Biomedical Signal Processing and Control 106 (2025) 107681

Contents lists available at ScienceDirect

Biomedical Signal Processing and Control

journal homepage: www.elsevier.com/locate/bspc

Subject-specific modeling by domain adaptation for the estimation of subglottal pressure from neck-surface acceleration signals

Emiro J. Ibarra [a], Julián D. Arias-Londoño [b], Juan I. Godino-Llorente [b], Daryush D. Mehta [c], Matías Zañartu [a,*]

[a] Department of Electronic Engineering and Advanced Center for Electrical and Electronic Engineering, Universidad Técnica Federico Santa Maria, Valparaiso, 2390123, Chile
[b] ETSI Telecomunicación, Universidad Politécnica de Madrid, Madrid, 28040, Spain
[c] Center for Laryngeal Surgery and Voice Rehabilitation Laboratory, Massachusetts General Hospital–Harvard Medical School, Boston, MA, United States

ARTICLE INFO

Dataset link: sederosa@partners.org

Keywords:
Voice
Vocal folds
Neck-surface vibration
Neural networks
Subglottal pressure
Domain adaptation
Transfer learning
Voice disorders
Synthetic voice production model

ABSTRACT

Subglottal air pressure is a critical physiologically-based parameter that reveals fundamental pathophysiological processes in patients with voice disorders. However, its assessment in both laboratory and ambulatory settings presents significant challenges due to the necessity for specialized instruments, invasive procedures, and the impracticality of direct measurement in ambulatory contexts. This study expands upon previous efforts to estimate subglottal pressure from portable, lightweight neck-surface acceleration signals using a physiologically relevant model of voice production combined with machine learning techniques. The proposed approach employs a neural network architecture initially trained with numerical simulations from the voice production model, which is subsequently refined through a domain adaptation strategy from synthetic data to *in vivo* laboratory data. This proposed method provides a means to create subject and group-specific refinements of the original neural network. For comprehensive comparisons with previous methods reported in the literature, the proposed approach is applied to both normal and disordered voices, including cases of unilateral vocal fold paralysis and phonotraumatic and non-phonotraumatic vocal hyperfunction. The study is divided into two datasets, encompassing a total of 135 participants. The *in vivo* recordings consist of synchronous measurements of oral airflow, intraoral pressure, and signals from a microphone and a neck-surface accelerometer. Each participant was asked to utter/p/-vowel syllable gestures with variations in loudness, vowels, pitch, and voice quality. Compared to previously reported approaches, the proposed method results in subject-specific models that achieve over a 21% improvement in the estimation of subglottal pressure, as measured by root mean square error. These findings underscore the effectiveness of a non-linear, subject-specific regression approach in enhancing the estimation of subglottal pressure from neck-surface vibration signals.

1. Introduction

Subglottal air pressure (P_s) is a primary factor in initiating and maintaining vocal fold oscillation [1], adjusting loudness [2], and contributing to the control of the fundamental frequency [3]. Thus, the clinical management of voice disorders would benefit significantly from an ambulatory evaluation of this physiological characteristic. Furthermore, previous studies have shown that measures of P_s are associated with phonatory efficiency [4], vocal effort [5], and can serve as a mechanism to differentiate between normal and disordered voices [6–10]. Consequently, P_s provides valuable information to improve methods for the treatment, diagnosis, and prevention of voice disorders.

Several methods have been used to measure P_s directly or indirectly. Direct approaches include inserting a needle through the trachea [11,12] or passing miniature pressure transducers transorally [13,14], while indirect methods use esophageal balloons [15,16]. Nonetheless, these techniques are rarely applied in clinical scenarios due to their cumbersome and invasive nature, as well as the need for expensive and specialized equipment. Consequently, the most widely accepted method involves interpolating P_s as the average value of two consecutive intraoral pressure plateaus, recorded during repetitions of /p/-vowel syllable utterances [17]. These plateaus result from the combined closure of the lips and the opening of the glottis before and after each vowel segment, occurring just before and following the /p/ sounds. This method relies on the assumption that intraoral pressure during /p/

* Corresponding author.
E-mail address: matias.zanartu@usm.cl (M. Zañartu).

https://doi.org/10.1016/j.bspc.2025.107681
Received 16 September 2024; Received in revised form 31 January 2025; Accepted 7 February 2025
Available online 26 February 2025

occlusion results from the equilibration of subglottal pressure produced during the vowel.

Recent studies have investigated methods for estimating P_s using neck-surface vibrations (NSV) recorded with an accelerometer (ACC). These approaches leverage the existing correlation between P_s and the amplitude of the ACC signal. Their primary advantage lies in the use of an affordable and non-intrusive sensor, a crucial requirement for future integration into wearable systems designed for ambulatory assessment of P_s [18]. Initial investigations of this method have proposed calibrated subject-specific linear regressors to map the amplitude of the ACC signal to P_s values [19]. Subsequent efforts yielded a strong correlation between these two parameters, particularly during variations in vocal effort. However, it is important to note that this correlation undergoes changes between different intensity levels in certain individuals [20] due to the specificities of the subject.

Furthermore, the relationship between the amplitude of the ACC signal and P_s is notably affected by non-modal phonations and by the presence of voice disorders [21,22]. To address these challenges, the aforementioned methods were extended to include further measurements of vocal function based on ACC. This extension aimed to offset the effects of non-modal phonation and thus improve the accuracy of P_s estimation [23]. Notably, this extended method exhibits superior predictive performance in estimating P_s based on NSV features for both normal and pathological voices [24].

Alternative research lines have employed numerical models of voice production to estimate essential clinical parameters, such as P_s. Among these approaches, the literature reports optimization-based voice inversion methods [25–31], approaches based on Bayesian estimation [32–39], and machine learning tools integrated with voice production models [40,41]. Based on this last approach, in [42], we proposed a method using non-linear regression to estimate P_s, the activation levels of two intrinsic laryngeal muscles, and the vocal fold collision pressure, all from ACC signals. For this purpose, we trained a Neural network (NN) regressor using thousands of simulations obtained from a Triangular Body Cover Model (TBCM) controlled by the coordinated activation of five intrinsic laryngeal muscles [43]. Validation against synthetic data showed high performance. However, when estimating P_s from *in vivo* laboratory data —including both normal and pathological voices—, the performance decreased with respect to subject-specific calibrated linear regression approaches, as those in [24].

The aforementioned suggests the need to align the generic characteristics of synthetic models to the specificities of the subject (or speaker), or cohort (i.e., control group or pathological condition). In this respect, domain adaptation methods aim to mitigate the distribution shift between domains by aligning the feature spaces (or probability distributions) of the source (synthetic) and target (subject or cohort) domains. Different domain adaptation techniques are prone to be used. For this purpose, transfer learning (TL) is a simple but useful domain adaptation technique commonly used [44]. TL harnesses knowledge, specifically weights and biases, from a pre-trained model to enhance the performance of a new one [44]. TL reduces the reliance on large datasets to model intricate nonlinear relationships, thus facilitating improved performance on smaller datasets by fine-tuning models that were originally trained for specific tasks using extensive datasets [45]. This technique has proven highly effective in various areas [46], including speech applications such as voice disorder classification [47], Parkinson's disease detection [48] and assessment [49], and Alzheimer's disease detection [50]. Based on this, we hypothesize that a domain adaptation using TL can compensate for errors arising from differences in the feature space and distribution between synthetic and *in vivo* voice recordings when estimating P_s from ACC signals using an NN-based framework.

On the other hand, we highlight that in simulations using numerical voice production models, the parameters are adjusted over an extensive range without considering anatomical differences. This suggests that the voice model primarily simulates an individual subject under various conditions, not being able to capture the expected inter-subject variability commonly observed in laboratory settings. Thus, new techniques are required to adapt the models to a much broader population or to the specific characteristics of each individual subject.

In our ongoing effort towards the ambulatory assessment of vocal function, we propose a domain adaptation to the speaker using TL to establish a more robust non-linear mapping between a set of ACC-based features and P_s. Initially, for this purpose, we utilized an NN regressor trained with data obtained from a synthetic voice production model. This NN takes aerodynamic and acoustic features as inputs and returns P_s as output. Later, a domain adaptation was carried out using TL to fine-tune the NN weights using laboratory recordings. We applied TL in two scenarios: firstly, by training a single model to estimate P_s across multiple subjects; and, secondly, by proposing subject-specific models. In the second approach, TL adjusts the source to a target domain for each individual subject. This subject-specific method is feasible since the TL technique enables the recalibration of the NN model using fewer training samples.

The novelty and contribution of this work consist of two main aspects: first, we support the idea that TL is a good strategy for domain adaptation from synthetic to *in vivo* data; second, we provide an improved subject-specific method to estimate P_s from ACC signals.

The rest of the document is structured as follows: Section 2 provides an introduction to the materials and methods employed in system development; Section 3 offers the results; Section 4 is dedicated to the discussion of the results; and Section 5 outlines the principal conclusions drawn from the current study.

2. Materials and methods

The illustration in Fig. 1 presents an overview of the method to improve the estimation of P_s from the NSV recorded with an ACC using an NN regressor. The upper block of the scheme represents the baseline model, which comprises an NN trained with thousands of simulations from a numerical voice production model as in [42]. This baseline model effectively maps various aerodynamic and acoustic input features to the subglottal pressure. In the first stage, aerodynamic features such as the fundamental frequency (f_o), Maximum Flow Declination Rate (MFDR), Open Quotient (OQ), Speed Quotient (SQ), Amplitude of Unsteady Glottal Airflow (ACFL), and spectral tilt (measured as the log-magnitude difference between 1st and 2nd harmonics, $H_1 - H_2$) are computed from the simulated Glottal Volume Velocity (GVV) signal. The Sound Pressure Level (SPL) is computed from the radiated pressure simulated at the lips (P_{out}). These features have already been shown to correlate with P_s in [23,24,41,42,51]. For detailed descriptions of each feature, see Table 1.

