INTEGRATING MACHINE LEARNING AND PHYSIOLOGICAL MODELING TOOLS FOR THE ASSESSMENT OF VOCAL FUNCTION USING NECK SURFACE ACCELERATION

BY

Emiro Jose Ibarra Sulbaran, M.S.c.

A dissertation submitted

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

in Electronic Engineering

Universidad Técnica Federico Santa María

Valparaíso, Chile

May 2024

"Integrating Machine Learning and Physiological Modeling Tools for the Assessment of Vocal Function Using Neck Surface Acceleration," a dissertation prepared by Emiro Jose Ibarra Sulbaran in partial fulfillment of the requirements for the degree, Doctor of Philosophy, has been approved and accepted by the following:

Matías Zañartu
Thesis Advisor

Mauricio Araya
Chair of the Examining Committee

Date

Committee in charge:

Dr. Matías Zañartu

Dr. Mauricio Araya, Chair

Dr. Alejandro Weinstein

Dr. Daryush Mehta

DEDICATION

I dedicate this work to my wife and daughter, my beautiful Moons. Their unconditional support and boundless love made this journey possible. Thank you for being my constant inspiration and for always believing in me.

ACKNOWLEDGMENTS

I am profoundly grateful to my advisor, Professor Matías Zañartu, whose guidance and vast knowledge in speech, acoustic modeling, and signal processing were pivotal throughout my doctoral studies. His patience, continuous guidance, and support not only made this thesis possible but also transformed my doctoral journey at UTFSM into a truly memorable experience. Thank you, Professor Zañartu, for providing me with the opportunity and confidence to develop this research and for instilling in me the rigor and passion that define your own work. Your enthusiasm and constant encouragement have left an indelible mark on my career and personal growth.

I extend a special thank you to Professor Juan I. Godino Llorente for inviting me to the Bioengineering and Optoelectronics Laboratory (ByO), and to Professor Julian Arias for his contributions. My time at ByO has been immensely fruitful and enriching, made possible by their support and the collaborative atmosphere they promote.

I extend my thanks to the members of my thesis committee: Professors Dr. Mauricio Araya, Dr. Alejandro Weinstein, and Dr. Daryush Mehta. Special thanks go to Dr. Daryush Mehta, whose clinical work and advice were highly beneficial to the goals of this thesis.

I am grateful to my fellow members of the Voice Production Lab at UTFSM—

Dr. Gabriel Alzamendi, Jesus Parra, Josue Martinez, Dr. Juan P. Cortes, and Christian Castro—for their friendship during my time at UTFSM. I am thankful for the fruitful discussions that influenced many aspects of my work, especially Dr. Gabriel Alzamendi, who made a significant impact on this work through his valuable contributions.

I gratefully acknowledge the financial support provided by the scholarships from ANID Beca Doctorado Nacional 21190074, the UTFSM PIIC programs N° 020/2021 and N° 009/2022, the Beca de Término de Tesis, and the National Institutes of Health (NIH) National Institute on Deafness and Other Communication Disorders grant P50 DC015446.

On a personal note, I extend my deepest gratitude to my wife, Maryori, and my daughter, Luna, for their support, patience, and unconditional love, which made this entire journey possible. To my mother, whose hard work laid the foundation of my education—I owe you everything. I would also like to thank the Orellana Coronel family for their unwavering support and guidance throughout my doctoral studies. To my friends in Chile, the Cedeño Alvear family, for the good times shared. Additionally, to my friends in Spain, the Useche Perez family, thank you for your support during my doctoral internship. I am grateful for the wonderful moments we shared together.

ABSTRACT

INTEGRATING MACHINE LEARNING AND PHYSIOLOGICAL MODELING TOOLS FOR THE ASSESSMENT OF VOCAL FUNCTION USING NECK SURFACE ACCELERATION

Emiro Jose Ibarra Sulbaran, MSc.

Doctor of Philosophy

Universidad Técnica Federico Santa María

Valparaíso, Chile, 2024

Dr. Matías Zañartu Salas, Chair

This thesis is dedicated to advancing the ambulatory assessment of vocal function by utilizing a neck-surface accelerometer attached directly to the skin surface of the neck. The motivation lies in the fact that a fully developed ambulatory method, capable of precisely identifying the underlying pathophysiological characteristics of both normal and pathological vocal functions, could revolutionize clinical practices in monitoring, evaluating, and treating common voice disorders. Accordingly, this work exploits the advantages of a low-order voice production model to introduce a non-invasive technique for estimating relevant vocal function metrics, such as subglottal pressure, vocal fold collision pressure, and intrinsic laryngeal muscle activation of the cricothyroid and thyroarytenoid muscles, based on signals from an accelerometer sensor. In the first stage, a Bayesian framework based on a constrained extended Kalman filter is proposed to link a low-order voice production model with either a glottal area waveform extracted from high-speed video recordings or glottal airflow estimated from Rothenberg mask measurements. The results provide new insights into the capacity of the

selected voice production model to replicate different phonation conditions and highlight the feasibility of using this method to estimate clinical measures that are difficult to ascertain in a clinical setting. The second stage of the thesis focuses on an alternate solution: a neural network trained exclusively with simulations from a voice production model. This nonlinear regressor maps seven input features, which can be extracted from an accelerometer signal, to the target measures of vocal function. The efficacy of this method, particularly in terms of subglottal pressure, was validated through *in vivo* recordings, which included synchronous measurements of oral volume velocity, intraoral pressure, microphone, and accelerometer. This method was applied to healthy and disordered voices (unilateral vocal fold paralysis and both phonotraumatic and nonphonotraumatic vocal hyperfunction). Participants were prompted to articulate /p/-vowel syllable strings, varying loudness, vowels, pitch, and voice quality. The neural network, trained with synthetic data, demonstrated subglottal pressure estimation comparable to that of previous studies for subjects without voice disorders. However, this nonlinear mapping was found to be less robust in cases of pathology. In the search for more accurate subject-specific models, the final research stage focuses on refining the neural network regressor, initially trained solely with simulations from a synthetic voice production model. This refinement is carried out by employing a domain adaptation strategy from synthetic to *in vivo* laboratory data, resulting in an improved estimate of subglottal pressure. This method yielded a set of subject-specific models that provided the most accurate estimation of subglottal pressure to date for both normal and disordered voices using an accelerometer. Additionally, through a case study—which, alongside the previously mentioned *in vivo* synchronous measurements, also incorporates fine-wire laryngeal electromyography—it is demonstrated that the performance of the subject-specific regressor in estimating subglottal pressure is maintained while concurrently estimating muscle activation of the cricothyroid and thyroarytenoid muscles. Overall, this thesis advances the field of vocal function assessment through a series of significant contributions. The proposed Bayesian framework reduces the need for multiple observations while yielding robust and reliable estimates of features that are difficult to measure in clinical practice. It also innovatively combines machine learning techniques with the voice production model to estimate physiologically relevant features such as subglottal pressure, vocal fold collision pressure, and laryngeal muscle activation from neck-surface accelerometers. Furthermore, this work introduces a subject-specific nonlinear regression enhanced by transfer learning, significantly improving the estimation of subglottal pressure from neck-surface vibration signals, with promising potential for application to other vocal function parameters.

Contents

List of Tables

List of Figures

ABBREVIATIONS

ACC	Neck-skin **ACC**elerometer (sensor or signal)
ACFL	**A**lternating **C**urrent **Fl**ow or AC-flow
ANOVA	**An**alysis **o**f **V**ariance
BCM	**B**ody-**C**over **M**odel
CEKF	**C**onstrained **E**xtended **K**alman **F**ilter
CT	**C**rico**t**hyroid muscle
EGG	**E**lectro**g**lotto**g**raphy
EKF	**E**xtended **K**alman **F**ilter
EMG	**E**lectro**m**yo**g**raphy
GAW	**G**lottal **A**rea **W**aveform
GVV	**G**lottal **V**olume **V**elocity
$H_1 - H_2$	**H**armonic **1** to **H**armonic **2** ratio
HPF	**H**igh **P**ass **F**ilter
HSV	**H**igh-**S**peed **V**ideoendoscopy
HSD	**H**onestly **S**ignificant **D**ifference
IA	**I**nterarytenoid
IBIF	**I**mpedance-**B**ased **I**nverse **F**iltering
IOP	**I**ntra-**O**ral **P**ressure
LCA	**L**ateral **C**rico**a**rytenoid muscle

LIG	**Lig**ament
LPF	**L**ow **P**ass **F**ilter
LSTM	Long Short-Term Memory
MAE	**M**ean **A**bsolute **E**rror
MAP	**Ma**ximum **P**osterior
MAPE	**M**ean **A**bsolute **P**ercentage **E**rror
MIC	**Mic**rophone
MFDR	**M**aximum **F**low **D**eclination **R**ate
MSE	**M**ean **S**quared **E**rror
MUC	**Muc**osa
NN	**N**eural **N**etwork
NPVH	**N**on-**P**honotraumatic **V**ocal **H**yperfunction
OQ	**O**pen **Q**uotient
OVV	**O**ral **V**olume **V**elocity
PCA	**P**osterior **C**rico**a**rytenoid muscle
PGO	**P**osterior **G**lottal **O**pening
PVH	**P**honotraumatic **V**ocal **H**yperfunction
RMS	**R**oot **M**ean **S**quared
RMSE	**R**oot **M**ean **S**quared **E**rror
SD	**S**tandard **D**eviation

SPL	**S**ound **P**ressure **L**evel
SQ	**S**peed **Q**uotient
TA	**T**hyro**a**rytenoid muscle
TBCM	**T**riangular **B**ody-**C**over **M**odel
TL	**T**ransfer **L**earning
UVFP	**U**nilateral **V**ocal **F**old **P**aralysis
VF	**V**ocal **F**old
VH	**V**ocal **H**yperfunction

Chapter 1
Introduction

1.1 Motivation

Laryngeal voice disorders significantly affect various aspects of life, including communication, financial stability, social interactions, work-related functions, and psychological well-being [1]. These disorders are prevalent globally and pose a significant occupational health concern. They affect approximately 7.7% of the adult population annually [2]. In the United States, roughly 30% of adults have experienced these disorders at some point in their lives [3]. Specifically, in Chile, approximately 75% of school teachers encounter voice-related problems, ranking these disorders as the second most common occupational health issue among the working-age population [4]. This high prevalence underscores the urgent need for effective management and intervention strategies for voice disorders, especially in professions that place high demands on vocal use.

Several prevalent vocal pathologies are usually preceded by detrimental patterns of daily behavior and abuse of voice, known as vocal hyperfunction (VH)

[5]. VH manifests in two forms: Phonotraumatic VH (PVH), associated with the formation of benign vocal fold lesions, and non-phonotraumatic VH (NPVH), which causes dysphonia and vocal fatigue without vocal fold (VF) tissue trauma or other conditions affecting phonation [6]. While the general concept of VH is widely accepted and applied in clinical practice, the underlying etiological and pathophysiological mechanisms remain unclear. This lack of clarity hinders its effective prevention, diagnosis, and treatment [5, 7].

Substantial evidence indicates that key clinical assessment measures, including subglottal pressure (P_s), vocal fold collision pressure (P_c), and laryngeal muscle activation, are essential for comprehensively understanding the characteristics and behaviors of various voice pathologies. For instance, Espinoza et al. demonstrated that measurements of P_s can effectively distinguish patients with vocal hyperfunction from vocally typical control speakers [8]. Similarly, research by Zeitels et al. revealed that significant changes in P_s serve as indicators of post-surgical outcomes in patients with unilateral vocal fold paralysis (UVFP) [9] and laryngeal cancer [10]. Chhetri et al. found that in laryngeal dystonia, intrinsic laryngeal muscles exhibit hyperfunction, while in conditions like paresis and paralysis, they are hypofunctional [11]. Furthermore, Hillman et al. concluded that trauma-induced vocal fold lesions in PVH, such as nodules and polyps, often result from compensatory behaviors leading to increased vocal fold collision pressures [5]. They also noted that NPVH involves excessive activity of laryngeal muscles [7]. This com-

pilation of research underscores the multifaceted nature of voice pathologies and highlights the pivotal role of clinical vocal function assessments in their diagnosis and management.

Measuring vocal function features often requires cumbersome and invasive procedures, limiting their use in clinical settings. For example, subglottal pressure can be measured using direct methods such as tracheal puncturing [12, 13, 14] or by inserting miniature pressure transducers transorally (through the mouth) to reach the vocal tract [15, 16, 17, 18], as well as indirect methods involving esophageal balloons [19, 20]. Assessing vocal fold collision pressure necessitates a miniature sensor at the glottis, which presents significant challenges including the size of the probe, bandwidth, potential risks to vocal fold tissue, and patient tolerance during the examination. Only a few studies have successfully gathered contact pressure data in human subjects using this method [18, 21]. Additionally, measuring laryngeal muscle activity requires intramuscular electromyography, recorded using needles or hooked-wire electrodes [22, 23]. The infrequent use of these techniques in clinical settings is generally due to their invasive nature and the need for expensive and specialized equipment.

Ambulatory voice monitoring with a neck-surface accelerometer (ACC) enables the assessment of daily vocal function and has demonstrated potential in modifying vocal behaviors through ambulatory biofeedback [24, 25, 26, 27, 28]. Numerous features can be extracted from the recordings of the ACC signal, includ-

ing phonation duration, sound pressure level (SPL) [29], fundamental frequency (f_o) [30], vocal vibration-dose measures [31, 32], spectral and cepstral measures [6, 33], subglottal pressure [34, 35, 36, 37, 38], and aerodynamic measures [39, 40]. These measures have been utilized to differentiate daily voice use in patients with vocal hyperfunction from matched controls [30, 40, 41] and to monitor changes related to surgical and voice therapy treatments for hyperfunctional voice disorders [42, 43]. The current classification accuracy using these parameters ranges from 0.7 to 0.85.

On the other hand, recent lines of research have taken advantage of the physiological relevance of the numerical voice production model to propose non-invasive methods for estimating relevant measures that are difficult to obtain in clinical settings. Prominent among these approaches are optimization-based voice inversion methods [44, 45, 46, 47, 48, 49, 50], multi-parameter estimation frameworks based on Bayesian estimation [51, 52, 53, 54, 55, 56, 57], and machine learning-based methods [58, 59]. These prior studies have introduced schemes for incorporating numerical voice production models into clinical practice to develop non-invasive methods for estimating vocal function. However, these methodologies have yet to be adapted to estimate vocal function from ambulatory sensors like accelerometers.

It is argued that estimating relevant vocal function measures from long-term recordings can effectively capture underlying phenomena associated with both

healthy and pathological voices [7, 8]. The neck-surface accelerometer sensor is particularly valuable due to its non-invasive nature and the ability to be comfortably worn by speakers in various settings, including laboratories, clinical environments, and ambulatory situations. Additionally, methods based on numerical models present attractive alternatives, being suitable for representing a broad spectrum of phonation conditions and facilitating access to measures that are difficult to obtain experimentally. On one hand, Bayesian inference from the low-order voice production model method has demonstrated promising results in deriving vocal function parameters using *in vivo* measures from a case study [51]. On the other hand, machine learning techniques trained with a synthetic numerical model offer significant advantages for voice assessment features by providing accurate predictions and efficient implementations [59]. In this thesis, the benefits of numerical modeling of voice production are leveraged in two distinct contexts: within a Bayesian inference framework and through the use of machine learning tools. The aim is to predict relevant clinical parameters, such as subglottal pressure, vocal fold collision pressure, and intrinsic laryngeal muscle activation. The overarching goal of this work is to propose a non-invasive method for estimating vocal function in both normophonic and disordered voices, which could be applicable in various settings, from clinical to ambulatory.

1.2 Goals

1.2.1 General Aim

To develop frameworks based on a low-order voice production model to estimate advanced vocal function features from neck-surface vibrations in both normal and disordered voices.

1.2.2 Specific Aims

1. To investigate the capabilities of a low-order voice production model for mimicking the behavior of vocal function as observed in laboratory measurements in a Bayesian framework.

2. To evaluate the inverse mapping between accelerometer-based features and vocal function measures, such as subglottal pressure, vocal fold collision pressure, and intrinsic laryngeal muscle activation using a neural network.

3. To determine if domain adaptation methods that combine numerical models and clinical data can improve the performance of neural network mappings for assessing vocal function using an accelerometer sensor.

1.3 Hypotheses

- **Hypothesis 1**: If constraints derived from prior physiological knowledge of the phonation process are incorporated into the Bayesian framework of a low-order voice production model, then this model will estimate vocal function using only single observation measurements with accuracy comparable to that of recent Bayesian inference models, which required two observation measurements. This approach will enable the estimation of vocal functions that are difficult to obtain in clinical scenarios using laboratory measurements, without the need for simultaneous or multi-sensor recordings.

- **Hypothesis 2:** If a nonlinear regressor that maps features extracted from accelerometer signals to vocal function parameters (such as subglottal pressure, vocal fold collision pressure, and intrinsic laryngeal muscle activation) is trained using data from a numerical low-order voice production model, it will estimate subglottal pressure from accelerometer data with a root mean squared error (RMSE) that is lower than that obtained by a linear regression model in normal and pathological voices. The accuracy will be estimated using reference values of subglottal pressure, obtained from intraoral pressure waveforms during /p/-vowel syllables [34, 35]. This evaluation will assess the performance of the nonlinear regressor in comparison to reported

methods in the literature, aiming to confirm the proposed method is feasible as a non-invasive approach for assessing vocal function in both clinical and ambulatory settings.

- **Hypothesis 3:** If a nonlinear regression model, initially trained on simulated data from numerical voice production models to estimate vocal function parameters, is fine-tuned with individual laboratory recordings, it will significantly enhance the accuracy of estimating vocal function parameters from neck-surface vibration recordings. The efficiency of this approach will be quantified by comparing the regressor estimations of subglottal pressure with reference values obtained from intraoral pressure waveforms during /p/-vowel syllables, using RMSE for comparison. Mean squared error (MSE) will be analyzed across training and validation epochs to check for overfitting. This subject-specific tuning is expected to offer superior accuracy in subglottal pressure estimation compared to all current methods reported in the literature. Additionally, the validation of this subject-specific method will extend to estimations of two intrinsic laryngeal muscle activations through *in vivo* laboratory recordings, which include measurements of laryngeal electromyography (EMG) from a man.

1.4 Overview of the proposed methods

The general aim of this research hinges on the ability of the selected voice production model to replicate behaviors observed in laboratory recordings. A method that efficiently used a low-order model to estimate subglottal pressure, vocal fold collision pressure, and laryngeal muscle activation was the Bayesian framework with an extended Kalman filter (EKF) [51]. This framework successfully inferred vocal function measures, along with their corresponding confidence intervals, in an *in vivo* case involving simultaneous high-speed videoendoscopy (HSV) and oral volume velocity (OVV) recordings. However, the requirement for multiple simultaneous recordings limits their practical applicability. Consequently, the initial goal of this study is to develop a Bayesian framework capable of estimating vocal function using either HSV or OVV as the sole observational input. In this context, a constrained Bayesian scheme is proposed that effectively integrates physiological knowledge about subglottal pressure and intrinsic laryngeal muscle activation within the speech range. This new proposal was applied to laboratory datasets without simultaneous recordings of phonation in subjects with no voice disorders. This novel approach facilitated a qualitative study of the lumped-element model in contexts of normal voices, encompassing variations in pitch and loudness.

The method for achieving an optimal nonlinear mapping between ACC-based features and vocal function parameters, such as subglottal pressure, vocal fold

collision pressure, and laryngeal muscle activation, involves training a neural network (NN) with a thousand simulations from a low-order voice production model proposed in [60]. The ACC-based features include the amplitude of the unsteady glottal airflow (ACFL), maximum flow declination rate (MFDR), open quotient (OQ), speed quotient (SQ), spectral tilt measured as the log-magnitude difference between the first and second harmonics ($H_1 - H_2$), f_o, and SPL. The first six features are computed from unsteady glottal airflow signal derived from the ACC through the impedance-based inverse filtering (IBIF) proposed in [61, 40]. The SPL can be directly computed from the root mean square (RMS) of the microphone envelope signal, calibrated in pascals, or from the RMS magnitude of the ACC signal, as detailed in [29]. Predictions generated by this scheme were validated against numerical simulations. The estimates of subglottal pressure were then compared with reference measurements of mean subglottal pressure derived from the intra-oral pressure (IOP) sensor using the standard airflow interruption technique in the laboratory, both in control and pathological cases.

To efficiently transfer network parameters learned from voice model simulations, domain adaptation through transfer learning (TL) is proposed to establish a more robust nonlinear mapping between accelerometer-based features and subglottal pressure. For this purpose, the NN regression architecture that demonstrated the best performance in estimating P_s in the synthetic dataset was selected. TL is then employed to fine-tune the NN weights, using cross-validation based on

laboratory recordings. TL was applied in two scenarios: initially, by training a single model to estimate subglottal pressure across multiple subjects, and subsequently, by developing a subject-specific neural network. In this latter approach, TL adjusts the model parameters based on recordings from individual subjects. Subsequently, the domain adaptation method was employed to adapt a subject-specific NN to an *in vivo* laboratory case that includes measures of IOP and fine-wire EMG of the cricothyroid (CT) and thyroarytenoid (TA) muscles, allowing for preliminary validation of the method in terms of P_s and the two muscle activations.

1.5 Scientific contributions

In general, this work represents the first attempt to comprehensively adapt a physiology-based low-order voice production model to diverse estimation methods, aiming to advance clinical and ambulatory assessment of vocal function across a population of normal and pathological subjects.

Firstly, the proposed Bayesian framework reduces the necessity for numerous observations while yielding robust and reliable estimates of vocal function measures. This advancement enables the correlation of the low-order voice production model with the laboratory recordings of a specific subject, even in the absence of multi-sensor or simultaneous measurements.

Then, the innovative approach integrating machine learning tools with a voice production model provides, for the first time, the ability to access physiologically relevant model-based features, such as subglottal pressure, vocal fold collision pressure, and laryngeal muscle activation, directly from a neck-surface accelerometer signal.

Finally, this work demonstrates that incorporating transfer learning into this combined framework enhances the robustness of the nonlinear mapping for the *in vivo* assessment of vocal function. The subject-specific NN approach proposed in this work significantly improves the estimation of subglottal pressure from neck-surface accelerometer signals, demonstrating effectiveness in both control and pathological cases. This method represents an advancement in P_s estimation over previous techniques reported in the literature, thereby constituting a direct contribution to the state-of-the-art. Furthermore, preliminary results are presented that demonstrate the feasibility of extending estimations from the subject-specific NN to additional vocal functions, such as intrinsic laryngeal muscle activations.

The ability to estimate subglottal pressure using only accelerometer-based features, within short 50-ms time windows, highlights the potential of this approach for application in laboratory, clinical, and ambulatory settings for monitoring vocal function.

1.6 Publications

This thesis is supported by a set of publications in which the candidate is a joint author. Related publications made during the research period are also included.

1.6.1 Journals Papers

1. **Emiro J. Ibarra**, Julián D. Arias-Londoño, Matías Zañartu, and Juan I. Godino-Llorente. (2023). *"Towards a Corpus (and Language)-Independent Screening of Parkinson's Disease from Voice and Speech through Domain Adaptation"*. Bioengineering 10, no. 11: 1316. DOI: 10.3390/bioengineering10111316.

2. Juan P. Cortés, Jon Z. Lin, Katherine L. Marks, Víctor M. Espinoza, **Emiro J. Ibarra**, Matías Zañartu, Robert E. Hillman, and Daryush D. Mehta. (2022). *"Ambulatory Monitoring of Subglottal Pressure Estimated from Neck-Surface Vibration in Individuals with and without Voice Disorders"*. Applied Sciences 12, no. 21: 10692. DOI: 10.3390/app122110692.

3. **Emiro J. Ibarra**, Jesus Parra, Gabriel A. Alzamendi, Juan P. Cortés, Victor M. Espinoza, Daryush D. Mehta, Robert E. Hillman and Matías Zañartu. (2021). *"Estimation of subglottal pressure, vocal fold collision pres-*

sure, and intrinsic laryngeal muscle activation from neck-surface vibration using a neural network framework and a voice production model". Frontiers in Physiology, section Computational Physiology and Medicine, Vol 12, pp. 1419. DOI: 10.3389/fphys.2021.732244.

4. Daryush D. Mehta, James B. Kobler, Steven M. Zeitels, Matías Zañartu, **Emiro J. Ibarra**, Gabriel A. Alzamendi, Rodrigo Manriquez, Byron D. Erath, Sean D. Peterson, Robert H. Petrillo, and Robert E. Hillman. (2021). *"Direct Measurement and Modeling of Intraglottal, Subglottal, and Vocal Fold Collision Pressures during Phonation in an Individual with a Hemilaryngectomy"*. Applied Sciences 11, no. 16: 7256. DOI:10.3390/app11167256.

1.6.2 Journals papers in review

1. **Emiro J. Ibarra**, Julián D. Arias, Juan I. Godino, Daryush D. Mehta, and Matías Zañartu. *"Subject-specific modelling of the Subglottal Pressure Estimation From Neck-Surface Vibration Signals by Domain Adaptation"*. Submitted for publication.

2. **Emiro J. Ibarra**, Gabriel E. Galindo, Gabriel A. Alzamendi, Juan P. Cortes, Christian Castro, Rodrigo Manríquez, Alba Testart, and Matías Zañartu, *"Empirical distribution of glottal edges (EDGE): A statistical assessment of vocal fold kinematics using high-speed videoendoscopy,"*. Sub-

mitted for publication.

1.6.3 Conferences

1. **Emiro J. Ibarra**, Julián D. Arias, Juan I. Godino, Daryush D. Mehta, and Matías Zañartu. *"Improved subglottal pressure estimation from neck-surface vibrations using transfer learning of deep neural networks trained from voice production model"*. 13th International Conference on Voice Physiology and Biomechanics, July 2024. Erlangen, Germany.

2. **Emiro J. Ibarra**, Julián D. Arias-Londoño, Matías Zañartu, and Juan I. Godino-Llorente. *"Domain adversarial convolutional neural network for Parkinson's disease detection from speech"*. In Models and Analysis of Vocal Emissions for Biomedical Applications: 13th International Workshop, September, 12-13, 2023 (p. 69). Firenze, Italy.

3. **Emiro J. Ibarra**, Gabriel A. Alzamendi, and Matías Zañartu. *"Estimating the biomechanics of vocal function from glottal airflow measures using constrained extended Kalman filter and body cover model of the vocal folds"*. 52nd Annual Symposium: Care of the Professional Voice, May 31 – June 4, 2023. Philadelphia, US.

4. **Emiro J. Ibarra**, Gabriel A. Alzamendi, and Matías Zañartu, *"Constrained extended Kalman filter for improving Bayesian inference of vocal function*

from laryngeal high-speed videoendoscopy". Proceedings of 18th International Symposium on Medical Information Processing and Analysis, November, 9-11, 2022. Valparaiso, Chile. DOI: 10.1117/12.2669812.

5. **Emiro J. Ibarra**, Gabriel A. Alzamendi, and Matías Zañartu, *"Inferencia Bayesiana de la función vocal a partir del flujo de aire glotal"*. 78° Congreso Chileno de Otorrinolaringología, November, 9-11, 2022. Viña del Mar, Chile.

6. Jesús A. Parra, **Emiro J. Ibarra**, Carlos Calvache, and Zañartu Matías *"Estudio de la activación muscular desbalanceada mediante el uso de modelos de pliegues vocales"*. 78° Congreso Chileno de Otorrinolaringología, November, 9-11, 2022. Viña del Mar, Chile.

7. **Emiro J. Ibarra**, Gabriel A. Alzamendi, and Matías Zañartu, *"El uso de la Inteligencia Artificial en la estimación clínica y ambulatoria de la función vocal"*. I Congreso de la Sociedad Chilena de Fonoaudiología and IV Congreso de Fonoaudiólogos Investigadores, October, 21-22, 2022. Valparaiso, Chile.

8. **Emiro J. Ibarra**, Jesus A. Parra, Gabriel A. Alzamendi, Juan P. Cortés, Victor M. Espinoza, Matías Zañartu *"Método basado en inteligencia artificial para la estimación de la función vocal usando un sensor de aceleración"*. 77° Congreso Chileno de Otorrinolaringología, ORL 2021, November, 9-12, 2021. Online.

