF of the input parameters with high probabilities and then substituting them into the UKE, which constitute the generalized shape library. Fourthly, for each shape in the library, the ground truth values of the flow rate $Q$ and pressure distribution $P_i$ are obtained by solving the N-S equation. Finally, the mapping relationship between the input parameters together with the subglottal pressure (input features) and the corresponding flow rate and pressure distribution along the inferior-superior direction of the glottal shape (output targets) are established by the fully-connected DNN. With this reduced-order flow solver, for any glottal shape, the input features can be extracted from the UKE with the GA and then the flow rate and pressure distribution can be predicted with the trained DNNs. The implementation procedure of the reduced-order flow solver can be summarized in Algorithm 1.

The developed reduced-order flow model is then coupled with the FEM based solid dynamics solver for FSI simulation. The abstract workflow of the ROM for FSI simulation is illustrated in Figure 14. First, the flow rate $Q$ and pressure distribution $P_i$ of the glott-
FIG. 12. Comparison of the true and predicted pressure distribution.
tal shape $X$ at a certain time instant $t$ can be obtained by the present reduced-order flow solver, then the pressure load is fed into the FEM solid solver to calculate the corresponding deformation of the glottis $\Delta X$, finally the updated glottal shape $X + \Delta X$ is used as the initial shape of the glottis at the next time instant $t + \Delta t$. The reduced-order flow solver and FEM based solid solver are coupled in a weak manner.

**Algorithm 1:** Implementation of the reduced-order flow solver

1. Extract various shapes from converged Bernoulli-FEA FSI results;
2. Fit these extracted shapes with the UKE using the GA;
3. Obtain the PDF of the fitted parameters of the UKE: $\xi_m$, $\alpha$, $\beta$ and $\phi$;
4. Resample the PDF of $\xi_m$, $\alpha$, $\beta$ and $\phi$ for various $P_0$;
5. Substitute the resampled values into the UKE to generate the generalized shape library;
6. Obtain the ground-truth values of $Q$ and $P_i$ for each shape in the library;
7. Establish the mapping relationship Eq.(8) with a fully-connected DNN
V. EVALUATION OF THE PERFORMANCE OF THE GENERALIZED ROM FOR FSI SIMULATION

To evaluate the prediction performance of the present generalized ROM for FSI simulation, the ROM-FSI results are first compared with the FOM quasi-static (QS) results and the correlation and agreement between these results are analyzed, and then compared with the FOM-FSI results in terms of the voice quality-related parameters and CPU time. Detailed discussions are given as below.

A. Comparison with FOM-QS Results

A series of new subglottal pressure and material properties are simulated using the ROM-FSI model to generate the glottal shapes that are not in the shape library and evaluate the corresponding prediction performance. The values of the selected subglottal pressure and material properties are listed in Table V. The simulation setup is the same as described
in Subsection III A. An example of the converged time history of the flow rate \( Q \) at \( P_0 = 0.8\, Pa, k_{CL} = 4.75, k_B = 3.75 \) predicted by the ROM is illustrated in Figure 15. Note that some fluctuations at the end of the closing phase can be observed, and this is likely due to the unsatisfactory representation of these shapes by the UKE due to the contact issue.

TABLE V. Selected subglottal pressure and material properties for evaluation.

| \( P_0(\text{kPa}) \) | \( k_{CL} \) | \( k_B \) |
|-----------------|--------|--------|
| 0.625           |        |        |
| 0.7             | 1.75, 2.75, 3.75, 4.75 | 1.75, 3.75 |
| 0.8             |        |        |
| 0.875           |        |        |

Various glottal shapes are extracted from the converged FSI results of the cases listed in Table V. By excluding the fully-closed and nearly-closed shapes which may not be well represented by the UKE due to the contact issue, the total number of the extracted shapes for evaluation is 1582.