The bottom block fine-tunes the NN using a domain adaptation strategy based on TL and *in vivo* laboratory measurements. This process adapts the results to the specific characteristics of the cohort. For this purpose, we freeze the initial layers of the baseline model and retrain the subsequent hidden layers. In this second stage, the aerodynamic input features are computed from the glottal airflow sequence, which is obtained by processing the ACC signal using the subglottal Impedance-Based Inverse Filtering (IBIF) model [53,54]. The SPL is estimated from the microphone (MIC) signals, and the reference subglottal pressure is derived from intraoral sensor pressure (PRE) measurements.

2.1. Simulated voice production

As in [42], the selected numerical model for voice production is the TBCM for vocal folds, controlled by the coordinated activation of five intrinsic laryngeal muscles, as detailed in [43]. This physiologically-based model refines previous studies by integrating vocal fold posturing [57], rules for controlling low-order lumped models [58], and the TBCM itself [59]. Consequently, the model represents a symmetrical, low-order depiction of the vocal folds. It is regulated by adjustments in

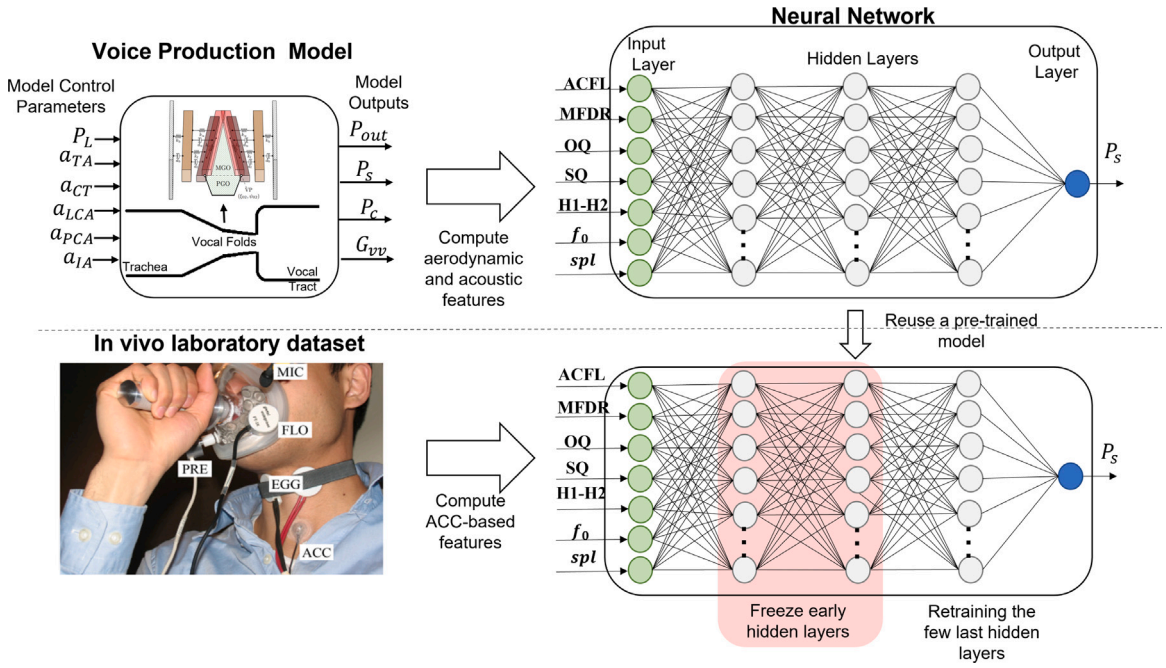

Fig. 1. Diagram outlining the TL procedure. Initially, a baseline regression NN is trained using synthetic data (i.e., generated from synthesizer simulations). This NN is then fine-tuned to accurately map ACC-based vocal features to clinical estimates of subglottal pressure.
Source: Adapted from [42,43,52].

Table 1
Descriptions of the aerodynamic and acoustic features.
Source: Adapted from [55].

Feature	Description	Units
f_o	Fundamental frequency	Hz
MFDR	The maximum flow declination rate is determined by the negative peak in the first derivative of the glottal airflow.	$\frac{l}{s^2}$
OQ	The ratio between the duration of the glottal opening within a vibratory cycle and the total period of that cycle.	%
SQ	The ratio between the opening and closing time of the glottal vibratory cycle.	–
ACFL	The range of the AC glottal airflow amplitude, computed as the difference between its peak and trough values within each glottal cycle.	$\frac{ml}{s}$
$H_1 - H_2$	Difference in level (in decibels) between the first harmonic (H_1) and the second harmonic (H_2)	dB
SPL	The sound pressure level can be directly computed from the root mean square (RMS) of the microphone signal's envelope, calibrated in pascals. Alternatively, it can be derived from the RMS magnitude of the ACC signal, as detailed in [56].	dB SPL

aerodynamic lung pressure (P_L), and the normalized levels of activation of several muscles: thyroarytenoid (a_{TA}), cricothyroid (a_{CT}), interarytenoid (a_{IA}), lateral cricoarytenoid (a_{LCA}), and posterior cricoarytenoid (a_{PCA}).

Furthermore, our numerical model accounts for interactions between tissue, fluid, and acoustics at the glottis by employing subglottal and supraglottal tract models [60]. The model also simulates sound wave propagation through the vocal tract to estimate the P_{out} [61]. For more in-depth information on the model, please refer to the comprehensive descriptions provided in [43,59].

From this synthetic voice production model, we generate thousands of sustained vowel simulations to replicate the physiological behavior of different phonatory conditions. These simulations are based on vocal tract area functions representative of typical male [62] and female [63] vocal tracts. For each vocal tract configuration, we vary the simulation

Table 2
Range and step settings of model control parameters used in simulations to create the synthetic dataset.
Source: Adapted from [42].

Parameters	Range	Step
a_{TA} and a_{CT}	0–1	0.1
a_{LCA} and a_{IA}	0.2–0.8	0.1
a_{PCA}	0–0.1	0.1
P_L (Pa)	500–2000	150

control parameters within the range and steps specified in Table 2. The range of the control parameters does not ensure a universal synthetic model but is considered wide enough for a wide range of phonation modes.

Subsequently, we isolated the final 50 ms of the simulated GVV signal to remove transient artifacts. We then low-pass filtered the signal using a 10th-order Chebyshev II filter with a cutoff frequency of 1100 Hz, and high-pass filtered it with a 4th-order Butterworth with a cutoff frequency of 60 Hz to match the typical frequencies of laboratory recording signals [10]. Filtering was applied bidirectionally to achieve zero-phase distortion. Following this, we computed the mean of six aerodynamic features listed in Table 1 and determined the SPL from the simulated sound pressure at the lips (P_{out}). This is calculated as SPL $= 20\log_{10}(P_{out}/2\times10^{-5})$. Samples that did not meet the clinical registry criteria, such as an ACFL below 30 ml/s or a f_o outside the 120–400 Hz range, were discarded [55]. In total, our synthetic dataset comprised 13,000 samples, (9000 for males and 4000 for females) with vocal tract configurations for the vowels /æ/and /ɑ/, in the same proportion.

2.2. *In vivo* laboratory recordings

The reference for P_s was derived from two datasets, each containing *in vivo* laboratory recordings. These recordings included data about Oral Airflow Volume Velocity (OVV), Intraoral Pressure (IOP), audio recorded with a microphone, and neck surface vibration collected with an ACC. The recordings of both datasets were made in a sound-treated environment to ensure signal integrity. The vocal health

diagnoses were provided by a board-certified laryngologist and a licensed speech-language pathologist in a team evaluation [9,21,22]. Clinical voice assessment included an auditory-perceptual evaluation, laryngeal stroboscopic examination, patient-reported quality of life ratings, and objective documentation of aerodynamic and acoustic voice characteristics. All participants signed an informed consent before their participation in the study. Descriptions of each of these databases are provided in the following.

2.2.1. Laboratory dataset 1 (LD1)

The participants in these laboratory recordings consisted of seventy-nine adult women, all without a history of vocal disorders and with normal vocal status verified by laryngeal endoscopic evaluation, auditory perception evaluation of voice, and auditory screening performed by a licensed speech-language pathologist. The mean age of this cohort was 29.6 years with a standard deviation of 13.0 years. Each participant was instructed to produce sequences of /pæ/syllables under distinct loudness conditions: soft, comfortable, and loud. LD1 has been used in several previous research studies, as mentioned in [9,10,42,52].

As detailed in [52], the laboratory equipment used for LD1 included: a microphone (model MKE104, Sennheiser®, Electronic GmbH, Wedemark, Germany), positioned 10 cm away from the mouth to capture acoustic pressure with a bandwidth ranging from 0 to 6000 Hz; a pneumotachograph mask (model PT-2E, Glottal Enterprises®, Syracuse, NY), equipped with circumferential vents and offering a bandwidth of approximately 1100 Hz to measure OVV; a low-bandwidth pressure sensor (model PT-25, Glottal Enterprises®), positioned inside the mouth via an oral catheter to measure IOP; and a BU-27135 sensor from Knowles Corp.®, Itasca, IL, USA, to capture NSV. The signals were subjected to a low-pass filter at 8000 Hz, employing the CyberAmp® Model 380, also from Axon Instruments®, Inc., and sampled at a rate of 20 kHz with 16 bits using Digidata® 1440 A equipment, also from Axon Instruments®, Inc. OVV, IOP, and MIC signals were calibrated into physical units —ml/s, cm H_2O, and Pa, respectively— as described in [9].

