Yang Zhang
Department of Mechanical Engineering,
University of Maine,
Orono, ME 04469
e-mail: yang.zhang@maine.edu

Weili Jiang
Department of Mechanical Engineering,
University of Maine,
204 Crosby Hall,
Orono, ME 04473
e-mail: weili.jiang@maine.edu

Luning Sun
Department of Aerospace and
Mechanical Engineering,
University of Notre Dame,
Notre Dame, IN 46556
e-mail: lsun7@nd.edu

Jianxun Wang
Department of Aerospace and
Mechanical Engineering,
University of Notre Dame,
Notre Dame, IN 46556
e-mail: JWANG33@ND.EDU

Xudong Zheng[1]
Department of Mechanical Engineering,
University of Maine,
Room 213 A, Boardman Hall,
Orono, ME 04473
e-mail: xudong.zheng@maine.edu

Qian Xue[1]
Department of Mechanical Engineering,
University of Maine,
Room 213, Boardman Hall,
Orono, ME 04473
e-mail: qian.xue@maine.edu

A Deep Learning-Based Generalized Empirical Flow Model of Glottal Flow During Normal Phonation

This paper proposes a deep learning-based generalized empirical flow model (EFM) 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 EFM is coupled with a finite element method (FEM)-based solid dynamics solver for fluid–structure–interaction (FSI) simulation of phonation. The EFM is evaluated by comparing the N-S solutions in both static glottal shapes and FSI simulations. The results demonstrate a good prediction performance in accuracy and efficiency.

[DOI: 10.1115/1.4053862]

1 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 [1]. 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 *was* the Bernoulli equation, which *simplified* the flow as a one-dimensional inviscid flow [2–4]. By coupling with lumped-mass or continuum vocal fold models, the model *provided* important understandings of the dynamics of FSI during voice production [5–13]. Yet, the inviscid assumption *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 [14–16]. To improve the accuracy, research efforts have been made to incorporate various viscous loss terms into the Bernoulli equation [7,14,17,18]. 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 two-dimensional configurations to complex three-dimensional 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 [19–22], 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 [23–26]. 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 empirical flow model (EFM) for glottal flow was proposed in our previous study [27]. 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

[1] Corresponding authors.

Manuscript received March 10, 2021; final manuscript received February 10, 2022; published online March 24, 2022. Assoc. Editor: Sarah Vigmostad.

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. Can we find a generalized model to represent the vibration pattern of the vocal fold so that we could use for fast and accurate prediction of the underlying flow variables? To answer this question, we perform some preliminary exploration and propose a deep learning-based generalized EFM of the glottal flow during normal phonation in this paper.

The outline of the paper is organized as follows: the overall methodology is presented in Sec. 2; the three-dimensional shape of the vocal fold during vibration, including the prephonatory geometry and universal kinematics equation (UKE), is introduced in Sec. 3; the process of building up the generalized glottal shape library is elaborated in Sec. 4; details about the implementation and evaluation of the DNN model are discussed in Sec. 5; implementation and evaluation of the performance of the present EFM for FSI Simulation are discussed in Sec. 6; finally, the conclusions and limitations are presented in Sec. 7.

2 Overall Methodology

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 [28]. A number of past studies showed that the vocal fold vibration in normal phonation is dominated by two modes [23–25,28]. Therefore, in this 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 UKE can be obtained. To efficiently verify this hypothesis, Bernoulli-finite element method (FEM) FSI simulations with various vocal fold material properties and subglottal pressures are employed as the fast shape generators, and the UKE is examined by generating a large number of glottal shapes from FSI simulations and fitting the glottal shapes with the UKE using the genetic algorithm (GA) [29–31]. We choose GA for the shape fitting because it can be abstracted as a constrained optimization problem with bounded variables. 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 [32] 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). We choose DNN because there is no need to care about the details of the mathematical relationship between the input and output, and the flow variables for any glottal shape that not in the shape library can be well predicted by virtue of the interpolation capability of the trained DNN. K-fold cross validation is performed to fine-tune the architecture and hyperparameters and evaluate the prediction performance of the DNN. The developed empirical 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 (EFM) is evaluated by comparing to the N–S solutions (FOM) in both static glottal shapes and FSI simulations.

3 Three-Dimensional Shape of Vocal Fold During Vibration

3.1 Prephonatory Geometry. The prephonatory geometry of the vocal fold (right half) is shown in Fig. 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 α equals to arctan0.5. 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. The above dimensions are selected in the range typical for adult humans [5,19,28]. The vocal fold model is discretized with 10,810 tetrahedral elements, the mesh density is comparable to our previous three-dimensional simulations of similar configurations [21,33,34] where grid convergence studies were performed.

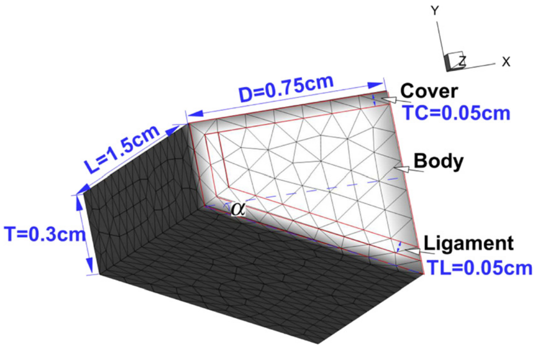


Fig. 1 Prephonatory geometry of the vocal fold

3.2 Universal Kinematics Equation. Past in vivo and ex vivo studies have shown that vocal fold vibrations are dominated by a few vibratory modes in real physiological conditions [23–26]. Following the surface–wave approach in Ref. [12], the kinematics of the medial surface of the vocal fold 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 medial–lateral motion and (1,1) represents the convergent–divergent motion [12,28]. The displacement of the medial surface over time can be represented by a linear combination of the modal displacement of these two modes

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

where the subscripts (1,0) and (1,1), respectively, refer to modes (1,0) and (1,1), and α is the weight coefficient of mode (1,0). An equivalent equation exists for the left-half vocal fold. Note that in our study, to simplify the model, only the lateral (x) vibration is allowed and the vertical (y) motion is fixed. This treatment is the same as that adopted in Refs. [12] and [28].

In Ref. [28], based on the surface–wave approach and small-angle approximation [12], the modal displacement of the medial surface of the vocal fold at any instant in time was defined as

$$\xi(y,z,t)_{(m,n)} = \xi_m \sin\left(\frac{m\pi z}{L}\right)\left[\sin\omega t - n\left(\frac{\omega}{c}\right)(y-y_m)\cos\omega t\right] \tag{2}$$

where ξ_m is the modal displacement amplitude, y_m is the inflection point for the vertical half wavelength, ω is angular frequency, and c is the speed of the mucosal wave [28].

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

$$\xi(y,z,t) = \xi_m \sin\left(\frac{\pi z}{L}\right)\left[\sin\omega t - (1-\alpha)\left(\frac{\omega}{c}\right)(y-y_m)\cos\omega t\right] \tag{3}$$

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

Table 1 Estimated physiological range of the parameters in the UKE

Parameters	Range
ξ_m	[0, 0.1 cm]
α	[0, 1]
β	[0, $T/2$]
ϕ	[0, 24]
f	[100 Hz, 250 Hz]

$$y_m = T - \beta\left(\sin\frac{\pi z}{L} + 1\right) \tag{4}$$

where $0 \le \beta \le$ 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. Equation (3) is also termed as the UKE in this paper.

4 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 ξ_m, weight coefficient of mode (1,0) α, inflection point factor β, phase $\phi = 12\omega t/\pi$, and ratio between the angular frequency and mucosal wave speed ω/c, which is related to the vibration frequency f. The estimated physiological range of these parameters for normal phonation [28] is listed in Table 1. It is worth pointing out that the variation in terms of the length of the vocal fold is not considered, which simplifies the transverse isotropic model with a constant ligament stiffness.