For each FSI case \( n \) in Table V, at each time step of the steady-cycle ROM-FSI result, the flow rate \( Q_{ROM}^{n,k} \) and pressure distribution \( P_{i,ROM}^{n,k} \) are respectively extracted, and the corresponding reference values of \( Q_{FOM}^{n,k} \) and \( P_{i,FOM}^{n,k} \) can be computed by the FOM, where \( k \) is the index of the time step for each case. The time-averaged error of \( Q \) and \( P_i \) for each FSI case, designated as \( E_Q^n \) and \( E_{P}^n \), can be calculated as follows:
FIG. 15. Example of the converged time history of the predicted flow rate $Q$ at $P_0 = 0.8kPa$, $k_{CL} = 4.75$, $k_B = 3.75$.

\begin{align*}
E_Q^n &= \frac{1}{n_t Q_{FOM}} \sum_{k=1}^{n_t} \left| Q_{FOM}^{n,k} - Q_{ROM}^{n,k} \right| \\
E_P^n &= \sum_{k=1}^{n_t} \sum_{i=1}^{N_P} \frac{P_{i,FOM}^{n,k} - P_{i,ROM}^{n,k}}{P_0}
\end{align*}

where $n_t$ and $\bar{Q}_{FOM}^n$ are the number of extracted time instants and the time-averaged reference values of the flow rate for each case, respectively.
The overall average error of $Q$ and $P_i$, designated as $E_Q$ and $E_P$, can be calculated as:

$$E_Q = \frac{1}{n_c} \sum_{n=1}^{n_c} E_Q^n$$

$$E_P = \frac{1}{n_c} \sum_{n=1}^{n_c} E_P^n$$

where $n_c$ is the number of cases listed in Table V. The overall average error of $Q$ and $P_i$ are $7.87\%$ and $1.68\%$, respectively.

Additionally, the correlation and agreement between the true and predicted $Q$ and $P_i$ for the extracted 1582 glottal shapes are quantified. In terms of $Q$, the Pearson correlation coefficient (Freedman et al., 2007) between $Q_{FOM}$ and $Q_{ROM}$ is excellent ($0.993$, $P < 0.0005$). The scatter and correlation plots are also depicted in Figure 16, where the horizontal and vertical axes correspond to the true ($Q_{FOM}$) and predicted ($Q_{ROM}$) values, respectively. The Bland-Altman plot (Altman and Bland, 1983) is used to analyze the agreement between $Q_{FOM}$ and $Q_{ROM}$. The result is plotted in Figure 17. As can be seen from this figure, the mean difference between $Q_{FOM}$ and $Q_{ROM}$ is $-2.784mL/s$, and the 95% limits of agreement (LoA) between them is from $-12.505mL/s$ to $6.936mL/s$. The 95% confidence interval (CI) of the mean difference, upper LoA and lower LoA between $Q_{FOM}$ and $Q_{ROM}$ is $[-3.0288mL/s, -2.5401mL/s]$, $[6.5177mL/s, 7.3539mL/s]$ and $[-12.9288mL/s, -12.0866mL/s]$, respectively. The number of the outliers is 38, and the percentage of the outliers is $2.40\%$.

Similarly, in terms of $P_i$, the Pearson correlation coefficient between $P_{i,FOM}$ and $P_{i,ROM}$ is excellent ($0.997$, $P < 0.0005$). The scatter and correlation plots are also depicted in Figure 18, where the horizontal and vertical axes correspond to the true ($P_{i,FOM}$) and
predicted ($P_{i,ROM}$) values, respectively. The Bland-Altman analysis between $P_{i,FOM}$ and $P_{i,ROM}$ is plotted in Figure 19. From this figure, we can observe that the mean difference between $P_{i,FOM}$ and $P_{i,ROM}$ is 0.006 kPa, and the 95% LoA between them is from $-0.011kPa$ to $0.023kPa$. The 95% CI of the mean difference, upper LoA and lower LoA between $P_{i,FOM}$ and $P_{i,ROM}$ is $[0.0053kPa, 0.0062kPa]$, $[0.0218kPa, 0.0232kPa]$ and $[-0.0117kPa, -0.0103kPa]$, respectively. The number of the outliers is 87, and the percentage of the outliers is 5.50%.

The above correlation and agreement analysis results between the true and predicted $Q$ and $P_i$ for various glottal shapes indicate that the present ROM-FSI results agree very well with the corresponding FOM-QS results.