9. **Emiro J. Ibarra**, Jesus A. Parra, Gabriel A. Alzamendi, Juan P. Cortés, Victor M. Espinoza and Matías Zañartu *"A Machine Learning Framework for Estimating Subglottal Pressure during Running Speech from Glottal Airflow measures"*. 14th International Conference on Advances in Quantitative Laryngology, Voice and Speech Research, AQL 2021, June, 7-10, 2021. Online.

10. Jesus A. Parra , **Emiro J. Ibarra**, Gabriel A. Alzamendi, Juan P. Cortés and Matías Zañartu. *"Discovering Underlying Physical Parameters From Daily Phonotrauma Index Distributions using Monte Carlo Simulations of a Low-Dimensional Voice Production Model"*. 14th International Conference on Advances in Quantitative Laryngology, Voice and Speech Research, AQL 2021, June, 7-10, 2021. Online.

Chapter 2
Background

This research is centered on utilizing a numerical low-order voice production model to estimate vocal function measures, which are typically challenging to assess in clinical settings. In this context, this chapter provides a detailed background of recent inverse methods that have leveraged the physiological relevance of numerical voice production models for voice assessment purposes. This overview aims to justify the selection of the voice production model and the methodology employed in this research. Additionally, this chapter describes the theoretical foundations underpinning the chosen voice production model.

2.1 Inverse methods for vocal function estimation

Inverse modeling methods use actual results from measurements of observable parameters to infer the actual input values of the model parameters [62]. Previous studies in voice research have employed this technique to propose non-invasive al-

ternatives for estimating relevant clinical features. These features include, among others, subglottal pressure, vocal fold collision pressure, and intrinsic laryngeal muscle activation. Some inverse methodologies are based on lumped or finite element models of voice production and include optimization-based voice inversion [44, 45, 46, 48, 50, 49, 47], Bayesian estimation [51, 56, 55, 54, 63, 52, 57], and machine learning tools [58, 59, 64]. In contrast, other successful methodologies are directly based on the correlation between measurements and target parameters [65, 34, 35, 36, 37, 38]. Detailed descriptions of each of these methods follow.

2.1.1 The optimization-based voice inversion methods

The optimization-based voice inversion method estimates vocal function properties by adjusting the parameters of a voice production model, aiming to minimize the difference between experimental and simulated signals, as exemplified in Figure 2.1.

Dollinger et al. [44, 45] implemented the first fully automatic optimization method. They utilized the Nelder–Mead algorithm to optimize the parameters of a two-mass spring-coupled model, replicating human vocal fold dynamics recorded by HSV during sustained phonation. Their method involved varying three parameters (vibrating mass, stiffness, and subglottal pressure) to minimize discrepancies in the periodic oscillatory components between trajectories of the model and those extracted from HSV recordings.

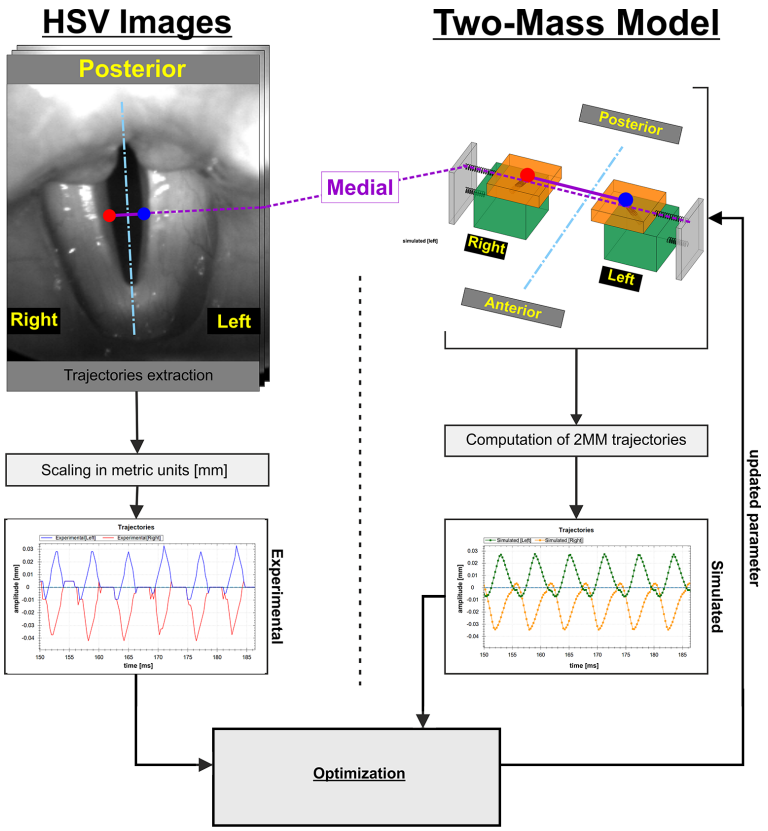


Figure 2.1: General scheme of optimization-based voice inversion methods, adapted from [46].

Afterward, Dollinger et al. [46] expanded their optimization-based voice inversion method to analyze age and gender-related differences, utilizing vocal fold dynamics recorded with high-speed endoscopic imaging. They adapted the numerical two-mass model to match the recorded vocal fold dynamics, employing three different optimization algorithms: Nelder–Mead, Particle Swarm Optimization, and Simulated Bee Colony, in combination with three cost functions. This

optimized parameterization effectively quantified biomechanical differences in dynamic symmetry, subglottal pressure, vocal fold masses, and stiffness across gender (male-female) and ages.

The optimization process was also successfully employed using a genetic algorithm. For example, Schwarz et al. [48] utilized a genetic algorithm to adjust the parameters of a biomechanical model of the vocal folds, aiming to replicate the trajectories in patients with unilateral vocal fold paralysis. Similarly, Tao et al. [50] employed this inversion procedure to extract mass, spring, and damper constants from high-speed videos. Then, Pinheiro et al. [49] combined genetic algorithms with a quasi-Newton method to optimize these same parameters from the glottal area time series extracted from HSV, successfully reproducing vocal fold dynamics in three subjects without voice disorders.

Recently, Gomez et al. [47] implemented a lumped-mass model in an inverse problem to match experimental trajectories of *ex vivo* porcine larynges. They tested three optimization algorithms: Nelder–Mead, particle swarm optimization, and genetic algorithms. The results showed that the model successfully replicated the porcine vocal fold dynamics. Additionally, they compared the subglottal pressure estimated from the optimization process with experimental measurements, observing an average absolute error of 2.90 cm H_2O.

These studies have shown that low-order voice production models can mimic the oscillatory dynamics of vocal folds. They represent initial efforts using nu-

merical VF models to estimate relevant laryngeal parameters such as subglottal pressure. However, validating these methods in clinical or laboratory settings has been limited due to the lack of simultaneous HSV and subglottal pressure measurements. Another major limitation is the necessity for extensive simulations in the voice inversion process to find the optimal solution. Given these challenges, these methods fall outside the aims of this research.

2.1.2 Bayesian inference from voice production model

The Bayesian frameworks are statistical inference approaches wherein all parameters and measurements are treated as random variables. The principal objective of these frameworks is to examine how uncertainty propagates from the observations to the model parameters. Bayesian estimation theory is used to estimate the probability density function of a random signal, facilitating statistical inference [66]. This involves assuming an unknown random variable X that produces noisy data Y, with the aim of determining $P(X|Y)$, the probability of X given Y. This process is formalized as:

$$\hat{X} = \arg\max_{X} P(X|Y) \tag{2.1}$$

where $\hat{X}$ represents the estimated value of X. Bayes' theorem elucidates this further:

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)} \tag{2.2}$$

In Bayesian terms, $P(X)$ is known as the *prior* information, $P(X|Y)$ as the *posterior*, $P(Y|X)$ as the *likelihood*, and $P(Y)$ as the *evidence*.

Cataldo et al. [67] introduced Bayesian inference to the voice production model. They utilized a particle filter methodology to reconstruct the posterior distributions of reduced-order model parameters from synthetic observation data. Subsequently, in [68], they presented two experimental validations using data from one subject with a normal voice and another with vocal fold pathology. They compared the probability density function of fundamental frequency between model simulations and the experimental data, achieving high similarity. Following this, Hadwin et al. [56] expanded the particle filter method to estimate non-stationary model parameters using synthetic data. This work marked the first implementation of particle filters for estimating time-varying parameters in a low-order voice production model. They showed that this non-stationary technique improves model-based estimations and reduces uncertainty, even when estimating time-invariant parameters. However, this approach came with an increased computational effort.

In a recent study, Drioli et al. [52] employed a particle filtering scheme to estimate the parameters and states of a biomechanical model using high-speed video endoscopic data. They linked the displacement of vocal fold edges in a low-

dimensional glottal model with asymmetry in the left/right plane to information extracted from an HSV-based digital kymogram. To reduce the computational limitations of the method, they updated the model parameters at each glottal cycle and computed the state of the model only at the sampling rate. This approach achieved higher computational efficiency without significantly compromising the effectiveness of parameter estimation. The method was validated with recordings featuring different types of phonations. The results demonstrated that the Bayesian inference of vocal fold motion closely matched the video recordings.

Hadwin et al. [55] implemented another Bayesian technique, proposing the EKF-based approach for estimating non-stationary vocal fold parameters. They inferred the evolution of the cricothyroid muscle activation parameter over time from simulated noisy glottal area measurements, demonstrating computational efficiency while maintaining accuracy compared to particle filters. However, this initial study was limited to using only synthetic signals. Consequently, Deng et al. [63] investigated the effects of HSV imaging variables — frame rate, spatial resolution, and viewing angle — on the uncertainty of the estimates. They recorded the trajectories of a physical VF model controlled by the symmetric body-cover model. From these videos, they estimated subglottal pressure and cricothyroid activation using glottal area waveforms. Their findings revealed that an offset in viewing angles leads to biased parameter estimates due to underestimation of glottal width. Furthermore, the frame rate and spatial resolution directly affected

the uncertainty of the parameter estimation.

On the other hand, Hadwin et al. [54] applied an EKF-based methodology to a two-dimensional finite element model of the VF, using HSV recordings of self-oscillating silicone VFs as observation data. They demonstrated that Bayesian estimation from finite element modeling is an effective tool for estimating the material properties of vocal folds from high-speed video recordings. More recently, they examined the impact of flow model selection on the accuracy and uncertainty of parameter estimates in their finite element model [53]. Although these studies showed the feasibility of integrating the finite element model of the VF into the Bayesian framework, they are constrained by high computational costs.

Recently, the EKF-based method was further developed by Alzamendi et al. [51], who incorporated the symmetric three-mass body-cover model with posterior glottal opening and a three-way interaction among tissue, flow, and sound. In this method, the model state vector to be inferred included the positions and velocities of the upper, lower, and body masses, posterior glottal opening, normalized activation levels for the CT and TA muscles, subglottal pressure, and the vocal fold contact pressure. The observation vector consisted of the glottal area waveform (estimated from image segmentation from HSV) and the glottal volume velocity (measured via inverse filtering of oral volume velocity using a circumferentially vented Rothenberg mask). This setup is schematized in Figure 2.2. Alzamendi et al. validated this method by comparing non-stationary estima-

tions of subglottal pressure with measurements obtained from repeated consecutive /pæ/ syllable pronunciations using an intraoral pressure sensor in a subject without voice disorders. This proof-of-concept illustrated that a Bayesian framework with a lumped-element model of phonation can successfully integrate data from HSV and glottal airflow signals to produce meaningful estimates of clinically relevant variables that are difficult to measure directly. Subsequently, Mehta et al. [18] applied this method to estimate subglottal and vocal fold collision pressures, comparing them with measurements obtained using a transoral dual-sensor in an individual with a hemilaryngectomy. The study exhibited good agreement between the model-based estimations and the sensor measurements.

The promising results of the EKF method open an avenue for applying the numerical lumped-element voice production model in clinical settings. However, its optimal performance was achieved using an observation vector that includes recordings of both the glottal area waveform and the glottal volume velocity. This requirement poses challenges for research and experimental validation in individuals with typical and disordered phonation due to the limited availability of datasets featuring simultaneous HSV recordings and multi-parametric sensor data.

In this research, prior efforts [51] were revisited and a constrained extended Kalman filter (CEKF) scheme that integrates a priori physiological information was proposed. This approach aims to reduce the required observations while still

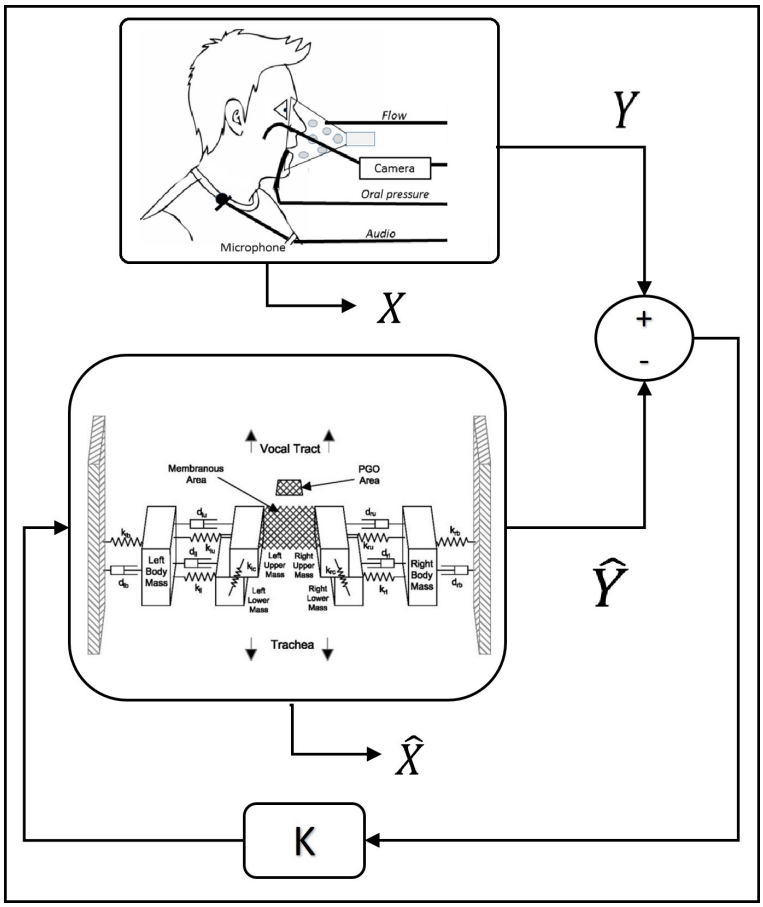


Figure 2.2: General scheme of the Extended Kalman Filter for voice inversion, adapted from [69].

providing robust and reliable estimates of vocal function measures, as detailed in Chapter 3. A specific goal of this research is to explore the feasibility of using CEKF-based subject-specific simulations to access clinically challenging features, such as muscle activation and vocal fold collision pressure, from solely glottal volume velocity observations via inverse filtering of oral volume velocity measured using a circumferentially vented Rothenberg mask.

2.1.3 Machine learning tools and voice production model

These approaches combine machine learning tools and a numerical voice production model to create a regression model that can predict vocal function measures. This regression model is trained using simulated data from a numerical voice production model and directly maps voice production output to the physiological parameters of the vocal system.

Gomez et al. [58] introduced this method using a long short-term memory network (LSTM) to estimate subglottal pressure from vocal fold trajectory recordings obtained via HSV. The scheme is illustrated in Figure 2.3. They trained the network exclusively with synthetic data from a numerical two-mass VF model. Then, they tested it using experimental data from 288 high-speed *ex vivo* video recordings of porcine vocal folds. The results showed that their network achieved performance similar to that of an optimization-based voice inversion method [47] in predicting subglottal pressure from HSV in excised porcine vocal folds.

Recently, Zhang [59] utilized data generated from a three-dimensional continuum model of voice production [70] to train neural networks for estimating vocal fold geometry, stiffness, position, and subglottal pressure. In contrast to Gomez et al. [58], who used vocal fold trajectory data to train their neural network, Zhang employed extracted features of voice production as inputs for network training. He hypothesized that carefully selected features would enable the network to ac-

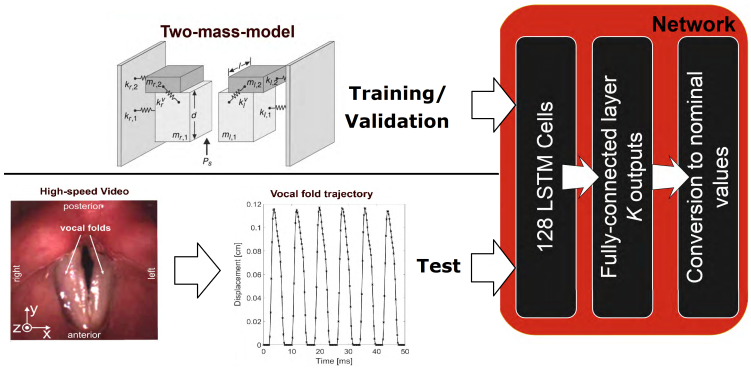


Figure 2.3: Scheme of combined low-order vocal fold model and LSTM network [47].

curately capture the relationship between model control parameters and voice production. This method was validated using voice feature data extracted from an excised human larynx, achieving reasonably good estimation accuracy. Later, Zhang [64] explored the potential for improving the estimation accuracy of physiological control parameters by including voice outcome features that characterize vocal fold vibration. He also identified voice feature sets that optimize estimation accuracy and robustness against measurement noise. However, this subsequent study relied solely on synthetic data.

This emerging method is beginning to capitalize on the advantages of numerical voice production models for training processes. Such models are beneficial as they can represent a wide range of conditions and provide access to relevant measures that are difficult to obtain experimentally. Once trained, machine learning methods for voice inversion could estimate relevant vocal function parameters

without generating new model simulations, making this methodology computationally more efficient than those previously described. In this context, this technique aligns with the overarching goal of this research proposal. The aim is to develop a new framework based on this methodology for *in vivo* estimation of vocal function, utilizing airflow-based features derived from inverse filtering of neck surface acceleration.

2.1.4 Linear regression models

The intraoral pressure measurement during controlled speech gestures is the most common indirect and noninvasive method for subglottal pressure assessment. This technique involves using the production of a bilabial stop consonant to estimate subglottal pressure during the production of an adjacent vowel (e.g., a string of /p/ + vowel syllables). During a sufficiently long consonant, this anatomical configuration leads to pulmonary pressure equalization above and below the glottis, thus allowing for the indirect estimation of subglottal pressure from a pressure measurement in the oral cavity [71, 72]. Subglottal pressure estimated through this method has been utilized to develop linear regression models based on variables such as sound pressure level [65] or subglottal neck-surface vibration [38].

Björklund and Sundberg [65] employed regression analysis to determine the relationship between subglottal pressure and sound pressure level. They measured oral pressure during the /pæ/ sound, with gradually increasing or decreasing vocal

loudness, produced by sixteen female and fifteen male healthy voices. The results indicated that SPL mostly has a logarithmic relationship to subglottal pressure, with an average correlation of 0.83. These results align with earlier research by Titze et al. [31], who developed an empirically derived, straightforward formula to calculate subglottal pressure:

$$P_s[\text{kPa}] = 0.14 + 0.06\left(\frac{f_o}{f_{oN}}\right)^2 + 10^{\left(\frac{\text{SPL}-88.5}{27.3}\right)} \tag{2.3}$$

This formula uniquely requires measurements of SPL and the fundamental frequency; f_{oN} represents the nominal speaking f_o value, which is 120 Hz for males and 190 Hz for females.

Other researchers have explored the relationships between accelerometer signal properties and subglottal pressure. Fryd et al. [38] were pioneers in identifying and quantifying the relationship between subglottal pressure and subglottal neck-surface vibration. They conducted experiments using simultaneous recordings of intraoral air pressure, neck-surface acceleration, and radiated acoustic pressure from ten vocally healthy speakers. The participants produced repetitions of three different /p/–vowel gestures at three pitch levels in the modal register, ranging from loud to soft. Then, using the subglottal pressure estimated indirectly from the intraoral air pressure signal, they performed subject-specific linear regressions to map the amplitude of the ACC signal to subglottal pressure. This analysis revealed a high degree of correlation between these features.

Subsequently, several studies have investigated the relationship between the magnitude of neck-surface vibration and intraoral estimates of subglottal pressure in various voice conditions, encompassing a broad cohort of healthy and pathological speakers. For instance, the study by Mckenna et al. [37] involved twelve vocally healthy adults who produced strings of /pæ/ syllables in three intensity conditions while increasing vocal effort. They discovered that the relationship between ACC-signal amplitude and subglottal pressure varied across different intensities in some individuals. In a related study, Marks et al. [36] observed significant changes in this relationship when speakers produced nonmodal phonation. Furthermore, in a subsequent study, Marks et al. [35] applied this method to participants with voice disorders such as PVH, NPVH, and UVFP. Their findings indicated that the relationship between ACC signal magnitude and subglottal pressure was statistically different in patients with voice disorders compared to vocally healthy controls.

The observed variation in the baseline regression line between accelerometer RMS level and subglottal pressure, especially for higher vocal efforts, modal phonation, and pathological voices, highlighted the need to incorporate other ACC-derived predictors to enhance the robustness of the method. In this context, Lin et al. [34] developed a subject-specific stepwise regression model. This model combined ACC RMS with cepstral peak prominence, f_o, and glottal airflow measures [61]. This newly proposed method was validated with twenty-six

vocally healthy adult speakers during non-modal phonation, leading to a 25% improvement in subglottal pressure estimation.

These research efforts represent the initial attempts to estimate relevant vocal function measures using accelerometer data. Their findings demonstrated a strong correlation between subglottal pressure and signals from the ACC sensor. These results support the second hypothesis of this research proposal, which posits the potential to access vocal function measures from features derived from neck surface acceleration. This work aims to explore whether this correlation can be extended to other relevant parameters, such as VF collision pressures or muscle activation.

2.2 Voice production model

Several numerical implementations have been proposed in the literature to mimic voice production. These models vary in their physiological accuracy, the complexity of mathematical formulation, and computational requirements [73, 74, 75]. Among these, lumped element models stand out for their utility in reproducing VF behavior related to voice production [76, 77, 78, 79], generating natural-sounding voices [80, 81], and modeling vocal fold pathologies [82]. These models typically consist of non-linear mechanical components such as coupled mass-spring-damper systems [76] and incorporate aspects like biomechanical properties of vocal fold tissues [83], transversal mucosal wave dynamics [73], aeroacoustic

interactions [84], glottal geometry [85], and laryngeal muscle control [86]. This research focuses on this particular numerical voice production model due to its ability to effectively balance computational efficiency with physiological relevance.

In this research, two types of lumped element models were utilized. First is the symmetric three-mass body-cover model (BCM) with posterior glottal opening [82], selected for its simplicity and suitability for the proposed Bayesian framework. The second model is the Triangular Body-Cover Model (TBCM) of the vocal folds, which includes coordinated activation of the five intrinsic laryngeal muscles [60]. This model is more physiologically representative and aligns with the objective of estimating intrinsic laryngeal muscle activity.

2.2.1 Vocal fold models

Story and Titze extended the classic two-mass model of the vocal fold [77] to develop the BCM [78]. The BCM incorporates a third mass to provide a more realistic representation of the body-cover vocal fold structure. The two cover masses are laterally coupled to a body mass through nonlinear springs and viscous damping elements. The body mass, representing muscle tissue, is laterally coupled to a rigid wall, using a similar configuration of a nonlinear spring and a damping element. Further developments were made by Zañartu et al. [82], who expanded the BCM by incorporating a posterior glottal opening (PGO), as is illustrated in Figure 2.4. Their research focused on examining the effects of the PGO on

tissue dynamics, energy transfer, acoustic interactions, and glottal aerodynamics. These enhancements have significantly improved the representation of vocal fold physiology in the model.

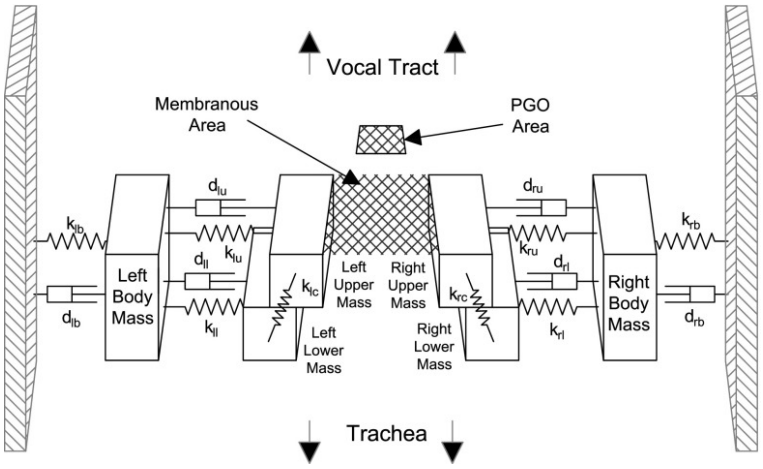


Figure 2.4: Three-dimensional representation of the body cover model showing the posterior glottal opening, the vocal fold masses, and the membranous area [82].

Subsequently, Galindo et al. [85] proposed the TBCM. This model consists of three interconnected mass systems, each representing a distinct element of the body-cover systems as detailed in [78]. Configured in a triangular anatomical shape, the TBCM mirrors the structure of the glottis [87]. Furthermore, it advances the vocal fold collision model by introducing a gradual, zipper-like incomplete glottal closure. This enhancement is crucial for accurately describing the time-varying vocal fold collision pressure during phonation, providing a more comprehensive understanding of vocal mechanics.

The equations of motion for the three masses within these lumped models are derived based on the interaction between the coupling forces, which link the masses together, and the external driving forces acting on each mass. These equations are formulated as follows:

$$F_u = m_u \ddot{x}_u = F_{k,u} + F_{d,u} - F_{kc} + F_{e,u} + F_{Col,u}, \tag{2.4a}$$

$$F_l = m_l \ddot{x}_l = F_{k,l} + F_{d,l} + F_{kc} + F_{e,l} + F_{Col,l}, \tag{2.4b}$$

$$F_b = m_b \ddot{x}_b = F_{k,b} + F_{d,b} - \left[F_{k,u} + F_{d,u} + F_{k,l} + F_{d,l}\right], \tag{2.4c}$$

where m represents mass and its subscripts u, l, and b denote the upper, lower, and body masses, respectively. x indicates the medial-lateral displacement over time, and F is the force component for each block. Force subscripts k, d, e, and kc correspond to the mechanical forces produced by the springs, dampers, flow pressures, and elastic coupling between the upper and lower masses, respectively. An additional spring force, F_{Col}, is introduced during vocal fold collision to capture the effects of impact among opposite upper and lower cover masses. The detailed definitions of these forces for the BCM can be found in [78], and for the TBCM in the appendix of [85].

Both the BCM and the TBCM are regulated through muscle activation based on the empirical rules introduced by Titze and Story [83]. The normalized activation levels of the cricothyroid (a_{CT}), thyroarytenoid (a_{TA}), and lateral cricoarytenoid (a_{LCA}) muscles are utilized to control the mechanical properties of the VF

models including linear stiffness, mass distribution, glottal convergence, thickness, and depth.

2.2.2 TBCM controlled with five intrinsic laryngeal muscles

Alzamendi et al. [60] proposed a multi-physics scheme featuring a low-order model of the vocal folds, which facilitates the coordinated activation of all five intrinsic laryngeal muscles: CT, TA, lateral cricoarytenoid (LCA), interarytenoid (IA), and posterior cricoarytenoid (PCA). This approach expands on previous research that established rules for controlling low-order models [83], vocal fold posturing [86], and the TBCM of the vocal folds [85]. Figure 2.5 provides a schematic of the TBCM model. This model also enables the investigation of the role of antagonistic muscle pairs in phonation, particularly in scenarios of normal and poorly regulated muscle activation, as is hypothesized to occur with NPVH [7].

