

Estimation of Physiological Vocal Features From Neck Surface Acceleration Signals Using Probabilistic Bayesian Neural Networks

Joaquín Sepúlveda ¹, Jesús A. Parra ², Emiro J. Ibarra ³, Mauricio Araya ⁴, Patricio De La Cuadra ⁵, and Matías Zañartu ⁶, *Senior Member, IEEE*

Abstract—This study presents a novel application of a Probabilistic Bayesian Neural Network (PBNN) for estimating vocal function variables and enhancing non-invasive ambulatory voice monitoring by addressing aleatoric and epistemic uncertainties in regression tasks. The proposed PBNN allows for estimating key physiological parameters including subglottal pressure, vocal fold contact pressure, thyroarytenoid, and cricothyroid muscle activations, from seven aerodynamic and acoustic features. The PBNN is trained on the Triangular Body-Cover Model (TBCM) of the vocal folds to produce a non-linear inverse mapping between its inputs and outputs. Furthermore, the selected aerodynamic and acoustic features can be obtained in ambulatory settings, thus enhancing the practical applicability of the proposed method. Transfer Learning is then applied to integrate real voice data into the initially synthetic-trained network to refine subglottal pressure estimations. The confidence intervals generated by the PBNN illustrate its ability to identify uncertain estimations, as the results show correlations between prediction errors and the estimated aleatoric and epistemic uncertainties. This correlation is advantageous because it shows that the network can effectively predict potential inaccuracies in its estimations. Increased uncertainty is mainly observed at operating points where the TBCM is likely to exhibit non-linear behaviors, at higher subglottal pressures. This suggests that the selected input features may not be robust enough for capturing the nonlinear effects in the TBCM. These results highlight the potential for future research to assess the viability of incorporating new features and additional measurements that could better capture non-linear responses.

Index Terms—Accelerometers, bayes methods, bayesian networks, speech analysis, speech processing, vocal folds.

I. INTRODUCTION

VOICE disorders are estimated to affect 30% of adults in the United States at some stage in their lives [1]. These disorders are significantly prevalent and impactful among those whose professions are vocally demanding. The consequences of such disorders are considerable, often resulting in financial, social, occupational, and psychological issues [2]. These problems raise the need and opportunity to look for effective ways to assist speech therapy with accurate measurements of vocal features. To achieve this, it is essential to have a comprehensive understanding of the phonatory system and to develop reliable tools that can analyze and simulate the behavior of the vocal folds and the elements that constitute the larynx.

Measuring critical vocal features, such as laryngeal muscle activation, subglottal pressure, and collision pressure, is key to addressing these challenges and advancing the assessment of vocal function. However, a significant limitation is that these features are difficult to measure in practice, particularly in non-invasive or ambulatory settings. Muscle laryngeal activity, for instance, plays a crucial role in controlling pitch and loudness; nevertheless, its study is often restricted to ex-vivo experiments, such as excised larynxes [3], [4], [5], or limited in vivo scenarios [6], [7]. Subglottal pressure, an indicator of the aerodynamic forces driving phonation, is estimated through in-lab techniques like intraoral pressure sensors combined with specific phonatory gestures such as /pae/ [8], [9]. Regression-based approaches, however, have demonstrated that aerodynamic features can improve subglottal pressure estimation in non-invasive or ambulatory settings [10], [11], [12]. Similarly, vocal fold collision pressure quantifies the impact of phonation on tissue and provides information about vocal effort. Directly measuring this parameter requires complex methods, such as placing sensors between the folds, making it impractical for routine clinical use [13]. To address these challenges, recent studies have proposed numerical physics-based models using Bayesian frameworks for estimating collision pressure. These models estimate collision pressure by tracking vocal fold motion from high-speed video endoscopy and have been tested in laboratory recordings with humans and experiments involving silicone vocal fold models [13], [14],

Received 10 May 2024; revised 27 January 2025; accepted 25 February 2025. Date of current version 10 April 2025. This work was supported in part by the National Institutes of Health (NIH) National Institute on Deafness and Other Communication Disorders under Grant P50 DC015446, in part by the Agencia Nacional de Investigación y Desarrollo (ANID) under Grant FONDECYT 1201551, Grant FONDECYT 1230828, and Grant Basal AFB240002. The associate editor coordinating the review of this article and approving it for publication was Prof. Mark Hasegawa-Johnson. (*Corresponding author: Joaquín Sepúlveda.*)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the governing institutional review board (IRB) from the Massachusetts General Hospital in IRB Protocol 2011P002376.

Joaquín Sepúlveda and Patricio De La Cuadra are with the Department of Electrical Engineering, Pontificia Universidad Católica de Chile, Santiago 78204362, Chile (e-mail: jnsepulveda2@uc.cl; pcuadra@uc.cl).

Jesús A. Parra, Emiro J. Ibarra, Mauricio Araya, and Matías Zañartu are with the Department of Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso 23400003, Chile, and also with the Advanced Center for Electrical and Electronic Engineering, Universidad Técnica Federico Santa María, Valparaíso 2340000, Chile (e-mail: jesus.parrap@sansano.usm.cl; emiro.ibarra@sansano.usm.cl; mauricio.araya@usm.cl; matias.zanartu@usm.cl).

Digital Object Identifier 10.1109/TASLPRO.2025.3552938

2998-4173 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See <https://www.ieee.org/publications/rights/index.html> for more information.

[15], [16]. While promising, further validation is required to establish their reliability and applicability in broader clinical contexts.

To overcome the limitations of direct measurement, numerical models offer an alternative by creating controlled environments that simulate the physics of phonation. These models enable the study of difficult-to-measure parameters and provide methodologies to estimate them indirectly by relating them to more easily measured characteristics, such as aerodynamic signals [10], [11] or video-based features [13], [15], [16]. A detailed approximation of the physiology of the phonatory process can be achieved with lumped element models of the vocal folds [17]. These models provide simplified, low-order representations of the mechanical, aerodynamic, and acoustic interactions within the phonatory system, modeling the vocal folds as discrete masses coupled by springs and dampers [18], [19], [20], [21]. They are particularly effective for analyzing the effects of various biomechanical parameters on phonation, especially those that are challenging to measure directly in experimental setups. These numerical models take an input vector comprising lung pressure (P_L) and the activation levels of the five intrinsic laryngeal muscles: cricothyroid (a_{CT}), thyroarytenoid (a_{TA}), lateral cricoarytenoid (a_{LCA}), posterior cricoarytenoid (a_{PCA}), and interarytenoid (a_{IA}). These inputs drive the oscillation of the discrete masses, leading to aerodynamic and acoustic interactions that generate glottal airflow, sound pressure waves, subglottal pressure (P_S), and vocal fold contact pressure (P_C) as model outputs. In this study, the vocal features of interest are P_S , P_C , a_{TA} , and a_{CT} .

The integration of numerical models and machine learning techniques has significantly advanced the estimation of vocal features. Machine learning methods trained on simulated data from numerical models can predict clinically relevant parameters with high accuracy. For instance, recurrent neural networks (RNNs) have been used to estimate subglottal pressure from high-speed video (HSV) in excised porcine vocal folds [22], [23]. Similarly, feedforward neural networks (FFNNs) have been employed to estimate geometrical and mechanical properties of the vocal folds and subglottal pressure from accelerometer- and microphone-based features [24], [25], [26]. While time-series approaches like RNNs are often suitable for parameter estimation, feedforward architectures have shown similar results with the advantage of requiring fewer data points. By leveraging expert knowledge for feature selection, these methods are particularly advantageous in clinical contexts, where datasets are often sparse and limited in size [27].

The studies by [25], [26] serve as a baseline for the current work. In these studies, a non-linear regression based on neural network technology was employed to address the challenges of estimating vocal function variables that are difficult to measure directly. This approach combined lumped-element models with machine learning techniques, utilizing more accessible aerodynamic vocal features to make estimations. For ambulatory measurements, a neck-surface accelerometer was used as the primary tool [12], [28], [29], allowing the estimation of glottal airflow [30]. This data, in turn, served as input to a Multi-layer Perceptron Neural Network trained to predict P_S ,

P_C , a_{TA} , and a_{CT} using aerodynamic features. The flexibility of the Triangular Body Cover Model (TBCM) [20], [21] was instrumental in generating a comprehensive dataset large enough to train the neural network. The neural network was designed to solve the inverse mapping of the TBCM model, allowing the estimation of internal vocal function variables. In [25], the inverse mapping was initially validated with a clinical dataset containing pressure values from multiple subjects, demonstrating its feasibility. A subsequent study [26] introduced a transfer learning step using a larger clinical dataset before testing with clinical data. This transfer learning process facilitated subject-specific adjustments, improving the model's ability to generalize across groups of subjects, including those grouped by pathology. This adaptation exploited the complex relationships captured by the TBCM-generated dataset and aligned them with clinical datasets through domain adaptation techniques. These efforts demonstrated the feasibility of combining numerical models and machine learning for vocal feature estimation, but left key challenges unaddressed.

One critical aspect overlooked in previous efforts is the quantification of uncertainty in solving inverse problems and parameter estimations within neural networks. This study introduces a Probabilistic Bayesian Neural Network (PBNN) to address this gap. The PBNN framework allows for the quantification of both aleatoric and epistemic uncertainties inherent in the regression task [31], [32]. Aleatoric uncertainty, associated with the inherent noise in the data, is addressed by setting the output layer of the neural network to be the parameters of Gaussian distributions, which enables the estimation of the mean and variance of each output variable given a specific input vector. On the other hand, epistemic uncertainty, which originates from the lack of knowledge about the model parameters, is captured by evaluating the network multiple times for the same input vector and calculating the variance of the means of the estimated output distributions. This variance is originated by the stochastic nature of the weights of the network, and it provides a measure of the uncertainty of the model in the parameter space, revealing the confidence in the predicted outputs given the observed data [31], [32].