B. Comparison with FOM-FSI Results

FSI simulations at $P_0 = 0.8kPa$, $k_{CL} = 1.75$, $k_B = 3.75$ (Case 1) and $P_0 = 0.875kPa$, $k_{CL} = 3.75$, $k_B = 3.75$ (Case 2) from Table V are conducted by using both the ROM-FSI model and FOM-FSI model. The comparison of the phase-averaged time history of the flow rate $Q$ for both cases are illustrated in Figure 20. From this figure, we can observe that the peak flow rate, mean flow rate and the fundamental frequency are close to each other while the skewing of the waveform is different. Several important voice quality-related parameters (Xue et al., 2014) are computed from Figure 20 for both of the cases and the corresponding phase-averaged values are listed in Table VI. It can be seen from this table that the overall agreement between the values obtained by the ROM-FSI and FOM-FSI is satisfactory. The relative errors $\delta$ of $F_0$, $Q_{max}$, $Q_{mean}$ and $\xi_m$ between the ROM-FSI and
FIG. 16. Scatter and correlation plot of $Q$. 

(a) Scatter plot

(b) Correlation plot
FIG. 17. Bland-Altman analysis plot of $Q$.

FOM-FSI for both cases are within 10%, while the relative errors $\delta$ of $\tau_0$ and $\tau_s$ between the ROM-FSI and FOM-FSI for both of the cases are relatively larger. The difference could come from two sources: (a) in the GA optimization process, although the desired location and value of the optimized minimum cross-section area are preset to be equal to the target one (Eq. (7)), the actual optimized location of the minimum cross-section area may be shifted and the corresponding value may be changed especially for the divergent shape, which may affect the profile of the flow rate at the flow decreasing phase. To remedy this, further improvements on the UKE model may be considered, and (b) the ROM-FSI model
FIG. 18. Scatter and correlation plot of $P_1$. 
FIG. 19. Bland-Altman analysis plot of $P_i$.

is a quasi-steady model while the FOM-FSI is a fully unsteady model. The quasi-steady assumption might also contribute to the differences between the two models.

The comparison of the phase-averaged pressure distribution $P_i$ for both cases are illustrated in Figure 21. Note that $T_0$ is the duration of the glottal open phase probed from Figure 20 for each case. The overall agreement is good except at the flow decreasing phase. The glottal vibration patterns at the correspondent phases for both cases obtained by the ROM-FSI and FOM-FSI are compared in Figure 22. From the figure, we can see that the glottal vibration patterns obtained by the ROM-FSI agree well with those obtained by the
TABLE VI. Comparison of voice quality-related parameters.

|                | ROM-FSI | FOM-FSI | δ₁  | ROM-FSI | FOM-FSI | δ₂  |
|----------------|---------|---------|-----|---------|---------|-----|
|                | Case 1  | Case 1  |     | Case 2  | Case 2  |     |
| $F_0$ (Hz)     | 210.8   | 216.3   | 2.5%| 212.0   | 222.5   | 4.7%|
| $Q_{\text{max}}$ (mL/s) | 117.0   | 107.3   | 9.0%| 138.5   | 129.5   | 6.9%|
| $Q_{\text{mean}}$ (mL/s) | 54.8    | 56.6    | 3.2%| 63.6    | 59.5    | 6.9%|
| $\tau_0$       | 0.30    | 0.45    | 33.3%| 0.53    | 0.46    | 15.2%|
| $\tau_s$       | 0.49    | 0.55    | 10.9%| 0.26    | 0.49    | 46.9%|
| $\xi_m$ (cm)   | 0.060   | 0.055   | 9.1%| 0.069   | 0.063   | 9.5%|

$F_0$ is the fundamental frequency; $Q_{\text{max}}$ and $Q_{\text{mean}}$ are the peak and mean glottal flow rate of the open quotient, respectively; $\tau_0$ is the open quotient, defined as $\tau_0 = T_0/T$, where $T_0$ is the duration of the glottal open phase and $T$ is the cycle period; $\tau_s$ is the skewing quotient, defined as $\tau_s = T_p/T_n$ where $T_p$ is the duration of the flow increasing phase and $T_n$ is the duration of the flow decreasing phase (Xue et al., 2014); $\xi_m$ is the vibration amplitude; $\delta_1$ and $\delta_2$ are the absolute value of the relative error between the ROM-FSI and FOM-FSI results for Case 1 and Case 2, respectively.

FOM-FSI except at the flow decreasing phase. The discrepancies of the pressure distribution
FIG. 20. Comparison of the phase-averaged time history of the flow rate.
as well as the vibration patterns are consistent with those of the phase-averaged flow rate shown in Figure 20.