2.2.2. Laboratory dataset 2 (LD2)

This dataset contains speakers divided into four distinct cohorts: ten patients with non-phonotraumatic vocal hyperfunction (NPVH), ten with phonotraumatic vocal hyperfunction (PVH), ten with unilateral paralysis of a vocal fold (UVFP), and 26 individuals with no history of voice disorders (control group). Table 3 provides detailed demographic information for each cohort. Participants were instructed to produce /p/-vowel syllables repetitively, modulating their loudness from loud to soft, in three distinct vowel contexts: /pa/, /pi/, and /pu/. In contrast to LD1, the method of eliciting /p/-vowel pairs with progressively decreasing loudness facilitated a more comprehensive collection of the spectrum of P_s [4]. Details regarding this dataset can be found in [21,23] for the control groups and in [22,24] for the pathological cohorts.

For these dataset recordings, the laboratory equipment is detailed in [21,22] and summarized as follows: a head-mounted condenser microphone (model ME 102, Sennheiser®, Electronic GmbH, Wennebostel, Germany) placed 15 cm away from the participants' lips; a pneumotachograph mask made by Glottal Enterprises® (Syracuse, NY, USA), with dedicated sensors PT-2E for measuring OVV and PT-75 for IOP; and an ACC sensor (BU-27135, Knowles Corp.®, Itasca, IL, USA) securely fastened halfway between the suprasternal notch and the thyroid prominence using double-sided hypoallergenic tape (Model 2181, 3M®, Maplewood, MN, USA). The ACC signal was collected using a sampling frequency of 11,025 Hz and 16 bits of quantization, facilitated by an Android® smartphone. The remaining signals were sampled and low-pass filtered in a manner similar to those in LD1, employing Digidata 1440 A and CyberAmp Model 380, respectively, both from Axon Instruments®. Ultimately, the signals were calibrated into physical units, adhering to the methodology outlined in [24].

Table 3
Comparative demographic statistics for the different cohorts in LD2.
Source: Adapted from [22].

Cohorts	Speakers		Mean (SD)	Age
	Female	Male	Age	Range
Control	18	8	31 (13)	19-50
PVH	10	0	29 (18)	18-62
NPVH	7	3	35 (11)	19-64
UVFP	6	4	45 (15)	22-60

2.3. Laboratory data pre-processing

First, the sequences obtained from the microphone, pneumotachograph mask, and ACC were segmented.

For LD1, signals were segmented at the vowel boundaries following the criteria outlined in [9]. In LD2, segmentation was carried out by identifying sounding/silent intervals in the acoustic signal recorded with the MIC using Praat® v.6.0.30 [64], as detailed in [23]. The reason for the differences in the segmentation procedure for both corpora lies in the different acoustic materials recorded. Additionally, in both datasets, each vowel segment of the OVV signal was filtered using a 10th-order Chebyshev II low-pass filter, adjusted to the bandwidth of the pneumotachograph mask (1100 Hz) [9]. The IOP signal was filtered using a 5th-order Butterworth low-pass filter with a cutoff frequency of 80 Hz [21].

Glottal airflow based on OVV was obtained through standard inverse filtering, applied to a stable window of 50 ms in the middle of vocalic segments of a given /pæ/syllable. The filtering technique uses a single-notch filter, characterized by a unity gain in DC and a pair of complex conjugate zeros, specifically targeting the first resonance frequency of the vocal tract, as detailed in [65,66]. In LD1, the OVV signal was obtained only for the syllable /pæ/closest to the mean SPL value for each loudness condition, using the same criterion as in [9]. For LD2, it was obtained for each vowel segment present in every phonation sequence.

Following this, the IBIF model was employed to derive ACC-based glottal airflow from the same 50 ms segments of the vowels. The IBIF method employs mechano-acoustic impedance representations [53] and incorporates a calibration process to determine subject-specific parameters, including neck-skin surface characteristics, tracheal length, and ACC placement, as detailed in [53,55,67]. A Particle Swarm Optimization approach [68] was used to determine model parameters by minimizing the error between the waveforms of the OVV-based glottal airflow and those obtained from the inverse filtered ACC signal.

The aerodynamic and acoustic features listed in Table 1 were computed from the glottal airflow signal derived using IBIF. The SPL was obtained from the calibrated acoustic signal and aligned to the same vowel segments of the ACC signal. The reference values for P_s were extrapolated as the mean value of two consecutive IOP plateaus, occurring just before and after each vowel segment. Finally, we obtained a total of 237 tokens for LD1 and 15,160 tokens among the four groups of LD2.

2.4. Domain adaptation from simulated voice production to in vivo recordings

TL is a powerful domain adaptation technique used in machine learning that offers an alternative approach to traditional training methods. Instead of training a model from scratch using domain-specific data, TL leverages previously trained models on different domains, tasks, or distributions. This technique was formally defined in [44] as:

"Given a source domain D_S and learning task $\mathcal{T}_S$, and a target domain D_T and learning task $\mathcal{T}_T$, TL aims to help improve the learning of the target

predictive function $f(\cdot)$ in $\mathcal{D}_T$ using the knowledge in $\mathcal{D}_S$ and $\mathcal{T}_S$, where $\mathcal{D}_S \neq \mathcal{D}_T$ or $\mathcal{T}_S \neq \mathcal{T}_T$."

Here, the domain is defined by a tuple composed of a feature space $\mathcal{X}$ and a marginal probability distribution $Q(X)$, regarding the target variable, represented as $\mathcal{D} = \{\mathcal{X}, Q(X)\}$. Complementary, the learning task comprises the label space $\mathcal{Y}$ and the conditional probability distribution $Q(Y|X)$, represented as $\mathcal{T} = \{\mathcal{Y}, Q(Y|X)\}$ [44]. According to the former TL definition, the condition $\mathcal{D}_S \neq \mathcal{D}_T$ implies that $\mathcal{X}_S \neq \mathcal{X}_T$ or $Q_S(X) \neq Q_T(X)$. Similarly, $\mathcal{T}_S \neq \mathcal{T}_T$ implies that either $\mathcal{Y}_S \neq \mathcal{Y}_T$ or $Q_S(Y|X) \neq Q_T(Y|X)$.

In our regression problem, the synthetic and laboratory data represent $\mathcal{D}_S$ and $\mathcal{D}_T$, respectively. The subglottal pressure estimation is the learning task, thus ($\mathcal{T}_S = \mathcal{T}_T$). Previously [42,51], it was assumed that $\mathcal{D}_S = \mathcal{D}_T$ and $\mathcal{T}_S = \mathcal{T}_T$, treating the problem as a traditional machine learning approach. However, simulations from the numerical voice production model are approximations of the real three-way interaction (sound, flow, and vocal fold tissue) at the glottal level. Consequently, since $\mathcal{X}_S \neq \mathcal{X}_T$ and $Q_S(X) \neq Q_T(X)$, it follows that $\mathcal{D}_S \neq \mathcal{D}_T$. Therefore, our regression problem aligns with the TL definition.

The approach utilized in this study, referred to as fine-tuning or parameter transfer [46,69], entails substituting the final or several layers of a base model with tailored layers for the desired task. Throughout training, the parameters of the pre-trained model are refined via ongoing back-propagation. This fine-tuning procedure enables the model to better conform to the nuances of the target learning objective.

In this study, the TL framework is applied to two distinct scenarios, aiming to demonstrate improvements in the estimation of P_s compared to state-of-the-art results. Firstly, we refine a single regression model to estimate P_s for various subjects, as explored in [42]. Secondly, we implement a subject-specific adaptation of the model, aligning our approach with the methodologies described in [21–23]. These two approaches aim to enhance the accuracy and applicability of subglottal pressure estimation from the neck-surface acceleration in diverse subject settings.

2.5. Baseline NN architecture and fine-tuning strategy

The baseline model is a regressor based on a Multilayer Perceptron NN. The input layer comprises seven features, including aerodynamic (f_o, MFDR, OQ, SQ, ACFL, $H1-H2$) and acoustic (SPL) features. The output of the NN is P_s. Each interconnected hidden layer comprises a rectified linear unit and a dropout layer. The search space for the hyperparameters of the baseline model was tuned using Talos® [70] following a 5-fold cross-validation strategy on the synthetic dataset (Table 4). The best-performing model comprises three hidden layers with 256, 128, and 64 neurons, respectively, each with a dropout rate of 0.1, and was trained using a batch size of 64. This configuration was selected as the baseline for all subsequent experiments.

The baseline regression model was adapted following a TL strategy by sequential layer freezing and additional training with laboratory datasets. Initially, we aimed to develop a general model capable of estimating P_s for a universal population and for certain cohorts. In the second stage, we focused on developing subject-specific models. The optimal performance of the TL strategy was achieved by sequentially freezing the hidden layers. This approach enabled the system to retain high-level features from the source domain of the baseline model while adapting to the target domain by retraining the unfrozen layers. The effectiveness of these models was validated through cross-validation.