The need to capture the uncertainty comes from two main reasons. The first is the inherent randomness in the phonation process itself, explained by the turbulent glottal airflow and instabilities of oscillations, and the noise from the data acquisition that can lead to slight variations in the input data of the network. The second and arguably more relevant source of aleatoric uncertainty is that obtaining P_S , P_C , a_{TA} , and a_{CT} from the aerodynamic features derived from the output of the TBCM model constitutes an ill-posed inverse problem [17]. Ill-posed problems are characterized by solutions that do not meet the criteria of existence, uniqueness, and stability. In the context of this work, the lack of a unique solution and sensitivity to data perturbations make the task of accurately estimating desired quantities challenging.

The contributions of this work are twofold. First, it introduces a PBNN framework specifically designed for vocal parameter estimation, integrating uncertainty quantification—both aleatoric and epistemic—into the estimation process. This integration

TABLE I
AERODYNAMIC AND ACOUSTIC FEATURES

Feature	Description
ACFL	AC glottal airflow
MFDR	Maximum Flow Declination Rate
OQ	Open Quotient
SQ	Speed Quotient
f_o	Fundamental Frequency
H1-H2	Difference between the first and second harmonics
SPL	Sound Pressure Level

provides confidence intervals for predictions and deeper insights into the relationships between aerodynamic features and key vocal parameters such as P_S , P_C , a_{TA} , and a_{CT} . Second, the framework offers new perspectives on the capability of low-order numerical models to replicate the complexity of voice production phenomena based on clinical data. By addressing uncertainties, this work marks a significant advancement in non-invasive methodologies that integrate machine learning and numerical voice production models for assessing vocal function and establishes new avenues for future research and clinical applications.

II. MATERIALS AND METHODS

A. Data Acquisition Process

Similarly to previous machine learning approaches [22], [24], [25], the data used in this research comes from two domains. On one side we have the synthetic data, that is generated by the TBCM, that is used for the first stage of training. And on the other side, there are clinical data, used for validation and for the final stage of state-of-the-art comparison

1) *Clinical Data*: The clinical dataset [27], [33] contains measurements of the aerodynamic and acoustic vocal features listed in Table I. These features were derived from signals obtained noninvasively using synchronous recordings of a neck surface accelerometer (ACC), pneumotachograph mask for capturing oral airflow volume velocity (OVV), an intraoral catheter for measuring intraoral pressure (IOP) and a microphone for obtaining the radiated sound pressure signal. The OVV signal is used as a reference to calibrate the parameters of an Impedance Based Inverse Filtering (IBIF) method [30]. This inverse filtering method is capable of estimating the Glottal Volume Velocity (G_{vv}) signal from the neck surface acceleration signal ACC, which enables the application of the framework in ambulatory settings.

The instruments employed by [33] for making the data acquisition are the following:

- 1) A Sennheiser MKE104 acoustic microphone, positioned 10 cm from the participant's lips (Sennheiser, Electronic GmbH, Wennebostel, Germany).
- 2) EG-2 electroglottograph electrodes placed on the thyroid cartilage to assess variations in laryngeal impedance (Glottal Enterprises, Syracuse, NY, USA).
- 3) A BU-27135 accelerometer, mounted at the neck's base to capture surface vibrations (Knowles Corp., Itasca, IL, USA).

- 4) A PT-2E pneumotachograph face mask, equipped with circumferential vents, used to gather high-bandwidth aerodynamic data (Glottal Enterprises).
- 5) A low-bandwidth PT-25 air pressure sensor, connected to a thin tube inserted into the mouth through the lips, for pressure measurements (Glottal Enterprises).

The research dataset consisted of 79 adult females, none of whom had a recorded history of voice disorders. They had an average age of 29.6 years with a standard deviation of 13.0 years. A licensed speech-language pathologist confirmed their vocal health status via interviews, laryngeal videostroboscopic exams, and the Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) assessment [34]. Each participant provided informed consent. The research protocols, both experimental and clinical, were approved by the institutional review board at Mass General Brigham of Massachusetts General Hospital. Data collection took place in a sound-treated room. Participants were asked to articulate the syllables /pæ/ at three different loudness levels: comfortable, loud, and soft. Although participants were guided to keep pitch and volume consistent during each syllable string, they were allowed to choose levels they felt were natural, without adhering to strict absolute pitch and volume requirements.

It is important to note that the clinical dataset does not contain the 4 desired output variables previously mentioned, but only contains P_S . The measurement of contact pressure and the acquisition of muscle activation data is a still challenging problem and there are not large enough datasets available for validation.

2) *Synthetic Datasets*: Two synthetic datasets were utilized in this work, both derived from thousands of simulations of the TBCM model [21]. The TBCM consists of paired three-mass body-cover systems with a muscle-controlled model of all five intrinsic laryngeal muscles [19], [35]. The forces on the body-cover masses, including the collision force, are computed using damped oscillator motion equations as described in [20]. Pressure wave propagation, including subglottal pressure, is simulated using the wave reflection analog approach [36]. In this work, pressure units are expressed in cmH_2O ¹ to facilitate comparisons with related literature.

Dataset 1 (D_1), originally generated by [25], consisted of 13.000 simulations where the model's input parameters were varied in discrete, fixed steps, as detailed in Table II. Dataset 2 (D_2), a newly created dataset, was designed to provide a more comprehensive and representative sampling of muscle activation variables. It consists of 80.000 Monte Carlo simulations, with input variables sampled within their operational ranges, as detailed in Table III. The differences in parameters ranges between the two datasets are due to the inability to achieve phonation with low levels of LCA or IA muscle activation [3], as well as the extension of the simulated subglottal pressure range to capture the very high-pressure values observed in clinical data [27], [33].

Another problem that was aimed to be addressed by generating the D_2 dataset is the 3.37 cmH_2O offset of P_S of synthetic

¹ cmH_2O is a standard unit used in phonatory research to facilitate clinical interpretability. $1 \text{ cmH}_2\text{O} = 98.0665\text{Pa}$

TABLE II
PARAMETER SETTINGS FOR THE ORIGINAL
SIMULATION D_1 [25]

Parameter	Range	Step
a_{CT}	0–1	0.1
a_{TA}	0–1	0.1
a_{LCA}	0.2–0.8	0.1
a_{PCA}	0–0.1	0.1
a_{IA}	0.2–0.8	0.1
P_L	5–20	1.5

Pressure in cmH₂O. a_{CT} : cricothyroid muscle activation. a_{TA} : thyroarytenoid muscle activation. a_{LCA} : lateral cricoarytenoid muscle activation. a_{PCA} : posterior cricoarytenoid muscle activation. a_{IA} : interarytenoid muscle activation. P_L : lung pressure.

TABLE III
PARAMETER SETTINGS FOR THE NEW SIMULATION
 D_2 , WITH RANDOM SAMPLING WITHIN
THE VARIABLE RANGE

Parameter	Range
a_{CT}	0–1
a_{TA}	0–1
a_{LCA}	0.4–0.8
a_{PCA}	0–0.1
a_{IA}	0.4–0.8
P_L	3–25

Pressure in cmH₂O. a_{CT} : cricothyroid muscle activation. a_{TA} : thyroarytenoid muscle activation. a_{LCA} : lateral cricoarytenoid muscle activation. a_{PCA} : posterior cricoarytenoid muscle activation. a_{IA} : interarytenoid muscle activation. P_L : lung pressure.

data in relation to clinical data. This offset was reported by [25] and it was then solved by directly subtracting 3.37 cmH₂O to the P_S vector, but an actual solution to the cause of the problem remained unresolved. The way this was addressed was by reducing the attenuation factor of the supraglottal tract from $3.8 \cdot 10^{-3}$ to $2.0 \cdot 10^{-3}$, allowing for lower P_S simulations to produce higher SPL. This worked as an initial experimental solution to the reported offset, but it still would be valuable to develop more research towards the clinical accuracy of the TBCM.

The two synthetic datasets described above were used for different purposes. D_1 was employed to perform the comparisons with the deterministic regression proposed in [25], [26], where an initial training stage utilized synthetic data, followed by a testing stage with clinical data [25] or a readjustment stage [26]. This setup allowed a comparison of P_S estimation using PBNN against state-of-the-art methods, while introducing a new probabilistic approach applicable to both synthetic and clinical scenarios. Notably, this analysis excluded the estimation of P_C , a_{TA} and a_{CT} . Conversely, D_2 , was exclusively used to make estimations in the synthetic setting, without subsequent clinical evaluation. Its purpose was to achieve improved regression of muscle activation parameters, which are not available in the clinical dataset. Furthermore, the application of PNN for estimating these parameters enables a detailed discussion and analysis of the uncertainty associated with the estimation

TABLE IV
AIC VALUES FOR THE FIT OF VARIOUS THEORETICAL DISTRIBUTIONS
TO THE CLINICAL DATA OF SUBGLOTTAL PRESSURE

Distribution name	AIC	$\exp((AIC_{\min} - AIC_i)/2)$
Gamma	1208.16	1.00
Beta	1209.26	0.58
Lognorm	1210.78	0.27
Norm	1246.09	0.00
Exp	1300.05	0.00

AIC: Akaike information criterion.

of these challenging-to-measure parameters through inverse regression from aerodynamic features.

B. Data Preprocessing

1) *Clinical Dataset Preprocessing*: The processing made by [27], [33] consisted in the following. The OVV signal was low-pass filtered to 1100 Hz and was subsequently downsampled to 8192 Hz. This procedure was primarily designed to isolate the first formant during inverse filtering, thus preventing antiresonance at approximately 1500 Hz, which is the typical frequency response of the pneumotachograph mask. The IOP signal was a low-pass filtered at 80 Hz. Additionally, the microphone signal was rectified, subjected to low-pass filtering at 80 Hz, and then downsampled to 256 Hz to produce a root-mean-square (RMS) envelope. To avoid any phase distortion and to guarantee alignment with other physiological signals, all filtering processes were implemented in both the forward and reverse directions.