This model is based on the understanding that the intrinsic musculature, along with the passive contribution of the vocal ligament (LIG) and VF mucosa (MUC), produces the stress imbalances responsible for the movement and deformation of laryngeal structures [88, 89, 90]. Accordingly, the internal stress resulting from muscle activation is represented by normalized muscle activity, ranging from

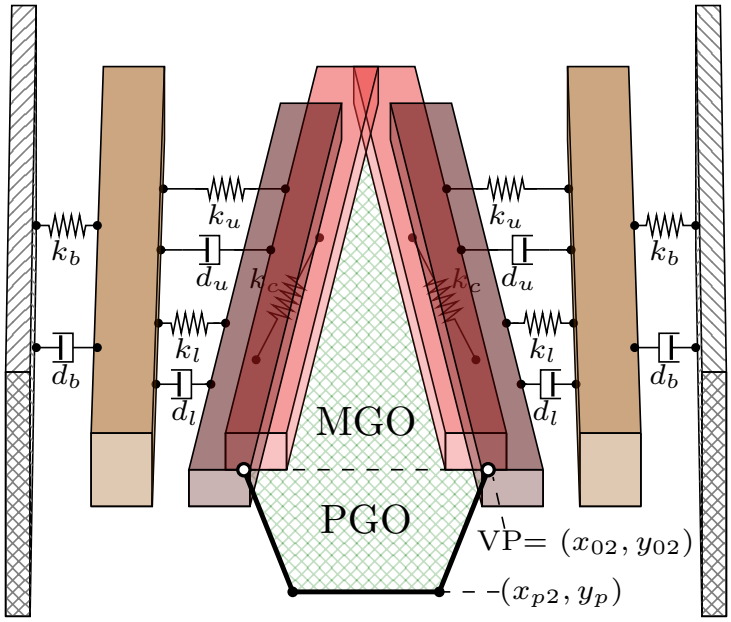


Figure 2.5: Schematic of the triangular body-cover model of the vocal folds [60].

$0 \leq a_i \leq 1$, for $i \in \{\text{LCA, IA, PCA, CT, TA}\}$. Coordinated movements of translation and rotation about the fixed cadaveric cricoarytenoid joint center (x_{CAJ}, y_{CAJ}), mimic the complex movements of arytenoid accommodation. Let (ξ, ψ) and θ respectively represent the Cartesian displacement coordinates and the angle of rotation of the right arytenoid cartilage relative to the cricoarytenoid

joint center. The equations of motion are as follows:

$$\ddot{\xi} = \frac{1}{M_a}\sum_{i\in\mathcal{I}} \alpha_i F_i - k_x\xi - d_x\dot{\xi} \tag{2.5a}$$

$$\ddot{\psi} = \frac{1}{M_a}\sum_{i\in\mathcal{I}} \beta_i F_i - k_y\psi - d_y\dot{\psi} \tag{2.5b}$$

$$\ddot{\theta} = \frac{1}{I_a}\sum_{i\in\mathcal{I}} \gamma_i F_i - \kappa\theta - \delta\dot{\theta} \tag{2.5c}$$

where F_i represent the forces, and α_i, β_i, and γ_i are the directional cosines for the Cartesian displacements and the directional moment arm, applicable to laryngeal tissue $\mathcal{I}$ = {LCA, IA, PCA, CT, TA, LIG, MUC}. Moreover, the parameters M_a and I_a denote the mass and moment of inertia of the arytenoid cartilage, respectively; k_y, k_x, and κ are the translational and rotational stiffnesses, while d_x, d_y, and δ are the translational and rotational damping coefficients, respectively.

As the vocal process is structurally linked to the arytenoid cartilage, the Cartesian coordinates of the vocal process position (x_{02}, y_{02}) are computed as follows [86]:

$$x_{02} = x_{CAJ} - (x_{CAJ} - \bar{x}_0)\cos\theta + y_{CAJ}\sin\theta + \xi, \tag{2.6a}$$

$$y_{02} = y_{CAJ}(1-\cos\theta) - (x_{CAJ} - \bar{x}_0)\sin\theta + \psi + L_0(\epsilon_r + \epsilon_t). \tag{2.6b}$$

where x_0 represents the cadaveric horizontal position of the vocal process, L_0 denotes the rest length, ϵ_r and ϵ_t correspond to the rotational and translational movements around the cricothyroid joint, respectively.

Based on VF adduction and the symmetry along the midsagittal plane, the

glottal area for the upper (A_u) and lower (A_l) cover masses in this model are calculated as follows:

$$A_u = 2(1-\alpha_u)L_g\big(\tilde{x}_u + 0.5(1+\alpha_u)x_{01}\big), \tag{2.7a}$$

$$A_l = 2(1-\alpha_l)L_g\big(\tilde{x}_l + 0.5(1+\alpha_l)x_{02}\big), \tag{2.7b}$$

where $\tilde{x}_u = x_u - x_{u,0}$ and $\tilde{x}_l = x_l - x_{l,0}$ are block displacements relative to the rest positions, and L_g is VF length. Additionally, α_u and α_l are the proportions of mass length for the upper and lower blocks undergoing collision at the solution time, where $0 \leq \alpha_u, \alpha_l \leq 1$. As a result, the area for the membranous glottal opening is $A_{\text{MGO}} = \min\{A_u,\, A_l\}$. Also, the model simulates the effects of laryngeal posture on the posterior cartilaginous portion of the glottis, assuming that the area of the posterior glottal opening has a trapezoid shape [88]. Thus:

$$A_{\text{PGO}} = \max\{0,\, \min\{(x_{p1} + x_{01}),\, (x_{p2} + x_{02})\}(y_{02} - y_p)\}, \tag{2.8}$$

where x_{p1} is the posterior wall half-width at the bottom, x_{p2} is the posterior wall half-width at the top, and y_p is the posterior wall position along the longitudinal axis. The total glottal area comprises both the membranous and the cartilaginous parts:

$$A_g = A_{\text{MGO}} + A_{\text{PGO}}. \tag{2.9}$$

In addition to controlling intrinsic muscle activation, the model facilitates the adjustment of aerodynamic lung pressure, denoted as P_L. The aerodynamic forces

exerted on the vocal fold cover layer are computed using the resulting subglottal pressure P_s, and supraglottal pressure, P_e, according to [91].

2.2.3 Interactions at the glottis and acoustic wave propagation

The three-way interaction between sound, flow, and VF tissue is critical to accurately capture the physics of human phonation. For this purpose, glottal airflow was computed from the acoustic driving pressures impinging on the combined membranous and posterior portions of the glottal area, as detailed in [92, 88, 82]. The simulation of time-varying acoustic wave propagation was achieved using the wave reflection analog scheme. In this scheme, the subglottal and supraglottal tracts were modeled as a series of short uniform acoustic cylinders with variable cross-sectional areas. The effects of the boundary condition at the lips were approximated by including an inertive radiation impedance, producing the reflected pressure wave and the radiated sound wave, P_{out}. Losses from viscosity, moving walls, and other factors are accounted for by an exponential attenuation factor in the propagation through these cylindrical sections [93, 79]. Additionally, vocal tract area functions representing a typical male [94] and female [95] can be selected, corresponding to vowels /i, ɪ, e, ɛ, æ, ʌ, ɑ, ɔ, o, ʊ, u/, along with a representative subglottal tract [82].

2.3 Chapter conclusions

This background review highlights the most innovative methodologies based on voice production models reported in the literature for refining vocal function assessment. Three relevant methodologies were identified: optimization-based voice inversion, Bayesian estimation, and machine learning tools. The majority of these methodologies have been primarily tested with synthetic and *ex vivo* experimental data [47, 58, 56, 55, 54, 63, 59, 64]. Other research has utilized *in vivo* data with HSV recordings [44, 45, 46, 48, 50, 49, 52], as this method provides the most direct link to the voice production model through the analysis of vocal fold trajectories. However, this approach limits their possible application to only clinical scenarios.

The optimization-based voice inversion methods have focused on linking the low-order vocal fold model to the dynamic behavior observed in HSV, without considering the three-way interactions between sound, flow, and vocal fold tissue, thereby limiting their extension to aerodynamic or acoustic signals [44, 45, 46, 48, 50, 49, 47]. Additionally, the principal challenge of this approach is the computational cost, which is due to the high number of simulations required during the voice inversion process. In this context, this methodology was excluded from the research objectives.

Bayesian inference, based on the EKF [55], allows for computational efficiency

in linking a low-order voice production model with combined HSV and airflow measures, showing promising results in the clinical assessment of vocal function [51]. In this context, it is aimed to study this technique under the hypothesis that considering prior information about the phonatory process as a constraint would extend its applicability in scenarios where multi-sensor recordings are not available. This approach points to measurements that could be obtained in ambulatory scenarios, such as glottal airflow.

The framework combining machine learning tools with the voice production model, as proposed by Zhang [59], shows promising results in estimating vocal function parameters from acoustics and glottal flow waveforms. Although this technique has not been validated *in vivo*, it represents a more direct approach to achieving the primary objective of this thesis: proposing a non-invasive method that improves the assessment of vocal function estimation, suitable for both clinical and ambulatory settings.

On the other hand, subject-specific regression models [65, 34, 35, 36, 37, 38] reinforce the underlying hypothesis of this research: that advanced vocal function metrics can be obtained through accelerometer data analysis. These methods have revealed a significant correlation between subglottal pressure and ACC sensor signals. However, the challenge lies in determining if this correlation extends to additional vocal function features, such as vocal fold collision pressures or muscle activation.

In this chapter, the relevant aspects of selected low-order voice production models are also presented. Their biggest advantage is highlighted: the effective balance between computational efficiency and physiological relevance. The BCM offers a simpler representation of the phonatory process [82]. Its implementation requires fewer parameter settings, making it ideal for application in Bayesian inference. A more comprehensive model is the TBCM, controlled by five intrinsic muscles [60], which better aligns with the interest in estimating muscle activation and vocal fold collision pressure. The TBCM still has low computational demands, and its complexity does not constitute a hindrance within a machine learning framework. Once synthetic data is generated through multi-setting TBCM parameterization, the nonlinear regression can be trained and tested without necessitating new simulation settings for the TBCM.

Chapter 3

Estimation of vocal function measures using constrained extended Kalman filter

This chapter examines how Bayesian inference in the voice production model can replicate vocal function behavior observed in laboratory measurements, thereby aligning with the first aim of the research. In the previous chapter, the potential of combining an extended Kalman filter with a muscle-controlled biomechanical model of the vocal folds was described. This method has shown promise in estimating physiologically relevant measures of glottal function, such as subglottal pressure, laryngeal muscle activation, and vocal fold collision pressure, using synchronized recordings from calibrated high-speed videoendoscopy and oral airflow. Building on this, a constrained extended Kalman filter scheme is introduced in this research, which simplifies the experimental requirements by employing either the glottal area waveform or glottal airflow as the sole observational input. This modified approach incorporates constraints based on physiological information about subglottal pressure and muscle activation ranges observed in normal speech. Ac-

cordingly, this chapter presents the theoretical foundations of the constrained extended Kalman filter and their validation through four experiments using *in vivo* laboratory recordings.The method and preliminary results (Experiment 1) detailed in this chapter were presented in [96], and the remaining experiments are currently being prepared for publication in a peer-reviewed journal.

3.1 Discrete state-space model of phonation

As in previous studies [51, 56, 55], a discrete state-space model is used to describe the dynamics of VF oscillations, considering a symmetric BCM of vocal fold described in Chapter 2, Subsection 2.2.1. It is assumed that the latent, time-varying properties of glottal dynamics can be summarized at time t_k by a state vector, $\mathbf{x}_k$, while the clinical multi-sensor data are represented in an observation vector, $\mathbf{y}_k$. Thus, the discrete state-space model takes the form [55]:

$$\mathbf{x}_{k+1} = f(\mathbf{x}_k, \boldsymbol{\theta}_{k+1}, t_{k+1}) + \mathbf{u}_{k+1}, \tag{3.1}$$

$$\mathbf{y}_k = g(\mathbf{x}_k, \boldsymbol{\theta}_k, t_k) + \mathbf{v}_k, \tag{3.2}$$

where subscript k denotes the index for time t_k, with $k = 0, 1, \ldots, K$. Here, f is a nonlinear state transition function that describes the evolution of states $\mathbf{x}$ for successive time steps, $\boldsymbol{\theta}_k$ is a vector of deterministic model parameters, $\mathbf{u}_k$ is the (unbiased) state noise vector with covariance matrix $\boldsymbol{\Gamma}_\mathbf{u}$, g is the nonlinear measurement function that relates the state $\mathbf{x}$ to the measurements $\mathbf{y}$ at time step

t_k, and $\mathbf{v}_k$ is the zero-mean observation noise vector with covariance matrix $\mathbf{\Gamma}_\mathbf{v}$.

The state estimations of this dynamical system are derived from analyzing measurements of an observable process as it evolves over time. Previous studies have explored two different observation models [51]. Initially, the glottal area waveform (GAW), obtained from HSV data, served as the sole clinical observation. Subsequently, a combination of GAW and glottal airflow, the latter estimated from OVV and denoted as U_g, was incorporated into the observation vector. This prior study underscored the importance of multi-sensor observations in generating robust and reliable estimates of glottal variables.

However, simultaneously recording HSV and OVV in a clinical setting necessitates specialized facilities. This includes an adapted Rothenberg mask to accommodate a flexible fiberscope, alongside careful management by a trained health professional during endoscopy. The absence of such specialized capabilities limits the collection of multi-sensor clinical data, thereby constraining the exploration of the proposed Bayesian approach.

Consequently, one of the primary objectives is to examine the impact on estimated phonation states when either GAW or U_g is the only clinical measurement available. To this end, two distinct observation models have been developed, which are detailed in the following sections.

3.1.1 Glottal area waveform as observation

Assuming that the state vector consolidates the primary glottal variables that characterize the voice production process:

$$\mathbf{x}_k = (x_u[k], v_u[k], x_l[k], v_l[k], x_b[k], v_b[k], A_{PGO}[k], P_s[k], a_{CT}[k], a_{TA}[k], P_C[k], U_g[k])^{\mathrm{T}}, \quad (3.3)$$

x represents the positions of the upper (u), lower (l), and body (b) masses, while v denotes their corresponding velocities. A_{PGO} is the area of the posterior glottal opening, P_s denotes the subglottal pressure, and a_{CT} and a_{TA} are the normalized activation levels for the cricothyroid and thyroarytenoid muscles, respectively. Additionally, P_C refers to the VF collision pressure, and U_g represents the glottal airflow.

The state transition function in Eq. (3.1) is structured as follows: The motion of the three body-cover masses is described by states $(x_\alpha[k], v_\alpha[k])^T$ with $\alpha \in \{u, l, b\}$. The equations for the dynamics of each mass are modeled by a truncated Taylor series approximation [97]:

$$\begin{pmatrix} x_\alpha[k+1] \\ v_\alpha[k+1] \end{pmatrix} = \begin{pmatrix} x_\alpha[k] + \Delta v_\alpha[k] + \frac{\Delta^2}{2m_\alpha} F_\alpha(\mathbf{x}_k, \boldsymbol{\theta}_{k+1}, t_{k+1}) \\ v_\alpha[k] + \frac{\Delta}{m_\alpha} F_\alpha(\mathbf{x}_k, \boldsymbol{\theta}_{k+1}, t_{k+1}) \end{pmatrix} + \begin{bmatrix} \xi_{x_\alpha} \\ \xi_{v_\alpha} \end{bmatrix}, \quad (3.4)$$

where F_α represents the net nonlinear force applied to mass α, arising from the springs, dampers, collision forces, and the aerodynamic pressures on both the sub- and supra-glottal tracts [97]. The errors in the motion equation, denoted as $[\xi_{x_\alpha}, \xi_{v_\alpha}]^T$, are assumed to be jointly zero-mean Gaussian with a non-diagonal

covariance matrix:

$$K_\alpha \mathbf{Q}_\alpha = K_\alpha \begin{pmatrix} \frac{\Delta T^3}{3} & \frac{\Delta T^2}{2} \\ \frac{\Delta T^2}{2} & \Delta T \end{pmatrix}, \tag{3.5}$$

K_α represents a tuning parameter, and ΔT denotes the sampling period.

BCM parameters $\boldsymbol{\theta}$ are controlled via normalized muscle activation levels a_{CT}, a_{TA} and a_{LCA} using a set of physiologically-inspired rules [83]. In this work, the a_{LCA} was set at 0.5 anchoring the VF to a "just touching" configuration [83]. The activation levels a_{CT} and a_{TA} must be in the range of [0, 1] (from no activation to full activation). They were modeled as unknown deterministic constants, with transition rules:

$$a_{CT}[k+1] = a_{CT}[k], \tag{3.6}$$

$$a_{TA}[k+1] = a_{TA}[k]. \tag{3.7}$$

Initial conditions $a_{CT}[0]$ and $a_{TA}[0]$ were selected using a grid-search process, which aimed at minimizing the RMSE between the Kalman estimate and the measured reference signal. The grid was constructed by varying $a_{CT}[0]$ and $a_{TA}[0]$ within the range of [0.1, 0.9], inclusive, with steps of 0.1, resulting in 81 simulation cases. Simulations where the estimates converged to extreme values (0 or 1) were discarded, as these muscle activation levels are not physiologically representative of typical voices.

Area A_{PGO} was assumed to be a constant geometric parameter. Consequently,

the transition function became:

$$A_{PGO}[k+1] = A_{PGO}[k], \tag{3.8}$$

where the initial state, $A_{PGO}[0]$, is defined as the minimum value of the glottal area obtained from HSV when available; otherwise, zero is assumed.

In turn, a random walk model [98] was considered for the subglottal pressure:

$$P_s[k+1] = P_s[k] + \xi_{P_s}, \tag{3.9}$$

where $\xi_{P_s} \sim \mathcal{N}(0, \sigma^2_{P_s})$.

The nonlinear collision pressure P_c was computed from the collision forces introduced in the BCM [78] to account for the overlap of the left and right cover masses during closure. The tissue-fluid-acoustic interactions at the glottis, as described in Chapter 2, Subsection 2.2.3, are also considered. Consequently, the glottal airflow U_g is estimated as a function of the glottal area A_g, resulting from VF displacements, the subglottal pressure P_s, and the acoustic pressure waves impinging on the glottis [92, 78]. In the transition equation for the states P_c and U_g, additive zero-mean Gaussian noises ξ_{P_c} and ξ_{U_g}, respectively, were included

The state covariance matrix $\mathbf{\Gamma_u}$, representing the model uncertainty and degrees of freedom in the state transition as detailed in Eq. (3.1), is structured as follows:

$$\mathbf{\Gamma_u} = \text{diag}(K_u\mathbf{Q}_u, K_l\mathbf{Q}_l, K_b\mathbf{Q}_l, \sigma^2_{A_{PGO}}, \sigma^2_{P_s}, \sigma^2_{a_{CT}}, \sigma^2_{a_{TA}}, \sigma^2_{P_c}, \sigma^2_{U_g}), \tag{3.10}$$

where σ^2 denotes the variance associated with the states in subscripts, and submatrices $K_\alpha \mathbf{Q}_\alpha$ represent the uncertainty related to the positions and velocities for each mass.

In this case, the observation vector is given by the glottal area (A_g):

$$\mathbf{y}_k = A_g[k], \tag{3.11}$$

In this context, the measurement function, as defined in Eq. (3.2), is represented by the determination of the glottal area, employing a similar approach to that used in the BCM. The A_g is derived from the positions of the masses. By assuming symmetrical dynamics for the left and right vocal folds, this area is calculated as follows:

$$A_u = \max(0, x_u \ell_g), \tag{3.12a}$$

$$A_l = \max(0, x_l \ell_g) \tag{3.12b}$$

where A_u and A_l represent the upper and lower glottal areas, respectively, and ℓ_g is the effective length of the glottis. Consequently, the total glottal area is:

$$A_g = \min\{A_u, A_l\}, \tag{3.13}$$

3.1.2 Glottal airflow as observation

In this case, the glottal airflow is the observation state instead of the glottal area, therefore the state vector is:

$$\mathbf{x}_k = (x_u[k], v_u[k], x_l[k], v_l[k], x_b[k], v_b[k], A_{PGO}[k], P_s[k], a_{CT}[k], a_{TA}[k], P_C[k], A_g[k])^{\mathrm{T}}, \tag{3.14}$$

and the measurement vector is:

$$\mathbf{y}_k = U_g[k], \tag{3.15}$$

The state transition functions presented here are similar to those described in the previous section. The primary difference is that, in this case, the determination function of the glottal area (Eq. 3.13) is incorporated into the transition functions, while the determination of the glottal airflow serves as the measurement function.

3.2 Kalman filter

Bayesian estimation, such as the Kalman filter, predicts the unknown system state variable $x(t)$ when both the evolution and observation models are linear in the state variable [55]:

$$\mathbf{x}_{k+1} = \mathbf{F}_{k+1}\mathbf{x}_k + \mathbf{u}_{k+1}, \tag{3.16}$$

$$\mathbf{y}_k = \mathbf{G}_k\mathbf{x}_k + \mathbf{v}_k, \tag{3.17}$$

where $\mathbf{F}$ and $\mathbf{G}$ are linear state and observation evolution models.

The Kalman filter algorithm operates in two steps: prediction and updating. In the prediction step, it generates a probability density $p(\mathbf{x}_k|\mathbf{y}_1, ..., \mathbf{y}_k)$, forecasting the state value and uncertainty at the next time step based on previous observations up to t_{k-1}. Upon receiving the measurement at t_k, the algorithm

updates by computing the likelihood density $p(\mathbf{y}_k|\mathbf{x}_k)$ to infer information about the state $\mathbf{x}_k$ [66].

The Kalman filter treats all distributions as Gaussian, which is beneficial mathematically since linear combinations of Gaussian variables are also Gaussian. This ensures that when the evolution and measurement models are linear in the state variable $\mathbf{x}(t)$, their likelihood, evolution, and posterior distributions are Gaussian and fully defined by their mean and covariance. The filter tracks the evolution of these parameters based on observations, estimating the posterior density at each time step. In this linear context, closed-form solutions exist for the maximum posterior (MAP) estimate and covariance of the posterior density. The prediction step in the Kalman filter involves evaluating the evolution model with the previous MAP estimate [66].

$$\hat{\mathbf{x}}_k = \quad f(\mathbf{x}_{k-1}^{MAP}, \boldsymbol{\theta}_k, t_k), \tag{3.18}$$

and the uncertainty of the prediction is:

$$\hat{\boldsymbol{\Gamma}}_k = \mathbf{F}_k \boldsymbol{\Gamma}_{k-1} \mathbf{F}_k^T + \boldsymbol{\Gamma}_{\mathbf{u}}, \tag{3.19}$$

where $\boldsymbol{\Gamma}_{k-1}$ is the covariance estimate associated with $\mathbf{x}_{k-1}^{MAP}$. From this, the MAP estimate at time t_k is:

$$\mathbf{x}_{k-1}^{MAP} = \hat{\mathbf{x}}_k + \mathbf{K}_k(\mathbf{y}_k - \mathbf{G}_k \hat{\mathbf{x}}_k), \tag{3.20}$$

which has the associated covariance matrix that models the uncertainty of the MAP estimate

$$\mathbf{\Gamma}_k = (\mathbf{I} - \mathbf{K}_k\mathbf{G}_k)\hat{\mathbf{\Gamma}}_k, \tag{3.21}$$

where $\mathbf{I}$ is the identity matrix, and $\mathbf{K}_k$ is a matrix called the Kalman gain at time t_k, which is given by

$$\mathbf{K}_k = \hat{\mathbf{\Gamma}}_k\mathbf{G}_k^T(\mathbf{G}_k\hat{\mathbf{\Gamma}}_k\mathbf{G}_k^T + \mathbf{\Gamma}_v)^{-1}, \tag{3.22}$$

3.2.1 Extended Kalman filter

The voice production model is a non-linear, time-varying system [78, 83, 82]. Consequently, the EKF formulation must be employed to linearize the state equations [66]. When dealing with nonlinear evolution and observation models, the posterior may not be Gaussian. The EKF addresses this issue by approximating the posterior as Gaussian, using the MAP estimate as the mean, and determining the covariance matrix using linear approximations of the models. Essentially, the EKF serves as an approximation to the Kalman filter for nonlinear models.

The application of the EKF mirrors the steps outlined in previously in for the Kalman filter, but it replaces the linear models $\mathbf{F}_{k+1}$ and $\mathbf{G}_k$ with the Jacobians of the nonlinear functions f and g to account for uncertainty. The prediction step employs the full nonlinear model as described in Eq. 3.18 and the uncertainty is

computed as:

$$\hat{\mathbf{\Gamma}}_k = \mathbf{J}_k^f \mathbf{\Gamma}_{k-1} (\mathbf{J}_k^f)^T + \mathbf{\Gamma}_\mathbf{u}, \tag{3.23}$$

with

$$\mathbf{J}_k^f = \left. \frac{\partial f(\mathbf{x}, \boldsymbol{\theta}_k, t_k)}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}_{k-1}^{MAP}} \tag{3.24}$$

Once an observation y_k is made, the updated estimate and uncertainty can be computed as:

$$\mathbf{K}_k = \hat{\mathbf{\Gamma}}_k (\mathbf{J}_k^g)^T (\mathbf{J}_k^g \hat{\mathbf{\Gamma}}_k (\mathbf{J}_k^g)^T + \mathbf{\Gamma}_v)^{-1}, \tag{3.25}$$

$$\mathbf{J}_k^g = \left. \frac{\partial g(\mathbf{x}, \boldsymbol{\theta}_k, t_k)}{\partial \mathbf{x}} \right|_{\mathbf{x}=\mathbf{x}_k}, \tag{3.26}$$

3.2.2 Constrained extended Kalman filter

The constrained Kalman filters are Bayesian processors that incorporate a priori information about a physical process, e.g., knowledge about equality or inequality constraints in the model states. Then, a Kalman filter scheme can be modified to exploit this additional information to improve its performance [99]. Among the available alternatives, soft constraints are instrumental and easy to implement in scenarios involving heuristic knowledge or when the constraint functions have uncertainty [99].

In the voice production model, additional information regarding subglottal pressure can be derived from peaks in intraoral air pressure recordings of /pæ/

syllable gestures. Espinoza et al. [8, 100] reported an average subglottal pressure of approximately 800 Pa for 40 normal subjects at a comfortable loudness level, with a standard deviation of around 260 Pa. This a priori information can be integrated into the proposed state-space model of phonation as a soft constraint. Specifically, it can be applied as an approximate equality constraint in the state-space model, where the subglottal pressure, measured in Pascals, is constrained following the guidelines provided by Simon [99]. The constraint is formulated as follows:

$$P_s[k] \approx P'_s[k] + \eta_{P'_S}, \tag{3.27}$$

where P'_s is a reference subglottal pressure level, and $\eta_{P'_s} \sim \mathcal{N}(0, \sigma^2_{P'_s})$ is a fictitious observation noise.

In this model, when the glottal area, denoted as A'_g and extracted via HSV, is available, the observed state is represented by both an extended measurement vector and a covariance matrix for the observation noise. These components are defined as follows:

$$\mathbf{y}_k = \left(A'_g[k], P'_s[k]\right)^T \tag{3.28}$$

$$\mathbf{\Gamma_v} = \text{diag}(\sigma^2_{A'_g}, \sigma^2_{P'_s}). \tag{3.29}$$

In the scenario where the glottal airflow, estimated from Rothenberg mask measurements and denoted as U'_g, is the observed state, the corresponding equations

are:

$$\mathbf{y}_k = \left(U'_g[k], P'_s[k]\right)^T \tag{3.30}$$

$$\mathbf{\Gamma_v} = \text{diag}(\sigma^2_{U'_g}, \sigma^2_{P'_s}). \tag{3.31}$$

On the other hand, additional constraint conditions could also be imposed based on muscle activation. As described in Chapter 2, the BCM is regulated through muscle activation, following the empirical rules introduced by Titze and Story [83]. This enables the model to operate differently across a wide range of a_{CT} and a_{TA} activations, yielding multiple solutions based on convergence rules. By utilizing physiological knowledge, the range of possible solutions can be narrowed down.