Baseline model training and fine-tuning for domain adaptation were carried out using the Adam optimization algorithm with mean squared error (MSE) as a loss function. The PyTorch® Lambda learning rate schedule was used, initialized at 0.001. For all experiments, the synthetic and laboratory data were Min–Max normalized. The regression models in this study were trained and tested on a Google Colab® virtual machine, powered by two Intel® Xeon® CPUs @ 2.00 GHz, using Python® (v.3.6.9) and the PyTorch library (v.2.1.0).

Table 4
Search space for the hyperparameters of the baseline model.

Hyperparameters	Values
Hidden layer	2, 3, 4, 6
Neurons by layer	32, 64, 128, 256
Dropout rate	0.1, 0.2
Batch size	8, 16, 32, 64

3. Results

The regression performance for estimating P_s, both in general and in subject-specific cases, is measured using several metrics: the coefficient of determination (R^2), the root-mean-squared error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE). These metrics were chosen for their ability to provide a comprehensive assessment of model accuracy and to provide a quantitative comparison with previous work. In all experiments, we compare model performance with and without domain adaptation in order to investigate if this new approach leverages pre-trained knowledge for enhancing accuracy and efficiency compared to models trained without pre-existing knowledge.

3.1. General adapted model for the estimation of P_s

The initial set of experiments was designed to evaluate the interest in applying a domain adaptation strategy to get a universal model and/or a model for each cohort (i.e., applicable to every potential subject belonging to a certain pathological condition). Thus, in this step, models are supposed to be subject-agnostic.

Initially, we adapted the baseline NN model that was trained using synthetic data to the domain of LD1 (which includes data from 76 subjects). The adaptation was evaluated using a 10-fold stratified subject-independent cross-validation strategy, ensuring that the speakers did not overlap between different folds. This adaptation was strategically chosen to test the model's ability to generalize across diverse subjects. Subsequently, in a second adaptation phase, we adapted the NN for each specific cohort within LD2. Although LD2 has a larger number of samples, it involves a smaller number of subjects. Therefore, to assess the performance of the model for each cohort in LD2, we used a leave-one-subject-out cross-validation approach. This methodology is particularly advantageous for datasets with a limited number of subjects, as it facilitates exhaustive testing and validation on an individual subject basis, maximizing the utilization of the available data.

3.1.1. Estimation of P_s from LD1

The results in Table 5 show the error metrics obtained (for the estimation of P_s) when the model is trained from scratch and fine-tuning the baseline model —originally trained using the physiological voice synthesizer— via TL with sequential freezing of its hidden layers. This process leads to a universal, general-adapted model, which is agnostic to the specificities of the subject and/or of the cohort. Optimal performance was observed when only the first hidden layer was frozen. Under this condition, there was a decrease in all error metrics and an increase in the coefficient of determination. Notably, increasing the number of frozen hidden layers correlated with elevated error metrics, indicating the disparities between domains (i.e., synthetic signals and *in vivo* laboratory recordings). Furthermore, the improvements achieved through TL, in contrast to training the model from scratch, suggest that freezing the first hidden layer of the baseline model contributes significantly to the robustness of our non-linear regression estimation.

The optimal performance results (achieved with only the first hidden layer frozen) demonstrate an improvement over our previous work [32], which used a simple NN with 2 hidden layers and 4 neurons per layer, trained solely on a synthetic dataset. In that study, for the same laboratory data, the estimation metrics of P_s yielded an RMSE of 2.48 cm H_20, an MAE of 1.84 cm H_20, a MAPE of 24. 9%, and an

Table 5
General adapted universal model for the estimation of P_s. Error metrics for an NN training using a random initialization and a TL strategy from synthetic data with sequential frozen layers (FL)

TL	FL	RMSE (cm H_2O)	MAE (cm H_2O)	MAPE (%)	R^2
	–	2.51 ± 0.45	1.93 ± 0.36	23.27 ± 9.49	0.63
✓	0	2.59 ± 0.47	1.99 ± 0.37	24.37 ± 9.58	0.60
✓	1	**2.30 ± 0.41**	**1.77 ± 0.28**	**21.38 ± 8.64**	**0.69**
✓	2	2.65 ± 0.56	1.99 ± 0.40	23.34 ± 7.85	0.58
✓	3	4.64 ± 0.68	3.50 ± 0.47	41.30 ± 14.03	−0.26

Note: 1 $cmH_2O = 98.0665$ Pa.

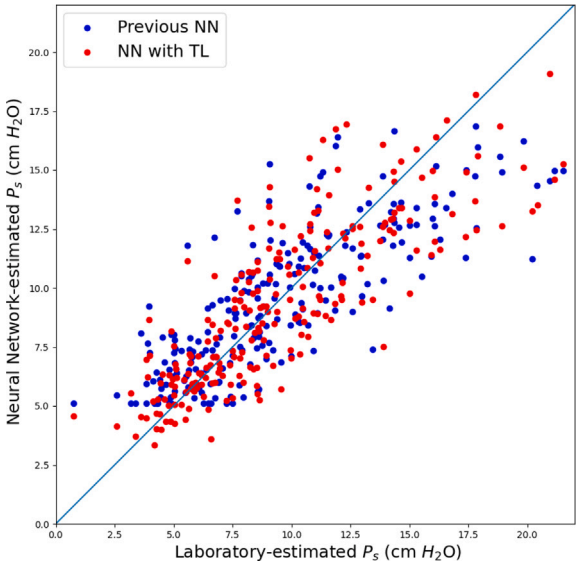

Fig. 2. Comparison between the estimates from the NN and the corresponding laboratory estimations of P_s, for the previous work [42] and using the current approach. The diagonal line represents a perfect matching in the theoretical sense (1:1).

R^2 of 0.65. In Fig. 2, a scatter plot contrasts the estimated P_s against the reference subglottal pressure, comparing our previous work (blue dots) with the universal model of the current approach (red dots). To ensure a comprehensive comparison, the plot compiles the results of the validation set using a 10-fold cross-validation. Generally, the results obtained with the TL strategy align more closely with the blue line, representing a one-to-one correspondence between the reference and the estimated P_s values. Importantly, the domain adaptation using TL yields estimations of P_s in ranges where our previous NN model was less effective, especially for P_s values below 5 cm H_20. This improvement is quantitatively manifested as an increase of 0.04 absolute points in the coefficient of determination when applying TL.

Despite these improvements, the similarity in the distribution of outlier points between the previous NN and the current TL-based approach reflects the inherent inter-subject variability present in the dataset. This highlights two critical aspects: first, while synthetic data can mimic certain population-level characteristics, it may fail to capture the full extent of variability in individual subjects; second, the stratified subject-independent cross-validation strategy in TL reveals limitations in generalization due to the relatively small population size of LD1 (79 subjects). These observations support the potential value of a subject-specific modeling strategy, which could better account for individual variability and further improve the estimation of P_s.

3.1.2. Estimation of P_s from LD2

Table 6 presents the results for the estimation of P_s for the control and pathological cohorts (i.e., PVH, NPVH and UVFP). When TL is applied to our model, we observe notable improvements in error metrics.

Table 6
Error metrics for the general adapted models for the estimation of P_s, trained with and without TL, and for cohorts: Control, PVH, NPVH, and UVFP.

Group	TL	RMSE (cm H_2O)	MAE (cm H_2O)	MAPE (%)	R^2
Control		2.12 ± 0.92	1.66 ± 0.70	23.64 ± 8.72	0.57
	✓	**2.03 ± 0.81**	**1.60 ± 0.61**	**23.16 ± 9.43**	**0.61**
PVH		3.46 ± 1.64	2.58 ± 1.28	30.45 ± 12.16	0.38
	✓	**3.16 ± 1.57**	**2.40 ± 1.15**	**30.26 ± 12.57**	**0.47**
NPVH		3.46 ± 1.83	2.81 ± 1.62	33.51 ± 14.29	0.20
	✓	**3.19 ± 1.61**	**2.57 ± 1.34**	**33.08 ± 14.60**	**0.33**
UVFP		5.18 ± 2.45	4.43 ± 2.40	59.12 ± 43.64	−0.14
	✓	**4.68 ± 2.23**	**4.02 ± 2.15**	**57.04 ± 42.44**	**0.04**

For example, in the case of the average RMSE, we observe a 4.2% improvement for the control group, while for the pathological cohorts, the improvement ranges from 7.8% to 9.7%. These results support the idea that maintaining some parameter learning from synthetic data provides better generalization compared to training the model from random initialization. It is worth noting that the error metrics in the control group are generally lower compared to the pathological cases. These differences in the estimations are consistent with the observations made in our previous work [24] and can be attributed to the high intrapathological variability. Specifically, patients with UVFP exhibited the highest RMSE, which may be due to effect of the pathology. For example, during laboratory recordings, patients with UVFP experienced more difficulty maintaining a steady pitch, changing pitch, and managing breath control [22]. Furthermore, a recent study employing high-speed video analysis revealed that a patient with UVFP exhibited chaotic behavior in the vibration dynamics of the vocal folds [71].

Fig. 3 contrasts the RMSE of the estimation of P_s using TL with two methods reported in the literature. Method 1 consists of an empirically derived formula proposed by Titze et al. in [72] that P_s computed using only SPL measurements and f_o. Method 2 is NN-trained using only synthetic data. These bar graphs show a reduction in both the mean and standard deviation of the RMSE for the TL-based approach, particularly in the control group. Applying a one-way analysis of variance (ANOVA) to the control group revealed an **F**-value of 6.15 (p = 0.0033), indicating a significant difference among the methods. Subsequently, for the same control group, a post-hoc analysis using Tukey's Honestly Significant Difference (HSD) test [73] identified statistical differences between Method 1 and the TL-based one (p = 0.0066); as well as between the Method 2 and the TL-based one (p = 0.013). The Cohen's d values for these comparisons were −0.37 and −0.28, respectively, suggesting small to medium effect sizes. In pathological cohorts, the ANOVA analysis did not reveal significant differences between the methods, with all p-values above the conventional significance threshold of 0.05.