The estimation of subglottal pressure was derived from the peaks of the IOP signals [27], [33]. Given that the vocal gesture involved the repetition of /pæ/ syllables, there exists a moment during the cycle where glottal opening and lip closure occur, precisely preceding the airflow burst generated by the /p/ sound. This specific moment aligns accurately with the peaks observed in the IOP signal. The subglottal pressure is then estimated by calculating the mean of two consecutive peaks.

2) *Synthetic Dataset Preprocessing*: The generated datasets contain input variables that lie within the range of clinical dataset, specifically, ACFL values exceeding 30 mL/s and f_o within the 120–400 Hz range, as described in [25]. However, the clinical dataset is limited and does not necessarily follow the same distribution as the synthetic variables. This can be problematic as it can produce a biased training towards ranges that are not common in normal phonation.

To avoid this problem, a Maximum Likelihood Estimation was made to adjust the subglottal pressure in D_1 to the distribution followed by the clinical data. For achieving this, five common distributions were fitted to the data and the Akaike Information Criterion (AIC) was derived to identify the one with the best matching. Table IV shows the AIC values for each distribution, along with the calculation of $\exp((AIC_{\min} - AIC_i)/2)$, which represents the difference between the smallest AIC value and the AIC value for each distribution, and AIC_i is the AIC value for each distribution. This quantity indicates how likely the i_{th} model is to minimize the loss of information as the model with the lowest AIC value [37]. Fig. 1 shows the histogram of the clinical dataset along with the optimized P_S distributions

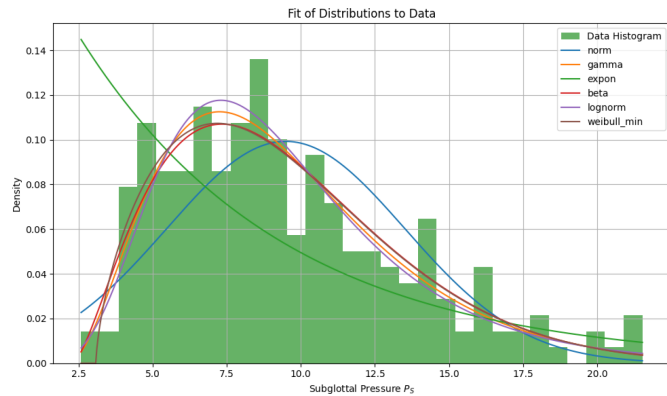


Fig. 1. Comparison of distribution fits to clinical subglottal pressure. Colors indicate probability density functions: blue for Normal, orange for Gamma, green for Exponential, red for Beta, and purple for Log-normal distributions. P_S : subglottal pressure.

for a visual comparison of the quality of the fit. The Gamma distribution had the best fit, thus it was used to sample 3177 simulations from the D_1 dataset.

C. Probabilistic Bayesian Neural Network

The tool for estimating the different types of uncertainties in the regression problem was a Probabilistic Bayesian Neural Network [32]. This kind of network incorporates the uncertainty in two different ways into its architecture. On one side, the weights of some of its nodes are treated as random variables, and on the other side, the output layer is set to estimate parameters of probability distributions, instead of point estimate of the desired variables.

The uncertainty in the weights serves two purposes: first, it is a kind of regularisation method (alternative to the more common Monte Carlo Dropout) that prevents overfitting to reduced datasets, improving the model's generalisation in non-linear regression problems; and second, it allows for the quantification of the epistemic uncertainty associated with the uncertainty of the model on its own estimated parameters [31]. By evaluating the output of the network multiple times for the same input vector, the weights are sampled from their estimated posterior distribution and thus the output value differs in each iteration, which allows to calculate the variance of the outputs. In this implementation, the network was evaluated 100 times for every test.

For calculating the posterior distribution of the weights, the Bayes By Backpropagation algorithm is used [31]. This algorithm involves finding the parameters θ that most appropriately define the variational posterior $q(w|\theta)$, which is an approximation to the distribution of the weights, given the data $P(w|\mathcal{D})$. This is achieved by using the Kullback-Leibler (KL) divergence as an approximated metric to be minimised to approach the posterior distribution. It involves defining prior distributions for the weights $P(w)$, which conveniently are commonly assumed to follow a standard Gaussian distribution. This distribution is suitable, as it constitutes a *non informative prior* and it serves a good starting point for the training. Then the distributions

given the data are updated to obtain the so-called posterior distribution. This configuration is capable of predicting outputs of unseen data by calculating the expectation $P(\hat{y}|\hat{x}) = \mathbb{E}_{P(w|\mathcal{D})}[P(\hat{y}|\hat{x}, w)]$. The parameters for the distributions of the weights are obtained by optimizing (1) [31].

$$\theta^* = \arg \min_{\theta} \text{KL}[q(w|\theta) || P(w)] - \mathbb{E}_{q(w|\theta)}[\log P(\mathcal{D}|w)] \quad (1)$$

The uncertainty in the output nodes, on the other hand, represents the irreducible inherent noise in the data. This type of noise can arise from the nature of the data itself, from its acquisition or from non uniqueness of solutions of the inverse problem, as stated previously. To address this uncertainty, the output layer of the implemented neural network consists in a tuple for each output variable, which constitute the estimated mean $\hat{\mu}_Y$ and standard deviation $\hat{\sigma}_Y$ for $\{P_S, P_C, a_{TA}, a_{CT}\}$. This way, the $-\mathbb{E}_{q(w|\theta)}[\log P(\mathcal{D}|w)]$ term in (1) represents the expected Negative Log-Likelihood for a Normal distribution with the estimated $\hat{\mu}_Y$ and $\hat{\sigma}_Y$ parameters.

Different network architectures were tested, from more simple networks with 4 neurons per hidden layer and 2 hidden layers, to more complex configurations with 128 neurons per hidden layer and 4 hidden layers. The general architecture was implemented as follows: 7 neurons in the input layer for the vocal features, a variable number of hidden layers with rectified linear units as activation function, where the first layer is a dense variational layer [38] with weights as random variables, and finally an output layer with 4 tuples of mean μ and standard deviation σ for the outputs. The optimization algorithm used was Adam, and the loss function was calculated as the Negative Log-Likelihood of the training labels. All neural networks in this study were implemented on Google Colaboratory.

D. Training and Testing of the PBNN Using Synthetic Data

The training schematic used is an adapted version of the one implemented by [25]. Fig. 2 shows the training schematic used. It first shows the TBCM model, from which the synthetic datasets were generated. The model outputs the P_S and P_C values of each simulation, as well as the simulated signals of the radiated sound pressure (P_{out}) and the Glottal Volume Velocity (G_{vv}). Then the aerodynamic features are extracted from the G_{vv} signal and the SPL is derived from the P_{out} signal. The seven features are then input into the PBNN and normal distributions are fitted to the four desired output variables. This way, the PBNN is used to perform a non-linear inverse mapping of the TBCM and getting confidence intervals of the input configurations that yield a given output.

In the synthetic context, D_2 was divided into 80% for training and 20% for testing. Three architectures were examined to evaluate the performance of PBNNs across different levels of complexity. The upper row of Fig. 3 illustrates the workflow for testing the Bayesian approach with the synthetic dataset D_2 , which includes: (1) dataset construction using TBCM, (2) PBNN training with D_2 , and (3) estimation of the four parameters of interest (P_S, P_C, a_{TA}, a_{CT}) using the testing subset. This

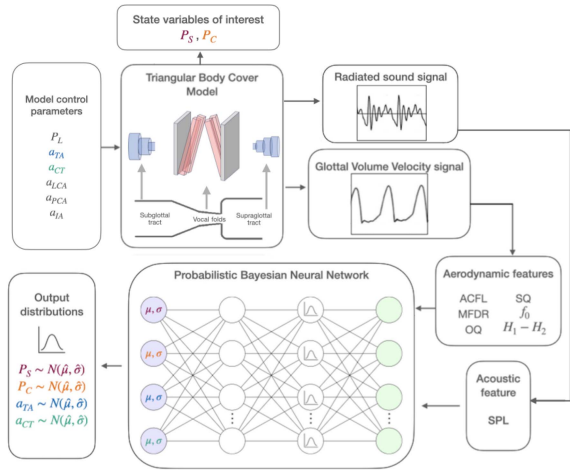


Fig. 2. Training schematic. Initially proposed by [25]. Deterministic neural network was replaced by a Probabilistic Bayesian Neural Network. Key abbreviations: a_{CT} : cricothyroid muscle activation. a_{TA} : thyroarytenoid muscle activation. a_{LCA} : lateral cricoarytenoid muscle activation. a_{PCA} : posterior cricoarytenoid muscle activation. a_{IA} : interarytenoid muscle activation. P_L : lung pressure. P_S : subglottal pressure. P_C : collision pressure. ACFL: AC glottal airflow. MFDR: maximum flow declination rate. OQ: open quotient. SQ: speed quotient. f_o : fundamental frequency. H1-H2: difference between the first and second harmonics. SPL: sound pressure level. N : normal distribution. μ : mean value. σ : standard deviation.

initial methodology stage provides a detailed analysis of the implications of uncertainty in parameter estimation.

E. Transfer Learning With Clinical Data

An extension of the initially proposed regression was made by [26] to better match the clinical data by applying transfer learning. This technique consists in taking the pre-trained neural network and make an additional stage of training using a fraction of the clinical data that was originally used for validation. Then, the other fraction is used for validation. This practice allows not only to yield better results on clinical validation, but also to let the network learn as much as necessary from the TBCM and not be so sensible to its imperfections, (e.g. the discrepancy of clinical and synthetic P_S reported by [25]).

In this work, the transfer learning method is re-implemented to leverage the additional information provided by the PBNN. To achieve this, the clinical dataset, consisting of 236 vocal feature vectors from 79 subjects, was divided into five folds based on subject identifiers, ensuring that all data points from the same subject were placed entirely in either the training set or the test set, but not both. This approach prevents the network's performance from being biased by prior knowledge of the data being predicted. The testing performance was then evaluated as the mean regression performance across the entire dataset, averaged over the five folds.