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 [33]. A large number of glottal shapes are extracted from the simulation results and used to fit the UKE by using the GA [29–31]. The fitting error is used to quantify the representative capability of the UKE. Finally, the PDF of each input parameter in the UKE is obtained and used to build the generalized glottal shape library through appropriate resampling.

4.1 Bernoulli-Finite Element Method Fluid–Structure–Interaction Simulation. The vocal fold tissue is modeled as the viscoelastic, transversely isotropic material. The baseline material properties of each layer of the vocal fold [5,34] are listed in Table 2.

Based on the baseline material properties listed in Table 2, the ranges of the material properties for each layer can be obtained by simultaneously multiplying the corresponding E^0_{pz} and G^0_{pz} 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 values 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 $P_0 = 0.5$ kPa, 0.75 kPa, 1.0 kPa can be respectively expressed as

Table 2 Baseline material properties of each layer of the vocal fold

	ρ (g/cm^3)	E_p (kPa)	υ_p	E^0_{pz} (kPa)	υ_{pz}	G^0_{pz} (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^0_{pz} are the transversal and longitudinal Young's Modulus, respectively; υ_p and υ_{pz} are the transversal and longitudinal Poisson ratio, respectively; G^0_{pz} is the longitudinal shear modulus [5,34].

$$k_{CL} = m\Delta k, \quad m = 1, 2, \ldots, 10 \tag{5}$$

$$k_B = n\Delta k, \quad n = 1, 2, \ldots, 10 \tag{6}$$

where the subscripts 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 1.145×10^{-3} g/cm^3 and $\nu = 1.655 \times 10^{-1}$ cm^2/s, respectively. The glottis is discretized with $N_S = 69$ uniformly spaced cross sections along the inferior–superior direction such that the spacing is 0.01 cm. Similar to the treatment adopted in Ref. [28], the contact surface is calculated as an average of the left and right surface coordinates. Note that this treatment is consistent in the subsequent EFM-FSI and FOM-FSI model. 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 $P_0 = 1.0$ kPa, $k_{\mathrm{CL}} = 1.0$, $k_B = 4.0$ is illustrated in Fig. 2, 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 five representative phases probed from the left subfigure. The vibration shows a typical alternative convergent-divergent glottal shape change.

4.2 Glottal Shape Fitting With the Genetic Algorithm. In this section, we aim to verify that those glottal shapes extracted from the Bernoulli-FEM FSI simulations 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 1 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 cross section 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

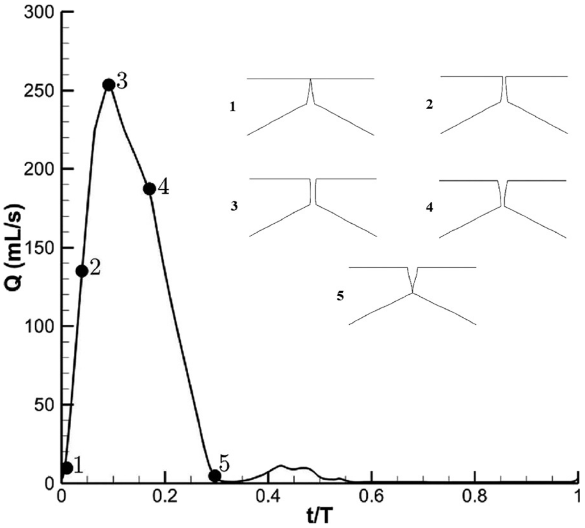


Fig. 2 Glottal flow rate and vocal fold vibration pattern during one cycle of a representative Bernoulli-FEM FSI simulation case at $P_0 = 1.0$ kPa, k_{CL}=1.0, $k_B = 4.0$

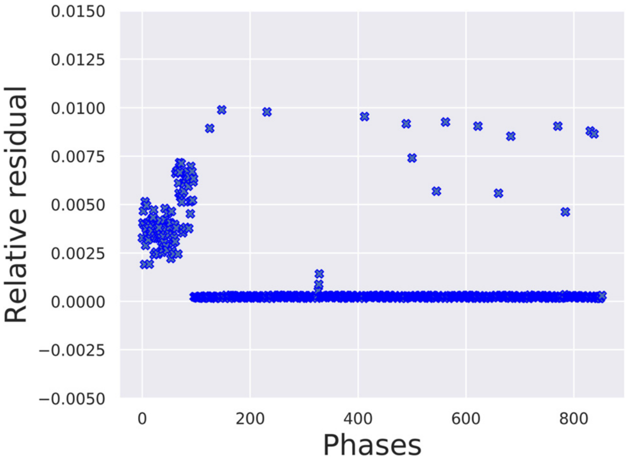


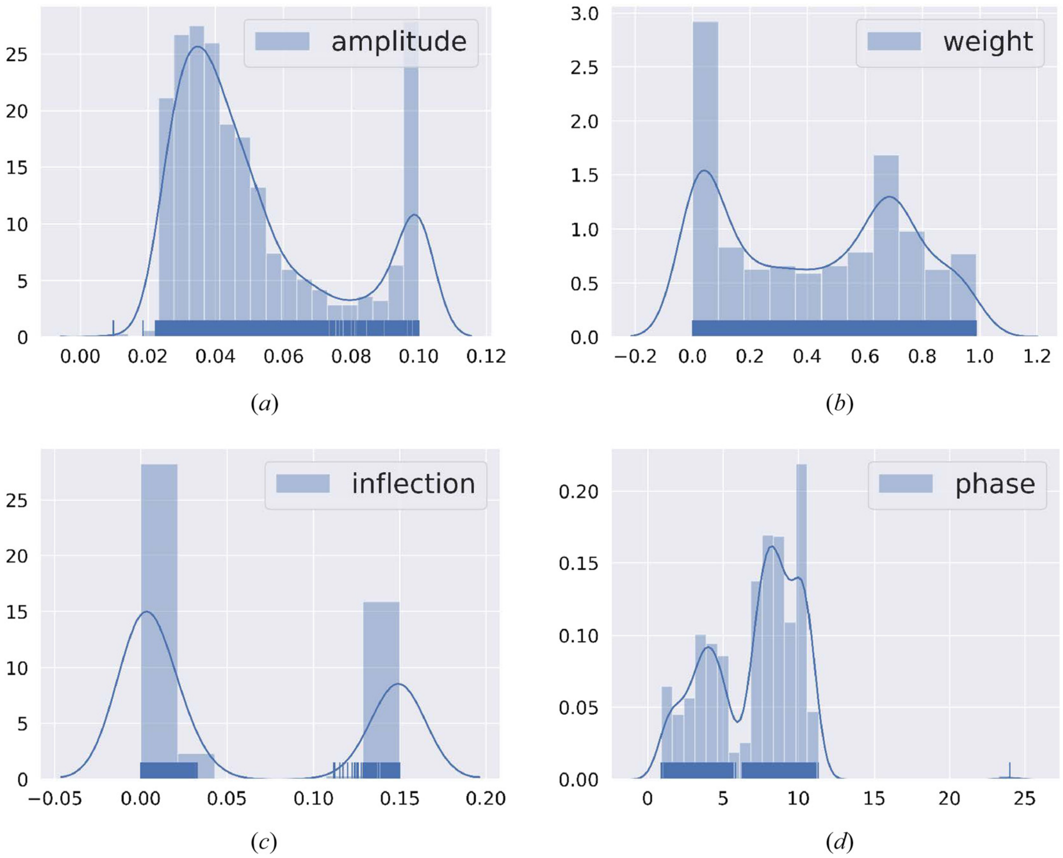
Fig. 3 Relative residual of the fitness function of GA