In "Principles of Voice Production" by Titze [101], the concept of a muscle activation map is detailed. This map specifically plots the activity range of the CT muscles against the TA muscles. Titze categorizes this map into four quadrants, each representing a different range of vocal registers: pressed, chest, speech/modal, and falsetto. Figure 3.1 displays a muscle activation map that illustrates these four quadrants. In this figure, the activation levels of the TA and CT muscles are shown, ranging from 0 (inactive) to 5 (maximally active). Based on these configurations, for typical phonation, constraining the muscle activation responses specifically to the speech quadrant (the left inferior quadrant) is suggested as follows:

$$(a_{TA}^2 - 0.3)^2 + (a_{CT}^2 - 0.3)^2 \approx 0 + \sigma_{act.}, \tag{3.32}$$

with

$$\sigma_{act.}^2 = (0.2^2)^2, \tag{3.33}$$

Equation 3.32 imposes a constraint on muscle activation, confining it within a

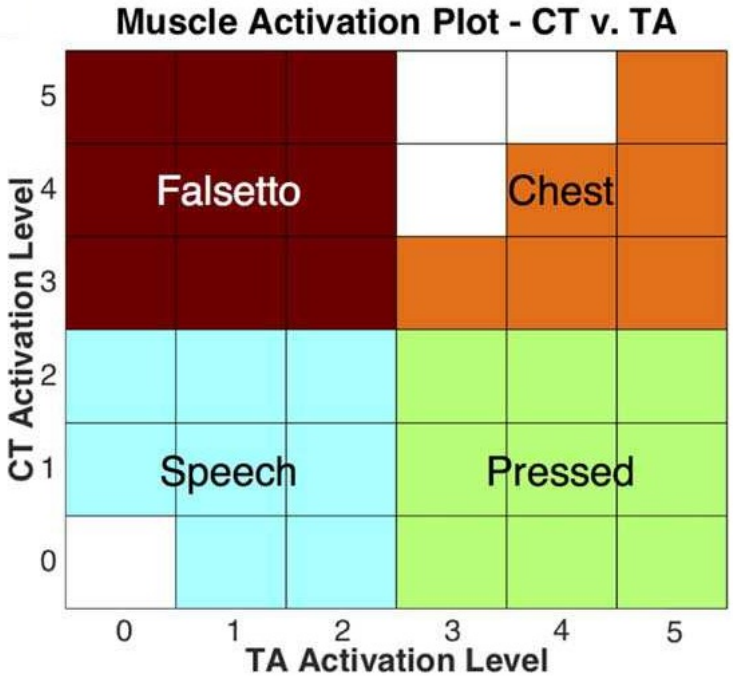


Figure 3.1: Muscle activation map of CT versus TA muscle showing the grouping of TA and CT muscle combinations. CT and TA muscle activation levels span from 0 (inactive) to 5 (maximum activation) [102].

circle that has a center at 0.3 and a radius of 0.2. This method ensures that, in normal voices, the voice production model maintains activations below 50%.

Therefore, when incorporating this constraint into the CEKF framework, which includes measurements of GAW, the observation state is structured as follows:

$$\mathbf{y}_k = \left(A'_g[k], P'_s[k], 0\right)^T \tag{3.34}$$

$$\mathbf{\Gamma_v} = \text{diag}(\sigma^2_{A'_g}, \sigma^2_{P'_s}, 0.2^2). \tag{3.35}$$

Alternatively, when the glottal airflow is observed, the model adjusts as:

$$\mathbf{y}_k = \left(U'_g[k], P'_s[k], 0\right)^T \tag{3.36}$$

$$\mathbf{\Gamma_v} = \text{diag}(\sigma^2_{U'_g}, \sigma^2_{P'_s}, 0.2^2). \tag{3.37}$$

3.3 Laboratory recordings

To qualitatively and quantitatively validate the proposed CEKF approach, three distinct datasets are utilized for varied analytical purposes. Dataset 1 is the most comprehensive, facilitating simultaneous measurements of laryngeal HSV, IOP, radiated acoustic pressure from the microphone (MIC), neck skin acceleration, electroglottography (EGG), and OVV. It is primarily used for the quantitative validation of model estimations, such as comparing glottal area and airflow states. Dataset 2, which is limited to HSV recordings, aids in the qualitative validation of muscle activation estimations during phonations of sustained vowels at varying pitch levels. Lastly, Dataset 3, detailed further in Chapter 4, includes synchronous recordings of IOP, OVV, MIC, and ACC from vocally healthy subjects.

This dataset is used for analyzing vocal fold collision pressure through utterances of /pæ/ syllables under different loudness conditions.

3.3.1 Dataset 1

This laboratory dataset includes time-synchronized *in vivo* recordings of laryngeal HSV, OVV, and IOP from a male participant with no history of voice disorders [103, 104]. The recording session employed a flexible endoscope for HSV, facilitating both aerodynamic assessment and normal articulation by the participant. A visual representation of this setup is illustrated in Figure 3.2. During the protocol, a specialist instructed the subject to produce two sustained vowels (/a/ and /i/) and to repeat /pæ/ syllables. All recordings were conducted in an acoustically treated room. The data collection was approved by the institutional review boards at Massachusetts General Hospital and the Massachusetts Institute of Technology.

To facilitate simultaneous HSV, OVV, IOP, and MIC measurements, the standard circumferentially-vented mask (model MA-1L, Glottal Enterprises) was modified to allow for flexible endoscope placement. This modification, as highlighted in [105], ensures both mobility and a proper seal. The mask was further redesigned to be self-supporting around the subject's head, accommodating the OVV sensor (model PT-series, Glottal Enterprises), an IOP sensor, and the MIC sensor (model MKE104, Sennheiser electronic GmbH & Co. KG). An electronics unit (model

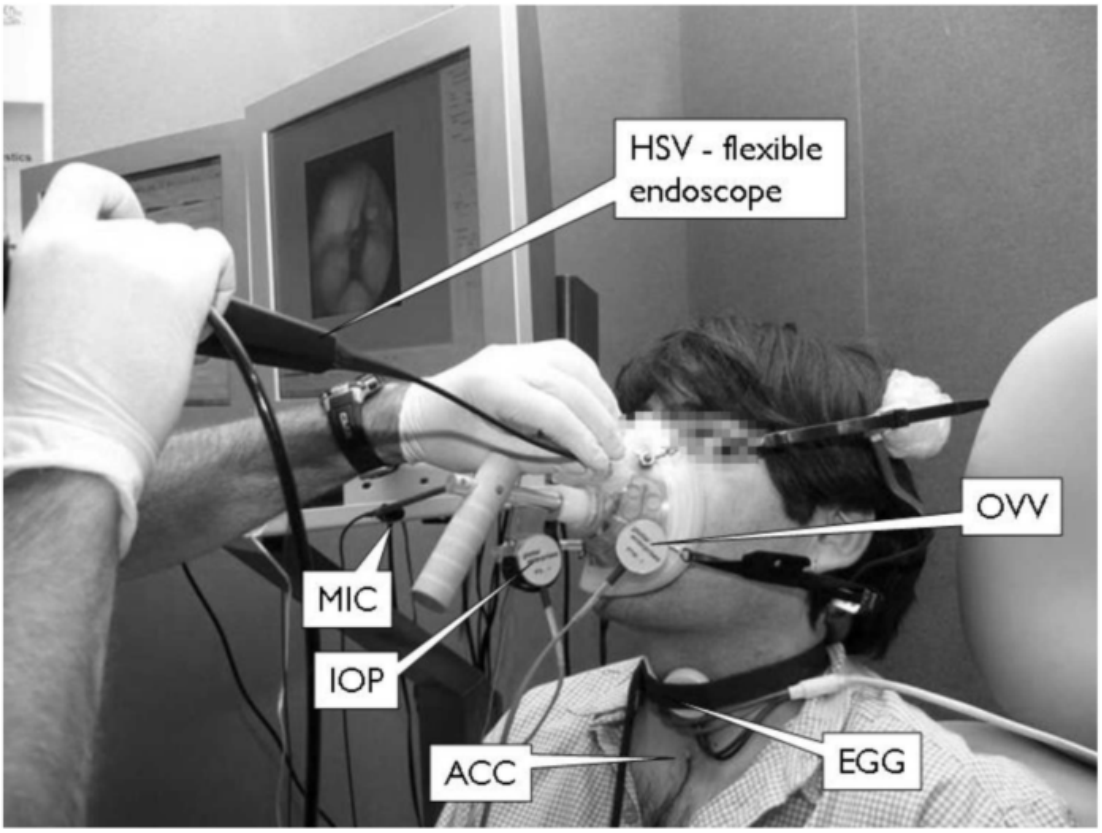


Figure 3.2: High-speed video measurement and data acquisition system. Flexible endoscopy through a modified CV mask [104].

MS-100A2, Glottal Enterprises) provided the necessary signal conditioning and gain for the OVV sensor prior to digitization.

For HSV recordings, a Vision Research Inc. Phantom v7.1 monochromatic camera was used. This setup was equipped with a KayPENTAX C-mount lens adapter and connected to a transnasal fiberscope (model FNL-10RP3, KayPENTAX) for enhanced flexibility in endoscopy. The recording rate was set at 4000

images per second, with a spatial resolution of 320 x 480 pixels. Additionally, the accelerometer (model BU-7135, Knowles) and EGG electrodes (model EL-2, Glottal Enterprises) were also integrated into the recordings to provide comprehensive data.

Reference values for subglottal pressure were obtained from IOP signals in /pa/ syllables [8]. Driving pressure was interpolated as the mean value of the two consecutive IOP plateaus, which were produced by the combined lip closure and glottis opening immediately before and after each vowel segment in the /p/ sounds.

The OVV-based estimated GVV signal was derived using a common inverse filtering technique. This technique employs a single-notch filter with a conjugate pair of zeros and unity gain at DC for the first vocal tract resonance [106, 107].

HSV recordings were digitally processed to detect the glottal contour on a frame-by-frame basis, enabling the segmentation of the GAW. The spatial calibration of the video-based GAW functions in physical units was accomplished by identifying a reference laryngeal landmark (e.g., blood vessels) near the glottis. The dimensions of this landmark were independently measured using a calibrated endoscope system [108].

3.3.2 Dataset 2

This dataset is part of OPENGLOT [109], a comprehensive collection of four data repositories designed for the evaluation of glottal inverse filtering algorithms. This research focused on Repository IV, which contains recordings from natural speakers. This repository features simultaneous recordings of MIC, EGG, and HSV signals. For the experiment, signals corresponding to normal loudness and two distinct pitch levels (low and high) were examined from three male speakers. During the recording protocol, these speakers were instructed to produce the vowel /i/ while positioning their tongues forward to ensure the clearest view of the glottis. However, due to the limitations imposed by the HSV endoscope, the actual vowel sounds produced varied, falling within the phonetic range between /æ/ and /œ/ [109].

The measurements were performed at Helsinki University Central Hospital using the KayPentax Color High-Speed Video System (model 9710) with a spatial resolution of 512 x 512 pixels and a temporal resolution of 2000 frames/s. EGG was acquired with a Glottal Enterprises electroglottograph (EG2-PCX2). A DPA omnidirectional headset microphone (model 4065-BL) was set 6.5 cm from the center of the speaker's mouth. The microphone signal and EGG were recorded using a MOTU UltraLite-mk3 Hybrid audio interface connected to a MacBook Pro running OS X (v. 10.9.5) and AudioDesk 4. To enable synchronization of the audio

signals with the video, a synchronization signal comprising binary frequency-shift keyed code at the beginning of each second was adopted, as detailed in [109].

The original HSV dataset underwent pre-processing to identify a 256x256 pixel region of interest, focusing exclusively on the vocal folds. This refined HSV dataset was then utilized with GlottalImageExplorer [110] for segmenting the glottal area on a frame-by-frame basis, facilitating the estimation of the GAW in pixels. In this dataset, instead of identifying a reference laryngeal landmark for calibrating the video-based GAW measurements in physical units, the glottis length at the maximum open phase of the phonatory cycle was determined, as illustrated in Figure 3.3. Subsequently, the GAW signal was normalized using the square of this glottis length in pixels, resulting in a dimensionless GAW signal.

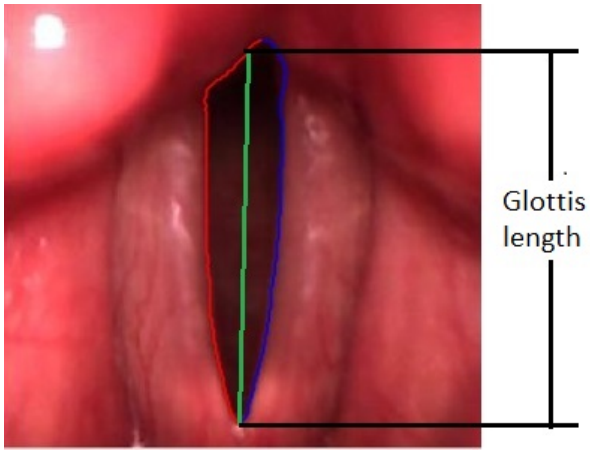


Figure 3.3: Pre-processing example from High-Speed Video Dataset 2 for obtaining GAW (Observed state). Visualization of left/right vocal fold edges (in blue/red) and defined glottis length for normalization.

3.4 Results

The results of the four simulation experiments conducted to validate the proposed CEKF-based framework are detailed next. In the first experiment, utilizing dataset 1, the Bayesian approach was implemented with GAW as the observation state. This simulation was validated by comparing the model-estimated glottal airflow against that obtained from the OVV signal. The second experiment, employing dataset 2, focused on qualitatively studying the response of the model to pitch changes using the estimated GAW as an observation state. In the third experiment, again with dataset 1, the observation state was the glottal airflow. Validation was achieved by contrasting the model-estimated glottal area against its HSV-based counterpart. In the final experiment, the CEKF was applied using glottal airflow as the observation state, specifically focusing on three female registers from dataset 3 (detailed in Chapter 4, Subsection 4.3.1). This experiment aimed to examine how the model responds to variations in loudness levels.

3.4.1 Experiment 1

In this experiment, the voice production model states were inferred by observing the glottal area and imposing constraints on the subglottal pressure state. To simulate sustained vowels /a/ and /i/, the P_s was constrained to 800 Pa with a variance of 100 Pa, following the expected values for normal subjects [8, 100].

For the /pæ/ string, the P_s constraints were set according to the reference values obtained from the IOP. In this experiment, it was not necessary to impose additional constraints on muscle activation. The covariance matrix for the state ($\mathbf{\Gamma_u}$) and measurements ($\mathbf{\Gamma_v}$) were empirically selected to ensure stable simulations.

Bayesian estimates of the states A_g, P_S, U_g, a_{CT}, a_{TA}, and P_C, as obtained by the CEKF-based model, are presented in Figure 3.4 for the vowels (/a/ and /i/), and in Figure 3.5 for the three middle segments of the /pæ/ string. For each case, the figures are organized as follows: the first row contrasts the glottal area extracted from HSV with the corresponding model approximation; the second row shows the estimated state Ps; the third row contrasts the estimated U_g state with that estimated from the OVV signal; the fourth row displays the muscle activation states a_{CT} and a_{TA}; and the fifth row illustrates the vocal fold collision pressure. These signals represent simulations over 50 ms, during which the state estimations of the model are stable. The shaded regions in the figures indicate the 95% confidence bounds.

Figures 3.4 and 3.5 demonstrate that the measured GAW aligns reasonably well with the glottal area derived from the CEKF model. However, some persistent mismatches are observed, such as small peaks in the closed phase of the glottal cycle and biased DC values. These discrepancies are likely induced by model approximations and oversimplifications. Nevertheless, the GAW extracted from HSV typically falls within the estimated uncertainty bounds. Estimates of

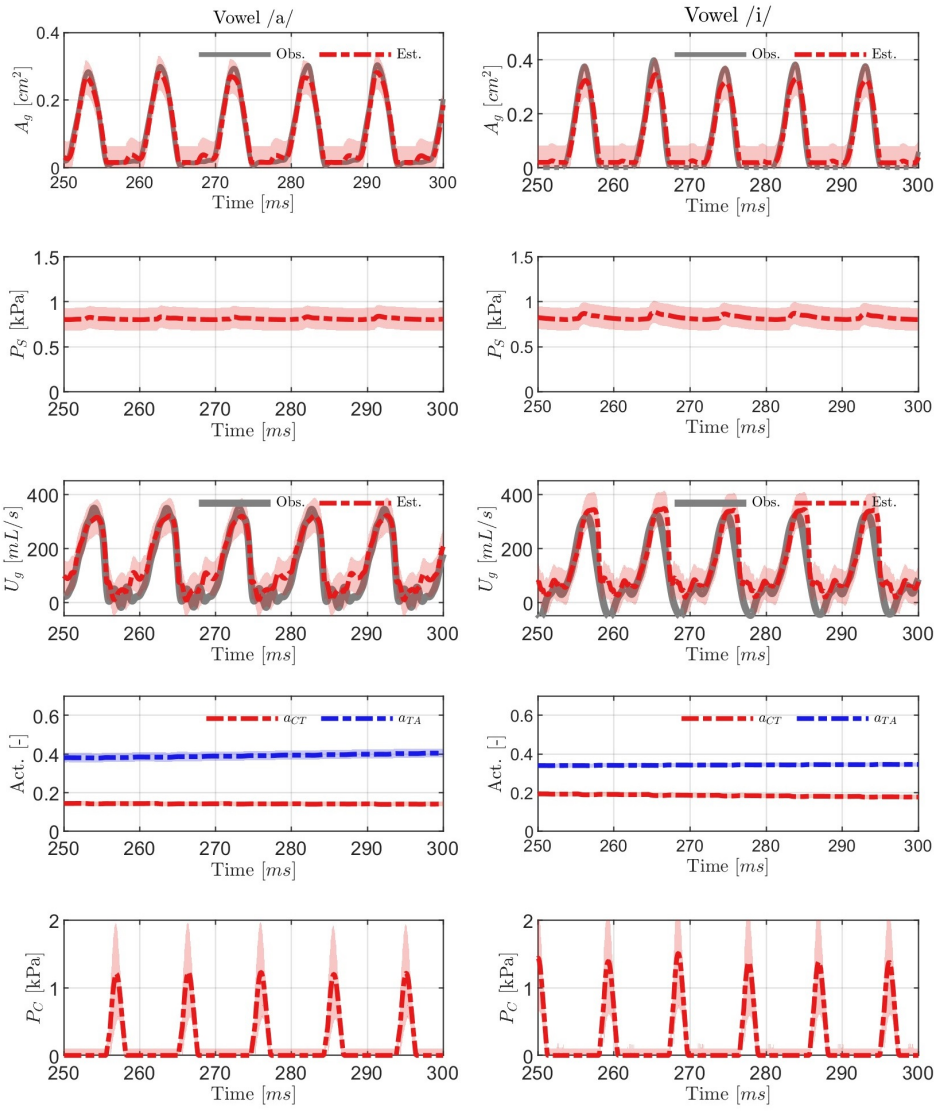


Figure 3.4: CEKF model estimations for a male from Dataset 1: vowel /a/ and /i/ signals using the glottal area as observation state. Observed data (solid gray line) and CEKF estimates (dash-dotted) for glottal area waveforms (first row), subglottal pressure (second row), glottal airflow (third row), muscle activation (fourth row), and vocal fold collision pressure (fifth row). Shaded areas represent 95% confidence intervals.

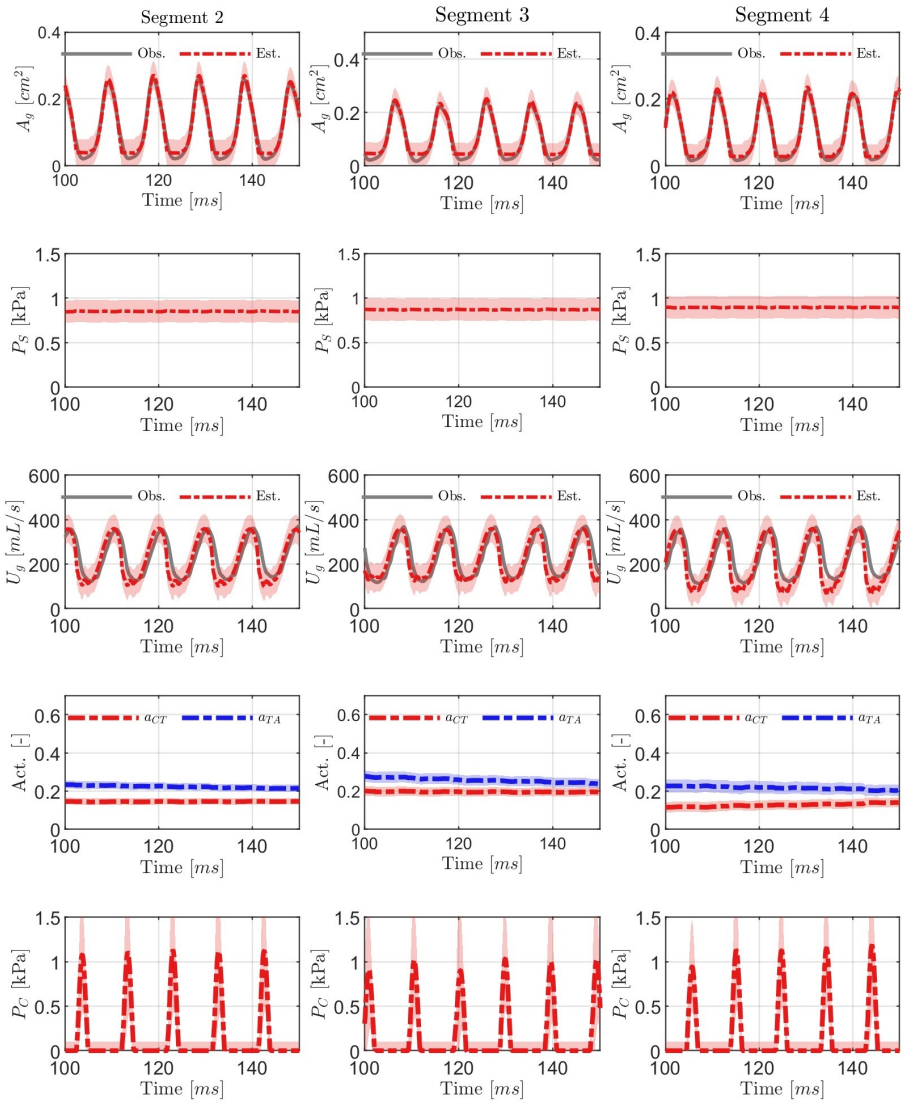


Figure 3.5: CEKF model estimations for a male from Dataset 1: Analysis of the three middle segments of /pæ/ strings signals using the glottal area as the observation state. Observed data (solid gray line) and CEKF estimates (dash-dotted) for glottal area waveforms (first row), subglottal pressure (second row), glottal airflow (third row), muscle activation (fourth row), and vocal fold collision pressure (fifth row). Shaded areas represent 95% confidence intervals.

subglottal pressure are nearly constant around the imposed constraint. It is worth noting that, although a fictitious constant value for subglottal pressure was imposed, the estimated states exhibited periodic fluctuations, as is notably observed in the simulation for the vowel /i/. This could be evidence of the modeled source-filter interaction at the glottis. However, further investigations are required to verify this hypothesis.

Furthermore, the model-based glottal airflow closely matches the observation-based measurements, showing a better fit during the open phase of the cycle in all cases. The muscle activations and vocal fold collision pressure exhibit similar behavior across both phonation vowels and in the three /pæ/ segments. The slight variability observed in these three features for /pæ/ segments is reasonably expected, given the changes in the amplitude of the observed state.

To quantify the errors in the CEKF model estimates compared to the measurement-based signals, the RMSE was computed between them. The results are presented in Table 3.1. It was observed that the RMSE for A_g was less than 0.03 cm^2, and for U_g it was less than 62.48 mL/s. These errors are comparable to those reported in previous work [51], where the RMSE ranged between $0.03 - 0.06$ cm^2 and 40-50 mL/s for A_g and U_g respectively. However, it is important to note that in this proposal, the advantage lies in the fact that glottal airflow was not included in the observation. These initial results suggest that by imposing constraints, the proposed scheme is capable of inferring the model states based solely on the GAW

Table 3.1: RMSE between CEKF model estimations and measurement-based observations for A_g and U_g for a male from Dataset 1 when glottal area waveform is used as the observation state.

Phonation	**Segment**	A_g (cm^2)	U_g (mL/s)
/a/	-	10.0x10$^-$3	41.10
/i/	-	28.4x10$^-$3	62.48
/pæ/	1	10.5x10$^-$3	50.15
	2	10.7x10$^-$3	50.67
	3	12.0x10$^-$3	46.50
	4	8.2x10$^-$3	58.75
	5	8.1x10$^-$3	58.30

extracted from HSV.

Table 3.2 summarizes the statistical information for the Bayesian estimates of muscle activations (a_{CT} and a_{TA}), subglottal pressure, and the peaks of vocal fold collision pressure. As expected, the estimated activation of the a_{CT} muscle is similar across the three cases, with a range of 0.077-0.169. This consistency is indicative of comfortable phonation with an almost constant fundamental frequency. In contrast, the estimates of a_{TA} activation show a notable dispersion, ranging from 0.415 to 0.206, which might be necessary to accommodate the ob-

served differences in the GAW signals using the biomechanical VF model. The mean values of the estimated VF collision pressure are similar in all three cases and analogous to the driving subglottal pressure. It is noted that even though GAW was the only real clinical signal used in the CEKF, the estimations reported in Table 3.2 are similar to those obtained in previous work [51], which combined GAW and U_g signals in the observation. These results suggest that imposing a constraint on the subglottal pressure aids in improving Bayesian inference of vocal function. However, it is challenging to assert the reliability and physiological relevance of the estimates for muscle activation and VF collision pressure, as measuring these latent variables of phonation is cumbersome, even under laboratory conditions.

3.4.2 Experiment 2

In this experiment, the CEKF model was implemented by observing the normalized glottal area obtained from dataset 2. Three male subjects with no medical history of voice disorders were analyzed. Their phonation consisted of a sustained /i/ vowel at two pitch levels (low and high). HSV was the only measurement available in these recordings. Additionally, due to the observed increase in fundamental frequency in the high-pitch recordings, a secondary constraint on muscle activation was imposed, as outlined in Equation 3.32, to ensure model convergence within the speech frequency quadrant for the a_{CT} vs. a_{TA} map. For the P_s, a

Table 3.2: Mean (standard deviation) of muscle activations (a_{CT}, a_{TA}), subglottal pressure (P_s), peak of VF collision pressure (P_c), estimated by the CEKF for a male from Dataset 1 when glottal waveform is used as the observation state.

Phonation	Segment	a_{CT}	a_{TA}	P_s (kPa)	P_c (kPa)
/a/	-	0.139 (0.005)	0.415 (0.012)	0.808 (0.062)	1.078 (0.093)
/i/	-	0.169 (0.006)	0.349 (0.005)	0.823 (0.063)	1.329 (0.099)
/pæ/	1	0,077 (0.008)	0.227 (0.011)	0.894 (0.063	1.080 (0.089)
	2	0.145 (0.009)	0.222 (0.010)	0.851 (0.063)	1.118 (0.100)
	3	0.195 (0.012)	0.257 (0.014)	0.870 (0.063)	1.020 (0.107)
	4	0.127 (0.013)	0.216 (0.016)	0.895 (0.063)	1.141 (0.096)
	5	0.167 (0.011)	0.206 (0.012)	0.893 (0.063)	1.116 (0.104)

typical value of 800 Pa with a variance of 100 Pa was assumed for all simulations.

Figures 3.6, 3.7, and 3.8 display the model estimates for subjects M01, M03, and M04, respectively. In each figure, the data for low pitch are shown in the left column, while high pitch data are on the right. The first row contrasts the normalized GAW estimated by the model with that obtained from HSV. The subglottal pressure, muscle activations (a_{CT} and a_{TA}), and vocal fold collision pressure are sequentially displayed from the second to the fourth row.