The base model was trained using 3,177 selected samples from D_1 , which demonstrated the best fit with the P_S distribution in the clinical dataset. The training followed the same scheme illustrated in Fig. 2. The model was then fine-tuned on the clinical dataset using a transfer learning approach with no frozen layers and a modest learning rate ($5 \cdot 10^{-4}$). In this case, D_1 was

used instead of D_2 to ensure a fair comparison with previous work [25], [26]. The lower row of Fig. 3 illustrates the workflow for testing the Bayesian approach with the clinic dataset. As shown here, the TBCM is configured to construct D_1 , which is then used to train a new PBNN. This PBNN is subsequently fine-tuned and tested on the clinical dataset via transfer learning within each fold of the five-fold cross-validation.

F. Performance Metrics

The performance metrics of the regression were evaluated from two perspectives. First, the estimated means, μ_Y , were assessed as point estimates of the variables, enabling an analysis equivalent to those performed by [22], [24], [25]. This evaluation involved calculating metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and the coefficient of determination (R^2), which measure how accurately the network predicts the variables' values; for the case of MAPE, a modified version, MAPE*, is used for an easy comparison with the other regression approaches. This modified metrics use the maximum of the range of values instead the real value in the denominator. Second, the estimated standard deviations, σ_Y , were used to derive additional performance metrics related to the predicted confidence intervals, including Prediction Interval Coverage (PIC) and Prediction Interval Average Width (PIAW). These metrics provide insights into the aleatoric uncertainty of the estimated variables. Furthermore, by running the network multiple times and computing $\text{Var}[\mu_Y]$, the variance of the predicted means can be analyzed to interpret the epistemic uncertainty associated with the model parameters.

The PIC quantifies the proportion of actual values that fall within the predicted confidence intervals, as defined in (2):

$$\text{PIC} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(\mu_i \in [\hat{\mu}_i - z \hat{\sigma}_i, \hat{\mu}_i + z \hat{\sigma}_i]) \quad (2)$$

The PIAW calculates the average width of the predicted 95% confidence intervals, as shown in (3):

$$\text{PIAW} = \frac{1}{N} \sum_{i=1}^N (z \hat{\sigma}_i) \quad (3)$$

Where N represents the total number of data points, y_i denotes the actual value, $\hat{\mu}_i$ is the predicted value, and $\hat{\sigma}_i$ is the estimated standard deviation of the prediction at x_i . The variable z corresponds to the quantile of the standard normal distribution used to define the confidence interval. In this study, a 95% confidence level was adopted, implying $z = 1.96$.

Given that the estimation involves different variables evaluated over an extended range, the coefficient of variation (CV) is used to quantify the relationship between uncertainty and the estimated value. The CV provides insight into whether uncertainty increases disproportionately across the range of the variable being estimated, beyond the increase in the estimated value itself. It is calculated as the ratio between the standard deviation (σ) and the mean (μ) for each prediction, allowing for

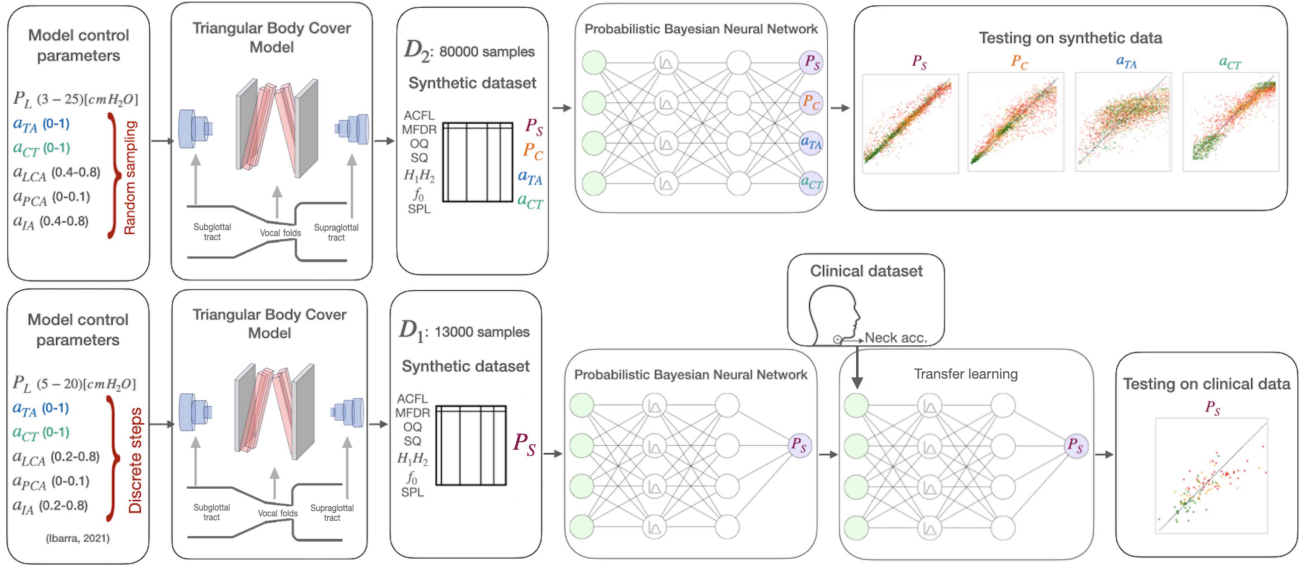


Fig. 3. Schematic of the Bayesian approach. From left to right, the diagram illustrates the construction of the synthetic dataset, the training of the Bayesian Neural Network, and the testing process. The upper row depicts the estimation of P_S , P_C , a_{TA} , and a_{CT} using the synthetic dataset (D_2), while the lower row shows P_S estimation in the clinical dataset using a transfer learning strategy. Scatter plots of testing data points are color-coded based on their standard deviation, with green representing the lower third, yellow the middle third, and red the upper third percentile. Key abbreviations: a_{CT} : cricothyroid muscle activation. a_{TA} : thyroarytenoid muscle activation. a_{LCA} : lateral cricoarytenoid muscle activation. a_{PCA} : posterior cricoarytenoid muscle activation. a_{IA} : interarytenoid muscle activation. P_L : lung pressure. P_S : subglottal pressure. P_C : collision pressure. ACFL: AC glottal airflow. MFDR: maximum flow declination rate. OQ: open quotient. SQ: speed quotient. f_0 : fundamental frequency. H1-H2: difference between the first and second harmonics. SPL: sound pressure level.

TABLE V
PERFORMANCE METRICS FOR DIFFERENT NETWORK CONFIGURATIONS. METRICS FOR P_S
AND P_C ARE IN [CMH₂O] (R-SQUARED AND PIC ARE DIMENSIONLESS)

Configuration	Variable	RMSE	MAE	MAPE*	R-squared	PIC	PIAW
4N 2HL	P_S	1.74	1.12	5.19	0.92	0.96	6.14
	P_C	3.03	2.14	4.77	0.85	0.96	11.18
	a_{TA}	0.18	0.14	14.05	0.34	0.97	0.70
	a_{CT}	0.10	0.07	7.01	0.87	0.96	0.36
32N 3HL	P_S	1.55	0.99	4.59	0.94	0.97	5.69
	P_C	3.11	1.97	4.39	0.84	0.96	10.62
	a_{TA}	0.16	0.13	13.04	0.45	0.97	0.63
	a_{CT}	0.09	0.06	6.01	0.91	0.97	0.31
128N 4HL	P_S	1.55	1.01	4.68	0.94	0.97	6.23
	P_C	2.94	1.92	4.28	0.86	0.97	10.59
	a_{TA}	0.16	0.13	13.04	0.48	0.97	0.63
	a_{CT}	0.08	0.06	6.01	0.91	0.98	0.36

Metrics for P_S and P_C are in [cmH₂O] (R-squared and PIC are dimensionless). RMSE: root mean square error. MAE: mean absolute error. MAPE*: mean absolute percentage error (respect of maximum). R-squared: coefficient of determination. PIC: prediction interval coverage. PIAW: prediction interval average width. N: neurons. HL: hidden layers. a_{CT} : cricothyroid muscle activation. a_{TA} : thyroarytenoid muscle activation. P_S : subglottal pressure. P_C : collision pressure.

a normalized evaluation of variability, as shown in (4):

$$CV_i = \frac{\sigma_i}{\mu_i} \quad (4)$$

III. RESULTS

A. Results in Synthetic Context

The evaluation of synthetic data D_2 , summarized in Table V, demonstrates how different network architectures perform across deterministic and probabilistic performance metrics. Deeper network configurations improved deterministic metrics such as RMSE, MAE, MAPE* and R-squared. For example, in P_S , RMSE decreased by 11% from the 4N 2HL to the 128N 4HL configuration. Similarly, RMSE for P_C decreased by 3%

, while the muscle activation variables a_{TA} and a_{CT} showed reductions in RMSE of 11% and 20%, respectively. These results highlight the capability of deeper networks to enhance prediction performance for these features and are consistent with results reported in [25]. A similar trend of improved performance with deeper architectures was observed for MAE across all variables. Furthermore, MAPE*, being a normalized metric, enables comparisons not only between different network architectures but also across the performance of different variables. The results indicate that MAPE* values slightly decrease with deeper PBNNs, while notable variations were observed among the variables. a_{TA} exhibited significantly higher MAPE* values (exceeding 13%) compared to a_{CT} (below 7%), whereas P_S and P_C demonstrated lower MAPE* values (less than 5%). This

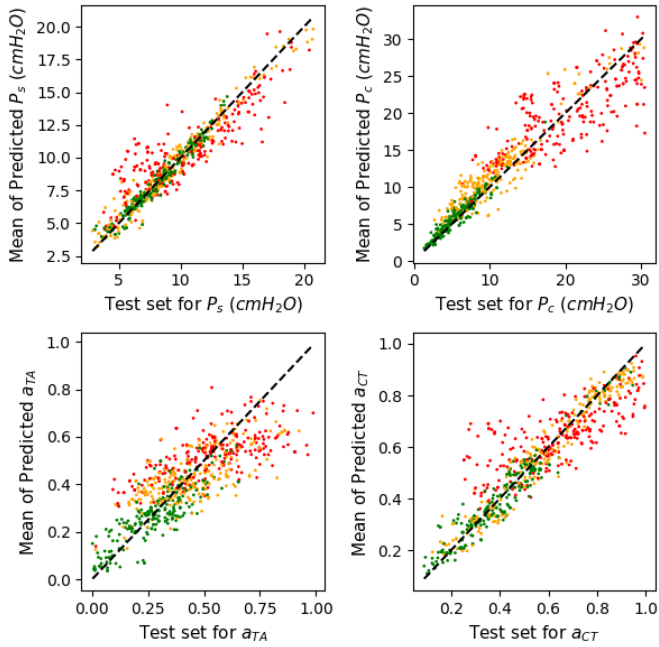


Fig. 4. Scatter plot of synthetic data with point estimations on the y-axis and the actual value of the data in the x-axis for best model trained with D_2 (128 neurons per hidden layer, 4 hidden layers). The dashed line represents the 1:1 ratio (ground truth). Data points are color-coded based on their standard deviation, with green representing the lower third, yellow the middle third, and red the upper third percentile. a_{CT} : cricothyroid muscle activation. a_{TA} : thyroarytenoid muscle activation. P_S : subglottal pressure. P_C : collision pressure.

reduced error in P_S and P_C is likely attributable to their strong correlation with SPL.