The proper orthogonal decomposition (POD) analysis (Liang et al., 2002) is employed to extract the most energetic empirical eigen-modes from the snapshots of ROM-FSI and FOM-FSI results for both cases. The three-dimensional and mid-coronal profile of the two most energetic empirical eigen functions at two extreme phases for both cases are illustrated in Figure 23. For both cases, these two modes contain around 98% of the total energy. To precisely quantify the similarity between the two modes obtained by the ROM-FSI and FOM-FSI, the dot-product between the corresponding normalized eigenmode (Xue et al., 2011) for both cases is computed and plotted in Figure 24. The dot-product of any two normalized modes is indicative of the similarity between the two modes with the value of one corresponding to an exact match, and zero indicating orthogonality. From the figure, we can observe that modes 1 and 2 obtained by the ROM-FSI are highly similar to the corresponding modes obtained by the FOM-FSI for both cases which indicates a good prediction performance of the present ROM for FSI simulation of the vocal fold vibration.

Furthermore, the average CPU time required for one vibration cycle of the ROM-FSI and FOM-FSI simulation is compared. In order to obtain one vibration cycle, the average CPU time required for the ROM-FSI simulation is 1.5 hours per CPU on a single CPU, while that required for the FOM-FSI simulation is 20 hours per CPU on a parallel computer with 64 CPUs, which indicating the high efficiency of the present ROM for FSI simulation of the glottal flow.
FIG. 21. Comparison of the phase-averaged pressure distribution.

(a) Case 1

(b) Case 2
FIG. 22. Comparison of the vibration pattern at different phases in the coronal view.

FIG. 23. Three-dimensional and mid-coronal profile of the most two energetic empirical eigenfunctions at two extreme phases (dash dot line: equilibrium position).
VI. CONCLUSION

A deep-learning based generalized reduced-order model (ROM) that can provide fast and accurate prediction of the dynamics of the glottal flow during normal phonations is proposed in this paper.

The approach is based on the assumption that the vocal fold kinematics can be approximated by a few vibration modes as described by the surface-wave approach. Therefore, the vibration of the vocal folds during normal phonations can be represented by a universal kinematics equation (UKE) which is a linear combination of the dominant two modes. To verify that the UKE can be used as a generalized equation to represent any glottal shape during...
normal phonation, A large number of glottal shapes are generated from Bernoulli-FEM FSI simulation under various subglottal pressure and material properties and are fitted with a UKE using the genetic algorithm (GA). Furthermore, the probability density function (PDF) for each fitting parameter is obtained and used to build the generalized glottal shape library by appropriately resampling the PDF of the parameters and substituting into the UKE. For each shape in the library, the ground truth value of the flow rate and pressure distribution are obtained from high-fidelity N-S solutions. A fully-connected deep neural network (DNN) is used to build the empirical mapping between input parameters (parameters in the UKE and subglottal pressure) and output parameters (flow rate and pressure distribution). K-fold cross validation is performed to fine tune the architecture and hyperparameters and evaluate the prediction performance of the DNN. The developed reduced order glottal flow model is therefore composed of two parts: (a) glottal shape parameterization using the UKE and GA, and (b) glottal flow rate and intraglottal pressure prediction using the trained DNN. The present reduced-order flow solver is directly coupled with a finite-element method (FEM) based solid dynamics solver for FSI simulation. The ROM-FSI results are compared with the full-order model (FOM) quasi-static (QS) and FSI results. For the comparison with the FOM-QS model, the ROM model shows an excellent agreement in terms of predicting the flow rate and pressure distribution. The average error of the prediction for the flow rate and pressure distribution are 7.87% and 1.68%, respectively. For the comparison with the FOM-FSI model, the ROM model shows a good agreement on the frequency, peak and mean flow rate and vocal fold vibration pattern with the relative errors less than 10%. The ROM model shows a relatively larger error in predicting the opening quotient and skewness
The comparison of the details of the intraglottal pressure distribution between the two models reflects that one of the reasons might be the inaccurate prediction of the location of the minimum area when the glottis has a divergent shape. It should be noted that the ROM-FSI model is a quasi-steady model while the FOM-FSI is a fully unsteady model. The quasi-steady assumption might also contribute to the differences between the two models. The overall good prediction performance of the present ROM in accuracy and efficiency indicates a great promise for future clinical use. The developed ROM can be further extended to predict the dynamics of the glottal flow during abnormal phonations with relative ease.