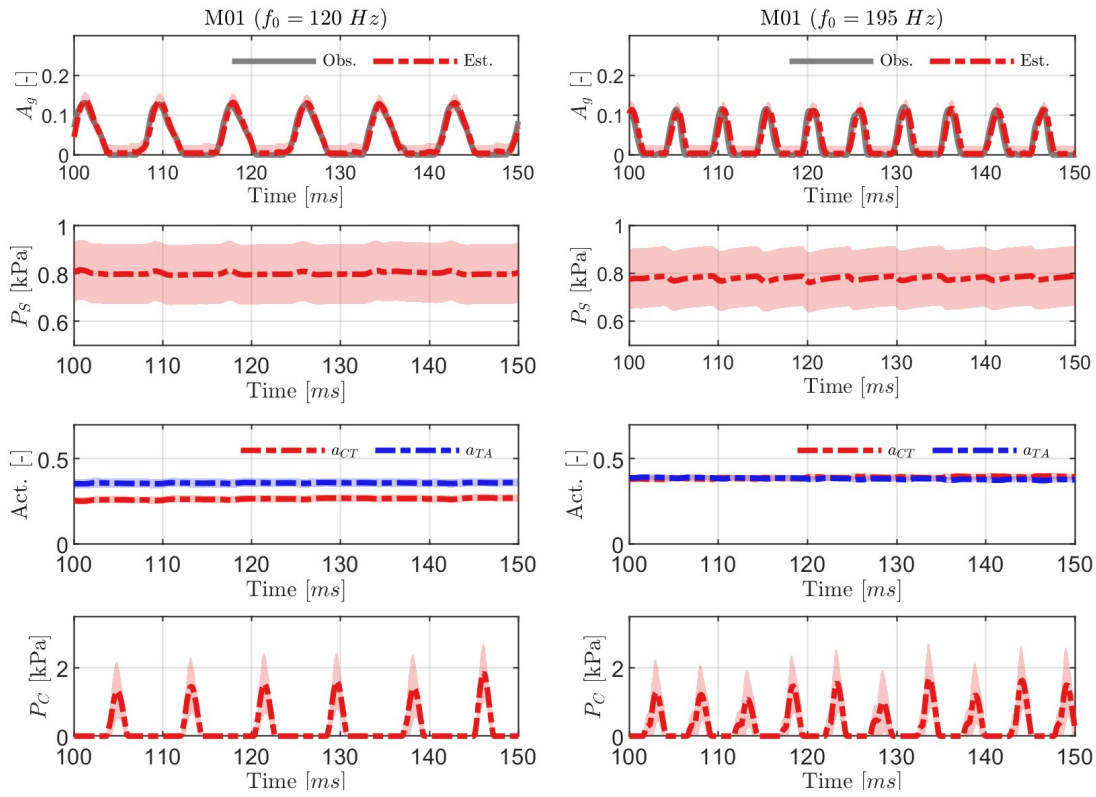


Figure 3.6: CEKF model estimations for subject M01 from dataset 2: sustained vowel in low and high pitch levels. Observed data (solid gray line) and CEKF estimates (dash-dotted) for glottal area waveform (first row), subglottal pressure (second row), muscle activation (third row), and vocal fold collision pressure (fourth row). Shaded areas represent 95% confidence intervals.

The model effectively captured the normalized glottal area waveforms for both pitch conditions in all three cases. As in the previous experiment, the GAW derived from HSV data generally aligns within the estimated uncertainty margins. These results are supported by the RMSE values presented in Table 3.3. In this

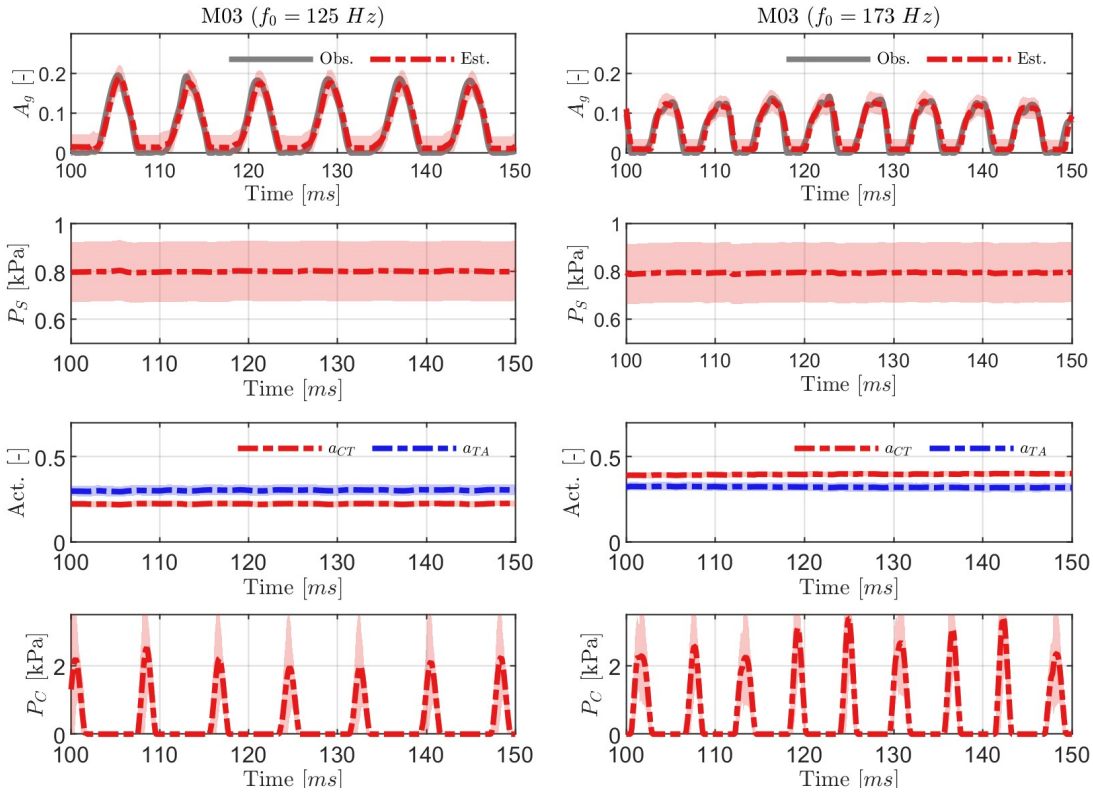


Figure 3.7: CEKF model estimations for subject M03 from dataset 2: sustained vowel in low and high pitch levels. Observed data (solid gray line) and CEKF estimates (dash-dotted) for glottal area waveform (first row), subglottal pressure (second row), muscle activation (third row), and vocal fold collision pressure (fourth row). Shaded areas represent 95% confidence intervals.

experiment, the RMSE between the CEKF model estimations and measurement-based observations for A_g is lower than 0.026 in all cases, which is similar to the errors computed in Experiment 1. Additionally, the subglottal pressure remained constant around the imposed constraints, with some simulations exhibiting the

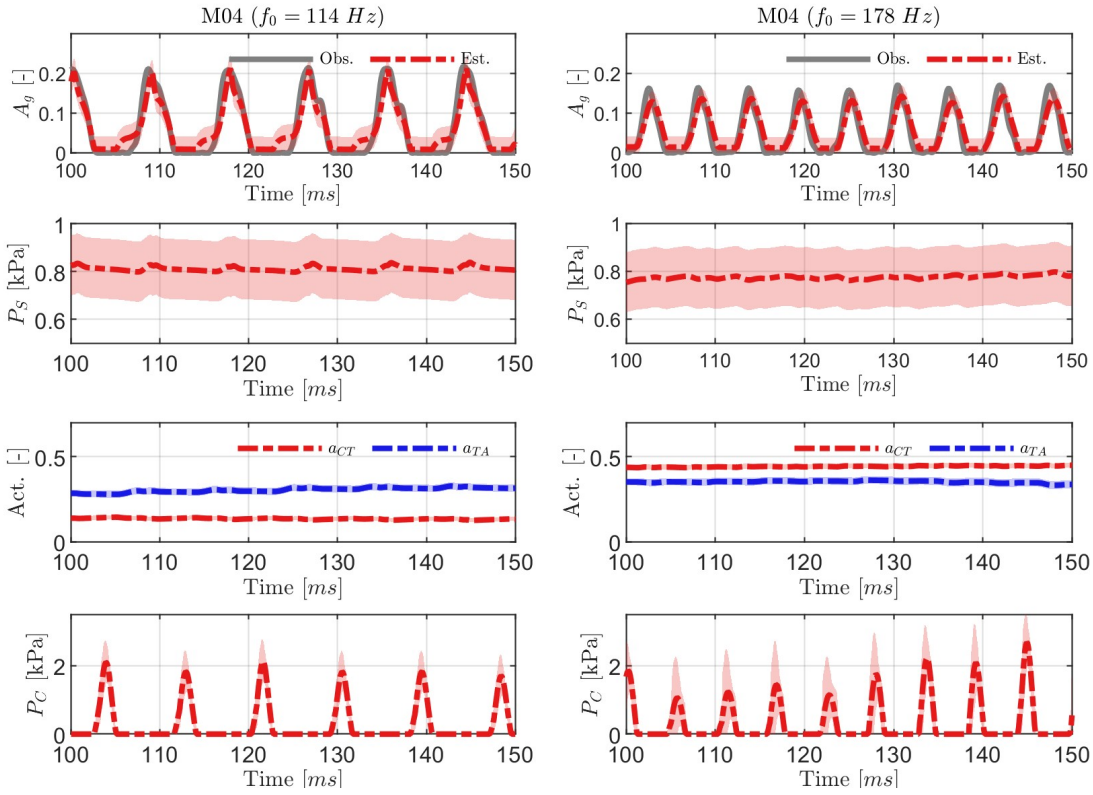


Figure 3.8: CEKF model estimations for subject M04 from dataset 2: sustained vowel in low and high pitch levels. Observed data (solid gray line) and CEKF estimates (dash-dotted) for glottal area waveform (first row), subglottal pressure (second row), muscle activation (third row), and vocal fold collision pressure (fourth row). Shaded areas represent 95% confidence intervals.

same periodic fluctuations observed in Experiment 1.

Significant changes are observed in muscle activation. It can be appreciated that from low to high pitch, the model adjusts a_{CT} to adapt to the new frequency condition, while a_{TA} shows minimal variation. These results align with literature

Table 3.3: RMSE between CEKF model estimations and measurement-based observations for A_g for the subjects in Dataset 2.

Subject	f_0	**RMSE**
M01	120	$8.0\text{x}10^{-3}$
	195	$16.2\text{x}10^{-3}$
M03	125	$15.9\text{x}10^{-3}$
	173	$13.1\text{x}10^{-3}$
M04	114	$25.2\text{x}10^{-3}$
	178	$24.7\text{x}10^{-3}$

reports, where it is generally accepted that activation of the CT muscle increases the fundamental frequency [111, 83]. Concerning P_c, it remains relatively stable despite changes in pitch, behavior that is expected given the assumption that the loudness condition remains constant. This behavior provides a qualitative validation analysis of the proposed model since the results are in accordance with the expected physiological behavior.

Table 3.4 summarizes the mean and standard deviation of muscle activations, subglottal pressure, and peak vocal fold collision pressure estimated in the 50 ms simulation of the CEKF model. In all three cases, it can be appreciated that changes in the fundamental frequency of more than 50 Hz are directly reflected

Table 3.4: Mean (standard deviation) of muscle activations (a_{CT}, a_{TA}), subglottal pressure (P_s) and peak of VF collision pressure (P_c), for the three subjects from Dataset 2, as estimated by the CEKF.

Subject	f_0 (Hz)	a_{CT}	a_{TA}	P_S (kPa)	P_c (kPa)
M01	120	0.264 (0.011)	0.358 (0.013)	0.801 (0.063)	1.523 (0.092)
	195	0.389 (0.010)	0.382 (0.009)	0.781 (0.063)	1.361 (0.170)
M03	125	0.224 (0.009)	0.303 (0.015)	0.801 (0.062)	2.178 (0.174)
	173	0.397 (0.008)	0.322 (0.013)	0.795 (0.063)	2.742 (0.225)
M04	114	0.137 (0.006)	0.306 (0.011)	0.812 (0.063)	1.901 (0.073)
	178	0.443(0.007)	0.352 (0.011)	776.42(0.063)	1.721(0.153)

in an increase of more than 0.1 in a_{CT}, while a_{TA}, P_s, and P_c remain almost constant. Although these simulations represent promising results, it is important to note that they are influenced by the imposed constraints. Therefore, more clinical measurements would be required to provide numerical validation that supports these estimated parameters.

3.4.3 Experiment 3

This experiment aims to validate the CEKF using glottal airflow as the observation state. For this purpose, constraints on the P_s and muscle activation were

necessary. Similar to Experiment 1, the P_s was set to 800 Pa with a variance of 100 Pa for simulating sustained vowels /a/ and /i/, and it was adjusted according to the reference values obtained from the IOP for the /pæ/ string. The muscle activations followed the constraint in Equation 3.32. In these simulations, the reference values of the state and measurement covariance matrix, which were obtained from Experiment 1, were readjusted to ensure convergence and stability in the simulations.

Figures 3.9 and 3.10 display the results for sustained vowels (/a/ and /i/), and the /pæ/ string, respectively. From the first to the last row in each figure, the following are presented: the glottal airflow estimate from the CEKF compared to the measurement, the constrained state P_s, the glottal area estimate from the CEKF compared to that obtained from HSV, the muscle activation model states (a_{CT} and a_{TA}), and the P_c model state. These figures illustrate 50 ms of simulation, and the shaded regions indicate the 95% confidence bounds.

In these simulations, the observed model state U_g coincides with the measurement-based glottal airflow. The RSMSE between these two signals is below 31 mL/s for all simulations, as shown in Table 3.5, representing a 50% reduction compared to the results obtained in Experiment 1. This improvement is expected since in this case the glottal airflow was directly observed by the CEKF approach during the simulations. The constrained state P_s also aligns with the fictitious constant value that was imposed as a constraint.

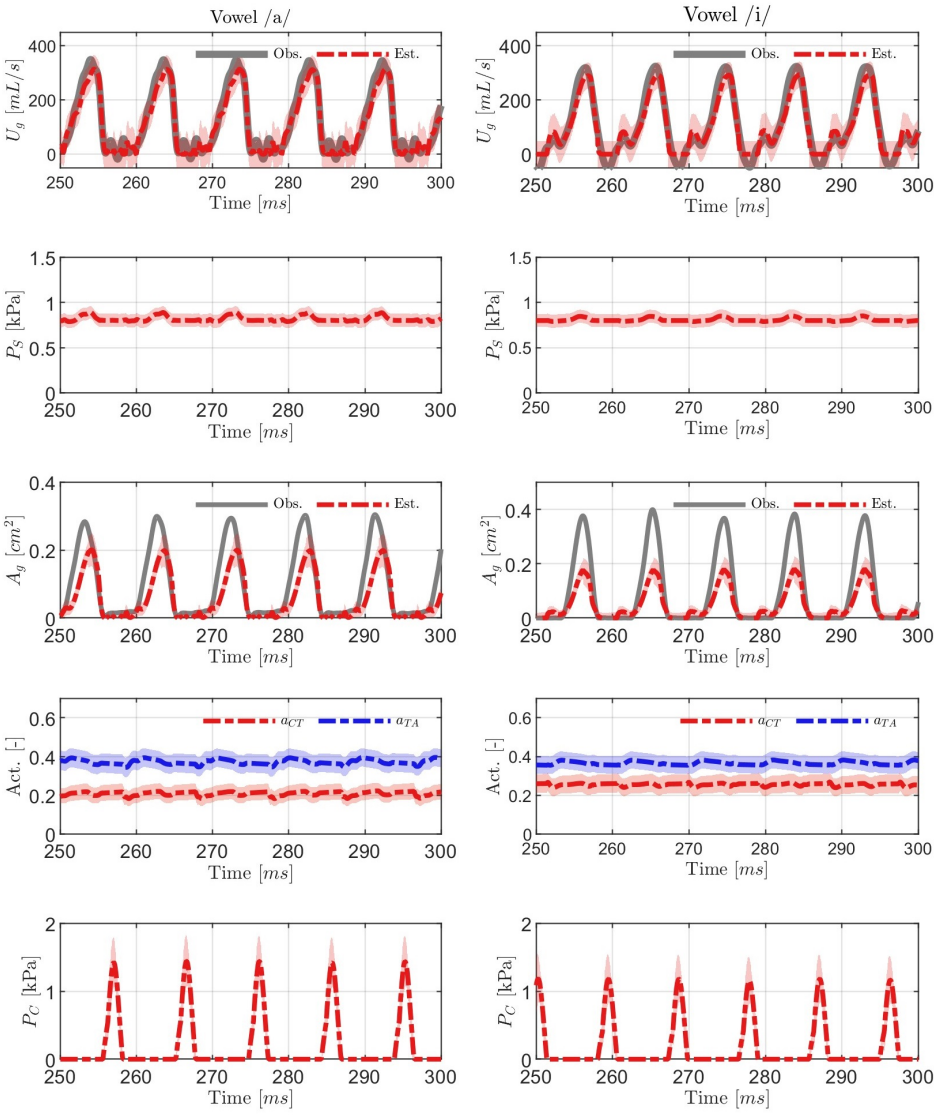


Figure 3.9: CEKF model estimations for a male from Dataset 1: vowel /a/ and /i/ signals using the glottal airflow as observation state. Observed data (solid gray line) and CEKF estimates (dash-dotted) for glottal airflow (first row), subglottal pressure (second row), glottal area waveforms (third row), muscle activation (fourth row), and vocal fold collision pressure (fifth row). Shaded areas represent 95% confidence intervals.

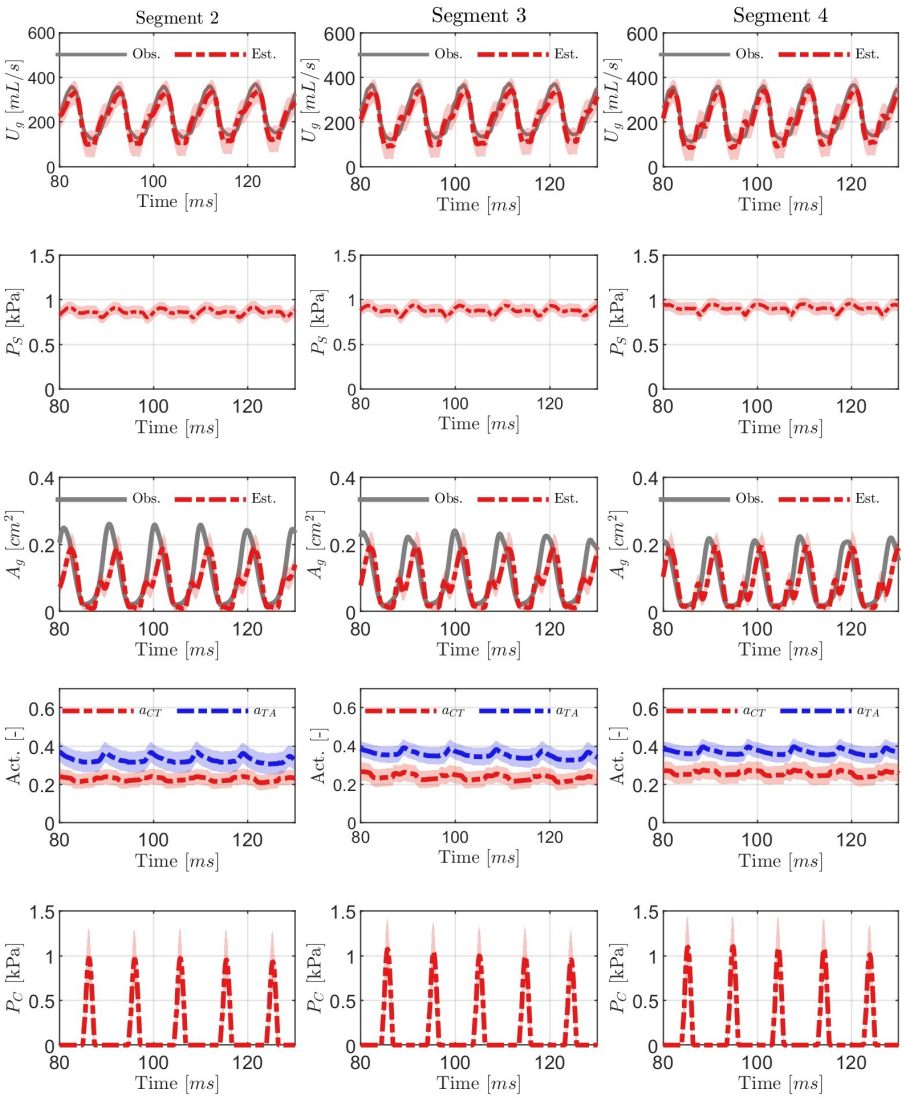


Figure 3.10: CEKF model estimations for a male from Dataset 1: Analysis of the three middle segments of /pæ/ strings signals using the glottal airflow as the observation state. Observed data (solid gray line) and CEKF estimates (dash-dotted) for glottal airflow (first row), subglottal pressure (second row), glottal area waveforms (third row), muscle activation (fourth row), and vocal fold collision pressure (fifth row). Shaded areas represent 95% confidence intervals.

The glottal area estimated by the model shows a notable amplitude difference compared to that obtained from HSV in simulations of sustained vowels, especially for the vowel /i/. This amplitude error is less perceptible in the /pæ/ string simulations. These differences are attributed to the simplicity and lack of anatomical representability of the BCM in computing the glottal area (see Equations 3.12 and 3.13). However, the periodicity, open phase time, and close phase time of the phonatory cycle observed in the A_g signal are well emulated by the model. Additionally, the RMSE between the model estimation and HSV-based GAW ranges from 0.04-0.09 cm^2, as shown in Table 3.5. Although these errors are higher than those in Experiment 1, or the range of 0.03-0.06 cm^2 reported in [51], it is important to note that in this CEKF implementation, the glottal airflow was the sole observed state. These results suggest that CEKF using glottal airflow as observation allows a reasonable estimation of glottal area.

The evolution of the states a_{CT}, a_{TA}, and P_c is similar to those observed when the GAW was the observed state. The mean values and standard deviation of the model states for each simulation of this experiment are summarized in Table 3.6. The muscle activations differ by an average of 0.1 for both a_{CT} and a_{TA} compared to Experiment 1, while P_c is within the same range. These results suggest that the proposed CEKF approach can estimate the model states using only the glottal airflow as the observation. However, the absence of clinical measurements with simultaneous recordings of a_{CT}, a_{TA}, and P_c limits an accuracy comparison study

Table 3.5: RMSE between CEKF model estimations and measurement-based observations for A_g and U_g for a male from Dataset 1 when glottal airflow is used as the observation state.

Phonation	**Segment**	A_g (cm^2)	U_g (mL/s)
/a/	-	65.8x10$^-$3	30.38
/i/	-	88.9x10$^-$3	30.54
/pæ/	1	63.7x10$^-$3	32.30
	2	70.6x10$^-$3	27.83
	3	61.7x10$^-$3	29.17
	4	60.1x10$^-$3	28.63
	5	49.7x10$^-$3	21.45

of both methodologies (observing GAW or U_g).

3.4.4 Experiment 4

In this final experiment, the CEKF was applied to analyze three female subjects without any clinical history of voice disorders. The recordings included the string of /pæ/ in three loudness conditions: soft, comfortable, and loud, with simultaneous measurements of OVV and IOP, but without HSV (see the description

Table 3.6: Mean (standard deviation) of muscle activations (a_{CT}, a_{TA}), subglottal pressure (P_s), peak of VF collision pressure (P_c), estimated by the CEKF for a male from Dataset 1 when glottal airflow is used as the observation state.

Phonation	Segment	a_{CT}	a_{TA}	P_s (kPa)	P_c (kPa)
/a/	-	0.209 (0.020)	0.372 (0.012)	0.817 (0.036)	1.439 (0.033)
/i/	-	0.256 (0.024)	0.366 (0.022)	0.808 (0.036)	1.181 (0.036)
/pæ/	1	0,234 (0.022)	0.365 (0.022)	0.907 (0.036)	1.001 (0.029)
	2	0.227 (0.020)	0.329 (0.027)	0.862 (0.036)	0.966 (0.029)
	3	0.237 (0.022)	0.354 (0.023)	0.881 (0.036)	1.002 (0.029)
	4	0.258 (0.021)	0.369 (0.025)	0.905 (0.036)	1.089 (0.029)
	5	0.288 (0.025)	0.361 (0.021)	0.898 (0.036)	1.000 (0.030)

of this dataset in Chapter 4, Subsection 4.3.1). In these simulations, the model was constrained using the reference subglottal pressure obtained from IOP and the condition specified in Equation 3.32, similar to the approach in Experiment 3. The state and measurement covariance matrices were adjusted for each subject.

The results obtained for each subject are presented in Figures 3.11, 3.12, and 3.13. Horizontally, the figures depict a 50 ms simulation of a /pæ/ segment at soft, comfortable, and loud levels of loudness, respectively. Vertically, from top to

bottom, they display the observed state U_g, the constrained state P_s, the muscle activation states, and the vocal fold collision pressure state.

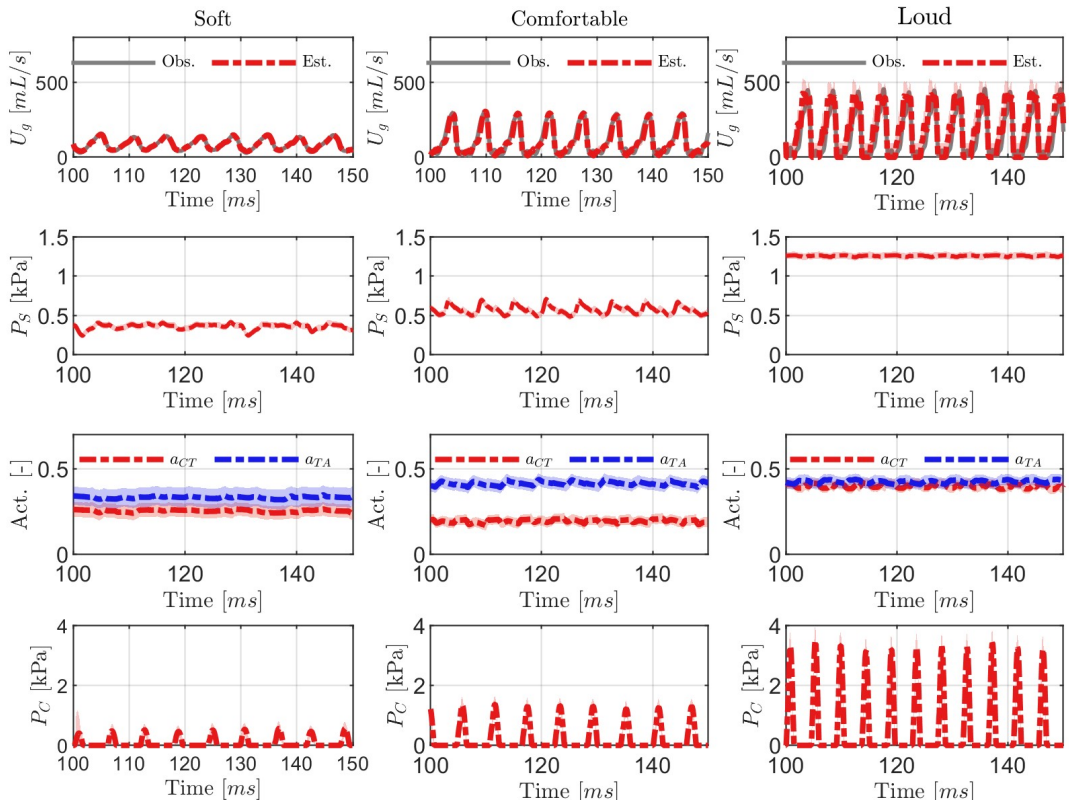


Figure 3.11: CEKF model estimations for subject NF01 from Dataset 3: a segment of /pæ/ string in three loudness levels. Observed data (solid gray line) and CEKF estimates (dash-dotted) for glottal airflow (first row), subglottal pressure (second row), muscle activation (third row), and vocal fold collision pressure (fourth row). Shaded areas represent 95% confidence intervals.