3.2. Subject-specific models for the estimation of P_s

In this step, we search for specific models adapted to a specific subject.

Subject-specific models were also developed following a TL strategy from the initial synthetic model. The mean results for all subjects by groups are shown in Table 7. Results were obtained using a 5-fold cross-validation as in [23]. The results show that for subject-specific models, the domain adaptation using TL also improves the estimation (in contrast to being trained from scratch). For all four pathological cohorts, it is evident that the error metrics are lower for the subject-specific models (compared to the general NN). This is directly associated with the fact that constrained data to unique subjects discards the inter-subject variability.

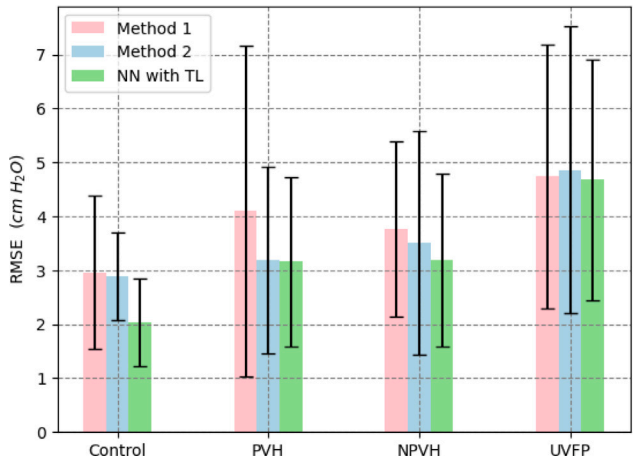

Fig. 3. Comparison of RMSE obtained for the general estimation of P_s and for the three methods used: proposed NN with TL; empirical equation (Method 1); and, NN trained exclusively with synthetic data (Method 2).

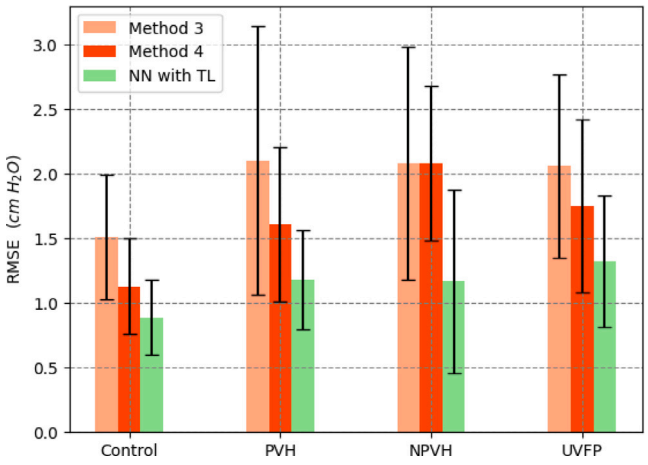

Fig. 4. Comparison of the mean RMSE for the best fold of subject-specific NN estimation among three methods: proposed NN with TL; linear regression model (Method 3); and, multi-linear regression model (Method 4).

In Fig. 4, we contrast the RMSE results of our subject-specific NN with TL against other subject-specific methodologies from the literature. Method 3, a linear regression model that estimates P_s from the RMS magnitude of the ACC signal, is detailed in [21]. Method 4, which uses a multilinear regression function that combines RMS with additional ACC-based features, is described in [23]. The bar plots in the figure illustrate that our subject-specific NN with TL provides a lower average RMSE in the estimation of P_s. It is important to highlight that, in Fig. 4, the bars indicate the mean of the best estimations for each fold and method to facilitate a direct comparison with the results reported in [24].

The two-way ANOVA performed on RMSE for the estimation of P_s, reported a cohort (i.e., Control, PVH, NPVH and UVFP) factor of $\mathbf{F} = 9.39$ ($p < 0.0001$), and a method (i.e., Method 3, Method 4, and subject-specific NN) factor of $\mathbf{F} = 21.07$ ($p < 0.0001$), which reveals significant differences in the average RMSE associated with each method for the different cohorts. Subsequent post-hoc analyses using Tukey HSD indicate that Method 3 and Method 4 have significantly higher error rates compared to the subject-specific NN, with mean differences of -0.75 ($p = 0.001$) and -0.37 ($p = 0.008$), respectively. Additionally, the effect sizes, measured by Cohen's d, show that the difference between the subject-specific NN and Method 3 is $d = -1.17$, representing a large effect size, while the difference between the subject-specific NN and Method 4 is $d = -0.64$, indicating a medium to large effect size. These statistical analyses support the notion that subject-specific NNs are more accurate and produce fewer errors compared to other methods based on subject-specific calibration.

4. Discussion

The integration of machine learning with a physiologically relevant voice synthesizer offers several advantages. Notably, it facilitates access to clinically hard-to-measure vocal features, such as subglottal pressure, muscle activation, and vocal fold contact pressure [42]. Its training process encompasses thousands of simulations, covering a comprehensive range of sustained vowel phonation. While numerical voice production models can provide a good representation of the phonatory process, the signals derived from these models are approximations to the intricate relationships between human vocal fold physiology and voice production.

However, it is essential to recognize that models trained with synthetic data rely on the assumption that the training domain (synthetic data from the numerical voice production model) and the target domain (laboratory data referenced to P_s) occupy the same feature space and share identical distributions. However, domain shifts are expected due to the cohort (i.e., control or pathological group), but also due to the speaker specificities. Thus, evidence suggests that domain adaptation is crucial to improving model performance.

Table 7
Error metrics for the estimation of P_s using subject-specific models, trained with and without TL, and for cohorts: Control, PVH, NPVH, and UVFP. Results are averaged for every subject-dependent model.

Group	TL	RMSE (cm H_2O)	MAE (cm H_2O)	MAPE (%)	R^2
Control		1.34 ± 0.43	1.04 ± 0.35	14.85 ± 3.97	0.78 ± 0.16
	✓	**1.24 ± 0.39**	**0.97 ± 0.32**	**14.17 ± 3.23**	**0.81 ± 0.14**
PVH		1.97 ± 0.80	1.50 ± 0.58	17.42 ± 4.96	0.71 ± 0.18
	✓	**1.88 ± 0.72**	**1.42 ± 0.50**	**17.13 ± 5.63**	**0.73 ± 0.18**
NPVH		1.99 ± 1.01	1.56 ± 0.86	18.55 ± 5.86	0.63 ± 0.22
	✓	**1.91 ± 0.98**	**1.51 ± 0.84**	**17.91 ± 6.36**	**0.68 ± 0.20**
UVFP		2.14 ± 0.75	1.68 ± 0.59	22.84 ± 14.07	0.48 ± 0.37
	✓	**2.02 ± 0.66**	**1.61 ± 0.52**	**22.42 ± 13.29**	**0.55 ± 0.28**

Our study shows notable advances in the estimation of P_s from NSV recorded using the ACC. We found that using NN initially trained on a synthetic voice production model and subsequently adapted using *in vivo* laboratory data significantly improves the estimation of P_s. This method outperformed existing approaches in both control subjects and those with voice pathologies, marking a substantial improvement over previous work. These findings are particularly promising for developing advanced, noninvasive assessment tools of P_s for clinical and ambulatory applications, which could offer new pathways for the diagnostic and therapeutic strategies of voice disorders.

The effectiveness of the methods that integrate NN with a numerical voice production model correlates with the ability of the model to accurately replicate laboratory data distributions [40]. In our previous work [42], we applied bias corrections to the synthetic P_s and SPL features to reduce the range and distribution discrepancies between clinical and synthetic datasets. This procedure forced $D_S = D_T$ to treat the problem as a traditional machine learning approach. In the current study, we found that domain adaptation using TL is able to address these discrepancies by fine-tuning the final layers of the NN. The results suggest that the relevant knowledge obtained from synthetic data is kept across the first hidden layers, and the refinement of subsequent layers with *in vivo* data allows an optimal domain adaptation. Consequently, we observed a 7% reduction in the RMSE for the estimation of P_s in LD1. This advance in our methodology not only demonstrates the efficacy of a domain adaptation but also more effectively harnesses

the physiological relevance of low-order lumped synthetic models of voice production combined with clinical recordings for vocal function analysis.

On the other hand, the results provide evidence that the subject-specific approach significantly improves the estimate of P_s compared to the general model (see Tables 6 and 7), with an improvement of more than 38% in the RMSE value between all groups.

While simulations using the synthetic voice production model provide a good representation of a wide range of prototypical sustained phonations, the simplicity of the synthetic low-order vocal folds model behind does not allow a representation of the variability of a universal population, failing to mimic the expected inter and intra-subject variability of real data. Under these circumstances, the domain adaptation performed improves the modeling capabilities to create more general models (universal or adapted to the cohort), although capturing the expected intersubject variability (i.e., the universal variability) with them would require a large number of clinical recordings from a very large number of subjects. In this work, this limitation is bypassed by developing subject-specific models adapted to the specific characteristics of each speaker. The improvement obtained with the subject-specific NN regressor (compared to previous techniques based on subject-specific calibration [24]), was greater than 21% in the RMSE values for the four cohorts, as demonstrated in Fig. 4. This highlights the contribution of our proposal to the state-of-the-art.