In addition to deterministic metrics, Table V also presents the probabilistic performance of the PBNN framework through the PIC and PIAW metrics, offering key insights into calibration and uncertainty estimation. The PIC metric consistently showed high values (0.96–0.98) across all configurations, indicating robust calibration regardless of network depth. For PIAW, P_S and P_C decreased by 7% and 5%, respectively, when comparing the 4N 2HL to the 32N 3HL configuration. Similarly, for the muscle activation variables a_{TA} and a_{CT} , PIAW showed reductions of 10% and 14%, respectively, under the same conditions. Beyond this point, further increases in network depth did not result in significant changes, suggesting that the PBNN framework achieves an optimal balance between model complexity and effective uncertainty estimation at intermediate configurations across all estimated parameters.

Fig. 4 visualizes the integration of the PBNN-estimated μ (predicted mean) and σ (predicted standard deviation) using a scatter plot of predicted vs. actual values for the four parameters of interest: P_S , P_C , a_{CT} , and a_{TA} . To enhance interpretability, the predicted standard deviation σ is color-coded into three categories: green for the lowest third of σ values, yellow for the middle third, and red for the highest third. Generally, most green dots (low-deviation estimates) align closely with the perfect prediction line, indicating minimal prediction error and lower uncertainty. In contrast, most red dots (high-deviation estimates) deviate further from the perfect prediction line, reflecting higher

TABLE VI
MEAN AND STANDARD DEVIATION OF COEFFICIENT VARIATION FOR EACH THIRDS OF THE TEST DATA

Third	P_S		P_C		a_{TA}		a_{CT}	
	mean	std	mean	std	mean	std	mean	std
Low	0.13	0.06	0.28	0.07	0.34	0.70	0.22	0.11
Medium	0.20	0.09	0.19	0.07	0.35	0.05	0.17	0.05
High	0.26	0.10	0.33	0.12	0.40	0.06	0.19	0.05

Metrics for P_S and P_C are in [cmH₂O]. a_{CT} : cricothyroid muscle activation. a_{TA} : thyroarytenoid muscle activation. P_S : subglottal pressure. P_C : collision pressure. std: standard deviation.

prediction errors and greater uncertainty. This behavior is particularly noticeable in P_S , P_C , and a_{CT} . However, the behavior of a_{TA} stands out as particularly distinct, as it exhibits the greatest dispersion compared to the other parameters. This result aligns with the deterministic metrics of a_{TA} , such as a low R^2 (< 0.5) and high MAPE* ($> 13\%$), underscoring its significant variability and the challenges associated with its prediction.

Fig. 4 highlights that red dots, representing high uncertainty predictions, tend to coincide with high-magnitude values, suggesting a potential dependency between uncertainty and magnitude. To assess this trend quantitatively, we analyzed the CV, which measures dispersion relative to the magnitude of the values. Table VI presents the mean and standard deviation of CV across the three uncertainty groups (low, medium, and high) for all parameters. For P_S , the mean CV increases from 0.13 in the low group to 0.26 in the high group, while for P_C , it rises from 0.28 to 0.33. Similarly, for a_{TA} , the mean CV grows from 0.34 in the low group to 0.40 in the high group, with the distinction that this parameter exhibited the highest variability and greatest prediction uncertainty compared to the other estimated parameters.

The observed trends for P_S , P_C , and a_{TA} indicate that the model's uncertainty predictions are not uniform across groups, reflecting a disproportionate relationship between prediction uncertainty and magnitude. If CV values had remained constant across groups, the observed differences in uncertainty could have been entirely attributable to prediction magnitude. However, the observed increase in CV suggests that additional factors contribute to uncertainty, potentially linked to the model's heightened sensitivity to higher magnitudes. This trend highlights limitations in the model's ability to handle extreme values, which may arise from the characteristics of the training data, revealing a lack of representativeness of the TBCM model at higher ranges.

In contrast, a_{CT} exhibits a different pattern: the mean CV decreases slightly from 0.22 in the low group to 0.17 in the medium group, followed by a moderate increase to 0.19 in the high group. This pattern suggests relative stability in CV across uncertainty groups, indicating that the uncertainty for a_{CT} is less influenced by variations in magnitude compared to the other parameters.

Additionally, a t-test revealed statistically significant differences in the mean CV values across the three groups for all variables, validating that the observed changes in CV are not due to random variability. Medium effect sizes (Cohen's $d > 0.5$) were observed for P_S and a_{CT} , large effect sizes ($d > 0.8$) for P_C , and small effect sizes ($d < 0.5$) for a_{TA} . These results reinforce the relationship between increased uncertainty and

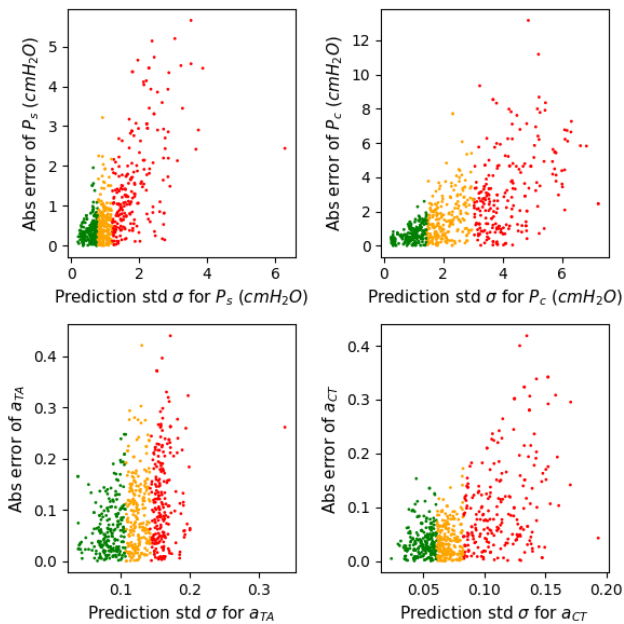


Fig. 5. Scatter plot of synthetic data with prediction errors on the y axis and the estimated standard deviations σ (aleatoric uncertainty) in the x axis for best model trained with D_2 (128 neurons per hidden layer, 4 hidden layers). a_{CT} : cricothyroid muscle activation. a_{TA} : thyroarytenoid muscle activation. P_S : subglottal pressure. P_C : collision pressure. σ : estimated standard deviation. std: standard deviation.

higher magnitudes for P_S , P_C , and a_{TA} , while emphasizing the distinct and more stable behavior of a_{CT} .

Another valuable observation is the relationship between prediction errors and estimated uncertainties. Fig. 5 illustrates this via a scatter plot where the absolute error ($e_i = |\mu\hat{Y} - y|$) is plotted against the predicted standard deviation (σ), using the same color code as in previous figures. Across all parameters, the plots exhibit a triangular shape, with uncertainty providing an upper bound on estimation error, characterizing random uncertainty. The slope of this relationship reveals that P_S estimates generally have lower errors than P_C for the same uncertainty, reflecting the differing difficulty in estimating these features and the effort invested in their estimation [11], [13]. For muscle activations, a_{TA} exhibits a steeper slope than a_{CT} , suggesting that while a_{CT} estimation is well-supported by acoustic and aerodynamic parameters, a_{TA} estimation requires further improvement, as corroborated by Table V and prior studies [3], [5], [25], [26].

Fig. 6 explores epistemic uncertainty by plotting prediction errors against the standard deviation of μ across multiple network iterations. For pressure parameters (P_S and P_C), the color trends for epistemic uncertainty are well-preserved, with green (low uncertainty) points clustering at lower error values and red (high uncertainty) points showing broader dispersion. For a_{CT} , the uncertainty pattern is less structured than for pressure parameters but more consistent compared to a_{TA} , suggesting that the model captures some of the underlying variability in a_{CT} but struggles more with a_{TA} . This is particularly evident in a_{TA} , where the lack of structure and greater dispersion suggest higher prediction uncertainty. Overall, the results indicate that

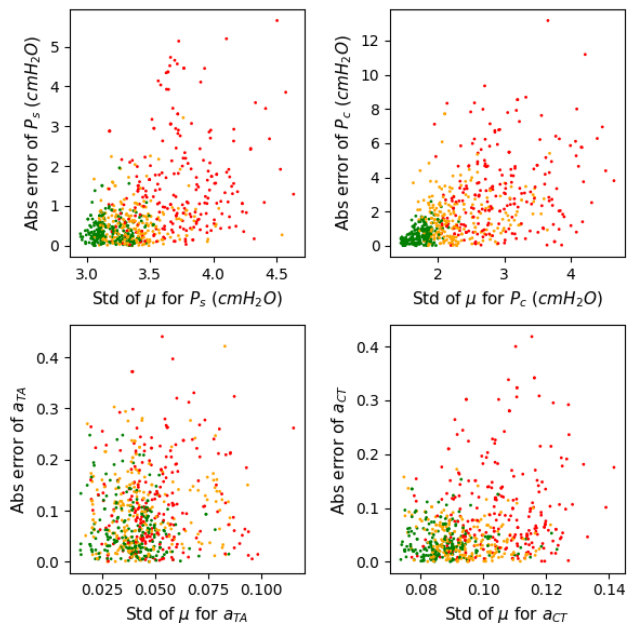


Fig. 6. Scatter plot of synthetic data with prediction errors on the y axis and the estimated standard deviations of μ (epistemic uncertainty) in the x axis for best model trained with D_2 (128 neurons per hidden layer, 4 hidden layers). Data points are color-coded based on their standard deviation, with green representing the lower third, yellow the middle third, and red the upper third percentile. a_{CT} : cricothyroid muscle activation. a_{TA} : thyroarytenoid muscle activation. P_S : subglottal pressure. P_C : collision pressure. μ : estimated mean. std: standard deviated.

the network provides better-calibrated uncertainty estimates for pressure parameters (P_S and P_C) compared to muscle activations, with a_{CT} showing moderate structure and a_{TA} presenting the greatest challenges.