$$\xi_m, \alpha, \beta, \phi, f = \text{argmin}\frac{\sum_{i=1}^{n}[\xi^i_{\text{optimized}}(\xi_m, \alpha, \beta, \phi, f) - \xi^i_{\text{target}}]^2}{n}$$

$$\text{Subject to argmin } A_j^{\text{optimized}} = \text{argmin} A_j^{\text{target}},\ (A_j^{\text{optimized}})_{\min} = (A_j^{\text{target}})_{\min} \tag{7}$$

where argmin refers to the argument of the minimum, the values of $\xi_m, \alpha, \beta, \phi, f$ are bounded by the corresponding ranges listed in Table 1, n is the number of nodal points of the glottis surface, and $A_j^{\text{optimized}}$ and A_j^{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 generations for the GA are chosen based on a trial-and-error experiment such that the optimization accuracy and efficiency is balanced. Specifically, the optimization is run for 6 times until the relative change of the fitness function doesn't show significant difference with a prescribed convergence criterion. For this case, the corresponding values are chosen as 160 and 100, respectively. The overall residual of the fitness function extracted from the Bernoulli-FEM FSI cases is plotted in Fig. 3. 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 [35] is used as a nonparametric way to estimate the PDF of the fitting parameters, and the corresponding PDF for $P_0 = 0.75$ kPa is plotted in Fig. 4. The PDF for $P_0 = 0.5$ kPa and $P_0 = 1.0$ 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 $f = 210$ Hz. Therefore, to reduce the number of redundant shapes, we fix the value of the optimized frequency to be $f = 210$ 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 resampled values of the input parameters under different subglottal pressure



(*a*)

(*b*)

(*c*)

(*d*)

Fig. 4 PDF of optimized shape parameters for $P_0 = 0.75$ kPa: (*a*) ξ_m, (*b*) α, (*c*) β, and (*d*) ϕ

Table 3 Resampled values of input parameters

P_0 (kPa)	ξ_m	α	β	ϕ
0.5	0.02, 0.03, 0.04, 0.1	0.0, 0.2, 0.4, 0.6, 0.8, 1.0	0.0, 0.015, 0.03, 0.135, 0.15	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11
0.75	0.025, 0.04, 0.055, 0.1			
1.0	0.035, 0.055, 0.075, 0.1			

conditions are listed in Table 3. Note that for different subglottal pressure values, only the amplitude ξ_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 3 into the UKE, and these shapes constitute the generalized glottal shape library which are used as the raw data for training the DNN in Sec. 5.2.

5 Implementation of the Deep Neural Network Model

For each shape in the generalized glottal shape library, the subglottal pressure P_0 and the parameters ξ_m, α, β, and ϕ 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:

$$Q,\ P_i = f(P_0,\ \xi_m,\ \alpha,\ \beta,\ \phi;\ \theta) \tag{8}$$

where f is the function representing the overall DNN, and θ 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.

5.1 N–S Solution of the Output Targets.

The fluid flow is governed by the incompressible N–S equations as follows:

$$\frac{\partial u_i}{\partial x_i} = 0 \tag{9}$$

$$\frac{\partial u_i}{\partial t} + \frac{\partial u_i u_j}{\partial x_j} = -\frac{1}{\rho_f}\frac{\partial p}{\partial x_i} + \upsilon_f \frac{\partial^2 u_i}{\partial x_j \partial x_j} \tag{10}$$

where u_i, ρ, p, and υ are the incompressible flow velocity, density, pressure, and kinematic viscosity, respectively. An in-house sharp-interface immersed-boundary N–S flow solver [22] is used to obtain the ground truth solution of the output targets. The size of the computational domain is 1.5 cm × 21.0 cm × 1.5 *cm* in the x (lateral), y (inferior–superior), and z (anterior–posterior) direction. The vocal folds are placed 3.2 cm and 17.0 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 Courant–Friedrichs–Lewy 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. From the results, 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 cm and 0.01 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 h on a parallel computer with 32 CPUs.

5.2 Implementation Details of the Deep Neural Network.

As mentioned above, the input features and corresponding output targets extracted from the shape library can be organized as a vector x and y, respectively,

$$\begin{aligned} x &= [P_0\ \xi_m\ \alpha\ \beta\ \phi]^{\mathrm{T}} \\ y &= [Q\ P_1\ P_2 \ldots P_{N_P}]^{\mathrm{T}} \end{aligned} \tag{11}$$

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 [32,36]. 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 $l = 1,\ldots, L-1$. Neurons in the hidden layer z_l have connections to all neurons of the previous layer z_{l-1}

$$z_l = \sigma_l(W_l^T z_{l-1} + b_l) \tag{12}$$

where W_l is the learnable weights, b_l is the additive bias, and σ_l is the nonlinear activation function.

The loss function J of the DNN is

$$J = \frac{1}{N}\sum ||z_L - y||_2^2 + \lambda ||W||_2 \tag{13}$$

where z_L is the predicted value, and λ is the regularization coefficient to prevent the overfitting of the DNN model and its value is taken as 0.001.

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 $\boldsymbol{x}$ are, respectively, mapped to the subsets of the output targets $\boldsymbol{y}$ (i.e., Q and P_i) with different architectures of the DNN.

The whole dataset from the shape library is randomly split into the training and test sets. To avoid the overfitting of the model, we use five-fold cross validation [32] 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 minibatch size, and the number of epochs [32]. 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.

Two separate networks are used for training the Q and P_i, denoted as DNN-Q and DNN-P, respectively. The input layer for both DNNs has five 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 [32] are used

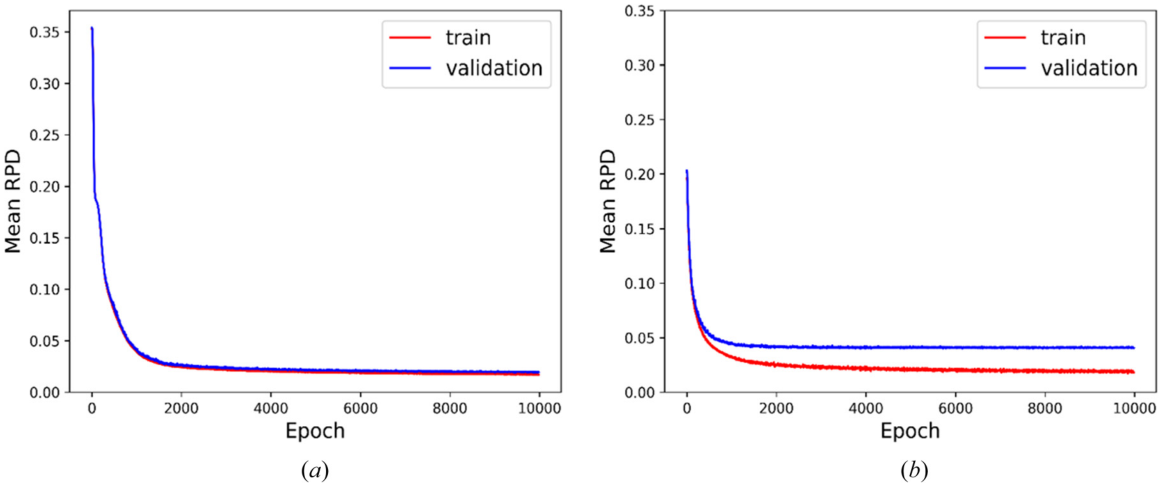


Fig. 5 Convergence history of the DNNs for flow rate and pressure using fivefold cross validation: (*a*) DNN-Q and (*b*) DNN-P

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 is 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 [32] 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) [37]. Both of the DNN models are trained with 10,000 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 [38] using TENSORFLOW [39] as the backend.