For the three subjects, the U_g state aligns well with the measurement-based glottal airflow, with RMSE values ranging from 3.96 ml/s to 88.51 ml/s, as shown

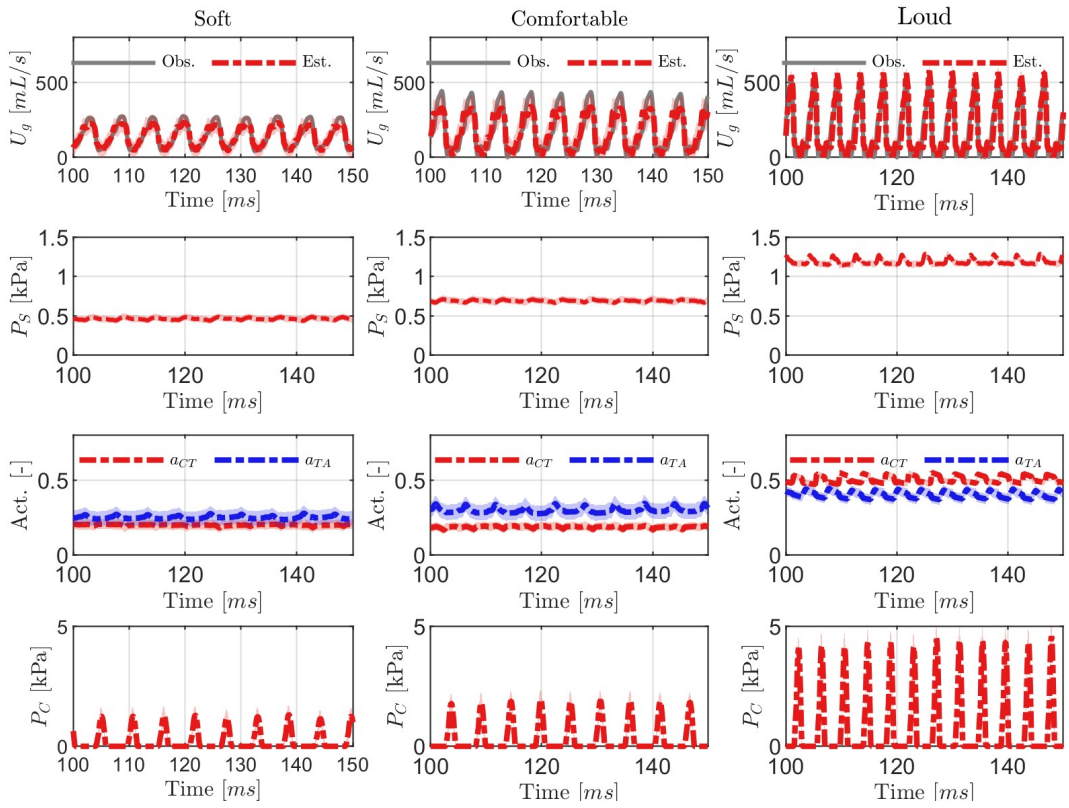


Figure 3.12: CEKF model estimations for subject NF02 from Dataset 3: a segment of /pæ/ string in three loudness levels. Observed data (solid gray line) and CEKF estimates (dash-dotted) for glottal airflow (first row), subglottal pressure (second row), muscle activation (third row), and vocal fold collision pressure (fourth row). Shaded areas represent 95% confidence intervals.

in Table 3.7. From these simulations, it is observed that loudness variability is directly reflected in the model states. The muscle activation adjusts to each loudness condition. From soft to loud, an increase in P_c is noticed, which is consistent with the physiological behavior of the phonatory process. The literature

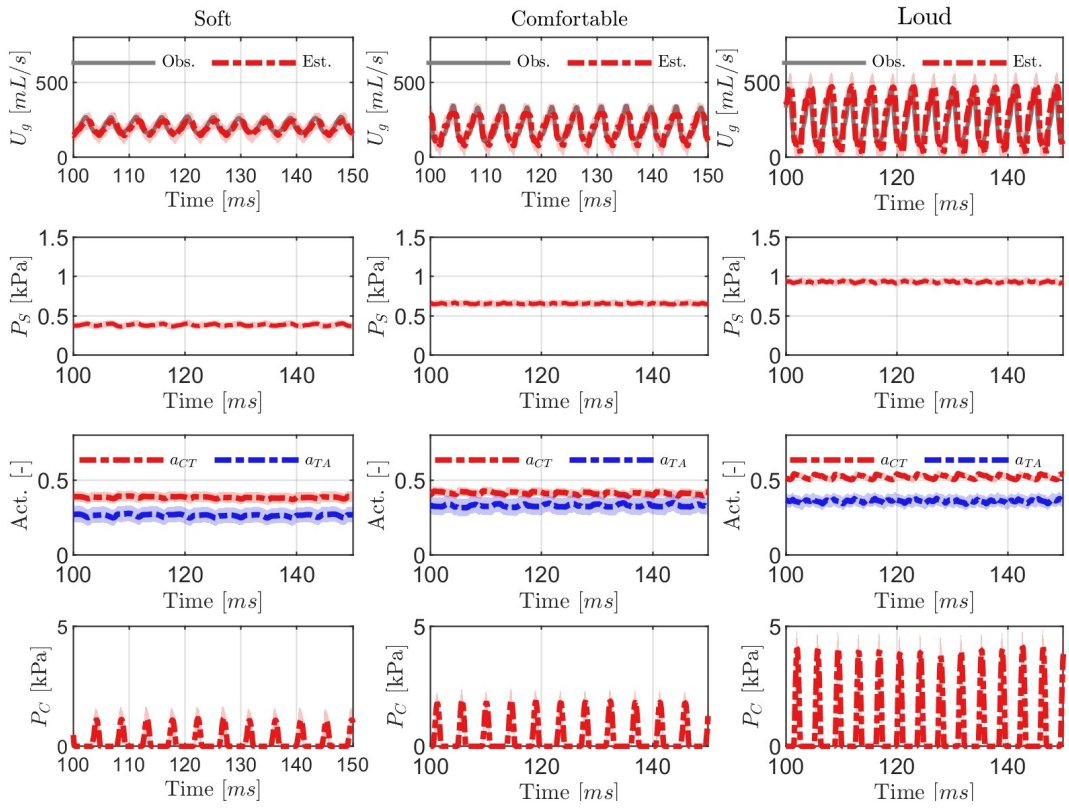


Figure 3.13: CEKF model estimations for subject NF03 from Dataset 3: a segment of /pæ/ string in three loudness levels. Observed data (solid gray line) and CEKF estimates (dash-dotted) for glottal airflow (first row), subglottal pressure (second row), muscle activation (third row), and vocal fold collision pressure (fourth row). Shaded areas represent 95% confidence intervals.

reports that the magnitude of the peak contact pressure primarily depends on the subglottal pressure used to produce the voice [112, 21]. Additionally, these qualitative analyses are aligned with the findings of Diaz et al. [57], who observed an increasing effect of collision pressure with the increment of loudness.

Table 3.7: RMSE between CEKF model estimations and measurement-based observations for U_g for the subjects in Dataset 3.

Subject	**Loudness**	**RMSE**
NF01	Soft	3.96
	Comfortable	22.77
	Loud	74.65
NF02	Soft	25.72
	Comfortable	63.52
	Loud	60.17
NF03	Soft	19.03
	Comfortable	31.94
	Loud	88.51

Table 3.8 presents the mean and standard deviation for all simulations conducted in Experiment 4. For the three subjects involved, there was a notable increase in muscle activations correlating with elevated loudness levels. This observation is consistent with Palaparthi et al. [113], who found in their fiber-gel finite element model that the trends of a_{CT} and a_{TA} muscle activations intensify with rising sound pressure levels. Furthermore, Luegmair et al. [114] have indi-

cated that TA muscle activation plays a pivotal role in regulating mean glottal flow and is particularly significant in maintaining airflow during periods of high subglottal pressure or when producing a loud voice. Additionally, it is important to note that the range of vocal fold collision pressures in this study varied from 0.5 kPa to 4.3 kPa, aligning with the values reported in prior research [57, 18].

Table 3.8: Mean (standard deviation) of muscle activations (a_{CT}, a_{TA}), subglottal pressure (P_s) and peak of VF collision pressure (P_c), for the three subjects from Dataset 3, as estimated by the CEKF. Loudness intensities levels: 1: Soft, 2: Comfortable, 3: Loud.

Subject	Level	f_0 (Hz)	a_{CT}	a_{TA}	P_S (kPa)	P_C (kPa)
NF01	1	170	0.253 (0.020)	0.332 (0.028)	0.357 (0.020)	0.502 (0.030)
	2	170	0.193 (0.015)	0.413 (0.018)	0.572 (0.021)	1.305 (0.026)
	3	220	0.400 (0.013)	0.424 (0.016)	1.259 (0.022)	3.292 (0.075)
NF02	1	180	0.199 (0.017)	0.250 (0.025)	0.463 (0.022)	1.250 (0.044)
	2	185	0187 (0.014)	0.299 (0.026)	0.691 (0.022)	1.870 (0.048)
	3	240	0.503 (0.010)	0.402 (0.017)	1.189 (0.022)	4.313 (0.064)
NF03	1	217	0.383 (0.018)	0.263 (0.026)	0.384 (0.022)	1.116 (0.088)
	2	225	0.412 (0.015)	0.329 (0.029)	0.656 (0.022)	1.861 (0.071)
	3	271	0.521 (0.011)	0.359 (0.022)	0.932 (0.022)	3.997 (0.108)

3.5 Chapter conclusions

This chapter introduced a proof of concept for using the CEKF as a viable method to link a low-order voice production model to subject laboratory measurements even when the simultaneous recording of phonation signals was not available. The study demonstrated that including additional information in the model as state constraints the Bayesian framework yields a performance similar to that achieved using simultaneous recordings of GAW and glottal airflow. This new approach expands the applicability of Bayesian inference to laboratory scenarios where only one of these recordings is available. The general trends observed in muscle activation and vocal fold collision pressure inferred by the model align with the physiological behaviors of voice production. However, confidently asserting the reliability of these estimations remains challenging, as measurements of these vocal function features are cumbersome to obtain in clinical practice.

The CEKF represents a significant contribution to the state of the art in Bayesian inference for numerical voice production models. This method efficiently correlated the low-order voice production model with laboratory recording datasets, using solely HSV in experiments 1 and 2, and GVV in experiments 3 and 4. In experiments 1 and 3, the inferred states U_g and A_g were respectively validated, with RMSE metrics comparable to previous work that used two simultaneous observation states [51]. Experiment 3 demonstrated a positive correlation

between a_{CT} and fundamental frequency variation, aligning with findings reported in the literature [111, 83]. In experiment 4, it was observed that vocal fold collision pressure directly correlated with P_s and increased with loudness increments, consistent with previous studies [57]. These results confirm the initial hypothesis that incorporating constraints based on prior physiological knowledge of the phonation process enhances Bayesian state inference in voice production models. Furthermore, they demonstrate the potential of the low-order model to reproduce different ranges of pitch or loudness through the adaptation of control variables such as muscle activation.

Although the promising results of the CEKF method pave the way for applying the numerical lumped-element voice production model in clinical settings, the direct application of Bayesian estimation from the ACC signal remains unfeasible. The current constrained extended Kalman filter approach faces several challenges when processing ambulatory data and relying solely on the ACC as the observation source. These challenges include the computational cost of processing large data volumes, the necessity for data fusion across different recording sessions, the requirement for online estimation of model covariance, and the integration of a time-domain neck skin model for the ACC sensor within the voice production model. In this context, the next chapters describe a more direct solution for the estimation of vocal function parameters from the ACC that uses machine learning and voice modeling tools.

Chapter 4

Estimation of vocal function from an accelerometer using a neural network

In this chapter, the main idea of the research is described: developing a novel method for estimating subglottal pressure, vocal fold collision pressure, and laryngeal muscle activation using neck-surface vibration signals. This approach integrates a physiologically relevant model of voice production with advanced machine learning tools. The process begins with the training and testing of a neural network regressor, which utilizes simulations from a voice production model based on a symmetric triangular body-cover model of the vocal folds. Subsequently, the effectiveness of the method is evaluated by comparing estimates of subglottal pressure with reference values from two laboratory datasets. These datasets include synchronous measurements of oral airflow, intraoral pressure, and signals from both a microphone and an accelerometer sensor. This study encompasses a range of voice types, including healthy voices and those affected by disorders such as phonotraumatic vocal hyperfunction, nonphonotraumatic vocal hyperfunction,

and unilateral vocal fold paralysis. Participants in the study were asked to articulate syllable strings of /p/-vowel combinations under various conditions of loudness, vowel context, pitch, and voice quality. Although the estimates of subglottal pressure are validated using numerical simulations and laboratory data, the estimates of vocal fold collision pressure and laryngeal muscle activation are limited solely to synthetic data test sets, due to the lack of laboratory recordings containing these measurements. The methodology and preliminary results detailed in this chapter were published in two peer-reviewed journals: first in [115], focusing on Dataset 3, and then in [116], which presents the results for pathological cases.

4.1 Proposal scheme for vocal function estimation from accelerometer

Figure 4.1 provides an overall schematic of the proposed method for estimating four vocal function measures from neck-surface vibration recorded using an ACC sensor. The first analysis block results in an estimate of the unsteady glottal airflow volume velocity signal using the IBIF model [61], which has demonstrated reliability in providing aerodynamic features for the classification of vocal hyperfunction in both laboratory [100] and ambulatory [40] settings.

The second analysis block computes six features from the glottal airflow signal

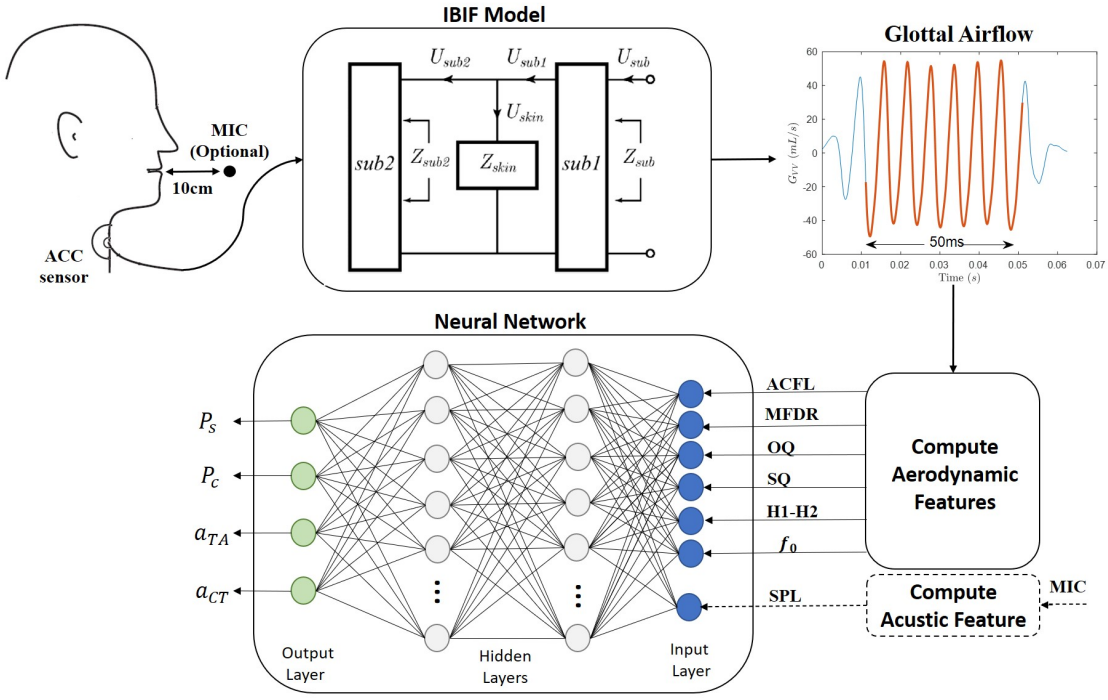


Figure 4.1: A schematic of the proposed method for the ambulatory vocal assessment based on processing the neck skin acceleration signal and a regression neural network.

(ACFL, MFDR, OQ, SQ, $H_1 - H_2$, and f_o) and an acoustic feature, SPL. This last feature can be estimated either directly using an acoustic MIC in the laboratory setting or through a log-log mapping between the root-mean-square magnitude of the ACC signal and SPL [29]. Descriptions of each feature can be found in Table 4.1.

These seven features are utilized as inputs for the NN, which was trained using a dataset based on a voice production model. This network estimates the vocal function parameters of interest, which include P_s, P_c, and the normalized muscle

activation levels (a_{CT} and a_{TA}).

Table 4.1: Description of aerodynamic features extracted from the glottal airflow signal and acoustic sound pressure level extracted from the microphone or accelerometer signal.

Feature	**Description**	**Units**
ACFL	The difference between the maximum and minimum amplitude of the AC glottal airflow (peak-to-peak) within each glottal cycle	mL/s
MFDR	Maximum flow declination rate: Negative peak of the first derivative of the glottal waveform	L/s^2
OQ	Open quotient: Ratio of the open time of the glottal vibratory cycle to the corresponding cycle period. Computed as in [40]	%
SQ	Speed quotient: Ratio of the opening time of the glottis to the closing time. Computed as in [40]	-
$H_1 - H_2$	Difference between the magnitude of the first two harmonics	dB
f_o	Fundamental frequency	Hz
SPL	Sound pressure level: dB from the RMS envelope of the acoustic signal	dB SPL

4.2 Voice production model-based dataset

The selected voice production model for the training stage is a multi-physics scheme that includes a triangular body-cover model of the vocal folds, controlled by the coordinated activation of five intrinsic laryngeal muscles. This model was proposed by Alzamendi et al. [60] and detailed in Chapter 2, Subsection 2.2.2. It was chosen due to its flexibility and its physically and physiologically relevant approach to covering various phonatory conditions, both normal and disordered [60]. This voice production model was employed to generate 110,000 Monte Carlo simulations of sustained phonation. The simulations included a wide variation of the model control parameters such as lung pressure (P_L) and muscle activation levels (a_{CT}, a_{TA}, a_{LCA}, a_{IA}, and a_{PCA}). The control model parameters and their respective variation ranges are detailed in Table 4.2.

Subsequently, from each simulation lasting 800 ms, the final 50 ms of the simulated glottal volume velocity (GVV) signal was isolated to remove transient artifacts. The signal was then filtered with a low-pass filter (LPF) at 1100 Hz and a high-pass filter (HPF) at 60 Hz to match the typical frequencies of laboratory recording signals. Following this, six aerodynamic features (ACFL, MFDR, OQ, SQ, $H_1 - H_2$, $f0$) were computed from the simulated glottal airflow signal. The simulated sound pressure at the lips ($Pout$) was used to determine SPL. Within the same 50 ms simulation windows, the mean of the P_s signal and the mean peak

Table 4.2: Range and increment step for control parameters in the numerical voice production model considered for building the synthetic dataset

Parameter	**Range**	**step**	**unit**
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	500 - 2000	150	Pa

of the P_c were computed. The selected normalized muscle activation levels of the a_{CT} and a_{TA} were also recorded. Samples that did not meet clinical registration criteria, such as an ACFL below 30 mL/s or f_0 outside the 120–400 Hz range, were discarded. In total, the synthetic dataset comprises 13,000 samples.

4.3 Laboratory recordings with reference to subglottal pressure

The reference values for subglottal pressure were derived from two databases, each containing *in vivo* laboratory recordings. These recordings included data from IOP, OVV, MIC, and ACC. Recordings in both laboratory databases were

conducted within a sound-treated environment to ensure signal integrity. A licensed speech-language pathologist rigorously assessed the vocal health status of the subjects through laryngeal videostroboscopic examination and a clinician-administered Consensus Auditory-Perceptual Evaluation of Voice assessment [117]. Informed consent was obtained from all the participants in this study, and the experimental and clinical protocols were approved by the Institutional Review Board of Mass General Brigham at Massachusetts General Hospital.

These datasets served as a testing platform for evaluating subglottal pressure estimates obtained through the regression neural network, utilizing accelerometer data. However, validation of vocal fold collision pressure and laryngeal muscle activation estimates was omitted due to the challenges of obtaining these measurements in a laboratory. It is important to note that, during this initial stage, the dataset was not used to train the neural network. Below, descriptions of each of these databases are provided.

4.3.1 Dataset 3

This *in vivo* laboratory dataset was previously analyzed in [6, 8, 100]. The data correspond to a group of participants composed of 79 adult females with no history of voice disorders. The mean age was 29.6 with a standard deviation (SD) of 13.0 years old. Study staff instructed each participant to repeat strings of /pæ/ syllables in three loudness conditions (comfortable, loud, and soft). Although

subjects were instructed to maintain a constant pitch and loudness within each syllable string, they were free to choose levels that were most natural for them without any prescribed levels of absolute pitch and loudness.

Recordings consisted of the simultaneous acquisition of acoustic pressure obtained with a condenser MIC (MKE104, Sennheiser, Electronic GmbH, Wedemark, Germany) placed 10 cm from the lips and having full bandwidth in the range of 0-6 kHz, OVV sensed by using a circumferentially vented pneumotachograph mask (PT-2E, Glottal Enterprises, Syracuse, NY) with a bandwidth of approximately 1.1 kHz, IOP measured with an oral catheter passed between the lips and connected to a low-bandwidth pressure sensor with an effective bandwidth of approximately 80 Hz [8], and ACC (BU-27135; Knowles Corp., Itasca, IL, USA) placed on the anterior neck surface halfway between the thyroid prominence and the suprasternal notch [61].

All signals were sampled at 20 kHz/16 bits (Digidata 1440A, Axon Instruments, Inc.), low-pass filtered at 8-kHz cutoff frequency (CyberAmp Model 380, Axon Instruments, Inc.), and calibrated to physical units [8]. Figure 4.2 displays a snapshot of synchronized in-laboratory waveforms from the /pæ/ task.

The preprocessing data consisted in:

- Digitally Filtered: Signals obtained from the ACC and pneumotachograph mask were low-pass filtered at 1100 Hz with a 10th-order Chebyshev Type II filter and decimated to 8192 Hz. Then, a fourth-order Butterworth HPF

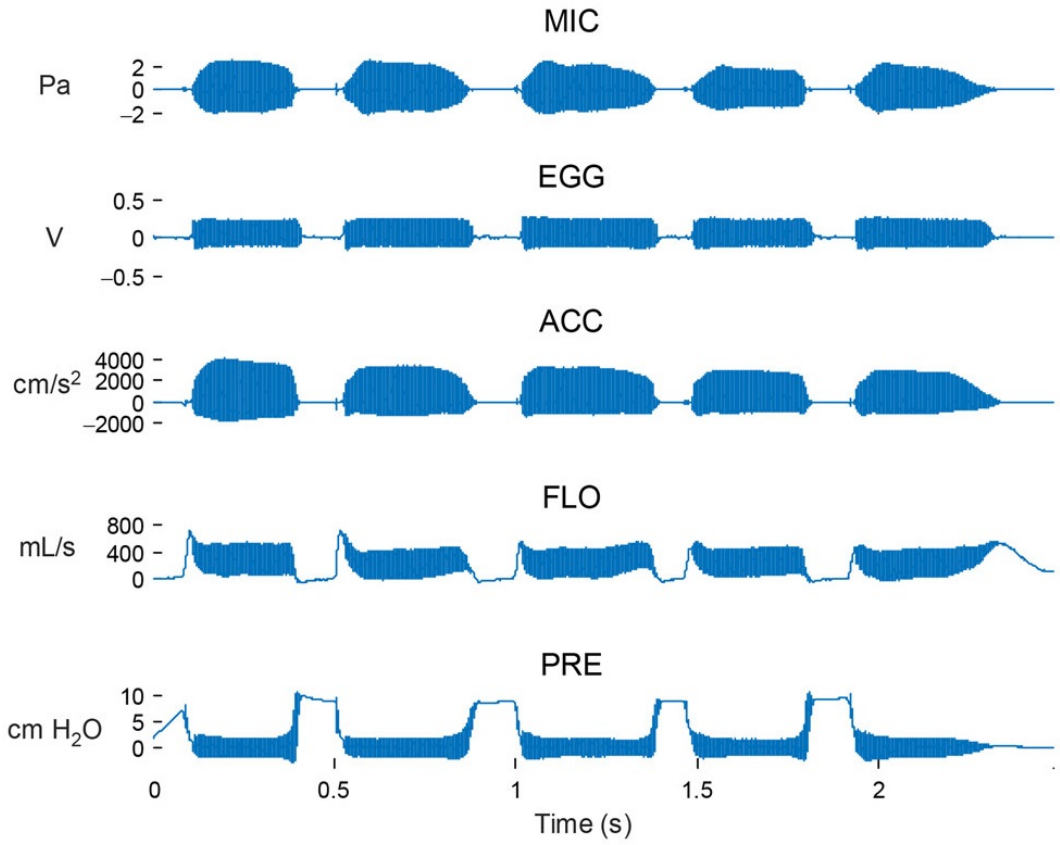


Figure 4.2: An example of the repeated /pæ/ gesture for one female participant in Dataset 3. Synchronized recordings are made of signals from an acoustic microphone (MIC), electroglottography electrodes (EGG), accelerometer sensor (ACC), high-bandwidth oral airflow (FLO), and intraoral pressure (PRE) [6].

with a cutoff frequency of 60 Hz was used to remove low-frequency components. The IOP signal was low-pass filtered at 80 Hz with a fifth-order Butterworth filter and then decimated to a 256 Hz sample rate. All filters were applied bidirectionally to achieve zero-phase distortion phase [118].

- Subglottal pressure: Reference values were obtained from IOP signals following [8]. The three middle syllables in each /pæ/ string were selected for the analysis so that the initial and final portions were disregarded to avoid any evident transient dynamics. The estimated subglottal pressure was the average of these three-syllable values. Three reference measures per participant for comfortable, loud, and soft loudness conditions were obtained. Thus, a total of 237 /pæ/ tokens were obtained from this dataset.

- OVV-Based Glottal Airflow: This signal was obtained using a common inverse filtering technique [107, 106]. Each single-notch filter was applied to a 50 ms stable portion of the middle /pæ/ string. The center frequency of the filter was determined following an optimization procedure developed by [8].

- ACC-Based Glottal Airflow: This signal was estimated using the IBIF scheme [61, 40]. This method uses an acoustic transmission line model and a calibration step to obtain a set of subject-specific parameters corresponding to the neck-skin surface, length of the trachea, and accelerometer position [61, 40, 104]. These parameters are determined by minimizing the waveform error between the OVV-based glottal airflow (reference signal described previously) and the inverse filtered neck-skin ACC signal via a particle swarm optimization [119].

- Aerodynamic Features: The middle 50 ms of the glottal airflow signal estimated from IBIF was selected to compute the six acceleration-based aerodynamic features (ACFL, MFDR, OQ, SQ, $H_1 - H_2$ and f_o) detailed in Table 4.1.

- Acoustic Feature: Even though SPL can be computed from the ACC signal using regression methods [29], this study utilized the synchronous microphone signal to avoid introducing any additional estimation errors at this stage.

4.3.2 Dataset 4

This laboratory study involved participants divided into four distinct groups: ten patients with PVH, ten with NPVH, ten with UVFP, and twenty-six individuals without any history of voice disorders (control group). Detailed demographic information for each group is provided in Table 4.3. Participants were instructed to produce /p/-vowel syllable strings, modulating their loudness from loud to soft, across three distinct vowel contexts: /pa/, /pi/, and /pu/. In contrast to laboratory database 1, this method of eliciting /p/-vowel pairs with progressively decreasing loudness facilitated a more comprehensive collection of the spectrum of P_s. Additional details regarding this database can be found in [36, 34] for the control groups and in [35, 116] for the pathological cases.

Table 4.3: Comparative demographic statistics across patient groups and vocally typical control group in laboratory dataset 2.

Group	**Female**	**Male**	**Mean (SD) Age** (Year)	**Age Range** (Years)
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

The acoustic signal was captured using a head-mounted condenser microphone (ME 102, Sennheiser Electronic GmbH, Wennebostel, Germany), positioned 15 cm from the lips of the participants. OVV and IOP signals were recorded via an aerodynamic assessment system, comprising a pneumotachograph mask (Glottal Enterprises, Syracuse, NY, USA) and dedicated sensors for oral airflow (PT-2E) and intraoral pressure (PT-75), both from Glottal Enterprises. These signals were sampled at 20 kHz with 16-bit quantization (Digidata 1440A, Axon Instruments), following their processing through an analog anti-aliasing LPF with an 8 kHz cut-off frequency (CyberAmp Model 380, Axon Instruments, Union City, CA, USA). The neck-surface vibration signal was recorded using a miniature accelerometer sensor (BU-27135; Knowles Corp., Itasca, IL, USA), secured halfway between

the thyroid prominence and the suprasternal notch using hypoallergenic double-sided tape (Model 2181, 3M, Maplewood, MN, USA). The ACC signal underwent sampling at 11,025 Hz with 16-bit quantization, facilitated by an Android smartphone. Ultimately, the signals were calibrated into physical units, adhering to the methodology outlined in [116]. Figure 4.3 shows example waveforms and spectrograms of oral airflow, intraoral pressure, acoustic microphone, and accelerometer signals, which were calibrated to units of milliliters per second (mL/s), centimeters of water (cm H_2O), pascals (Pa), and vibration acceleration (cm/s^2), respectively.

The microphone, accelerometer, and pneumotachograph mask signals were segmented by identifying sounding/silent intervals in the microphone signal using Praat version 6.0.30 [120]. Similar to dataset 3, each vowel segment of the OVV signal was low-pass filtered at 1100 Hz to match the bandwidth capacity of the pneumotachograph mask. Additionally, the IOP signal underwent low-pass filtering at 80 Hz using a fifth-order Butterworth filter to remove harmonic information.