B. Results for Clinical Validation

Table VII summarizes the network performance on clinical data for different training methods: 'Only clinical data' refers to a network trained exclusively on clinical data. 'Only synthetic data' corresponds to a network trained on synthetic D_1 and tested directly on clinical data. Finally, 'Transfer Learning' refers to a network pre-trained on synthetic data and fine-tuned on clinical data using the strategy described in Section II.E.

The PBNN architecture with 4 neurons, 2 hidden layers, and Transfer Learning significantly enhances performance across all configurations in terms of deterministic metrics. This configuration achieved an RMSE of 2.49, an MAE of 1.82, and an R^2 of 0.65, closely aligning with the metrics reported by [25] (RMSE = 2.48, MAE = 1.95, R^2 = 0.65) and performing slightly below the recently reported MPL fine-tuning with Transfer Learning in [26] (e.g., RMSE = 2.30, MAE = 1.77, R^2 = 0.69). While these deterministic results highlight the model's ability to effectively capture the non-linear relationships within vocal function variables, a deeper understanding of its performance can be obtained by analyzing the probabilistic metrics, which provide key insights into the model's calibration and uncertainty estimation.

TABLE VII
PERFORMANCE METRICS FOR P_S [cmH₂O] ON CLINICAL DATA FOR DIFFERENT CONFIGURATIONS AND TRAINING METHODS (R-SQUARED AND PIC ARE DIMENSIONLESS)

Configuration	Training Method	RMSE	MAE	MAPE*	R-squared	PIC	PIAW
4N 2HL	Only clinical data	3.37	2.72	12.21	0.34	0.97	15.65
	Only synthetic data	2.77	2.20	9.87	0.56	0.91	8.88
	Transfer Learning	2.49	1.82	8.17	0.65	0.96	9.37
32N 3HL	Only clinical data	3.70	2.95	13.23	0.31	0.97	18.65
	Only synthetic data	2.86	2.24	10.05	0.53	0.90	8.86
	Transfer Learning	2.94	2.30	10.32	0.56	0.97	11.85
128N 4HL	Only clinical data	3.97	3.29	14.76	0.13	0.98	20.00
	Only synthetic data	2.71	2.14	9.61	0.58	0.93	9.16
	Transfer Learning	2.70	2.20	9.87	0.61	0.96	13.52

RMSE: root mean square error. MAE: mean absolute error. MAPE*: mean absolute percentage error (respect of maximum). R-squared: coefficient of determination. PIC: prediction interval coverage. PIAW: prediction interval average width. N: neurons. HL: hidden layers. a_{CT} : cricothyroid muscle activation. a_{TT} : thyroarytenoid muscle activation. P_S : subglottal pressure. P_C : collision pressure.

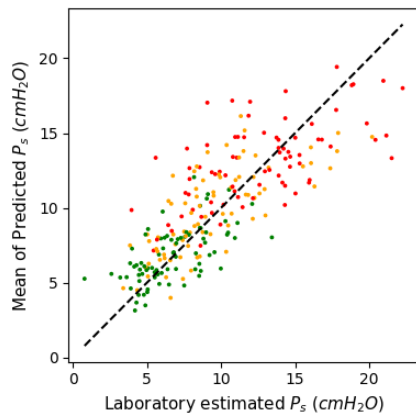


Fig. 7. Scatter plot of clinical data with point estimations on the y axis and the actual value of the data in the x axis for best model trained with D_1 (4 neurons per hidden layer, 2 hidden layers). The dashed line represents the 1:1 ratio (ground truth). Data points are color-coded based on their standard deviation, with green representing the lower third, yellow the middle third, and red the upper third percentile. P_S : subglottal pressure.

As shown by the probabilistic metrics in Table VII, Transfer Learning achieves consistently high PIC values (0.96–0.97) across configurations, comparable to the ‘Only clinical data’ method and superior to the ‘Only synthetic data’ method, which exhibits lower PIC values (0.91–0.93), potentially due to the inherent differences between synthetic and clinical data. However, for PIAW, the ‘Only synthetic data’ method generally produces narrower uncertainty intervals (8.86–9.16), whereas Transfer Learning results in slightly wider PIAW values (9.37–13.52). In contrast, the ‘Only clinical data’ method shows the widest PIAW values (15.65–20.00), reflecting greater uncertainty when training exclusively on limited clinical samples. These results suggest that while Transfer Learning ensures robust calibration (high PIC), it strikes a balance between calibration and uncertainty estimation compared to the other training methods.

For the PBNN 4N 2HL architecture with Transfer Learning, Fig. 7 presents the scatter plot of estimated vs. laboratory-measured P_S . The color coding is grouped into the same three categories as previously detailed, indicating the range of the estimated σ . The observed trend is consistent with that seen in synthetic data: higher uncertainty is associated with higher pressure values. To quantify this trend, a CV analysis was performed, yielding values of 0.26, 0.27, and 0.31 for the low, medium,

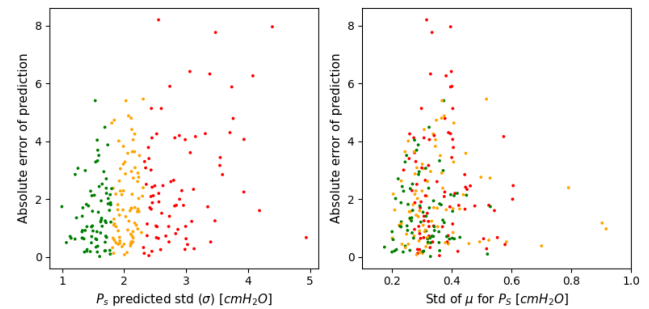


Fig. 8. Side-by-side scatter plots of clinical data for the best model trained with D_1 (4 neurons per hidden layer, 2 hidden layers). On the left: Prediction errors (y-axis) versus estimated standard deviations σ (x-axis), representing aleatoric uncertainty. On the right: Prediction errors (y-axis) versus the estimated standard deviations of μ (x-axis), representing epistemic uncertainty. Data points are color-coded based on their standard deviation, with green representing the lower third, yellow the middle third, and red the upper third percentile. P_S : subglottal pressure. μ : estimated mean. σ : estimated standard deviation. std: standard deviation.

and high groups, respectively. The CV exhibits a similar increasing trend from the low to high groups as seen in synthetic data but with a higher overall range, likely reflecting the increased variability and noise inherent to clinical measurements. A posteriori t-test analysis revealed no statistically significant difference between the first and second thirds of the data, but a significant difference was found between the second and third thirds, with a medium effect size ($d = 0.54$). This indicates that while Transfer Learning enhances the model’s performance in lower and medium ranges, it faces challenges in maintaining this performance for higher magnitudes.

For the same PBNN architecture, Fig. 8 shows the scatter plot of the absolute error of predicted P_S versus the two forms of uncertainty. On the left, aleatoric uncertainty is modeled by σ , while on the right, epistemic uncertainty is represented by the standard deviation of μ_{P_S} . The left plot (σ) shows a triangular shape, although the alignment between lower uncertainty (green points) and minimal prediction errors is less pronounced compared to synthetic data. Data points with higher uncertainty (red points) generally correspond to larger prediction errors, forming a boundary for error dispersion. This suggests that the model captures random variability in clinical data, although the relationship is less clearly defined than in synthetic data. On the other hand, the right plot (μ_{P_S}) exhibits a less distinct

clustering of the three uncertainty groups. The overlap between green, yellow, and red points is more pronounced in the clinical data compared to the estimations in synthetic data, suggesting that the model encounters greater challenges in distinguishing epistemic uncertainty levels in real-world scenarios. These findings highlight the need for additional clinical samples to further refine the model's epistemic uncertainty estimation and improve its ability to handle the inherent variability of clinical applications.

IV. DISCUSSION

Inverting the TBCM model using the PBNN proved to be a feasible approach for estimating P_S , P_C , a_{CT} , and a_{TA} in synthetic data, as well as P_S in clinical data. The results demonstrate satisfactory values for MAE, RMSE, and R^2 , comparable to those achieved by traditional MLP-based models [25], [26]. However, the most significant contribution of this work lies beyond performance metrics. The PBNN offers a distinct advantage by providing uncertainty quantification through the σ parameter, a feature not available in previous studies. This capability makes PBNNs particularly valuable for applications where understanding and quantifying uncertainty is critical. The integration of uncertainty information opens new possibilities for more informed decision-making and deeper model interpretation, as discussed in the following subsections.

A. Error and Uncertainty in Synthetic Data

In Fig. 5 it is observable that the scatter plot of estimation error vs aleatoric uncertainty shows a tendency of increasing the upper bound of the error along with the increasing value of σ . This is a very valuable finding, as it reveals that the model possesses an inherent capacity to anticipate the accuracy of its outputs. This means that the model not only provides a value prediction but also an estimation on how likely that punctual prediction really is.