5.3 Evaluation of the Trained Deep Neural Network 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 is as follows:

$$E_Q = \frac{|Q - \hat{Q}|}{\max(|Q|, |\hat{Q}|)} \tag{14}$$

$$E_P = \frac{\sum_{i=1}^{N_P} \frac{|P_i - \hat{P}_i|}{\max(|P_i|, |\hat{P}_i|)}}{N_P} \tag{15}$$

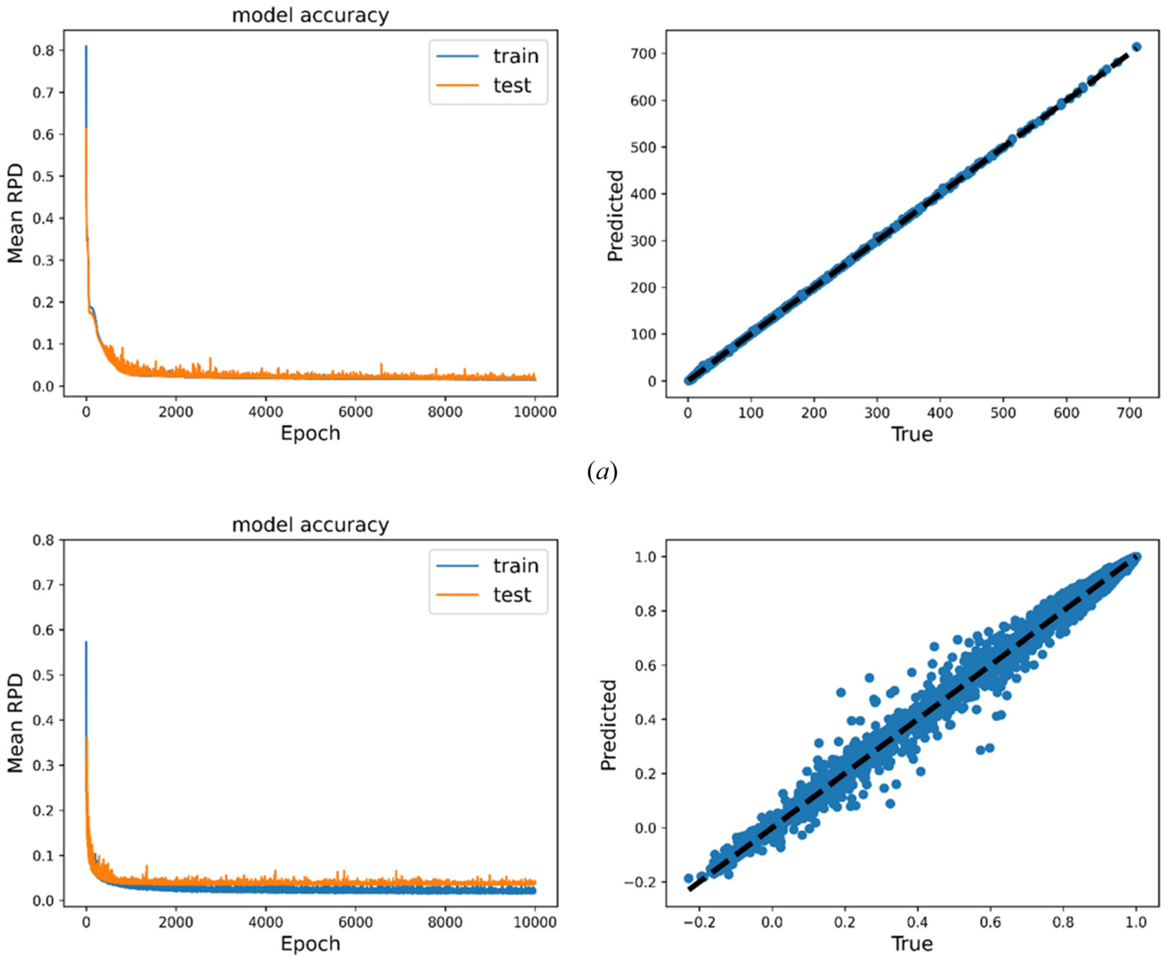
where Q, P_i and $\hat{Q}$, $\hat{P}_i$ are, respectively, the true and predicted outcomes.

The history of the fivefold cross validation results for DNN-Q and DNN-P is plotted in Fig. 5. The horizontal axis corresponds to the number of epochs, and the vertical axis corresponds to the mean RPD between the true and predicted outcomes. The comparison is between the training and validation sets. It took 10,000 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 Fig. 6. After running 10,000 epochs, the mean RPD on the test set converges at 1.74% and 3.52% for DNN-Q and DNN-P, respectively. The scatter plots of the true and predicted outcomes on the test set show a good prediction performance. Note that the plot of DNN-P is more scattered than that of DNN-Q. Although DNN-P has more neurons in the hidden layers than DNN-Q, given that the dimension of the output pressure distribution is much higher than the output flow rate as well as the input parameters, it's more challenging to predict the pressure distribution. Further improvements could be introducing more advanced neural network architectures (e.g., convolutional neural network [32], long short-term memory network [32]) and feeding inputs with higher dimensions into the neural networks. The final mean RPD on the training, validation and test sets for DNN-Q and DNN-P are summarized in Table 4.

Furthermore, six 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 are shown in Fig. 7. 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 empirical flow model is illustrated in Fig. 8. Concretely, it is divided into the following steps: first, various glottal shapes are extracted from 300 converged Bernoulli-FEM FSI results under different subglottal pressure and material properties. Second, these extracted shapes are fitted with the UKE using the GA and the PDF of the fitted input parameters of the UKE are determined. Third, 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. Fourth, 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 empirical flow model, 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 empirical flow model can be summarized in Table 5.

The developed empirical flow model is then coupled with the FEM based solid dynamics solver for FSI simulation. The abstract workflow of the EFM for FSI simulation is illustrated in Fig. 9. First, the flow rate Q and pressure distribution P_i of the glottal shape X at a certain time instant t can be obtained by the present empirical flow model, then the pressure load is fed into the FEM solid solver to calculate the corresponding deformation of the glottis ΔX, 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 empirical flow model and FEM based solid solver are coupled in a weak manner, i.e., they are solved sequentially/explicitly with only one fixed-point iteration required at each time-step.



(*a*)

(*b*)

Fig. 6 Performance of the trained DNN models on the test set: (*a*) DNN-Q and (*b*) DNN-P

Table 4 Mean RPD on the training, validation, and test sets

	Train	Validation	Test
Q	1.71%	1.89%	1.74%
P_i	1.97%	4.12%	3.52%

6 Evaluation of the Performance of the Generalized Empirical Flow Model for Fluid–Structure–Interaction Simulation

To evaluate the prediction performance of the present generalized EFM for FSI simulation, the EFM-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.

6.1 Comparison With Full-Order Model-Quasi-Static Results. A series of new subglottal pressure and material properties are simulated using the EFM-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 6. The simulation setup is the same with the Bernoulli-FEM FSI simulation. An example of the converged time history of the flow rate Q at $P_0 = 0.8\,\text{kPa}$, $k_{CL} = 4.75$, $k_B = 3.75$ predicted by the EFM is illustrated in Fig. 10. Note that some small 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 because of the contact issue (i.e., the contact surface is calculated by averaging the left and right surface coordinates, which may not strictly satisfy the UKE) and the intrinsic weak extrapolation capability of the DNN. However, since these values are very small, the whole prediction performance will barely be affected.