ACKNOWLEDGMENTS

The project was supported by Grant Number 5R21DC016428 from the National Institute on Deafness and Other Communication Disorders (NIDCD). Numerical simulations were performed using resources of the Extreme Science and Engineering Discovery Environment (XSEDE) (allocation Award Nos. TG-BIO150055 and TG-CTS180004).

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden,
P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). “TensorFlow: Large-scale machine learning on heterogeneous systems” https://www.tensorflow.org/, software available from tensorflow.org.

Alipour, F., Berry, D. A., and Titze, I. R. (2000). “A finite-element model of vocal-fold vibration,” The Journal of the Acoustical Society of America 108(6), 3003–3012.

Altman, D. G., and Bland, J. M. (1983). “Measurement in medicine: the analysis of method comparison studies,” Journal of the Royal Statistical Society: Series D (The Statistician) 32(3), 307–317.

Berry, D. A. (2001). “Mechanisms of modal and nonmodal phonation,” Journal of Phonetics 29(4), 431–450.

Berry, D. A., Herzel, H., Titze, I. R., and Krischer, K. (1994). “Interpretation of biomechanical simulations of normal and chaotic vocal fold oscillations with empirical eigenfunctions,” The Journal of the Acoustical Society of America 95(6), 3595–3604.

Chollet, F. et al. (2015). “Keras”.

Deverge, M., Pelorson, X., Vilain, C., Lagrée, P.-Y., Chentouf, F., Willems, J., and Hirschberg, A. (2003). “Influence of collision on the flow through in-vitro rigid models of the vocal folds,” The Journal of the Acoustical Society of America 114(6), 3354–3362.

Döllinger, M., Berry, D. A., and Berke, G. S. (2005). “Medial surface dynamics of an in vivo canine vocal fold during phonation,” The Journal of the Acoustical Society of America 117(5), 3174–3183.

Erath, B. D., Zañartu, M., Peterson, S. D., and Plesniak, M. W. (2011). “Nonlinear vocal fold dynamics resulting from asymmetric fluid loading on a two-mass model of speech,”
Forrest, S. (1996). “Genetic algorithms,” ACM Computing Surveys (CSUR) 28(1), 77–80.

Freedman, D., Pisani, R., and Purves, R. (2007). International student edition Statistics: Fourth International Student Edition (W.W. Norton & Company), https://books.google.com/books?id=mviJQgAACAAJ.

Geng, B., Xue, Q., and Zheng, X. (2016). “The effect of vocal fold vertical stiffness variation on voice production,” The Journal of the Acoustical Society of America 140(4), 2856–2866.

Goldberg, D. (2006). Genetic Algorithms (Pearson Education), https://books.google.com/books?id=6gzS07Sv9hoC.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning (MIT press).

Ishizaka, K., and Flanagan, J. L. (1972). “Synthesis of voiced sounds from a two-mass model of the vocal cords,” Bell System Technical Journal 51(6), 1233–1268.

Jiang, J. J., and Zhang, Y. (2002). “Chaotic vibration induced by turbulent noise in a two-mass model of vocal folds,” The Journal of the Acoustical Society of America 112(5), 2127–2133.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). “Deep learning,” Nature 521(7553), 436.

Liang, Y., Lee, H., Lim, S., Lin, W., Lee, K., and Wu, C. (2002). “Proper orthogonal decomposition and its applications?Part I: Theory,” Journal of Sound and Vibration 252(3), 527–544.

Luo, H., Mittal, R., Zheng, X., Bielamowicz, S. A., Walsh, R. J., and Hahn, J. K. (2008). “An immersed-boundary method for flow–structure interaction in biological systems with application to phonation,” Journal of Computational Physics 227(22), 9303–9332.
Mitchell, M. (1998). *An Introduction to Genetic Algorithms* (MIT press).

Mittal, R., Zheng, X., Bhardwaj, R., Seo, J. H., Xue, Q., and Bielamowicz, S. (2011). “Toward a simulation-based tool for the treatment of vocal fold paralysis,” Frontiers in Physiology 2, 19.