For the case of the scatter plot for epistemic uncertainty in Fig. 6, the expected result was that the contour of the dots followed a less evident pattern. Given that the network was tested on data that comes from the same nature and the same ranges of operation, epistemic uncertainty was not expected to present a significant issue in the synthetic context, and most of the test data should have been interpreted by the network as familiar. However, a similar tendency is observable on increasing the upper bound of the error along with the increase of the standard deviation of μ . These results show that the model is capable of telling how sure it is on the output that it is yielding by recognising priory if it will be prone to have these errors depending on the value range of the prediction.

This probabilistic analysis revealed that the estimation of a_{TA} still presents challenges. This difficulty is attributed to the inherent complexity of vocal physiology and the need to explore new approaches for better capturing the relationships between muscle activation and glottal airflow. The seven input features used for the regression appear insufficient to fully characterize the impact of the TA muscle on voice.

B. Robustness of Predictions in Clinical Data

We believe that estimating aleatoric uncertainty in a generalized model, as opposed to a subject-specific adjustment [39], [40], is a particularly appropriate strategy for addressing the complexities of phonation across diverse individuals. This approach acknowledges the inherent variability in how different people phonate. When applying the transfer learning stage, when estimating the aleatoric uncertainty, the network presents greater values of σ than in the purely synthetic context. This shows that it is able to capture the natural fluctuations and inconsistencies present in voice data coming from a diverse population, acknowledging the specific characteristics of individual subjects. This method ensures that the model remains broadly applicable and robust for diverse input data. However, this also presents a downside, as the prediction interval average width (PIAW) can escalate quickly in the transfer learning stage, which in a way is not desirable, because bigger confidence intervals are less informative for clinicians.

In relation with the scatter plot presented in Fig. 7, a tendency of an increase of the uncertainty along with the increase of the value is observable. In other words, green dots are more likely to be in the lower range of P_S , yellow dots are more likely to be in the middle section and red points in the higher end. Incorporating the CV into the analysis suggests that the vocal fold model better captures the phonation phenomenon at low and moderate intensities. This interpretation is further supported when the PBNN was trained using the transfer learning strategy: while the CV increases compared to synthetic data, reflecting greater variability among subjects in the clinical dataset, the transition across the first and second σ ranges remains consistent, as no statistically significant differences were observed between these groups. In contrast, a significant difference was observed between the second and third σ ranges, indicating that the model faces greater challenges at higher magnitudes, consistent with the trend of increased uncertainty.

Another significant factor to consider is the data imbalance: the preprocessing of synthetic data resulted in the training data being more densely populated in the lower range of P_S values, enabling the network to recognize a more defined trend and estimate less dispersion in that range. Finally, a third factor that could explain the higher uncertainty at greater values is the presence of non-linear behaviors, which are more likely to manifest when high subglottal pressure is employed for phonation [41]. These non-linearities may be more challenging for the network to accurately capture, especially considering that the seven input features used might not adequately represent the higher-frequency phenomena present when nonlinear systems operate at high levels. Consequently, greater uncertainty is required to minimize the network's loss function.

C. Error and Uncertainty in Clinical Data

In this clinical context, the relationship between errors and their respective uncertainties is less clear than in the purely synthetic data. However, for the case of the aleatoric uncertainty shown in the left side of Fig. 8, a similar tendency to the one of the synthetic case can be identified, as the upper bound of

the error increases as the standard deviation increases. This feature is particularly valuable in clinical applications, as it allows clinicians to discern how likely the point estimations of the model are.

For the case of the scatter plot of estimation error vs $\text{std}(\mu)$ presenting the epistemic uncertainty in the right side of Fig. 8, there is no observable pattern. Also, the range where the standard deviations of the point estimations lies within is very limited, suggesting that no significant differences in the estimated epistemic uncertainty can be derived on the clinical data after the transfer learning stage. This could be due to the characteristics of the input data from the clinical dataset. The patients were all performing the same gesture, in a comfortable modal voice, making the test cases known to the network. The model, having been fine-tuned with a specific subset of clinical data, is being tested with data that has similar characteristics to the data it was trained on.

A notable result that is relevant to highlight is the great support that the TBCM provides. Table VII presents the results obtained by training using only synthetic data perform significantly better than the models trained only with clinical data. And when combining the data from both natures, the best performance metrics are obtained. This outcome highlights the TBCM model as an essential component in the implemented methodology, contributing not only to the robustness of the PBNN but also to its accuracy and generalizability.

D. Limitations and Future Work

The applicability of the proposed methodology in clinical contexts hinges on the representativeness of the estimations derived from clinical cases. The incorporation of the PBNN, which provides an additional layer of uncertainty quantification, aims to enhance confidence in these estimations. Furthermore, the use of a clinical dataset for testing and the application of transfer learning were crucial steps toward improving model performance. However, the limited availability of large-scale clinical datasets highlights the need to explore alternative transfer learning techniques. While this study exclusively applied transfer learning using clinical data, the potential to combine synthetic and clinical datasets is hindered by the significant imbalance between them, with 3,177 synthetic samples compared to only 236 clinical samples. Future research could focus on a more extensive investigation of hybrid transfer learning strategies, leveraging the introduced PBNN and transfer learning framework to further solidify its relevance for clinical applications.

Aligned with the discussion on data limitations, expert knowledge was leveraged in this study through the selection of features as network inputs, bypassing the direct use of raw signals. This design choice naturally led to the adoption of an MLP architecture. While this work successfully integrates probabilistic modeling into the estimation process, valuable insights were primarily derived from uncertainty quantification within the MLP framework. A key avenue for future research involves extending the Bayesian formalism to more complex architectures such as

PB-CNN, PB-LSTM, or even the less-explored Bayesian Transformer. Such an exploration would enable a deeper examination of parameter uncertainty consistency across architectures, providing a broader perspective on the robustness of probabilistic approaches in the domain of vocal fold modeling.

Finally, while the presented approach is computationally efficient for inference, its real-time applicability requires further development. The proposed framework relies on aerodynamic parameters that can be derived from ambulatory sensors such as accelerometers, making real-time or ambulatory extension feasible. Once trained, the neural networks exhibit fast inference times, with the most complex network requiring approximately 20 minutes for training and the simplest one only 5 minutes. However, the main bottleneck for real-time implementation lies in the preprocessing stage. Current monitoring systems for aerodynamic parameters typically process recorded signals offline over extended periods [33]. Future work should focus on integrating real-time feature extraction with the neural network inference process, enabling a complete real-time pipeline for clinical or ambulatory applications.

V. CONCLUSION

This study demonstrated how the use of a PBNN enabled the estimation of vocal function variables critical in voice disorder diagnostics, adding a new perspective to what had been done in previous works. This approach successfully addressed both aleatoric uncertainty inherent in the data, and epistemic uncertainty arising from lack of knowledge of the network. The non-linear inverse mapping was performed in both synthetic and clinical contexts, and the outcomes were on the most part coherent with what was expected and comparable with previous efforts.

Regarding the outputs yielded in the purely synthetic context, the performance of the inverse mapping was satisfactory for P_S , P_C and a_{CT} , but the TA muscle activation a_{TA} presented clear issues when tried to be estimated. This suggests the need for further research to identify additional input features that could capture the effect of the TA muscle so the model accuracy could be enhanced.

For the case of subglottal pressure, the estimated confidence intervals were noticeably narrower in the synthetic context than when clinical data was introduced in the training stage. That phenomenon is a clear consequence of the differences that exist in the phonation process of different people. Each person has their own dimensions of their vocal folds and vocal tract, which makes every voice unique, unlike how the synthetic data was generated, with fixed physical properties. The TBCM provides consistent outputs all along its input parameter space, which works great as a starting point for training an inverse mapping model in a later transfer learning stage.

In clinical contexts, the implemented approach showed promising results when incorporating the transfer learning stage. The PBNN was able to capture natural fluctuations and inconsistencies in voice data. However, larger prediction interval widths in clinical data indicate a trade-off between model applicability

and the informativeness of confidence intervals, so further refinement would be best for clinical utility. An optimal way to address this would be using transfer learning for subject specific vocal function estimation, so the re-trained model could learn from the subject it is intended to be tested in [26]. This way the randomness of the differences of the characteristics of the population would not contaminate the confidence intervals and a narrower estimation would be more feasible.

An important insight that is possible to take with the information that the PBNN provides is on which ranges of values the inverse model struggles the most in terms of variability. This analysis can help us hypothesize that the input features that are used in the neural network are suitable for estimating only the lower end of subglottal pressure values, as these features do not capture appropriately the effects that the non-linearity of the TBCM has for higher pressures. These observations open the possibility for future work to re-evaluate which features are actually relevant and which other measurements could be incorporated for a more accurate representation of the non-linear responses.

This study provides a new perspective on what can be measured in ambulatory, non-invasive vocal function monitoring. It opens the possibility of refining probabilistic inverse mapping, as narrowing the confidence intervals presents a significant challenge that could lead to the development of a genuinely useful tool for vocal function estimation. Additionally, the acquisition of clinical data for other vocal function variables remains a challenge. Addressing this would extend the analysis to a much more complete understanding of vocal function and potential assessment on speech therapy.