Full-order model-QS is achieved by extracting various glottal shapes from the converged EFM-FSI results at different phases, and then feeding each extracted shape into the standalone N–S solver to obtain the corresponding ground-truth flow rate and pressure distribution. To this end, various glottal shapes are extracted from the converged FSI results of the cases listed in Table 6. 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 6, at each time-step of the steady-cycle EFM-FSI result, the flow rate $Q^{n,k}_{\text{EFM}}$ and pressure distribution $P^{n,k}_{i,\text{EFM}}$ are, respectively, extracted, and the corresponding reference values of $Q^{n,k}_{\text{FOM}}$ and $P^{n,k}_{i,\text{FOM}}$ 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^n_Q and E^n_P, can be calculated as follows:

$$E^n_Q = \frac{1}{n_t \bar{Q}^n_{\text{FOM}}} \sum_{k=1}^{n_t} \left| Q^{n,k}_{\text{FOM}} - Q^{n,k}_{\text{EFM}} \right| \tag{16}$$

$$E^n_P = \sum_{k=1}^{n_t} \sum_{i=1}^{N_P} \frac{\left| P^{n,k}_{i,\text{FOM}} - P^{n,k}_{i,\text{EFM}} \right|}{P_0} \tag{17}$$

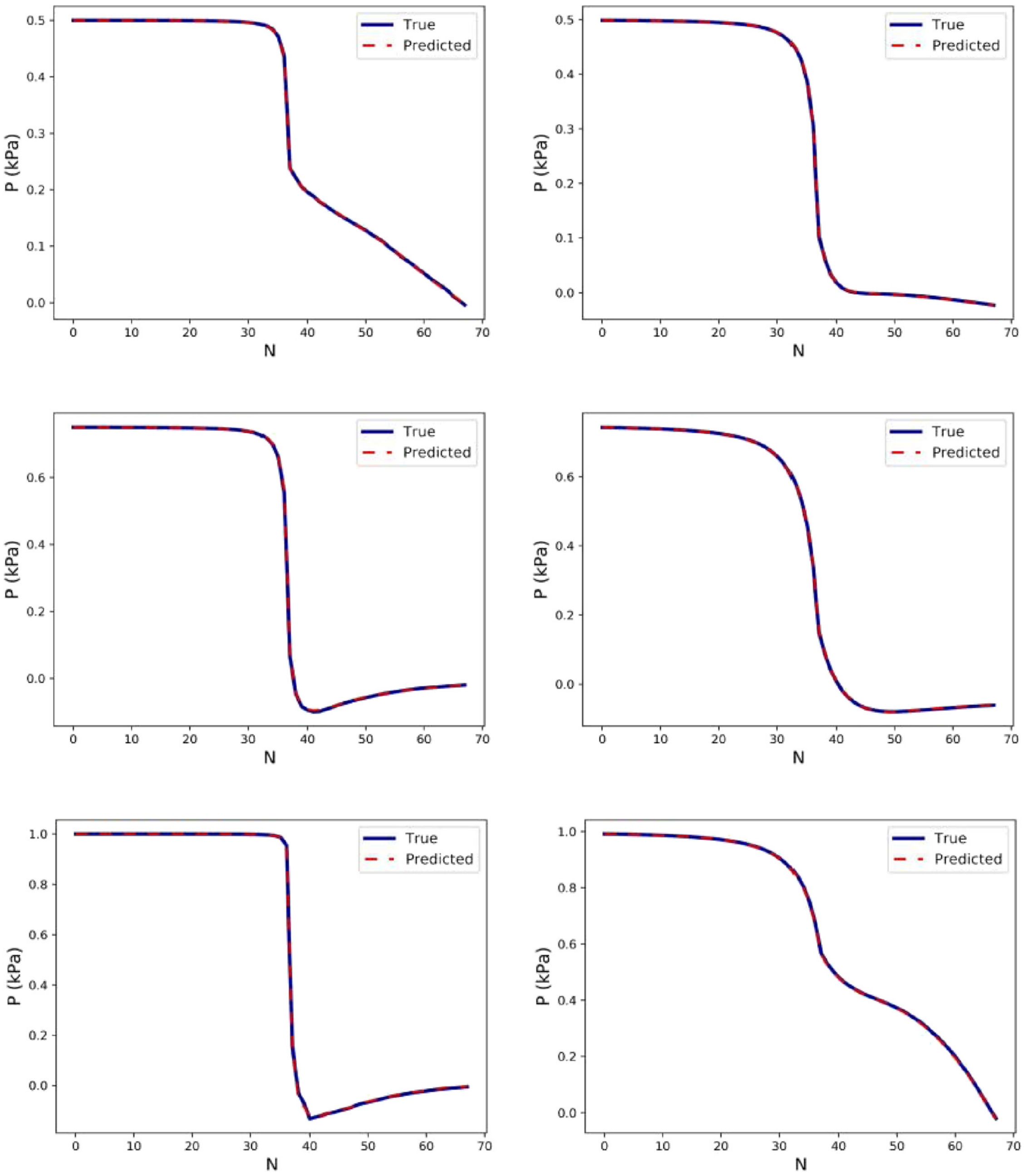


Fig. 7 Comparison of the true and predicted pressure distribution in six randomly selected glottal shapes

where n_t and $\bar{Q}^n_{\mathrm{FOM}}$ 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

$$E_Q = \frac{1}{n_c}\sum_{n=1}^{n_c} E^n_Q \tag{18}$$

$$E_P = \frac{1}{n_c}\sum_{n=1}^{n_c} E^n_P \tag{19}$$

where n_c is the number of cases listed in Table 6. 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 between Q_{FOM} and Q_{EFM} is excellent (0.993, $P < 0.0005$). The scatter and correlation plots are also depicted in Fig. 11, where the horizontal and vertical axes correspond to the true (Q_{FOM}) and predicted (Q_{EFM}) values, respectively. The Bland–Altman plot [40] is used to analyze the agreement between Q_{FOM} and Q_{EFM}. The result is plotted in Fig. 12. As can be seen from this figure, the mean difference between Q_{FOM} and Q_{EFM} is -2.784 mL/s, and the 95% limits of agreement (LoA) between them is from -12.505 mL/s to 6.936 mL/s. The 95% confidence interval of the mean difference, upper LoA and lower LoA between Q_{FOM} and Q_{EFM} is [-3.0288 mL/s, -2.5401 mL/s], [6.5177 mL/s, 7.3539 mL/s] and [-12.9288 mL/s, -12.0866 mL/s], respectively. The number of the outliers, which mainly come from the divergent glottal shapes at the closing phase, is 38, and the percentage of the outliers is 2.40%.

Similarly, in terms of P_i, the Pearson correlation coefficient between $P_{i,\mathrm{FOM}}$ and $P_{i,\mathrm{EFM}}$ is excellent (0.997, $P < 0.0005$). The

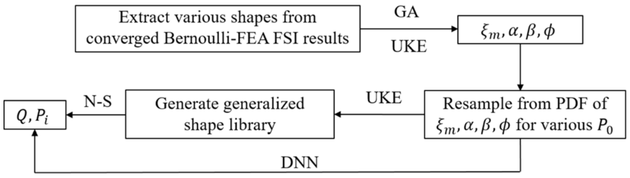

Extract various shapes from converged Bernoulli-FEA FSI results
GA
UKE
$\xi_m, \alpha, \beta, \phi$
Resample from PDF of $\xi_m, \alpha, \beta, \phi$ for various P_0
UKE
Generate generalized shape library
N-S
Q, P_i
DNN

Fig. 8 Diagram of the implementation of the empirical flow model

scatter and correlation plots are also depicted in Fig. 13, where the horizontal and vertical axes correspond to the true ($P_{i,\mathrm{FOM}}$) and predicted ($P_{i,\mathrm{EFM}}$) values, respectively. The Bland–Altman analysis between $P_{i,\mathrm{FOM}}$ and $P_{i,\mathrm{EFM}}$ is plotted in Fig. 14. From this figure, we can observe that the mean difference between $P_{i,\mathrm{FOM}}$ and $P_{i,\mathrm{EFM}}$ is 0.006 kPa, and the 95% LoA between them is from -0.011 kPa to 0.023 kPa. The 95% confidence interval of the mean difference, upper LoA and lower LoA between $P_{i,\mathrm{FOM}}$ and $P_{i,\mathrm{EFM}}$ is [0.0053 kPa, 0.0062 kPa], [0.0218 kPa, 0.0232 kPa] and [−0.0117 kPa, −0.0103 kPa], 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 EFM-FSI results agree very well with the corresponding FOM-QS results.