For this dataset as well, glottal airflow, based on OVV, was obtained through standard inverse filtering [107, 106], while ACC-based airflow was derived using the IBIF method. The six aerodynamic features and SPL were obtained for the middle 50 ms segment of the glottal airflow signal estimated from IBIF and the calibrated MIC signal, respectively. Reference values for subglottal pressure were derived from IOP signals, resulting in a total of 15,160 /pæ/ tokens collected

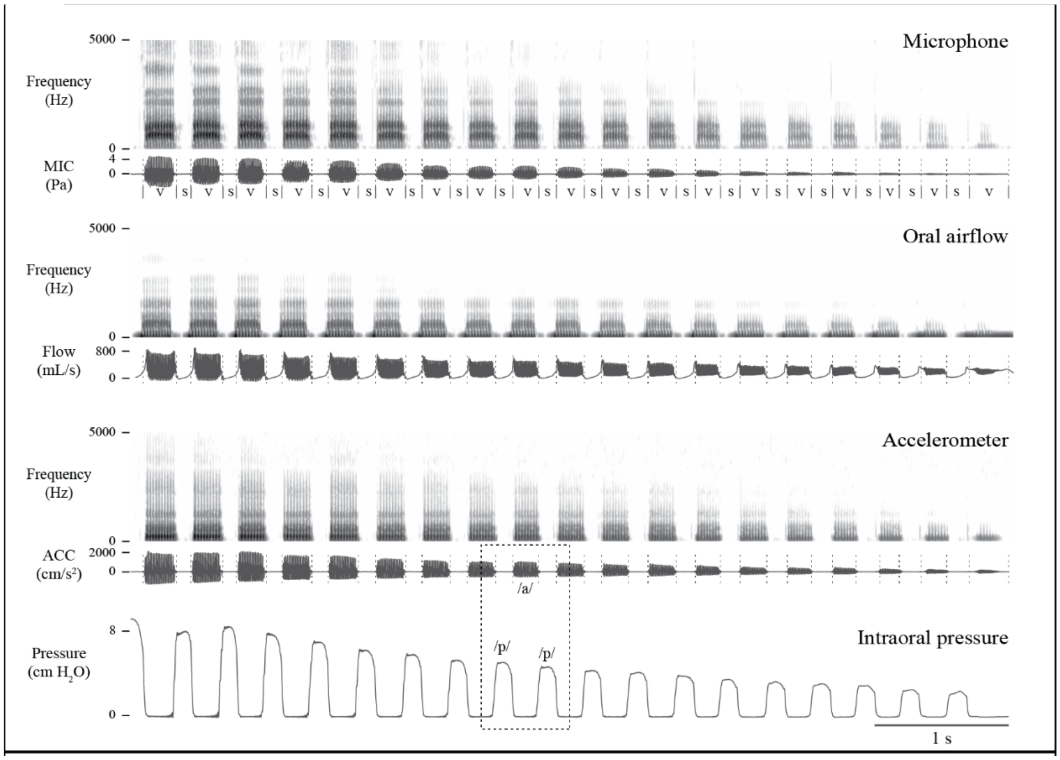


Figure 4.3: An example of the repeated /pæ/ gesture with descending loudness for one male participant in Dataset 4. Time-aligned signals from the acoustic microphone, neck-surface accelerometer, and intraoral pressure sensor, along with the Praat TextGrid tier (S = silence, V = vowel) [36].

among the four groups in the dataset.

4.4 Neural network architecture and training

A supervised machine learning framework for regression was implemented based on a multi-layer NN [121]. The network consisted of an input layer with

seven aerodynamic and acoustic features (ACFL, MFDR, OQ, SQ, $H_1 - H_2$, f_o, and SPL), an output layer comprising the four target vocal function measures (P_s, P_c, a_{TA} and a_{CT}), and interconnected hidden layers with a 10% dropout to avoid overfitting. Each neuron within the hidden layers had adjustable weight and bias parameters that combined with the outputs of the preceding layer to activate a rectified linear unit function; then, the resulting activation served as input for the next layer [122]. The number of neurons for each layer was investigated as a function of model performance against both numerical and experimental data. The training stage updates the weights and biases using the Adam optimization algorithm [123] with a learning rate of 0.001.

The NN regression models were trained following the scheme shown in Figure 4.4. For this purpose, the synthetic voice dataset described in Section 4.2 of this chapter was used. Similar approaches have been adopted by other authors using different sensing modalities, i.e., high-speed videoendoscopy [58] and MIC sensors [59] in *ex vivo* experimental validation platforms. Using synthetic data for training addressed the lack of comprehensive and massive *in vivo* human datasets with thousands or even millions of conditions.

As suggested by Gomez et al. [58], the training data should resemble the empirical distribution of the population-based aerodynamic and acoustic feature set. Figure 4.5 displays the normalized histogram of features for the synthetic data (in blue) and laboratory data (in red). It is noticeable that the feature ranges

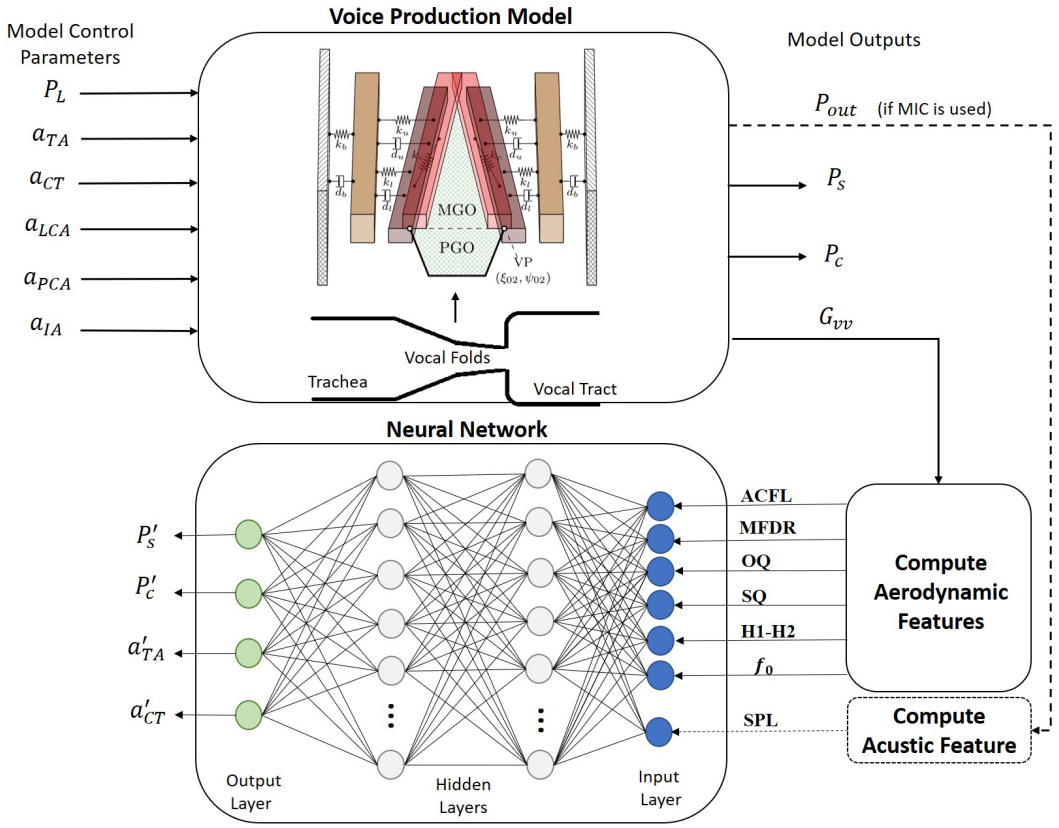


Figure 4.4: A schematic for the proposed training procedure. A regression neural network is built for mapping accelerometer-based vocal features into clinically relevant estimates for subglottal pressure, subglottal collision pressure, and laryngeal muscle activation levels of the TA and CT muscles. Training data are produced from a numerical voice production model.

and distributions for both clinical and synthetic datasets agree, except for SPL and P_s, where the ranges are noticeably dissimilar (as seen in the histograms with attenuated red color). Two bias corrections were considered for these components. First, as the SPL for the voice production model is obtained at the lips, the SPL

value was corrected to match the 10 cm mouth-to-microphone recording distance considered in the clinical recordings, yielding a -28.5 dB correction factor [124]. In addition, histograms of P_s suggest that the physiological voice synthesizer yields higher values for this measure. It is possible that sub and supra-glottal tract propagation losses and the losses at the glottal boundary were not sufficiently high, thus amplifying source-filter interactions and raising subglottal pressure. This bias has motivated subsequent exploration and model developments. However, to address the need to correct for the difference in P_s in this study, a bias correction was applied by taking the differences between the mean of clinical and synthetic P_s values, thus leading to a -3.37 cm H_2O offset.

4.5 Results

Validation with human data is the gold standard for assessing the ability of the neural network regression scheme to represent *in vivo* data. However, direct measurement of certain physiological measures of vocal function is not feasible. An advantage of using a voice production model to train a neural network is that it allows us to estimate vocal function measures that are difficult to measure in practice. This is the case for vocal fold collision pressure and intrinsic muscle activation. Therefore, the assessment of the estimates of subglottal pressure is described in terms of test sets from numerical simulations and laboratory data.

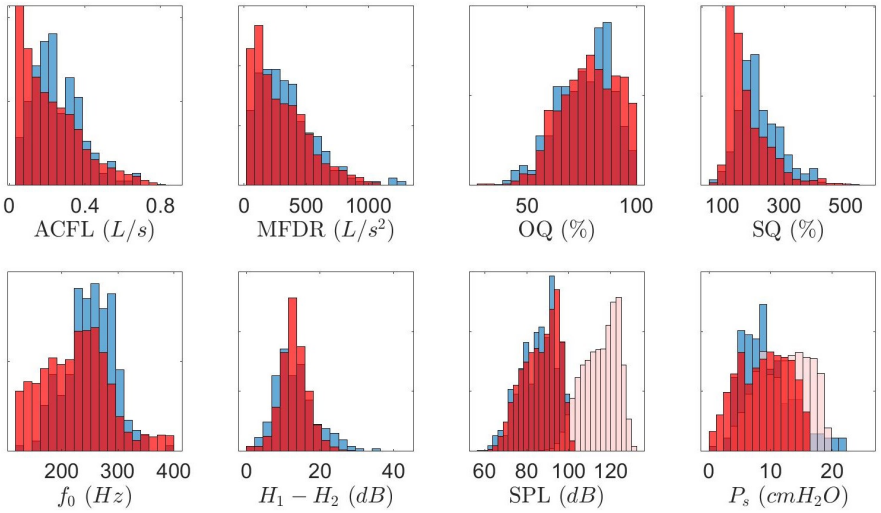


Figure 4.5: Normalized histogram illustrating vocal features from the clinical dataset (in blue). Superimposed are histograms for the synthetic dataset (in red), demonstrating model matching. Additional histograms (in light red) show bias correction for synthetic SPL and Ps.

In contrast, the estimates of vocal fold collision pressure and laryngeal muscle activation are evaluated using only a synthetic data test set.

4.5.1 Estimation of subglottal pressure in synthetic and laboratory Dataset 3

Several NN architectures, with varying numbers of neurons and hidden layers, were trained for two distinct cases. Case I involved six glottal aerodynamic features as the input layer for the NNs, specifically those described in Table 4.1:

ACFL, MFDR, OQ, SQ, f_o, and $H_1 - H_2$, all extracted solely from IBIF. In contrast, Case II employed all seven features listed in Table 4.1 for the NN input layer. To facilitate this, synthetic training data were subjected to min-max normalization and randomly selected, comprising 80% of the total simulations. The testing phase utilized the remaining 20% of synthetic data and the clinical dataset 3, aiming to identify the models that provide the most accurate estimation of subglottal pressure. To evaluate the regression performance during both the training and validation stages, and for comparison with prior studies, [47, 59, 125], the mean absolute error (MAE) and RMSE metrics were utilized.

It is important to note that all neural networks were trained using 100 epochs. This criterion was chosen to ensure the convergence of the models. Figure 4.6 displays the mean squared error versus epochs for both the training and validation phases of the simplest and the most complex architectural models. These curves demonstrate the convergence of the training process, with the simplest regression model showing more rapid convergence. However, at around 100 epochs, the error stabilizes, indicating that training progress for both architectures plateaus. A similar trend was observed across all tested configurations. Another point to emphasize is the absence of overfitting, as evidenced by the simultaneous and monotonic decrease in both training and validation errors. This suggests that the network effectively learns the structure of the observed data and can accurately infer from the validation data. An indication of overfitting would be a scenario

where the training error decreases while the validation error either remains constant or increases.

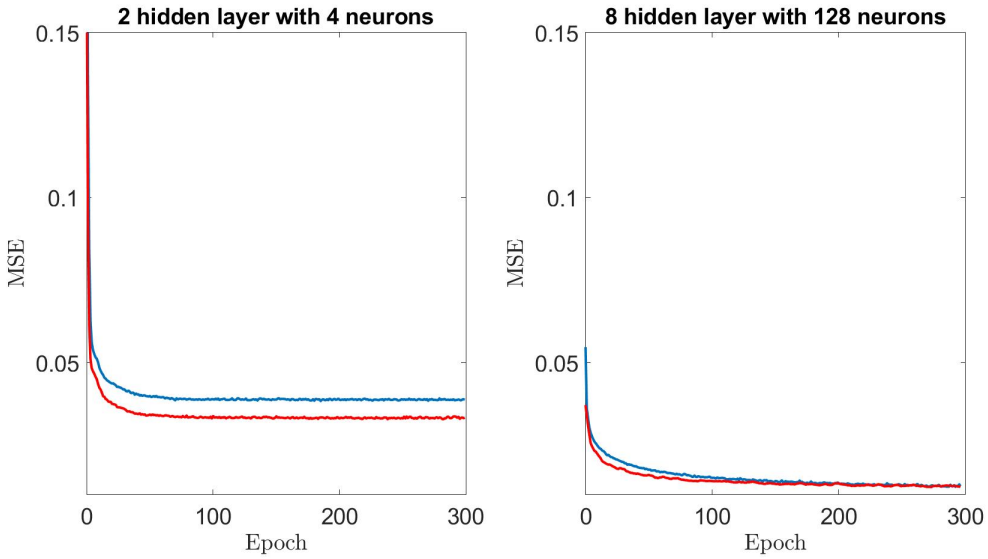


Figure 4.6: Mean Squared Error (MSE) versus epoch for training (in blue) and validation (in red) across two neural network architectures. On the left are two hidden layers with four neurons each. On the right are eight hidden layers with 128 neurons each.

The MAE and RMSE values describing P_s estimates for the different architectures are presented in Tables 4.4 and 4.5 for Cases I and II, respectively. Both tables show results from synthetic and clinical tests. In Cases I and II, the addition of more hidden layers and neurons per layer yields improved subglottal pressure estimation in synthetic data tests. For example, in Case I, the MAE decreased from 1.98 cm H2O to 0.93 cm H2O when comparing the simplest architecture (2 hidden layers with 4 neurons) to a more complex one (4 hidden layers with 128

neurons). Similarly, in Case II, the MAE decreased from 1.84 cm H2O to 0.78 cm H2O for architectures of comparable complexity. This represents a reduction of more than 50% in MAE for both cases. A similar trend was observed with RMSE. The improvement can be attributed to the fact that both the training and testing data were derived from the same voice production model. Therefore, more complex NN models seem to more efficiently capture the non-linear mechanisms of the model, as suggested by [59] in the context of training and testing with synthetic data from the same source. However, for NN architectures comprising more than six hidden layers with 128 neurons each, both the MAE and RMSE for synthetic data increased, indicating that a deeper NN does not necessarily enhance the estimation of subglottal pressure in this context.

On the other hand, for the laboratory validation of subglottal pressure, the opposite trend for MAE as a function of the NN architecture complexity was observed. In Case I, MAE increased from 2.23 cm H_2O to 3.17 cm H_2O for an increasing complexity from the 2 hidden layers with 4 neurons to 4 hidden layers with 128 neurons model. Case II also exhibited MAE increases from 1.95 cm H_2O to 3.23 cm H_2O for the same increasing complexity in the NN architecture. These results represent an increase of 42% and 66% in MAE for Case I and II, respectively, with similar trends for RMSE. Therefore, higher NN complexity was not adequate to represent sample distribution from the laboratory dataset.

Contrasting Tables 4.4 and 4.5, it is evident that the inclusion of SPL in the

Table 4.4: MAE and RMSE for the estimated P_s as obtained using the proposed NN regression model, compared with reference measures from synthetic and laboratory test data in Case I. The input aerodynamic features for Case I include ACFL, MFDR, OQ, SQ, f_o, and H_1-H_2. Errors are reported for various NN architectures, each characterized by different numbers of neurons (N) and hidden layers (HL)

N	**HL**	**Synthetic Data**		**Laboratory Data**	
		MAE (cm H_2O)	RMSE (cm H_2O)	MAE (cm H_2O)	RMSE (cm H_2O)
4	2	1.98	2.51	**2.23**	**2.82**
8	2	1.81	2.34	2.28	2.86
16	2	1.35	1.83	2.56	3.13
32	2	1.18	1.64	2.82	3.43
64	2	1.02	1.48	2.89	3.50
128	2	0.99	1.68	2.94	3.58
128	4	**0.93**	**1.33**	3.17	3.87
128	6	0.97	1.38	3.14	3.85
128	8	1.01	1.45	3.12	3.76

input feature vector improves the estimation of subglottal pressure for all tested NN architectures. Using the best architecture for laboratory validation, a 12% reduction in both MAE and RMSE was observed. Similarly, for synthetic valida-

Table 4.5: MAE and RMSE for the estimated P_s as obtained using the proposed NN regression model, compared with reference measures from synthetic and laboratory test data in Case II. The input aerodynamic features for Case II include ACFL, MFDR, OQ, SQ, f_o, $H_1 - H_2$, and SPL. Errors are reported for various NN architectures, each characterized by different numbers of neurons (N) and hidden layers (HL)

N	**HL**	**Synthetic Data**		**Laboratory Data**	
		MAE (cm H_2O)	RMSE (cm H_2O)	MAE (cm H_2O)	RMSE (cm H_2O)
4	2	1.84	2.42	**1.95**	**2.48**
8	2	1.87	2.43	1.97	2.52
16	2	1.27	1.74	2.42	2.98
32	2	1.13	1.58	2.55	3.17
64	2	0.99	1.42	2.88	3.45
128	2	0.90	1.30	2.98	3.58
128	4	**0.78**	**1.12**	3.23	3.87
128	6	0.87	1.21	3.04	3.71
128	8	1.00	1.38	3.08	3.70

tion, the best architecture exhibited a 16% reduction in MAE and RMSE when SPL was added. These results align with previous studies [8, 65, 31] that reported a strong correlation between subglottal pressure and acoustic SPL. Although not

reported here, no significant error differences were observed when estimating SPL from either the MIC or ACC sensor.

Based on these results, the NN model with the lowest error in the laboratory data validation set was selected, featuring 4 neurons in the hidden layers and all seven input features. Figure 4.7 presents a scatter plot of the NN-estimated subglottal pressure versus the reference subglottal pressure from the laboratory data. The dashed line indicates a 1:1 correspondence between the estimated and reference subglottal pressure. The coefficient of determination (R^2) is 0.65, and the mean absolute percentage error is 24.9%. It is important to note that although the IOP data was used as the ground truth for this assessment, differences in subglottal pressure estimates from IOP and direct measurement via tracheal puncture have been reported to be in the range of 5% [126]. Additionally, interpolation between the peaks of the pulses can lead to a 12% error [127].

The predicted subglottal pressure in this study is comparable to that obtained in previous studies. Table 4.6 summarizes the lowest MAE for predicted subglottal pressure from laboratory data and compares it with those from other studies. The first two studies reported mean absolute errors lower than 2 cm H_2O, similar to the NN implementation. However, it is important to highlight that NN predictions were obtained using a neck-surface accelerometer and tested against *in vivo* human data. In contrast, these studies utilized porcine [58] and human excised larynx experiments [59]. Notably, when the NN regressor proposed in

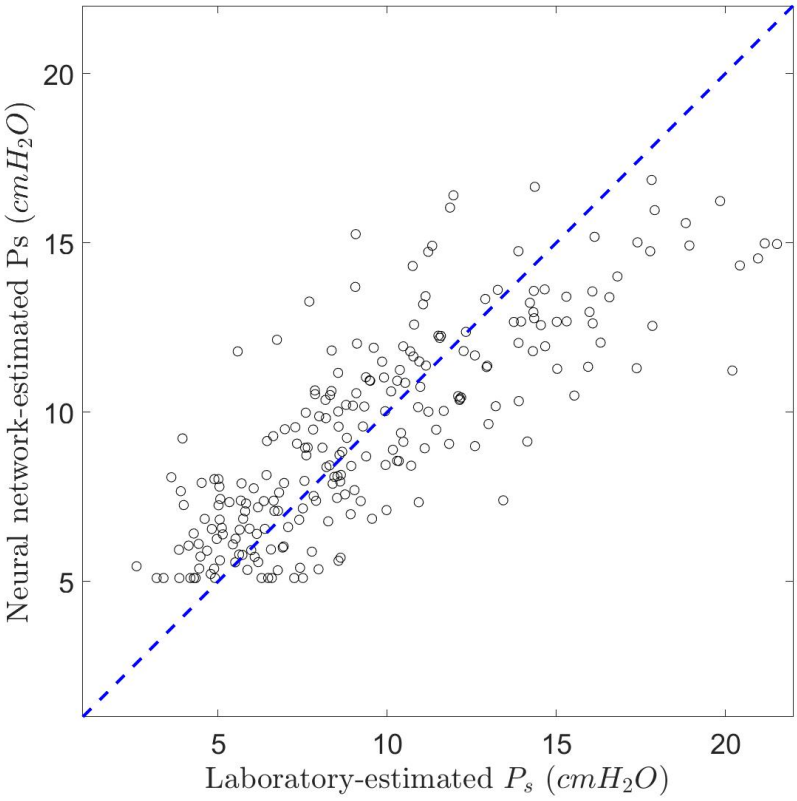


Figure 4.7: Comparison of laboratory-estimated subglottal pressure with corresponding estimates from the trained neural network (featuring 2 hidden layers, 4 neurons in each layer, and 7 voice features). The coefficient of determination (R^2) is 0.65. The dashed line represents the theoretical 1:1 perfect match.

[59] was applied to an *in vivo* single subject study [125], their MAE increased to more than double their preliminary results. The fourth method corresponds to an empirically derived formula (Eq. 2.3) [31]. This equation correlates subglottal pressure with measurements of SPL and *fo*. Applying this formula to the labo-

ratory data (237 tokens) in the current study, the mean absolute error was 2.11 cm H2O. The relatively good performance of such a simple formula supports the idea that simple regression architectures are adequate for predicting subglottal pressure under vocally typical conditions. However, the accuracy of these estimations is compromised when non-modal voice qualities are included, as suggested by [34, 35, 36].

Table 4.6: Comparison of the estimated P_s using the proposed NN regression model (2 hidden layers, 4 neurons in each layer, and 7 voice features) with those obtained in previous studies.

Method	**Materials**	MAE (cm H_2O)
LSTM Clasification [58]	Porcine excised Larynx vocal folds	1.98
NN Regression [59]	One excised human larynx	1.17
NN Regression [125]	Single-subject study	2.96
Empirically derived formula [31]	ACC signals from 79 normal subjects	2.11
Proposed Method	ACC signals from 79 normal subjects	**1.95**

4.5.2 Vocal fold collision pressure and laryngeal muscle activation estimation

Table 4.7 present the coefficient of determination and MAE (in physical units and as a percentage of the range) for synthetic data across four outputs (P_s, P_c, a_{CT}, a_{TA}), obtained using two NN architectures: the first with two hidden layers and four neurons, and the second with four hidden layers and 128 neurons. A comparison of both tables reveals that architectures with more layers demonstrated superior performance in estimating the target vocal function parameters for the synthetic data.

As observed earlier in the synthetic validation of subglottal pressure, increasing the complexity of the NN architecture enhances the accuracy of the estimates. This performance improvement holds for the estimates of vocal fold collision pressure and laryngeal muscle activation. However, the coefficient of determination for the a_{TA} estimation is significantly lower at 0.52, compared to R^2 values greater than 0.8 for the estimation of other measures using the NN with four hidden layers and 128 neurons. This finding suggests that certain measures, such as a_{TA}, may require even more complex NN architectures to achieve similar levels of performance. Additionally, these results may be associated with the increment step for muscle activation used in the TBCM simulation (see Table 4.2). It is suggested that decreasing the increment step for this parameter during simulation could

provide the necessary variability in muscle activation for the learning process in the NN.

Table 4.7: using a synthetic dataset with two NN architectures: 1) two hidden layers with 4 neurons each and 2) four hidden layers with 128 neurons each. The table reports values for R^2 and MAE (in physical units and as a percentage of the range). The input vector for this architecture includes seven aerodynamic measures.

NN	**Parameters**	Units	R^2	**MAE** (Units)	**MAE** (%)
1	P_s	cm H_2O	0.64	1.84	11.4
	P_c	cm H_2O	0.70	3.33	8.2
	a_{TA}	-	0.07	0.21	21.1
	a_{CT}	-	0.53	0.15	14.6
2	P_s	cm H_2O	0.93	0.74	4.7
	P_c	cm H_2O	0.92	1.70	4.2
	a_{TA}	-	0.52	0.13	13.3
	a_{CT}	-	0.84	0.07	7.1

4.5.3 Subglottal pressure estimation on pathological cases.

The NN architecture used in this experiment was selected based on its performance in Dataset 3 for estimating subglottal pressure. For Dataset 4, the

microphone was positioned 15 cm from the lips. Consequently, the SPL for the synthetic dataset was adjusted using the correction factor detailed in [124], which takes into account the SPL attenuation due to distance. Subsequently, a new NN with two hidden layers and four neurons was trained and employed to estimate subglottal pressure for this new dataset.

The RMSE and MAE in P_s estimates for the control group and each patient group are presented in Table 4.8. In this case, the errors for subjects without voice disorders increased compared to those obtained in Dataset 3. For instance, the RMSE increased from 2.48 to 2.89 cm H_2O, and the MAE rose from 1.95 to 2.31 cm H_2O. This increase in error is associated with the higher inter- and intra-subject variability in Dataset 4. As mentioned, Dataset 4 includes both female and male subjects, and for each subject, hundreds of samples were recorded under various conditions of /p/-syllable, loudness, and pitch. However, it is important to note that these results still outperformed the single-subject estimations reported in [125], and the estimation derived from the empirical formula proposed in [31]. For the latter method, the estimated subglottal pressure for the same control group was reported with an RMSE of 2.96 $\pm$ 1.42 cm H_2O [116].

In the case of the pathological groups, a significant increase in both metrics for P_s estimation is observed, with the highest error found in the UVFP group. This poorer estimation was expected since the NN was trained using the TBCM, which was primarily designed for simulating vocally typical conditions. The TBCM

Table 4.8: RMSE and MAE error metrics for subglottal pressure estimation using a neural network with two hidden layers and four neurons in Dataset 4. Results are categorized by participant groups: Control, Phonotraumatic Vocal Hyperfunction (PVH), Nonphonotraumatic Vocal Hyperfunction (NPVH), and Unilateral Vocal Fold Paralysis (UVFP).

Group	RMSE ($cm\ H_2O$)	MAE ($cm\ H_2O$)
Control	2.89 ± 0.82	2.31 ± 0.58
PVH	3.18 ± 1.73	2.43 ± 1.27
NPVH	3.51 ± 2.07	2.79 ± 1.65
UVFP	4.86 ± 2.66	4.17 ± 2.43

does not account for the anatomical and physiological changes associated with pathologies. For instance, it does not consider the asymmetrical muscle activation in NPVH and UVFP, the presence of polyps in PVH, or the asymmetrical vocal fold vibratory oscillation in UVFP [128]. These error increases were also observed with the empirical formula [31], which estimated the P_s with RMSE values of 4.10, 3.76, and 4.74 cm H_2O for PVH, NPVH, and UVFP, respectively [116].

In the results section of [116], comparisons were made between NN estimations and two subject-specific methodologies proposed in [36] and [34] (described in Chapter 2, Subsection 2.1.4). Both approaches yielded an RMSE under 2.08

cm H_2O, outperforming the NN predictions. The noted improvement in the accuracy of subject-specific subglottal pressure estimation was attributed to the elimination of inter-subject variability. This finding suggests that developing NN models, with parameters adjusted individually for each subject, could improve subglottal pressure estimation.