REFERENCES

- [1] N. Bhattacharyya, "The prevalence of voice problems among adults in the United States," *Laryngoscope*, vol. 124, no. 10, pp. 2359–2362, 2014.
- [2] National Institute on Deafness and Other Communication Disorders (NIDCD), U.S. Department of Health and Human Services, "NIDCD. 2017–2021 NIDCD strategic plan," National Institute on Deafness and Other Communication Disorders (NIDCD), U. S. Department of Health and Human Services, Bethesda, MD, USA, Tech. Rep., 2017.
- [3] D. K. Chhetri, J. Neubauer, E. Sofer, and D. A. Berry, "Influence and interactions of laryngeal adductors and cricothyroid muscles on fundamental frequency and glottal posture control," *J. Acoustical Soc. Amer.*, vol. 135, no. 4, pp. 2052–2064, 2014.
- [4] D. K. Chhetri and J. Neubauer, "Differential roles for the thyroarytenoid and lateral cricoarytenoid muscles in phonation," *Laryngoscope*, vol. 125, no. 12, pp. 2772–2777, 2015.
- [5] A. M. Vahabzadeh-Hagh, P. Pillutla, Z. Zhang, and D. K. Chhetri, "Dynamics of intrinsic laryngeal muscle contraction," *Laryngoscope*, vol. 129, no. 1, pp. E21–E25, 2018.
- [6] A. D. Hillel, "The study of laryngeal muscle activity in normal human subjects and in patients with laryngeal dystonia using multiple fine-wire electromyography," *Laryngoscope*, vol. 111, no. S97, pp. 1–47, 2001.
- [7] C. J. Poletto, L. P. Verdun, R. Strominger, and C. L. Ludlow, "Correspondence between laryngeal vocal fold movement and muscle activity during speech and nonspeech gestures," *J. Appl. Physiol.*, vol. 97, no. 3, pp. 858–866, 2004.
- [8] J. Van Den Berg, "Direct and indirect determination of the mean subglottic pressure; sound level, mean subglottic pressure, mean air flow, subglottic power and efficiency of a male voice for the vowel (a)," *Folia Phoniatrica et Logopaedica*, vol. 8, no. 1, pp. 1–24, 1956.
- [9] P. Ladefoged and N. P. McKinney, "Loudness, sound pressure, and subglottal pressure in speech," *J. Acoustical Soc. Amer.*, vol. 35, no. 4, pp. 454–460, 1963.
- [10] K. L. Marks, J. Z. Lin, J. A. Burns, T. A. Hron, R. E. Hillman, and D. D. Mehta, "Estimation of subglottal pressure from neck surface vibration in patients with voice disorders," *J. Speech, Language, Hear. Res.*, vol. 63, no. 7, pp. 2202–2218, 2020.
- [11] J. Z. Lin, V. M. Espinoza, K. L. Marks, M. Zanartu, and D. D. Mehta, "Improved subglottal pressure estimation from neck-surface vibration in healthy speakers producing non-modal phonation," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 2, pp. 449–460, Feb. 2020.
- [12] J. P. Cortés et al., "Ambulatory monitoring of subglottal pressure estimated from neck-surface vibration in individuals with and without voice disorders," *Appl. Sci.*, vol. 12, no. 21, 2022, Art. no. 10692.
- [13] D. D. Mehta et al., "Direct measurement and modeling of intraglottal, subglottal, and vocal fold collision pressures during phonation in an individual with a hemilaryngectomy," *Appl. Sci.*, vol. 11, no. 16, Art. no. 7256, 2021.
- [14] M. Motie-Shirazi, M. Zañartu, S. D. Peterson, D. D. Mehta, R. E. Hillman, and B. D. Erath, "Collision pressure and dissipated power dose in a self-oscillating silicone vocal fold model with a posterior glottal opening," *J. Speech, Language, Hear. Res.*, vol. 65, no. 8, pp. 2829–2845, 2022.
- [15] G. A. Alzamendi et al., "Bayesian estimation of vocal function measures using laryngeal high-speed videoendoscopy and glottal airflow estimates: An in vivo case study," *J. Acoustical Soc. America*, vol. 147, no. 5, pp. EL434–EL439, 2020.
- [16] E. J. Ibarra et al., "Constrained extended Kalman filter for improving Bayesian inference of vocal function from laryngeal high-speed videoendoscopy," in *Proc. 18th Int. Symp. Med. Inf. Process. Anal.*, C. M. D., Eds., International Society for Optics and Photonics. SPIE, 2023, vol. 12567, Art. no. 125671E. [Online]. Available: <https://doi.org/10.1117/12.2669812>
- [17] B. D. Erath, M. Zañartu, K. C. Stewart, M. W. Plesniak, D. E. Sommer, and S. D. Peterson, "A review of lumped-element models of voiced speech," *Speech Commun.*, vol. 55, no. 5, pp. 667–690, 2013. [Online]. Available: <https://doi.org/10.1016/j.specom.2013.02.002>
- [18] B. H. Story and I. R. Titze, "Voice simulation with a body-cover model of the vocal folds," *J. Acoustic Amer. Soc.*, vol. 97, no. 2, pp. 1249–1260, 1995.
- [19] I. R. Titze and E. J. Hunter, "A two-dimensional biomechanical model of vocal fold posturing," *J. Acoustical Soc. Amer.*, vol. 121, no. 4, p. 2254–2260, Mar. 2007.
- [20] G. E. Galindo, S. D. Peterson, B. D. Erath, C. Castro, R. E. Hillman, and M. Zañartu, "Modeling the pathophysiology of phonotraumatic vocal hyperfunction with a triangular glottal model of the vocal folds," *J. Speech, Language, Hear. Res.*, vol. 60, no. 9, pp. 2452–2471, Sep. 2017.
- [21] G. A. Alzamendi, S. D. Peterson, B. D. Erath, R. E. Hillman, and M. Zañartu, "Triangular body-cover model of the vocal folds with coordinated activation of the five intrinsic laryngeal muscles," *J. Acoustical Soc. Amer.*, vol. 151, no. 1, pp. 17–30, Jan. 2022.
- [22] P. Gómez, A. Schützenberger, M. Semmler, and M. Döllinger, "Laryngeal pressure estimation with a recurrent neural network," *IEEE J. Transl. Eng. Health Med.*, vol. 7, 2019, Art. no. 2000111.
- [23] J. Donhauser, B. Tur, and M. Döllinger, "Neural network-based estimation of biomechanical vocal fold parameters," *Front. Physiol.*, vol. 15, 2024, Art. no. 1282574.
- [24] Z. Zhang, "Estimation of vocal fold physiology from voice acoustics using machine learning," *J. Acoustical Soc. Amer.*, vol. 147, pp. EL264–EL270, 2020.
- [25] E. J. Ibarra et al., "Estimation of subglottal pressure, vocal fold collision pressure, and intrinsic laryngeal muscle activation from neck-surface vibration using a neural network framework and a voice production model," *Front. Physiol.*, vol. 12, 2021, Art. no. 732244.
- [26] E. J. Ibarra, J. D. Arias-Londoño, J. I. Godino-Llorente, D. D. Mehta, and M. Zañartu, "Subject-specific modeling by domain adaptation for the estimation of subglottal pressure from neck-surface acceleration signals," *Biomed. Signal Process. Control*, vol. 106, 2025, Art. no. 107681. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1746809425001922>
- [27] V. M. Espinoza, M. Zañartu, J. H. Van Stan, D. D. Mehta, and R. E. Hillman, "Glottal aerodynamic measures in women with phonotraumatic and nonphonotraumatic vocal hyperfunction," *J. Speech, Language, Hear. Res.*, vol. 60, no. 8, pp. 2159–2169, 2017.
- [28] D. D. Mehta, M. Zañartu, S. W. Feng, H. A. I. Cheyne, and R. E. Hillman, "Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 11, pp. 3090–3096, Nov. 2012. [Online]. Available: <https://doi.org/10.1109/TBME.2012.2207896>

- [29] M. Ghassemi et al., "Learning to detect vocal hyperfunction from ambulatory neck-surface acceleration features: Initial results for vocal fold nodules," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 6, pp. 1668–1675, Jun. 2014.
- [30] M. Zañartu, J. C. Ho, D. D. Mehta, R. E. Hillman, and G. R. Wodicka, "Subglottal impedance-based inverse filtering of voiced sounds using neck surface acceleration," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 21, no. 9, pp. 1929–1939, Sep. 2013.
- [31] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," in *Proc. Int. Conf. Mach. Learn.*, Lille, France, 2015, vol. 37, pp. 1613–1622.
- [32] K. Salama, "Building probabilistic bayesian neural network models with tensorflow probability," Jan. 2021. Accessed: Jan. 15, 2021. [Online]. Available: https://keras.io/examples/keras_recipes/bayesian_neural_networks/
- [33] D. D. Mehta et al., "Using ambulatory voice monitoring to investigate common voice disorders: Research update," *Front. Bioeng. Biotechnol.*, vol. 3, Art. no. 155, 2015.
- [34] G. B. Kempster, B. R. Gerratt, K. V. Abbott, J. Barkmeier-Kraemer, and R. E. Hillman, "Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol," *Amer. J. Speech- Lang. Pathol.*, vol. 18, pp. 124–132, 2009.
- [35] I. R. Titze and B. H. Story, "Rules for controlling low-dimensional vocal fold models with muscle activation," *J. Acoustical Soc. Amer.*, vol. 112, no. 3, pp. 1064–1076, 2002.
- [36] B. Story, "Physiologically-based speech simulation using an enhanced wave-reflection model of the vocal tract," Ph.D. dissertation, Dept. Speech Pathology Audiology, Univ. Iowa, Iowa, 1995.
- [37] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. Springer-Verlag, New York, 2002.
- [38] Tensorflow probability, "A Python library combining probabilistic models and deep learning for data science, statistics, and machine learning," Accessed: 2023. [Online]. Available: <https://www.tensorflow.org/probability>
- [39] P. J. Hadwin et al., "Non-stationary Bayesian estimation of parameters from a body cover model of the vocal folds," *J. Acoustical Soc. Amer.*, vol. 139, no. 5, pp. 2683–2696, 2016.
- [40] P. J. Hadwin and S. D. Peterson, "An extended Kalman filter approach to non-stationary Bayesian estimation of reduced-order vocal fold model parameters," *J. Acoustical Soc. Amer.*, vol. 141, no. 4, pp. 2909–2920, 2017.
- [41] I. R. Titze, "Nonlinear source-filter coupling in phonation: Theory," *J. Acoustical Soc. Amer.*, vol. 123, no. 5, pp. 2733–2749, 2008. [Online]. Available: <https://doi.org/10.1121/1.2832337>