6.2 Comparison With Full-Order Model-Fluid–Structure–Interaction Results.

FSI simulations at P_0= 0.8 kPa, $k_{CL} = 1.75$, $k_B = 3.75$ (case 1) and $P_0 = 0.875$ kPa, $k_{CL} = 3.75$, $k_B = 3.75$ (case 2) from Table 6 are conducted by using both the EFM-FSI model and FOM-FSI model. The comparison of the phase-averaged time history of the flow rate Q for both cases is illustrated in Fig. 15. From this figure, we can observe that the peak flow rate, mean flow rate, and fundamental frequency are close to each other while the opening quotient and skewing of the waveform are different. The phase-averaged values of these quantities are listed in Table 7. The relative errors of the F_0, $Q_{\max}$, and Q_{mean} between the EFM-FSI and NS-FSI simulations are below 11%, while it is as high as 17% and 48%, respectively, for the opening quotient and skewing quotient. The large errors in the opening quotient and skewing quotient 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, (b) the EFM-FSI model is a quasi-steady model while the FOM-FSI is a fully unsteady model. The quasi-steady assumption is known to affect the waveform of the glottal flow. Moreover, the consistent underestimation of the open and skew coefficients would point to a more symmetric waveform, which could result from a lack of higher harmonics. This might indicate the need for higher order modes. Therefore, further improvements on the UKE model may be considered.

Furthermore, the average computational time required for one vibration cycle of the EFM-FSI and FOM-FSI simulation is compared. In order to obtain one vibration cycle, the average time required for the EFM-FSI simulation is 1.5 h on a single CPU, while that required for the FOM-FSI simulation is 20 h on a parallel computer with 64 CPUs, which indicates the high efficiency of the present EFM for FSI simulation of the glottal flow.

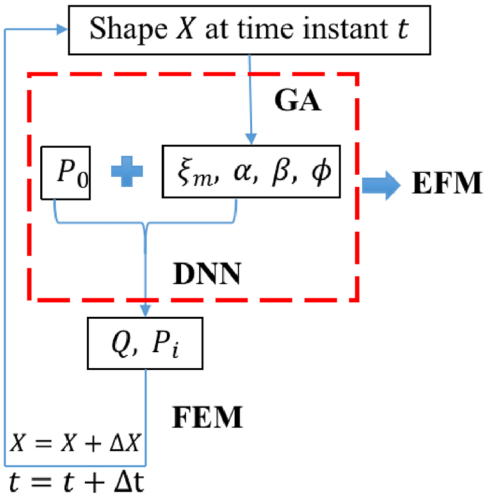


Fig. 9 Workflow of the empirical flow model for FSI simulation

Table 6 Selected subglottal pressure and material properties for evaluation

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

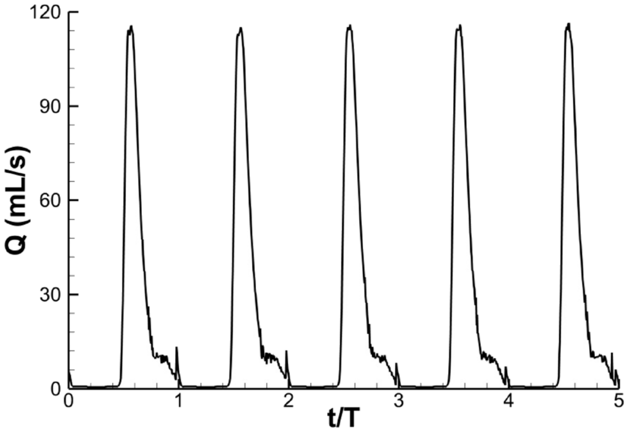
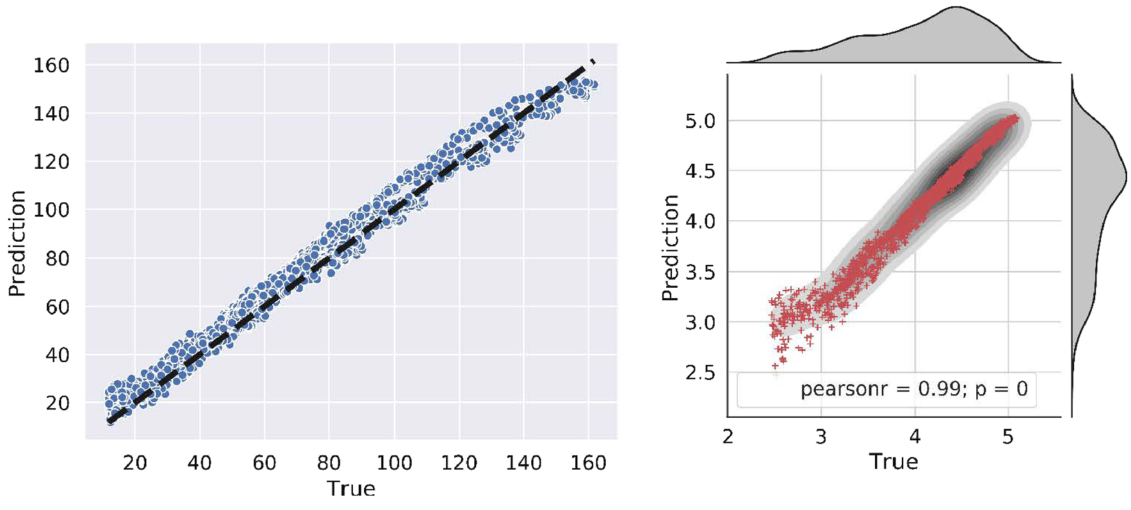


Fig. 10 Example of the converged time history of the flow rate Q predicted by EFM-FSI at P_0 = 0.8 kPa, k_{CL} = 4.75, k_B = 3.75

Table 5 Algorithm of the implementation of the empirical flow model

1 Extract various shapes from converged Bernoulli-FEM FSI results;
2 Fit these extracted shapes with the UKE using the GA;
3 Obtain the PDF of the fitted parameters of the UKE: ξ_m, α, β, and ϕ;
4 Resample the PDF of ξ_m, α, β, and ϕ 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.



(*a*) (*b*)

Fig. 11 Scatter and correlation plot of *Q* comparing EFM-FSI solutions to quasi-static N–S solutions: (*a*) scatter-plot and (*b*) correlation-plot

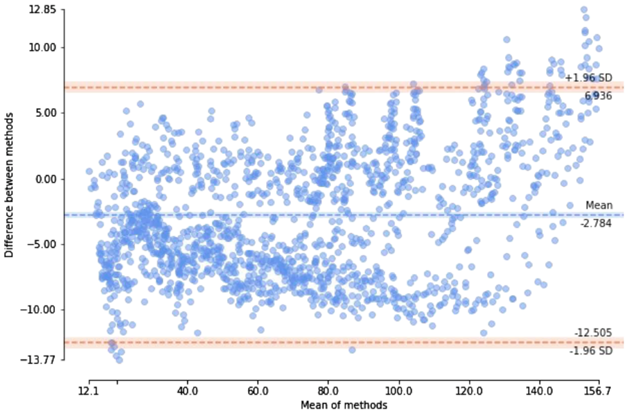


Fig. 12 Bland–Altman analysis plot of *Q* comparing EFM-FSI solutions to quasi-static N–S solutions