Pelorson, X., Hirschberg, A., Van Hassel, R., Wijnands, A., and Auregan, Y. (1994). “Theoretical and experimental study of quasisteady-flow separation within the glottis during phonation. application to a modified two-mass model,” The Journal of the Acoustical Society of America 96(6), 3416–3431.

Ruder, S. (2016). “An overview of gradient descent optimization algorithms,” arXiv preprint arXiv:1609.04747.

Ruty, N., Pelorson, X., Van Hirtum, A., Lopez-Arteaga, I., and Hirschberg, A. (2007). “An in vitro setup to test the relevance and the accuracy of low-order vocal folds models,” The Journal of the Acoustical Society of America 121(1), 479–490.

Scherer, R. C., Titze, I. R., and Curtis, J. F. (1983). “Pressure-flow relationships in two models of the larynx having rectangular glottal shapes,” The Journal of the Acoustical Society of America 73(2), 668–676.

Smith, S. L., and Titze, I. R. (2018). “Vocal fold contact patterns based on normal modes of vibration,” Journal of Biomechanics 73, 177–184.

Steinecke, I., and Herzel, H. (1995). “Bifurcations in an asymmetric vocal-fold model,” The Journal of the Acoustical Society of America 97(3), 1874–1884.

Story, B. H., and Titze, I. R. (1995). “Voice simulation with a body-cover model of the vocal folds,” The Journal of the Acoustical Society of America 97(2), 1249–1260.
Tao, C., and Jiang, J. J. (2008). “Chaotic component obscured by strong periodicity in voice production system,” Physical Review E 77(6), 061922.

Titze, I. R. (1988). “The physics of small-amplitude oscillation of the vocal folds,” The Journal of the Acoustical Society of America 83(4), 1536–1552.

Titze, I. R. (1994). *Principles of Voice Production* (Prentice Hall), https://books.google.com/books?id=m48JAQAAMAAJ.

Van den Berg, J., Zantema, J., and Doornenbal Jr, P. (1957). “On the air resistance and the bernoulli effect of the human larynx,” The Journal of the Acoustical Society of America 29(5), 626–631.

Wurzbacher, T., Schwarz, R., Döllinger, M., Hoppe, U., Eysholdt, U., and Lohscheller, J. (2006). “Model-based classification of nonstationary vocal fold vibrations,” The Journal of the Acoustical Society of America 120(2), 1012–1027.

Xue, Q., Mittal, R., Zheng, X., and Bielamowicz, S. (2012). “Computational modeling of phonatory dynamics in a tubular three-dimensional model of the human larynx,” The Journal of the Acoustical Society of America 132(3), 1602–1613.

Xue, Q., Zheng, X., Bielamowicz, S., and Mittal, R. (2011). “Sensitivity of vocal fold vibratory modes to their three-layer structure: Implications for computational modeling of phonation,” The Journal of the Acoustical Society of America 130(2), 965–976.

Xue, Q., Zheng, X., Mittal, R., and Bielamowicz, S. (2014). “Subject-specific computational modeling of human phonation,” The Journal of the Acoustical Society of America 135(3), 1445–1456.
Zanartu, M., Mongeau, L., and Wodicka, G. R. (2007). “Influence of acoustic loading on an effective single mass model of the vocal folds,” The Journal of the Acoustical Society of America 121(2), 1119–1129.

Zhang, L. T., and Yang, J. (2016). “Evaluation of aerodynamic characteristics of a coupled fluid-structure system using generalized bernoulli’s principle: An application to vocal folds vibration,” Journal of coupled systems and multiscale dynamics 4(4), 241–250.

Zhang, Y., and Jiang, J. J. (2008). “Nonlinear dynamic mechanism of vocal tremor from voice analysis and model simulations,” Journal of Sound and Vibration 316(1-5), 248–262.

Zhang, Y., Zheng, X., and Xue, Q. (2020). “A deep neural network based glottal flow model for predicting fluid-structure interactions during voice production,” Applied Sciences 10(2), 705.

Zheng, X., Xue, Q., Mittal, R., and Beilamowicz, S. (2010). “A coupled sharp-interface immersed boundary-finite-element method for flow-structure interaction with application to human phonation,” Journal of Biomechanical Engineering 132(11), 111003.