As expected, the results showed that the proposed domain adaptation (for both general and subject-specific approaches) was more efficient for the control group than for the pathological cohorts. Although the selected voice production model provides a flexible and physiologically relevant method to control both sustained vowels and time-varying glottal gestures, it has limitations in representing the physical mechanisms of the underlying disordered phonation. For instance, it does not encompass the asymmetric oscillatory vibration of the vocal folds seen in NPVH and UVFP, or the overall changes in mass and stiffness due to nodules in PVH groups. In this sense, the present findings could be significantly enhanced by further exploring numerical voice production models that more accurately mimic pathophysiological behavior [74–76], which could facilitate the transfer of knowledge in cases involving subjects with voice disorders.

In our effort to assess physiologically relevant metrics in the ambulatory setting, the versatility of ACC sensors offers the potential to extend this method into ambulatory settings. Subject-specific fine-tuning improves the ability of the NN to estimate P_s by operating only on ACC-based features within short 50 ms windows. This method not only provides increased confidence in the non-invasive estimation of P_s but also opens up the possibility for its application in clinical, laboratory, and ambulatory monitoring of vocal function during natural voice production The long-term goal is to develop algorithms for analyzing NSV-monitored using a smartphone device and to improve the diagnosis, prevention, and treatment of voice disorders through a deeper understanding of their underlying mechanisms.

Although the scope of the present study was to estimate P_s, the use of the aforementioned synthetic voice production model allows access to a set of additional phonatory measurements, such as the collision pressure of the vocal folds, and the activation of the cricothyroid and thyroarytenoid muscles. However, the approach followed is limited to estimating only P_s by the scarcity of clinical recordings that include such measures. In the future, we will explore transductive TL techniques to enable domain adaptation in scenarios where labeled data are abundant in the source domain but are unavailable in the target one [44].

5. Conclusions

The paper introduces a method for estimating subglottal pressure from neck-surface acceleration signals by combining a synthetic model of voice production with a neural network regressor, which is refined through TL using in vivo laboratory data. This approach enables the creation of subject- and group-specific refinements to the original neural network. The results demonstrate significant improvements in the estimates of P_s compared to previously reported techniques, achieving over a 21% reduction in RMSE. Consequently, this method sets a new standard for the estimation of subglottal pressure from neck-surface accelerometer signals.

In general, the findings highlight the effectiveness of subject-specific regression models based on domain adaptation to estimate P_s in individuals with normal and disordered voices. Our results illustrate that combining machine learning methods with numerical voice production models significantly improves the estimation of certain parameters of vocal function. This improvement is particularly evident in the accurate determination of P_s from the ACC, adapted to the subject using a domain adaptation strategy based on TL.

Although there is strong interest in developing generic models to estimate P_s, the scarcity of laboratory recordings that cover a broad population hinders the creation of a robust model capable of accounting for the complex variability between subjects. Therefore, subject-specific NN regressors represent the best alternative to improve the estimation of P_s with a reduced number of recordings. In this context, TL represents a viable solution for this purpose.

As such, our results can be considered the best estimates of P_s from ACC signals reported in the literature. Future efforts will focus on applying this method to gauge P_s during spontaneous speech within the realm of ambulatory monitoring and biofeedback. This will occur as individuals engage in their everyday routines in various settings such as home, work, and social environments.

CRediT authorship contribution statement

Emiro J. Ibarra: Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Julián D. Arias-Londoño:** Writing – review & editing, Supervision, Investigation, Conceptualization. **Juan I. Godino-Llorente:** Writing – review & editing, Visualization, Methodology, Formal analysis. **Daryush D. Mehta:** Writing – review & editing, Supervision, Conceptualization. **Matías Zañartu:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Matías Zañartu has a financial interest in Lanek SPA, a company focused on developing and commercializing biomedical devices and technologies. Dr. Zañartu's interests were reviewed and managed by Universidad Técnica Federico Santa María in accordance with its conflict-of-interest policies. Dr. Daryush Mehta has a financial interest in Inno-Voyce LLC, a company focused on developing and commercializing technologies for the prevention, diagnosis, and treatment of voice-related disorders. Dr. Mehta's interests were reviewed and managed by Massachusetts General Hospital and Mass General Brigham in accordance with their conflict-of-interest policies. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research was supported by the National Institutes of Health (NIH) National Institute on Deafness and Other Communication Disorders, United States grant P50 DC015446, R21 DC015877 and R33 DC011588; by the Universidad Técnica Federico Santa María, Chile grant DPP PIIC N° 020/2021; by ANID, Chile grants BASAL AFB240002, FONDECYT 1230828, and Beca de Doctorado Nacional 21190074. This work was also supported by the Ministry of Economy and Competitiveness of Spain under Grants PID2021-128469OB-I00 and TED2021-131688B-I00, and by Comunidad de Madrid, Spain. Julián D. Arias-Londoño was supported by Universidad Politécnica de Madrid, Spain through a María Zambrano UP2021-035 grant funded by European Union-NextGenerationEU, Spain. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Data availability

Mass General Brigham and Mass General are not allowed to give access to data without the Principal Investigator (PI) for the human studies protocol first submitting a protocol amendment to request permission to share the data with a specific collaborator on a case-by-case basis. This policy is based on very strict rules dealing with the protection of patient data and information. Anyone wishing to request access to the data must first contact Ms. Sarah DeRosa, Program Coordinator for Research and Clinical Speech-Language Pathology, Center for Laryngeal Surgery and Voice Rehabilitation, Massachusetts General Hospital: sederosa@partners.org.

References

[1] K. Verdolini, I.R. Titze, A. Fennell, Dependence of phonatory effort on hydration level, J. Speech Lang. Hear. Res. 37 (5) (1994) 1001–1007, http://dx.doi.org/10.1044/jshr.3705.1001.

[2] J. Sundberg, I. Titze, R. Scherer, Phonatory control in male singing: A study of the effects of subglottal pressure, fundamental frequency, and mode of phonation on the voice source, J. Voice 7 (1) (1993) 15–29, http://dx.doi.org/10.1016/S0892-1997(05)80108-0, The Voice Foundation's 22nd Annual Symposium.

[3] I.R. Titze, On the relation between subglottal pressure and fundamental frequency in phonation, J. Acoust. Soc. Am. 85 (2) (1989) 901–906, http://dx.doi.org/10.1121/1.397562.

[4] S. Björklund, J. Sundberg, Relationship between subglottal pressure and sound pressure level in untrained voices, J. Voice 30 (1) (2016) 15–20, http://dx.doi.org/10.1016/j.jvoice.2015.03.006.

[5] A.L. Rosenthal, S.Y. Lowell, R.H. Colton, Aerodynamic and acoustic features of vocal effort, J. Voice 28 (2) (2014) 144–153, http://dx.doi.org/10.1016/j.jvoice.2013.09.007.

[6] D.M. Hartl, S. Hans, J. Vaissière, M. Riquet, D.F. Brasnu, Objective voice quality analysis before and after onset of unilateral vocal fold paralysis, J. Voice 15 (3) (2001) 351–361, http://dx.doi.org/10.1016/S0892-1997(01)00037-6.

[7] S.M. Zeitels, R.A. Franco, R.E. Hillman, G.W. Bunting, Voice and treatment outcome from phonosurgical management of early glottic cancer, Ann. Otol. Rhinol. Laryngol. 111 (12_suppl) (2002) 3–20, http://dx.doi.org/10.1177/00034894021111S1202.

[8] E.B. Holmberg, P. Doyle, J.S. Perkell, B. Hammarberg, R.E. Hillman, Aerodynamic and acoustic voice measurements of patients with vocal nodules: variation in baseline and changes across voice therapy, J. Voice 17 (3) (2003) 269–282, http://dx.doi.org/10.1067/S0892-1997(03)00076-6.

[9] V.M. Espinoza, M. Zañartu, J.H.V. Stan, D.D. Mehta, R.E. Hillman, Glottal aerodynamic measures in women with phonotraumatic and nonphonotraumatic vocal hyperfunction, J. Speech Lang. Hear. Res. 60 (8) (2017) 2159–2169, http://dx.doi.org/10.1044/2017_JSLHR-S-16-0337.

[10] V.M. Espinoza, D.D. Mehta, J.H.V. Stan, R.E. Hillman, M. Zañartu, Glottal aerodynamics estimated from neck-surface vibration in women with phonotraumatic and nonphonotraumatic vocal hyperfunction, J. Speech Lang. Hear. Res. 63 (9) (2020) 2861–2869, http://dx.doi.org/10.1044/2020_JSLHR-20-00189.

[11] R.L. Plant, A.D. Hillel, Direct measurement of subglottic pressure and laryngeal-resistance in normal subjects and in spasmodic dysphonia, J. Voice 12 (3) (1998) 300–314, http://dx.doi.org/10.1016/S0892-1997(98)80020-9.

[12] J. Sundberg, R. Scherer, M. Hess, F. Müller, S. Granqvist, Subglottal pressure oscillations accompanying phonation, J. Voice 27 (4) (2013) 411–421, http://dx.doi.org/10.1016/j.jvoice.2013.03.006.

[13] B. Cranen, L. Boves, Pressure measurements during speech production using semiconductor miniature pressure transducers: Impact on models for speech production, J. Acoust. Soc. Am. 77 (4) (1985) 1543–1551, http://dx.doi.org/10.1121/1.391997.

[14] D.D. Mehta, J.B. Kobler, S.M. Zeitels, M. Zañartu, E.J. Ibarra, G.A. Alzamendi, R. Manriquez, B.D. Erath, S.D. Peterson, R.H. Petrillo, R.E. Hillman, Direct measurement and modeling of intraglottal, subglottal, and vocal fold collision pressures during phonation in an individual with a hemilaryngectomy, Appl. Sci. 11 (16) (2021) http://dx.doi.org/10.3390/app11167256.