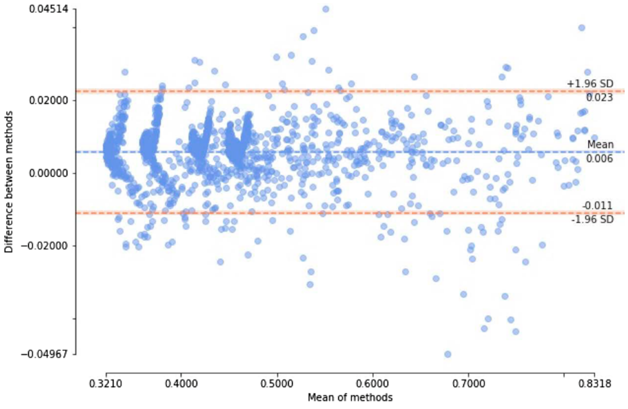
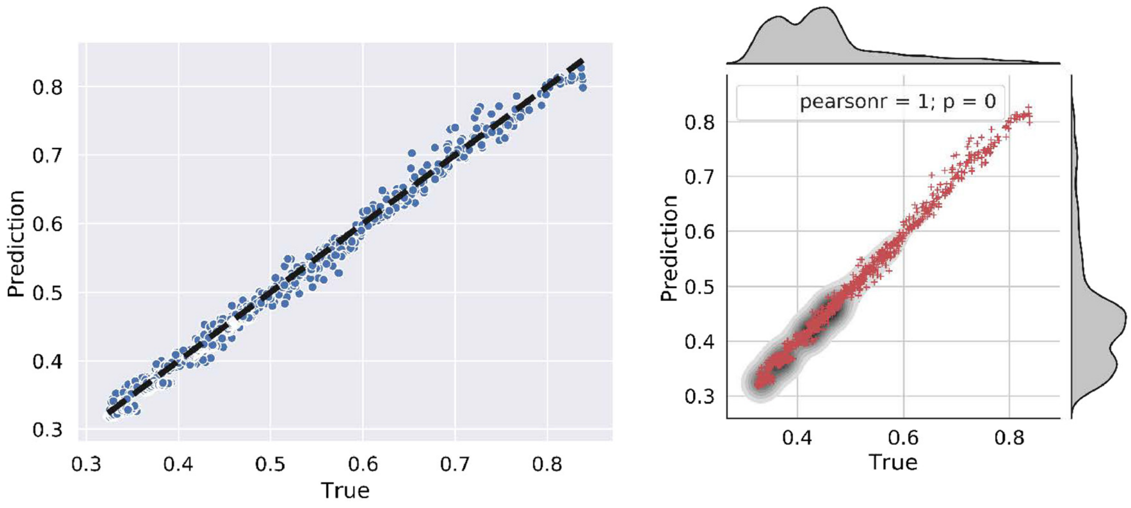
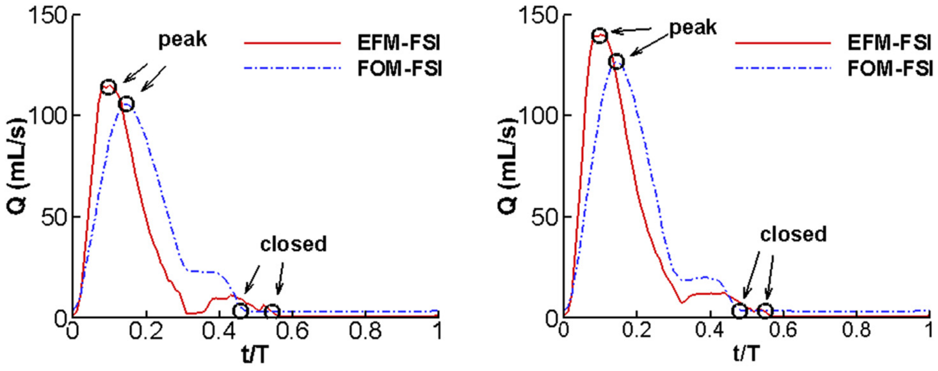


Fig. 14 Bland–Altman analysis plot of P_i comparing EFM-FSI solutions to quasi-static N–S solutions



(*a*) (*b*)

Fig. 13 Scatter and correlation plot of P_i comparing EFM-FSI solutions to quasi-static N–S solutions: (*a*) scatter-plot and (*b*) correlation-plot



(*a*)

(*b*)

Fig. 15 Comparison of the phase-averaged time history of the flow rate between EFM-FSI simulations and FOM-FSI simulations: (*a*) case 1 and (*b*) case 2

Table 7 Comparison of voice quality-related parameters between EFM-FSI simulations and FOM-FSI simulations

	EFM-FSI Case 1	FOM-FSI Case 1	δ_1(%)	EFM-FSI Case 2	FOM-FSI Case 2	δ_2 (%)
F_0 (Hz)	210.8	207.9	1.4	212.0	219.3	3.3
$Q_{\max}$ (mL/s)	115.0	105.5	9.0	140.0	126.7	10.6
Q_{mean} (mL/s)	54.8	53.0	3.4	63.6	58.9	7.9
τ_o	0.54	0.46	17.4	0.55	0.48	14.6
τ_s	0.23	0.44	48.1	0.22	0.41	46.0

F_0 is the fundamental frequency; $Q_{\max}$ and Q_{mean} are the peak and mean glottal flow rate of the open quotient, respectively; τ_o is the open quotient, defined as the ratio of the duration of the glottal open phase to the cycle period; τ_s is the skewing quotient, defined as the ratio of the duration of the flow increasing phase to the duration of the flow decreasing phase [21]; δ_1 and δ_2 are the absolute value of the relative error between the EFM-FSI and FOM-FSI results for cases 1 and 2, respectively.

7 Conclusion

A deep learning-based generalized EFM 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 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 GA. Furthermore, the 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 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 empirical 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 empirical flow model is directly coupled with a FEM based solid dynamics solver for FSI simulation. The EFM-FSI results are compared with the full-order model (FOM) QS and FSI results. For the comparison with the FOM-QS model, the EFM 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 is 7.87% and 1.68%, respectively. For the comparison with the FOM-FSI model, the EFM 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 EFM shows a relatively larger error in predicting the opening quotient and skewness quotient. 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 EFM-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 EFM in accuracy and efficiency indicates a great promise for future clinical use. The developed EFM can be further extended to predict the dynamics of the glottal flow during abnormal phonations with relative ease.

Nevertheless, we acknowledge that there are limitations for the present EFM which need to be addressed in the future work. The limitations are summarized as follows:

(1) Although the two-mode representation is reasonable for describing the glottal shapes during normal vocal fold vibration, it would fail when the vibration pattern becomes more complex, i.e., asymmetric vibration, anterior–posterior wave. For these cases, including higher order modes in the UKE would be necessary, which will be explored in future studies.

(2) The model is assumed to vibrate only along the lateral direction while the vertical motion is fixed. This limitation needs to be addressed by including the vertical motion in the UKE model in the future.

(3) Another complexity not included in this study is the initial shape of the glottis. The deformation modes describe the profile of the medial surface of the vocal fold, and the vocal folds of different materials/geometries can have the same medial surface profiles to be described by the same modes and the corresponding parameters. However, the initial shape of the glottis is related to the library. We assumed a fully closed prephonatory glottal shape. In realistic cases, various shapes could occur. This complexity also needs to be included in the future.

(4) The quasi-steady assumption we used in the present model is based on the work of Ref. [41], which demonstrated that the flow acceleration/deceleration term is an order smaller than other terms during the most of the vibration cycle and only significant during late closing stage. The model at the

current stage does not include unsteady effects, causing errors in FSI simulations, as can be seen from the deviated skewing of the flow rates in Fig. 15. In the future work, unsteady effects could be included in the EFM and the long short-term memory [42] network could also be employed for better and robust time series prediction.

Acknowledgment

The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIDCD or the National Institutes of Health (NIH).