[15] P. Lieberman, Direct Comparison of Subglottal and Esophageal Pressure during Speech, J. Acoust. Soc. Am. 43 (5) (2005) 1157–1164, http://dx.doi.org/10.1121/1.1910950.

[16] 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 Phoniatr. et Logop. 8 (1) (2009) 1–24, http://dx.doi.org/10.1159/000262725.

[17] M. Rothenberg, A new inverse-filtering technique for deriving the glottal air flow waveform during voicing, J. Acoust. Soc. Am. 53 (6) (2005) 1632–1645, http://dx.doi.org/10.1121/1.1913513.

[18] D.D. Mehta, M. Zañartu, S.W. Feng, H.A. Cheyne II, R.E. Hillman, Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform, IEEE Trans. Biomed. Eng. 59 (11) (2012) 3090–3096, http://dx.doi.org/10.1109/TBME.2012.2207896.

[19] A.S. Fryd, J.H.V. Stan, R.E. Hillman, D.D. Mehta, Estimating subglottal pressure from neck-surface acceleration during normal voice production, J. Speech Lang. Hear. Res. 59 (6) (2016) 1335–1345, http://dx.doi.org/10.1044/2016_JSLHR-S-15-0430.

[20] V. McKenna, A. Llico, D. Mehta, J. Perkell, C. Stepp, Magnitude of neck-surface vibration as an estimate of subglottal pressure during modulations of vocal effort and intensity in healthy speakers, J. Speech Lang. Hear. Res. 60 (2017) 1–13, http://dx.doi.org/10.1044/2017_jslhr-s-17-0180.

[21] K.L. Marks, J.Z. Lin, A.B. Fox, L.E. Toles, D.D. Mehta, Impact of nonmodal phonation on estimates of subglottal pressure from neck-surface acceleration in healthy speakers, J. Speech Lang. Hear. Res. 62 (9) (2019) 3339–3358, http://dx.doi.org/10.1044/2019_JSLHR-S-19-0067.

[22] K.L. Marks, J.Z. Lin, J.A. Burns, T.A. Hron, R.E. Hillman, D.D. Mehta, Estimation of subglottal pressure from neck surface vibration in patients with voice disorders, J. Speech Lang. Hear. Res. 63 (7) (2020) 2202–2218, http://dx.doi.org/10.1044/2020_JSLHR-19-00409.

[23] J.Z. Lin, V.M. Espinoza, K.L. Marks, M. Zañartu, D.D. Mehta, Improved subglottal pressure estimation from neck-surface vibration in healthy speakers producing non-modal phonation, IEEE J. Sel. Top. Signal Process. 14 (2) (2020) 449–460, http://dx.doi.org/10.1109/JSTSP.2019.2959267.

[24] J.P. Cortés, J.Z. Lin, K.L. Marks, V.M. Espinoza, E.J. Ibarra, M. Zañartu, R.E. Hillman, D.D. Mehta, Ambulatory monitoring of subglottal pressure estimated from neck-surface vibration in individuals with and without voice disorders, Appl. Sci. 12 (21) (2022) http://dx.doi.org/10.3390/app122110692.

[25] M. Dollinger, U. Hoppe, F. Hettlich, J. Lohscheller, S. Schuberth, U. Eysholdt, Vibration parameter extraction from endoscopic image series of the vocal folds, IEEE Trans. Biomed. Eng. 49 (8) (2002) 773–781, http://dx.doi.org/10.1109/TBME.2002.800755.

[26] M. Döllinger, T. Braunschweig, J. Lohscheller, U. Eysholdt, U. Hoppe, Normal voice production: Computation of driving parameters from endoscopic digital high speed images, Methods Inf. Med. 42 (2003) 271–276, http://dx.doi.org/10.1055/s-0038-1634360.

[27] M. Döllinger, P. Gómez, R.R. Patel, C. Alexiou, C. Bohr, A. Schützenberger, Biomechanical simulation of vocal fold dynamics in adults based on laryngeal high-speed videoendoscopy, PLoS ONE 12 (11) (2017) 1–26, http://dx.doi.org/10.1371/journal.pone.0187486.

[28] P. Gómez, A. Schützenberger, S. Kniesburges, C. Bohr, M. Döllinger, Physical parameter estimation from porcine ex vivo vocal fold dynamics in an inverse problem framework, Biomech. Model. Mechanobiol. 17 (3) (2018) 777–792, http://dx.doi.org/10.1007/s10237-017-0992-5.

[29] R. Schwarz, U. Hoppe, M. Schuster, T. Wurzbacher, U. Eysholdt, J. Lohscheller, Classification of unilateral vocal fold paralysis by endoscopic digital high-speed recordings and inversion of a biomechanical model, IEEE Trans. Bio- Med. Eng. 53 (2006) 1099–1108, http://dx.doi.org/10.1109/TBME.2006.873396.

[30] A.P. Pinheiro, D.E. Stewart, C.D. Maciel, J.C. Pereira, S. Oliveira, Analysis of nonlinear dynamics of vocal folds using high-speed video observation and biomechanical modeling, Digit. Signal Process. 22 (2) (2012) 304–313, http://dx.doi.org/10.1016/j.dsp.2010.11.002.

[31] C. Tao, Y. Zhang, J. Jiang, Extracting physiologically relevant parameters of vocal folds from high-speed video image series, IEEE Trans. Bio- Med. Eng. 54 (2007) 794–801, http://dx.doi.org/10.1109/TBME.2006.889182.

[32] E.J. Ibarra, G.A. Alzamendi, M. Zañartu, Constrained extended Kalman filter for improving Bayesian inference of vocal function from laryngeal high-speed videoendoscopy, in: J. Brieva, P. Guevara, N. Lepore, M.G. Linguraru, L. Rittner, E. Romero Castro M.D. (Eds.), 18th International Symposium on Medical Information Processing and Analysis, Vol. 12567, International Society for Optics and Photonics, SPIE, 2023, 125671E, http://dx.doi.org/10.1117/12.2669812.

[33] G. Alzamendi, R. Manríquez, P. Hadwin, J. Deng, S. Peterson, B. Erath, D. Mehta, R. Hillman, M. Zañartu, Bayesian estimation of vocal function measures using laryngeal high-speed videoendoscopy and glottal airflow estimates: An in vivo case study, J. Acoust. Soc. Am. 147 (5) (2020) EL434–EL439, http://dx.doi.org/10.1121/10.0001276.

[34] C. Drioli, G.L. Foresti, Fitting a biomechanical model of the folds to high-speed video data through Bayesian estimation, Informatics Med. Unlocked 20 (2020) 100373, http://dx.doi.org/10.1016/j.imu.2020.100373.

[35] P.J. Hadwin, B.D. Erath, S.D. Peterson, The influence of flow model selection on finite element model parameter estimation using Bayesian inference, JASA Express Lett. 1 (4) (2021) 045204, http://dx.doi.org/10.1121/10.0004260.

[36] P.J. Hadwin, M. Motie-Shirazi, B.D. Erath, S.D. Peterson, Bayesian inference of vocal fold material properties from glottal area waveforms using a 2D finite element model, Appl. Sci. 9 (13) (2019) http://dx.doi.org/10.3390/app9132735.

[37] P.J. Hadwin, S.D. Peterson, An extended Kalman filter approach to non-stationary Bayesian estimation of reduced-order vocal fold model parameters, J. Acoust. Soc. Am. 141 (4) (2017) 2909–2920, http://dx.doi.org/10.1121/1.4981240.

[38] P.J. Hadwin, G.E. Galindo, K.J. Daun, M. Zañartu, B.D. Erath, E. Cataldo, S.D. Peterson, Non-stationary Bayesian estimation of parameters from a body cover model of the vocal folds, J. Acoust. Soc. Am. 139 (5) (2016) 2683–2696, http://dx.doi.org/10.1121/1.4948755.

[39] M.E. Díaz-Cádiz, S.D. Peterson, G.E. Galindo, V.M. Espinoza, M. Motie-Shirazi, B.D. Erath, M. Zañartu, Estimating vocal fold contact pressure from raw laryngeal high-speed videoendoscopy using a hertz contact model, Appl. Sci. 9 (11) (2019) http://dx.doi.org/10.3390/app9112384.

[40] P. Gómez, A. Schützenberger, M. Semmler, M. Döllinger, Laryngeal pressure estimation with a recurrent neural network, IEEE J. Transl. Eng. Heal. Med. 7 (2019) 1–11, http://dx.doi.org/10.1109/JTEHM.2018.2886021.

[41] Z. Zhang, Estimation of vocal fold physiology from voice acoustics using machine learning, J. Acoust. Soc. Am. 147 (3) (2020) EL264–EL270, http://dx.doi.org/10.1121/10.0000927.

[42] E.J. Ibarra, J.A. Parra, G.A. Alzamendi, J.P. Cortés, V.M. Espinoza, D.D. Mehta, R.E. Hillman, M. Zañartu, 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. 12 (2021) http://dx.doi.org/10.3389/fphys.2021.732244.

[43] G.A. Alzamendi, S.D. Peterson, B.D. Erath, R.E. Hillman, M. Zañartu, Triangular body-cover model of the vocal folds with coordinated activation of the five intrinsic laryngeal muscles, J. Acoust. Soc. Am. 151 (1) (2022) 17–30, http://dx.doi.org/10.1121/10.0009169.

[44] S.J. Pan, Q. Yang, A survey on transfer learning, IEEE Trans. Knowl. Data Eng. 22 (10) (2010) 1345–1359, http://dx.doi.org/10.1109/TKDE.2009.191.

[45] L. Alzubaidi, J. Bai, A. Al-Sabaawi, J. Santamaría, A. Albahri, B. Al-dabbagh, M. Fadhel, M. Manoufali, J. Zhang, A. Al-Timemy, Y. Duan, A. Abdullah, L. Farhan, Y. Lu, A. Gupta, F. Albu, A. Abbosh, Y. Gu, A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications, J. Big Data 10 (2023) http://dx.doi.org/10.1186/s40537-023-00727-2.

[46] A. Ebbehoj, M.Ø. Thunbo, O.E. Andersen, M.V. Glindtvad, A. Hulman, Transfer learning for non-image data in clinical research: A scoping review, PLoS Digit. Heal. 1 (2) (2022) 1–22, http://dx.doi.org/10.1371/journal.pdig.0000014.

[47] Z. Yihua, Z. Xincheng, W. Yuanbo, Z. Xiaojun, X. Yishen, T. Zhi, Pathological voice detection using transfer learning methods, in: 2021 International Conference on Sensing, Measurement & Data Analytics in the Era of Artificial Intelligence, ICSMD, 2021, pp. 1–5, http://dx.doi.org/10.1109/ICSMD53520.2021.9670828.

[48] E.J. Ibarra, J.D. Arias-Londoño, M. Zañartu, J.I. Godino-Llorente, Towards a corpus (and language)-independent screening of Parkinson's disease from voice and speech through domain adaptation, Bioengineering 10 (11) (2023) http://dx.doi.org/10.3390/bioengineering10111316.

[49] J.D. Arias-Londoño, J.A. Gómez-García, Predicting UPDRS scores in Parkinson's disease using voice signals: A deep learning/transfer-learning-based approach, in: Automatic Assessment of Parkinsonian Speech: First Workshop, AAPS 2019, Cambridge, Massachussets, USA, September 20–21, 2019, Revised Selected Papers 1, Springer, 2020, pp. 100–123.

[50] A. Balagopalan, B. Eyre, J. Robin, F. Rudzicz, J. Novikova, Comparing pre-trained and feature-based models for prediction of Alzheimer's disease based on speech, Front. Aging Neurosci. 13 (2021) http://dx.doi.org/10.3389/fnagi.2021.635945.

[51] Z. Zhaoyan, Voice feature selection to improve performance of machine learning models for voice production inversion, J. Voice (2021) http://dx.doi.org/10.1016/j.jvoice.2021.03.004.

[52] D.D. Mehta, J.H. Van Stan, M. Zañartu, M. Ghassemi, J.V. Guttag, V.M. Espinoza, J.P. Cortés, H.A. Cheyne, R.E. Hillman, Using ambulatory voice monitoring to investigate common voice disorders: Research update, Front. Bioeng. Biotechnol. 3 (2015) 155, http://dx.doi.org/10.3389/fbioe.2015.00155.

[53] M. Zañartu, J.C. Ho, D.D. Mehta, R.E. Hillman, G.R. Wodicka, Subglottal impedance-based inverse filtering of voiced sounds using neck surface acceleration, IEEE Trans. Audio Speech Lang. Process. 21 (9) (2013) 1929–1939, http://dx.doi.org/10.1109/TASL.2013.2263138.

[54] A. Morales, J.I. Yuz, J.P. Cortés, J.G. Fontanet, M. Zañartu, Glottal airflow estimation using neck surface acceleration and low-order Kalman smoothing, IEEE/ ACM Trans. Audio Speech Lang. Process. 31 (5) (2023) 2055–2066, http://dx.doi.org/10.1109/TASLP.2023.3277269.

[55] J.P. Cortés, V.M. Espinoza, M. Ghassemi, D.D. Mehta, J.H. Van Stan, R.E. Hillman, J.V. Guttag, M. Zañartu, Ambulatory assessment of phonotraumatic vocal hyperfunction using glottal airflow measures estimated from neck-surface acceleration, PLoS ONE 13 (12) (2018) 1–22, http://dx.doi.org/10.1371/journal.pone.0209017.

[56] J.G. Švec, I.R. Titze, P.S. Popolo, Estimation of sound pressure levels of voiced speech from skin vibration of the neck, J. Acoust. Soc. Am. 117 (3) (2005) 1386–1394, http://dx.doi.org/10.1121/1.1850074.

[57] I.R. Titze, E.J. Hunter, A two-dimensional biomechanical model of vocal fold posturing, J. Acoust. Soc. Am. 121 (4) (2007) 2254–2260, http://dx.doi.org/10.1121/1.2697573.

[58] I.R. Titze, B.H. Story, Rules for controlling low-dimensional vocal fold models with muscle activation, J. Acoust. Soc. Am. 112 (3 Pt 1) (2002) 1064, http://dx.doi.org/10.1121/1.1496080.

[59] G.E. Galindo, S.D. Peterson, B.D. Erath, C. Castro, R.E. Hillman, M. Zañartu, Modeling the pathophysiology of phonotraumatic vocal hyperfunction with a triangular glottal model of the vocal folds, J. Speech Lang. Hear. Res. 60 (9) (2017) 2452–2471, http://dx.doi.org/10.1044/2017_JSLHR-S-16-0412.

[60] M. Zañartu, G.E. Galindo, B.D. Erath, S.D. Peterson, G.R. Wodicka, R.E. Hillman, Modeling the effects of a posterior glottal opening on vocal fold dynamics with implications for vocal hyperfunction, J. Acoust. Soc. Am. 136 (6) (2014) 3262–3271, http://dx.doi.org/10.1121/1.4901714.

[61] M. Zañartu, Influence of Acoustic Loading on the Flow-Induced Oscillations of Single Mass Models of the Human Larynx (Master's thesis), School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, 2006.

[62] B.H. Story, Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002, J. Acoust. Soc. Am. 123 (1) (2008) 327–335, http://dx.doi.org/10.1121/1.2805683.

[63] B.H. Story, I.R. Titze, E.A. Hoffman, Vocal tract area functions for an adult female speaker based on volumetric imaging, J. Acoust. Soc. Am. 104 (1) (1998) 471–487, http://dx.doi.org/10.1121/1.423298.

[64] P. Boersma, D. Weenink, Praat: Doing Phonetics by Computer. URL http://www.praat.org.

[65] H.A. Cheyne, Estimating glottal voicing source characteristics by measuring and modeling the acceleration of the skin on the neck, in: 2006 3rd IEEE/EMBS International Summer School on Medical Devices and Biosensors, 2006, pp. 118–121, http://dx.doi.org/10.1109/ISSMDBS.2006.360113.

[66] J.S. Perkell, E.B. Holmberg, R.E. Hillman, A system for signal processing and data extraction from aerodynamic, acoustic, and electroglottographic signals in the study of voice production, J. Acoust. Soc. Am. 89 (4) (1991) 1777–1781, http://dx.doi.org/10.1121/1.401011.

[67] M. Zañartu, Acoustic Coupling in Phonation and its Effect on Inverse Filtering of Oral Airflow and Neck Surface Acceleration (Ph.D. thesis), School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN, 2010.

[68] J. Kennedy, R.C. Eberhart, Particle swarm optimization, in: Proceedings of the IEEE International Conference on Neural Networks, 1995, pp. 1942–1948.

[69] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks? Adv. Neural Inf. Process. Syst. 27 (2014).

[70] Autonomio, Talos computer software, 2020, Retrieved from http://github.com/autonomio/talos. (Accessed 26 July 2023).

[71] E.J. Ibarra, G.E. Galindo, G.A. Alzamendi, J.P. Cortes, C. Castro, R. Manríquez, A. Testart, M. Zañartu, Empirical distribution of glottal edges (EDGE): A statistical assessment of vocal fold kinematics using high-speed videoendoscopy, IEEE J. Biomed. Heal. Inform. (2024) 1–14, http://dx.doi.org/10.1109/JBHI.2024.3462632.

[72] I.R. Titze, J.G. Švec, P.S. Popolo, Vocal dose measures: Quantifying accumulated vibration exposure in vocal fold tissues, J. Speech Lang. Hear. Res. 46 (4) (2003) 919–932, http://dx.doi.org/10.1044/1092-4388(2003/072).

[73] J.W. Tukey, Comparing Individual Means in the Analysis of Variance, Biometrics, 1949.

[74] J.A. Parra, C. Calvache, G.A. Alzamendi, E.J. Ibarra, L. Soláque, S.D. Peterson, M. Zañartu, Asymmetric triangular body-cover model of the vocal folds with bilateral intrinsic muscle activation, J. Acoust. Soc. Am. 156 (2) (2024) 939–953, http://dx.doi.org/10.1121/10.0028164.

[75] D.D. Mehta, M. Zañartu, T.F. Quatieri, D.D. Deliyski, R.E. Hillman, Investigating acoustic correlates of human vocal fold vibratory phase asymmetry through modeling and laryngeal high-speed videoendoscopya), J. Acoust. Soc. Am. 130 (6) (2011) 3999–4009, http://dx.doi.org/10.1121/1.3658441.

[76] B.D. Erath, M. Zañartu, S.D. Peterson, M.W. Plesniak, Nonlinear vocal fold dynamics resulting from asymmetric fluid loading on a two-mass model of speech, Chaos 21 (3) (2011) 033113, http://dx.doi.org/10.1063/1.3615726.