Funding Data

- National Institute on Deafness and Other Communication Disorders (NIDCD) (Grant No. 5R21DC016428; Funder ID: 10.13039/100000055).
- Extreme Science and Engineering Discovery Environment (XSEDE) (Award Nos. TG-BIO150055 and TG-CTS180004; Funder ID: 10.13039/100000001).

References

[1] Titze, I. R., 2000, *Principles of Voice Production*, National Center for Voice and Speech, Iowa City, IA.

[2] 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," J. Acoust. Soc. Am., **121**(1), pp. 479–490.

[3] Wurzbacher, T., Schwarz, R., Döllinger, M., Hoppe, U., Eysholdt, U., and Lohscheller, J., 2006, "Model-Based Classification of Nonstationary Vocal Fold Vibrations," J. Acoust. Soc. Am., **120**(2), pp. 1012–1027.

[4] Zañartu, M., Mongeau, L., and Wodicka, G. R., 2007, "Influence of Acoustic Loading on an Effective Single Mass Model of the Vocal Folds," J. Acoust. Soc. Am., **121**(2), pp. 1119–1129.

[5] Alipour, F., Berry, D. A., and Titze, I. R., 2000, "A Finite-Element Model of Vocal-Fold Vibration," J. Acoust. Soc. Am., **108**(6), pp. 3003–3012.

[6] 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," Chaos, **21**(3), p. 033113.

[7] Ishizaka, K., and Flanagan, J. L., 1972, "Synthesis of Voiced Sounds From a Two-Mass Model of the Vocal Cords," Bell Syst. Tech. J., **51**(6), pp. 1233–1268.

[8] Jiang, J. J., and Zhang, Y., 2002, "Chaotic Vibration Induced by Turbulent Noise in a Two-Mass Model of Vocal Folds," J. Acoust. Soc. Am., **112**(5), pp. 2127–2133.

[9] Steinecke, I., and Herzel, H., 1995, "Bifurcations in an Asymmetric Vocal-Fold Model," J. Acoust. Soc. Am., **97**(3), pp. 1874–1884.

[10] Story, B. H., and Titze, I. R., 1995, "Voice Simulation With a Body-Cover Model of the Vocal Folds," J. Acoust. Soc. Am., **97**(2), pp. 1249–1260.

[11] Tao, C., and Jiang, J. J., 2008, "Chaotic Component Obscured by Strong Periodicity in Voice Production System," Phys. Rev. E Stat. Nonlinear, Soft Matter Phys., **77**(6), pp. 1–8.

[12] Titze, I. R., 1988, "The Physics of Small-Amplitude Oscillation of the Vocal Folds," J. Acoust. Soc. Am., **83**(4), pp. 1536–1552.

[13] Zhang, Y., and Jiang, J. J., 2008, "Nonlinear Dynamic Mechanism of Vocal Tremor From Voice Analysis and Model Simulations," J. Sound Vib., **316**(1–5), pp. 248–262.

[14] 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," J. Acoust. Soc. Am., **114**(6), pp. 3354–3362.

[15] Pelorson, X., Hirschberg, A., van Hassel, R. R., Wijnands, A. P. J., and Auregan, Y., 1994, "Theoretical and Experimental Study of Quasisteady-Flow Separation Within the Glottis During Phonation. Application to a Modified Two-Mass Model," J. Acoust. Soc. Am., **96**(6), pp. 3416–3431.

[16] Scherer, R. C., Titze, I. R., and Curtis, J. F., 1983, "Pressure-Flow Relationships in Two Models of the Larynx Having Rectangular Glottal Shapes," J. Acoust. Soc. Am., **73**(2), pp. 668–676.

[17] Zhang, L., 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," J. Coupled Syst. Multiscale Dyn., **4**(4), pp. 241–250.

[18] van den Berg, J., Zantema, J. T., and Doornenbal, P., 1957, "On the Air Resistance and the Bernoulli Effect of the Human Larynx," J. Acoust. Soc. Am., **29**(5), pp. 626–631.

[19] 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," J. Comput. Phys., **227**(22), pp. 9303–9332.

[20] 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," Front. Physiol., **2**(19), pp. 1–15.

[21] Xue, Q., Zheng, X., Mittal, R., and Bielamowicz, S., 2014, "Subject-Specific Computational Modeling of Human Phonation," J. Acoust. Soc. Am., **135**(3), pp. 1445–1456.

[22] 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," ASME J. Biomech. Eng., **132**(11), p. 111003.

[23] 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," J. Acoust. Soc. Am., **95**(6), pp. 3595–3604.

[24] Berry, D. A., 2001, "Mechanism of Modal and Non-Modal Phonation," J. Phon., **29**(4), pp. 431–450.

[25] Döllinger, M., Berry, D. A., and Berke, G. S., 2005, "Medial Surface Dynamics of an In Vivo Canine Vocal Fold During Phonation," J. Acoust. Soc. Am., **117**(5), pp. 3174–3183.

[26] Neubauer, J., Mergell, P., Eysholdt, U., and Herzel, H., 2001, "Spatio-Temporal Analysis of Irregular Vocal Fold Oscillations: Biphonation Due to Desynchronization of Spatial Modes," J. Acoust. Soc. Am., **110**(6), pp. 3179–3192.

[27] Zhang, Y., Zheng, X., and Xue, Q., 2020, "A Deep Neural Network Based Glottal Flow Model for Predicting Fluid-Structure Interactions During Voice Production," Appl. Sci., **10**(2), pp. 1–18.

[28] Smith, S. L., and Titze, I. R., 2018, "Vocal Fold Contact Patterns Based on Normal Modes of Vibration," J. Biomech., **73**, pp. 177–184.

[29] Forrest, S., 1996, "Genetic Algorithms," ACM Comput. Surv., **28**(1), pp. 77–80.

[30] Goldberg, D. E., 2006, *Genetic Algorithms*, Pearson Education, Delhi, India.

[31] Mitchell, M., 1998, *An Introduction to Genetic Algorithms*, MIT Press, Cambridge, MA.

[32] Goodfellow, I., Bengio, Y., Courville, A., and Bengio, Y., 2016, *Deep Learning*, MIT Press, Cambridge, MA.

[33] Geng, B., Xue, Q., and Zheng, X., 2016, "The Effect of Vocal Fold Vertical Stiffness Variation on Voice Production," J. Acoust. Soc. Am., **140**(4), pp. 2856–2866.

[34] 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," J. Acoust. Soc. Am., **132**(3), pp. 1602–1613.

[35] Rosenblatt, M., 1956, "Remarks on Some Nonparametric Estimates of a Density Function," Ann. Math. Statist., **27**(3), pp. 832–837.

[36] LeCun, Y., Bengio, Y., and Hinton, G., 2015, "Deep Learning," Nature, **521**(7553), pp. 436–444.

[37] Ruder, S., 2016, "An Overview of Gradient Descent Optimization Algorithms," arXiv Preprint arXiv1609.04747.

[38] Gulli, A., and Pal, S., 2017, *Deep Learning With Keras*, Packt Publishing Ltd., Birmingham, UK.

[39] Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D. G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., and Zheng, X., 2016, "TensorFlow: A System for Large-Scale Machine Learning," Proceedings 12th USENIX Symposium Operating System Design Implementation, OSDI, **101**(C), Savannah, GA, Nov. 2–4, pp. 265–283.

[40] Altman, D. G., and Bland, J. M., 1983, "Measurement in Medicine: The Analysis of Method Comparison Studies," J. R. Stat. Soc. Ser. D Stat., **32**(3), pp. 307–317.

[41] Krane, M. H., and Wei, T., 2006, "Theoretical Assessment of Unsteady Aerodynamic Effects in Phonation," J. Acoust. Soc. Am., **120**(3), pp. 1578–1588.

[42] Hochreiter, S., and Urgen Schmidhuber, J., 1997, "Long Shortterm Memory," Neural Comput., **9**(8), pp. 1735–1780.