4.6 Chapter conclusions

The purpose of this study was to explore the combination of neural network regression networks with a voice production model to estimate physiologically relevant vocal measures, i.e., subglottal pressure, vocal fold collision pressure, and muscle activation (TA and CT) from a neck-surface vibration signal. Validation for this study was done both numerically and experimentally. Given that some of the predicted measures are difficult to obtain experimentally, only the estimates of subglottal pressure could be compared with reference estimates of mean subglottal pressure derived from the standard airflow interruption technique in the laboratory.

Both numerical and experimental validations involving normal subjects yielded reasonably accurate results. However, as the complexity of the NN architecture increased, a decrease in estimation error for synthetic data was observed, but an increase in error for laboratory data. This discrepancy can be attributed to how

the synthetic dataset was obtained. For simulations, the numerical model parameters were varied across a wide range, but without considering anatomical changes, effectively representing a reduced population under various conditions. This lack of inter-subject variability likely impacted the accuracy of subglottal pressure estimation. Furthermore, the adjustments made to correct for bias in subglottal pressure estimation suggest that losses in the subglottal and supraglottal tract models, as well as at the glottal boundary, were not adequately represented. This could potentially amplify source-filter interactions and affect subglottal pressure estimates. These observations indicate that the numerical model requires adjustments to more accurately reflect population behaviors and to assess their effects on the accuracy of the proposed approach. Despite its simplicity and these limitations, the triangular body-cover model provides a good general representation of typical sustained phonation across a wide range of subjects and conditions.

The error metrics for subglottal pressure estimations in pathological cases are noticeably higher compared to subjects without voice disorders. These results indicate that the robustness and reliability of the estimates from the proposed method depend on the ability of the selected voice production model to mimic the observed distributions in laboratory data. Even though the TBCM provides a close representation of typically sustained phonation, there is a significant gap between this numerical voice production model and the pathological laboratory dataset, especially in UVFP cases. Despite this, the results obtained in PVH

and NPVH are reasonable when contrasted with other techniques reported in the literature [125, 31]. However, compared to subject-specific linear regression approaches, the RMSE of the proposed estimation method is higher [116]. These results suggest that incorporating subject-specific adjustments into the NN could enhance the estimation of subglottal pressure.

On the other hand, the accuracy of estimates for muscle activation and collision pressure, which were assessed against synthetic data, was compromised by the simplicity of the rather shallow NN architecture initially chosen to match clinical data for subglottal pressure. When the complexity of the network is increased, the estimation accuracy for muscle activation and collision pressure improves. This finding is encouraging for the investigation of subject-specific models that can accommodate more complex neural network architectures while retaining the ability to accurately predict subglottal pressure.

These findings support that vocal function measures, such as subglottal pressure, vocal fold collision pressure, and intrinsic laryngeal muscle activation, can be estimated from accelerometer-based features using nonlinear regression. These initial results serve as a proof of concept, demonstrating that the proposed NN method is a viable option for estimating clinically relevant vocal measures that are challenging to measure directly in both laboratory and ambulatory settings. However, the model requires refinement to improve its estimations, especially in pathological cases. In this context, the next chapter introduces a methodology

based on subject-specific tuning with transfer learning. This approach aims to enhance subglottal pressure estimation from an accelerometer.

Chapter 5

Transfer learning for improving neural network estimation from a neck-surface accelerometer

In the previous Chapter, a promising method for estimating vocal function features from neck-surface vibration signals was introduced, leveraging a framework that combines a physiologically relevant model of voice production with neural network regression. This Chapter builds upon that initial work by incorporating a transfer learning strategy to refine neural network regression parameters using laboratory data. Initially, a baseline regression model was trained exclusively with simulations from a numerical voice production model, and subsequently fine-tuned through cross-validation with laboratory data. The results indicate a significant reduction in the average RMSE of subglottal pressure estimation compared to preliminary results from laboratory datasets 3 and 4. Furthermore, the use of transfer learning to develop subject-specific neural regression models is explored, applying transfer learning to fine-tune model parameters based on recordings from individual subjects. This method has resulted in an improvement of over 21% in

the average RMSE of subglottal pressure estimation, outperforming existing methods based on subject-specific calibration reported in the literature. Furthermore, the proposed subject-specific method facilitates a preliminary validation of muscle activation estimation using simultaneous recordings of OVV, IOP, MIC, and laryngeal electromyography from a case study. These findings highlight the efficacy of a nonlinear, subject-specific regression approach in improving subglottal pressure estimation from neck-surface vibration signals and illustrate the potential for extending this estimation to muscle activation. This approach aligns with the central objectives of this thesis, which aim to advance the assessment of vocal function through non-invasive techniques. The methodology, results, and discussion on subglottal pressure estimation presented in this chapter have been reported in a paper submitted to a peer-reviewed journal [129] and will be presented at the 13th International Conference on Voice Physiology and Biomechanics [130].

5.1 Proposal transfer learning scheme

Figure 5.1 presents a schematic of the proposed method for enhancing vocal function assessment from the neck skin acceleration signal through NN regression. The upper block of the schematic depicts the baseline model, consisting of an NN trained with thousands of simulations from a numerical voice production model, as detailed in Chapter 4. This baseline model effectively maps various aerodynamic

and acoustic input features (ACFL, MFDR, OQ, SQ, H1-H2, f_0, and SPL) to vocal function measures such as P_s, a_{TA}, a_{CT}, and P_c.

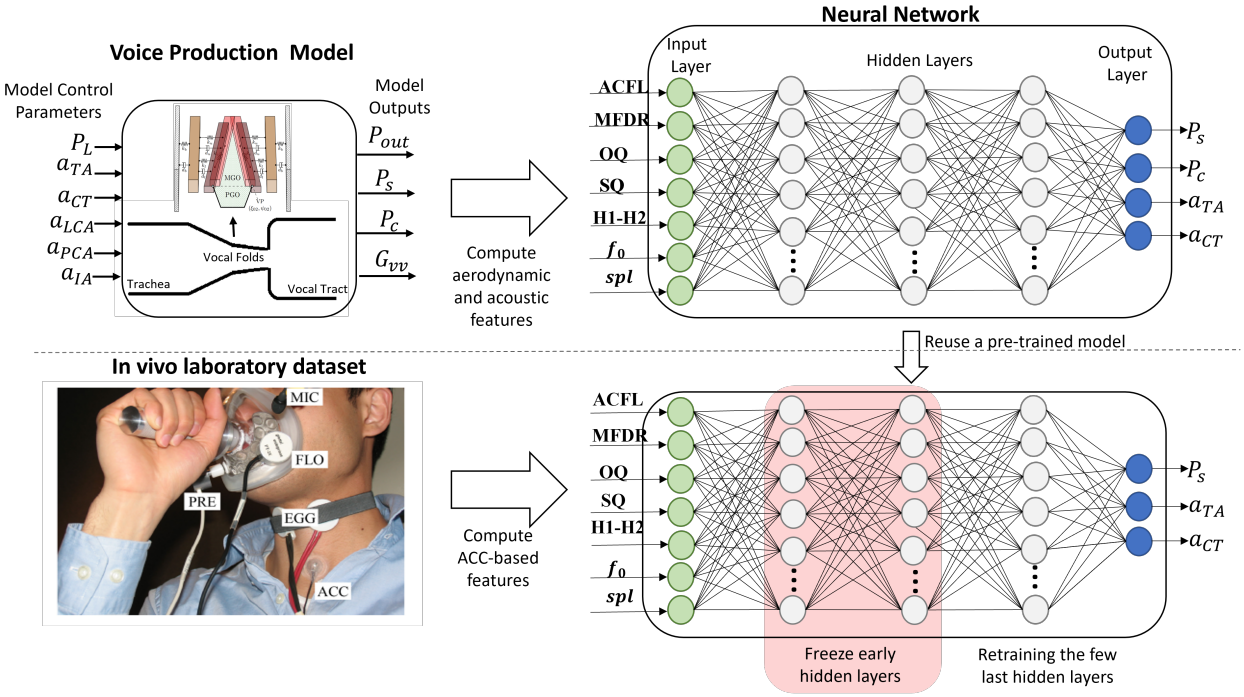


Figure 5.1: Scheme of the transfer learning procedure.

The analysis in the lower block involves fine-tuning the NN through transfer learning, utilizing *in vivo* laboratory measurements. For this process, the initial layers of the baseline model were frozen, and the subsequent hidden layers were retrained using cross-validation techniques. At this stage, the input aerodynamic features are computed from an unsteady glottal airflow signal, which is derived by applying the IBIF model to the ACC signal captured from the neck skin. Meanwhile, the SPL is derived from a calibrated MIC signal. The reference P_s is estimated based on IOP measurements, while the reference values for a_{TA} and

a_{CT} are estimated from laryngeal EMG. In this work, the validation is on these three outputs due to the absence of a laboratory dataset containing aerodynamic and acoustic measurements simultaneously with P_c measurements.

5.2 Transfer learning from simulated voice production to *in vivo* recordings

Transfer learning is a powerful machine learning technique 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 by Pan et al. [131] as:

"Given a source domain $\mathcal{D}_S$ and learning task $\mathcal{T}_S$, and a target domain $\mathcal{D}_T$ 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 given by the feature space ($\mathcal{X}$) and a marginal probability distribution ($P(X)$), represented as $\mathcal{D} = \{\mathcal{X}, P(X)\}$, and the target is composed of the label space ($\mathcal{Y}$) and the conditional probability distribution ($P(Y|X)$), represented as $\mathcal{T} = \{\mathcal{Y}, P(Y|X)\}$. Therefore, the condition $\mathcal{D}_S \neq \mathcal{D}_T$ implies that either $\mathcal{X}_S \neq \mathcal{X}_T$ or $P_S(X) \neq P_T(X)$. Similarly, $\mathcal{T}_S \neq \mathcal{T}_T$ implies that either

$\mathcal{Y}_S \neq \mathcal{Y}_T$ or $P_S(Y|X) \neq P_T(Y|X)$.

In the regression problem of this work, the synthetic and laboratory data represent $\mathcal{D}_S$ and $\mathcal{D}_T$, respectively. The P_s and muscle activations estimations are the learning task, thus ($\mathcal{T}_S = \mathcal{T}_T$). Previously [115, 64], 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 at the glottal level between sound, flow, and vocal fold tissue. Consequently, since $\mathcal{X}_S \neq \mathcal{X}_T$ and $P_S(X) \neq P_T(X)$, it follows that $\mathcal{D}_S \neq \mathcal{D}_T$. Therefore, the regression problem in this study aligns with the definition of transfer learning.

The transfer learning strategy employed in this work is known as fine-tuning, or parameter transfer [132, 133]. It involves replacing the last one or several layers of a baseline model with customized layers for the target task. During the training process, the weights of the pre-trained model are fine-tuned through continued back-propagation. The fine-tuning process adapts the model to be more specifically aligned with the details of the target learning task.

In this study, the transfer learning framework is applied in distinct scenarios. Firstly, a single regression model is refined to enable it to estimate P_s across various subjects, using laboratory datasets 3 and 4. These datasets are described in Chapter 4, specifically in Sections 4.3.1 and 4.3.2, respectively. Secondly, subject-specific NN tuning is implemented, aligning this approach with the methodologies

described in [36, 35, 34] and using laboratory dataset 4. These two approaches aim to enhance the accuracy and applicability of P_s estimation from an ACC across diverse domain adaptations. Additionally, the subject-specific tuning enables the validation of other vocal function measures, such as muscle activation, by fine-tuning the NN to a case study that includes laboratory recordings with laryngeal EMG measures for CT and TA muscles.

5.3 Laryngeal EMG Dataset 5: A case study

These laboratory recordings include time-synchronized *in vivo* recordings of OVV, ACC, IOP, and laryngeal EMG from an adult male subject with no history of voice disorders. The laryngeal EMG recordings capture muscle activity for the left and right CT, right TA, and right LCA; however, the LCA was not considered in this study. The participant gave his informed consent for this study, and the Institutional Review Board at Mass General Brigham, Massachusetts General Hospital, approved the protocol.

The in-laboratory protocol required the subject to perform various phonatory tasks at different levels of loudness and pitches, as detailed in Table 5.1, in addition to non-phonatory tasks such as coughing, swallowing water, and performing a chin press against resistance. This wide range of tasks enhances the understanding of how each intrinsic laryngeal muscle contributes to vocal fold movement during

tasks and aids in determining the maximum level of activation for each muscle. Pinpointing this maximum activation is crucial in the processing stage for normalizing muscle activation values, thereby making them comparable to those obtained in the numerical voice production model.

The equipment and sensors for MIC, ACC, and IOP used in this study were the same as those described for Dataset 4 in Chapter 4, Subsection 4.3.2. The electrodes for EMG recordings were bipolar hooked wires. The wireless EMG channels were grounded to a single large pad placed on the dorsal cervical spine. All signals were digitized at a rate of 20 kHz. The EMG signals for the TA and LCA muscles were processed through an additional CyberAmp (Model 380, Axon Instruments). The filter settings for these signals included a HPF at 1Hz, a LPF at 2000Hz, and a total gain of 5X. The EMG signals for the left and right CT muscles were collected using a 2-channel Bagnoli unit (wired), with the gain set to 1000X. This gain configuration comprised a 10X head-stage gain in addition to a 10X gain applied by the Bagnoli box. The amplified signals were then processed through the CyberAmp. Control of these channels was managed by the Axoscope software. The filter setting for these channels was an HPF at 10Hz, an LPF at 8000Hz, and a total gain of 5X.

The MIC, ACC, OVV and EMG were segmented by identifying sounding and intervals based on the frequency obtained by the microphone signal. As in Dataset 3, described in Chapter 4, Subsection 4.3.1, signals from the ACC and OVV were

Table 5.1: Phonatory tasks in laboratory Dataset 5.

Task	**Pitch**	**Loudness**
Sustained vowels /a/ and /i/ stopping with Valsalva maintaining tracheal pressure	Normal	Soft, 6. Syllables /hi, e, loud and very loud
Sustained vowels /a/	Normal	Soft, comfortable and loud
	Low and high	Comfortable
Pitch glides for vowels /a/ and /i/	Low chest to falsetto, falsetto to low chest and faster alternation of low-high	Comfortable
Plosives /pæ/	Normal	Comfortable
	Low, normal and high	Descending loudness
Syllables /hi/ on a single breath	Normal	Comfortable
Syllables /ifi/ and /afa/ on a single breath	Normal	Comfortable
Syllables /ifi/ and /afa/ strained	Normal	Comfortable
Words: counting one to ten	Normal	Comfortable

filtered using LPF at 1100 Hz using a 10th-order Chebyshev Type II filter, followed by HPF at 60 Hz with a fourth-order Butterworth filter to eliminate low-frequency components. The IOP signal was filtered at 80 Hz with a fifth-order Butterworth filter. All filtering processes were applied to the signals in both forward and reverse directions to yield zero-phase distortion and thus maintain time alignment with the other physiological signals.

The EMG pre-processing analysis involves HPF through a sixth-order Butterworth anti-alias filter at 20 Hz, followed by full-wave rectification, and smoothing using a zero-phase shift, sixth-order Butterworth filter with a cutoff frequency of 8 Hz, as referenced in [23]. Figure 5.2 depicts the resulting signal at each stage of processing for the right TA EMG, recorded during a series of/pæ/ syllables spoken at a comfortable loudness level.

The 95th percentile of the amplitude of the envelope EMG signal for each phonatory gesture was compared to identify the maximum activation level within each muscle, as shown in the bar plots of Figures 5.3 and 5.4. These figures illustrate the 95th percentile of the amplitude for the envelope EMG signals of right TA and right CT, respectively. It was observed that for the right TA, the highest value occurred during throat clearing and coughing, followed by the Valsalva maneuver. In contrast, for right CT, the greatest amplitudes were recorded for phonatory gestures associated with high pitch, aligning with measurements reported in [134]. The normalization value for activation level was then determined

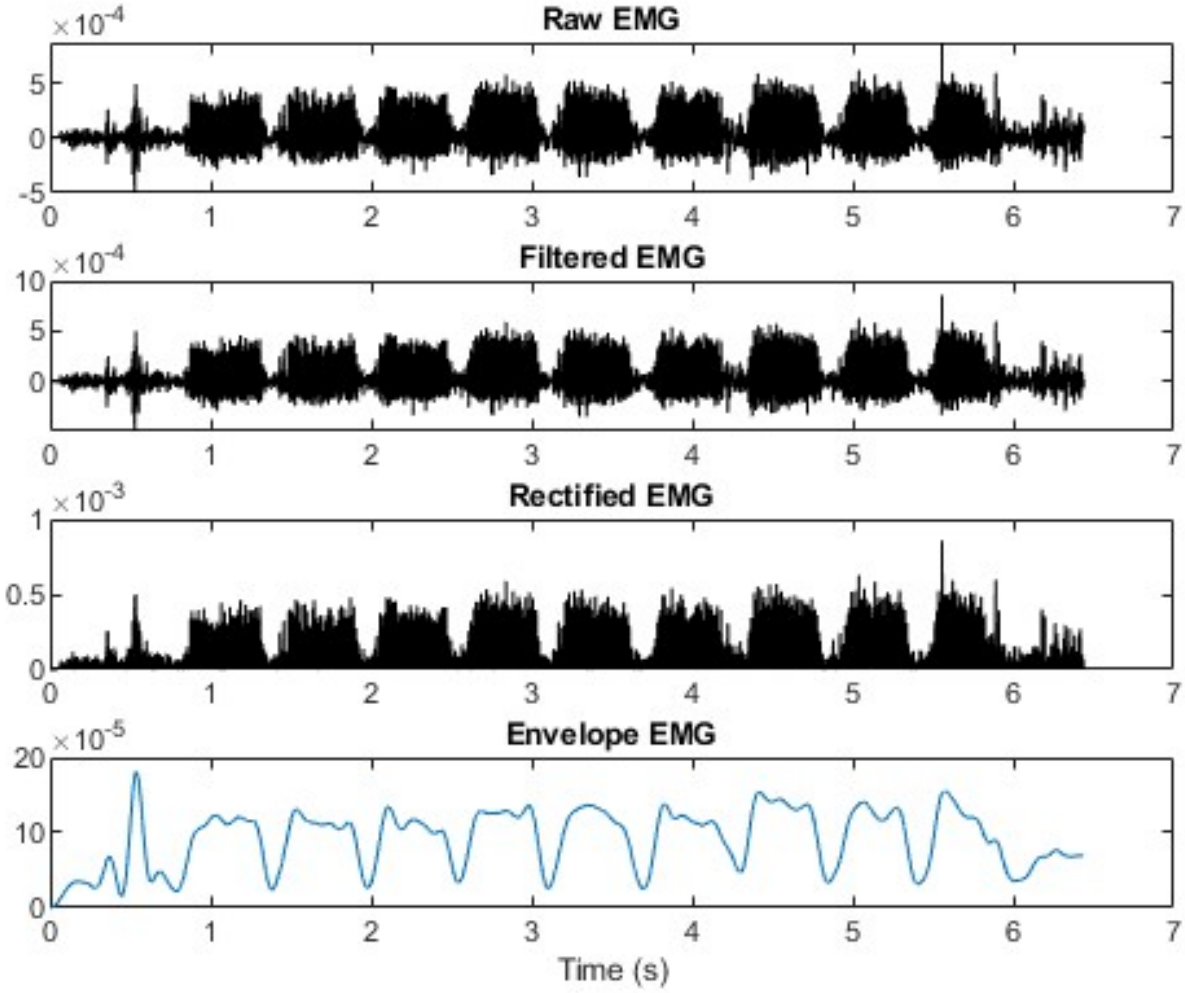


Figure 5.2: EMG signal processing before normalization for right TA during comfortable /pæ/.

as the 95th percentile across all phonation gestures for each muscle, as indicated by the red line in Figures 5.3 and 5.4. The 95th percentile was chosen over the maximum value to ensure robustness against outliers, consistency across various gestures, and statistical reliability.

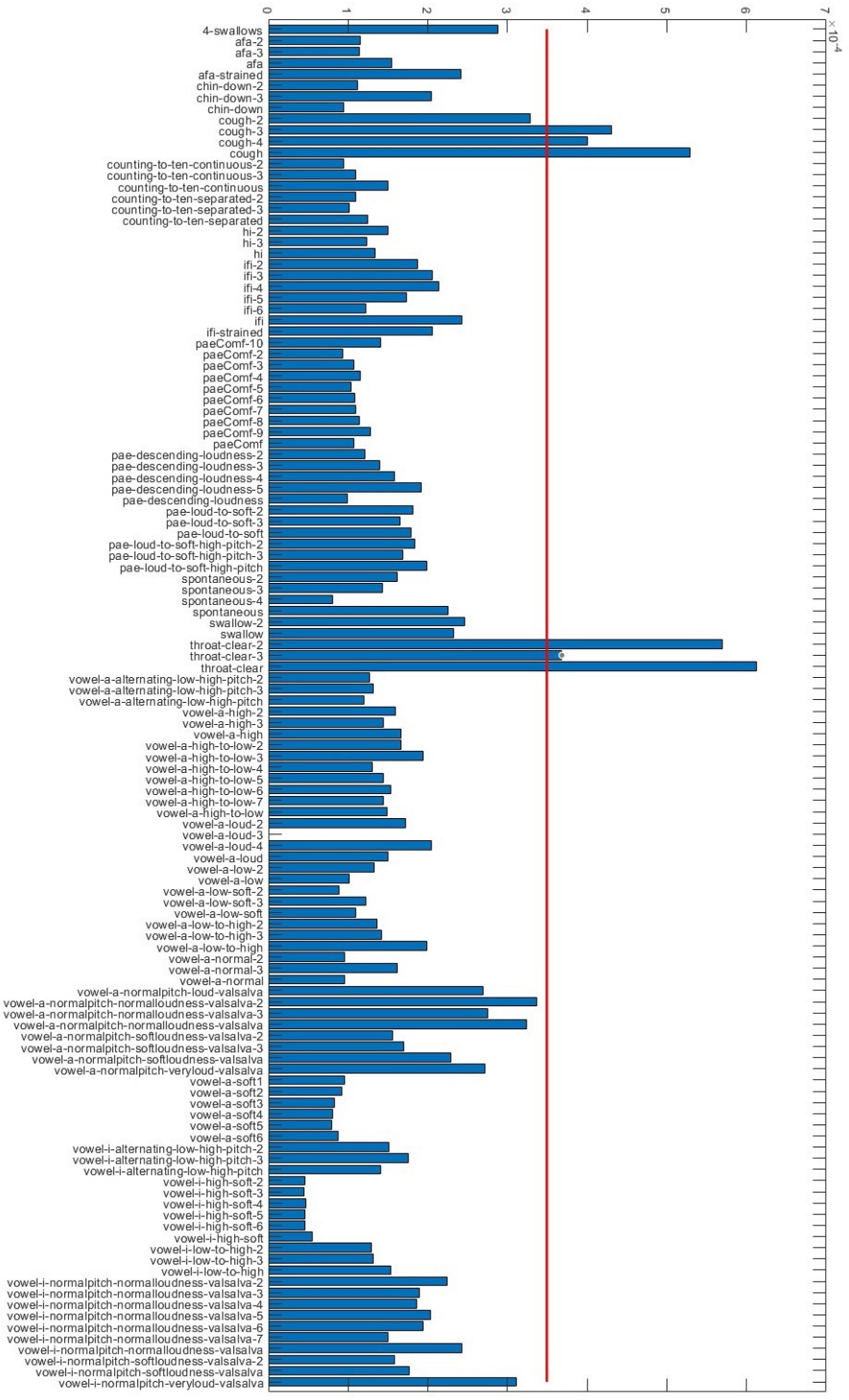


Figure 5.3: Percentile 95% for the amplitude of envelope EMG signals for right TA. The red line indicates the selected value for normalization.

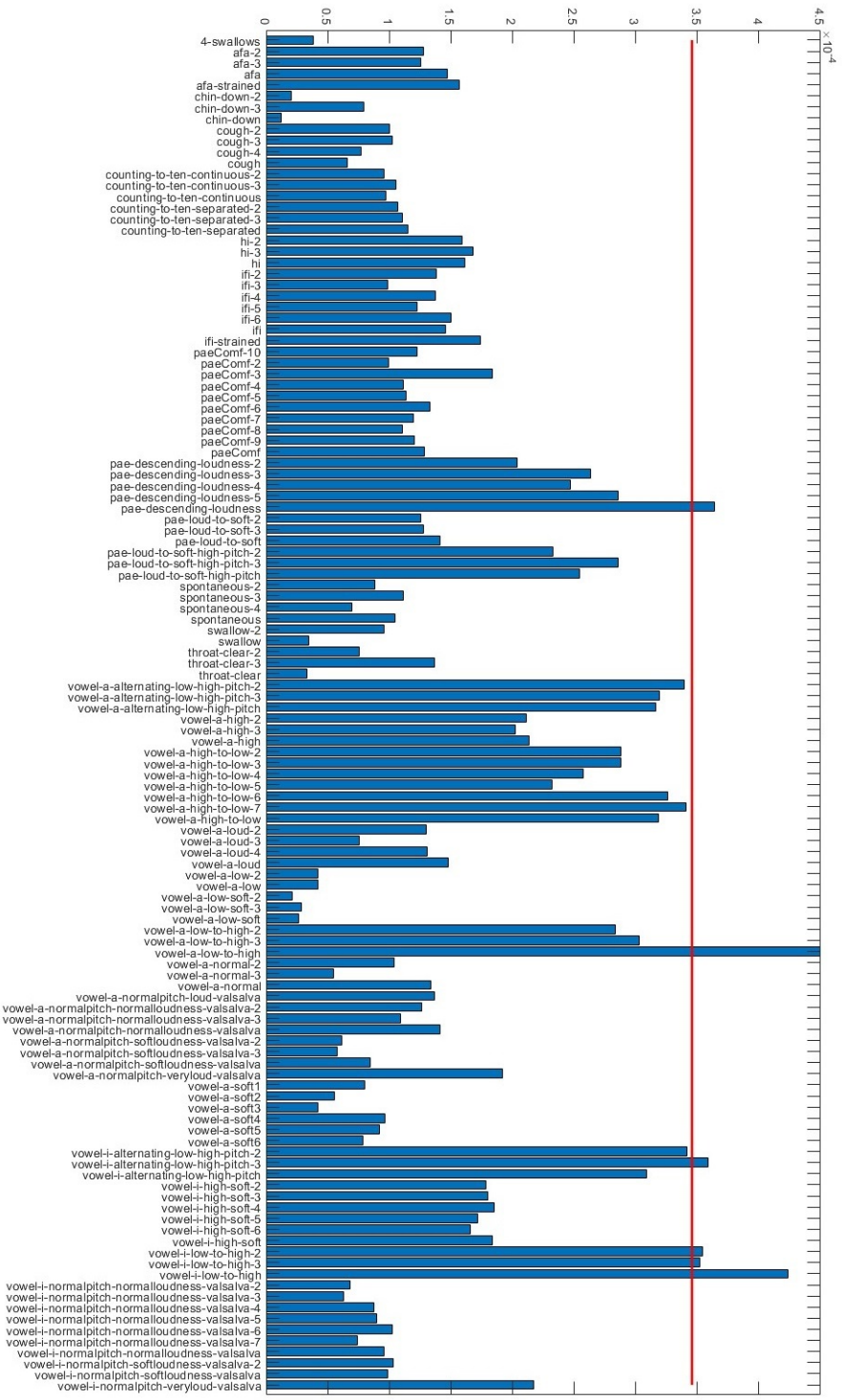


Figure 5.4: Percentile 95% for the amplitude of envelope EMG signals for right CT.

The red line indicates the selected value for normalization.

Figure 5.5 presents an example of the resulting normalized EMG signals, representing the left and right a_{CT} and the right a_{TA}. Upon normalization, it is observed that both the right and left a_{CT} exhibit similar behavior, as expected due to symmetric muscle activation in a normal subject [128]. This similarity was observed in other phonatory tasks as well; therefore, the left a_{CT} was chosen for subsequent analysis.

In the stable portion of the normalized EMG signals, specifically avoiding the onset and offset peaks observed in the vowel segment of the /pæ/ sequence, the mean value within a 50 ms window with a 50% overlap was selected to derive the muscle activation tokens for a_{CT} and a_{TA}. Then, employing windows of the same size and overlap, P_s, along with aerodynamic and acoustic features, was computed to assemble a dataset comprising 1015 samples that exclusively include /pæ/ gestures. Additionally, 1340 samples were obtained from pitch glide gestures, but these do not include reference values for P_s.

5.4 Neural network architecture and fine-tuning strategy

The baseline model, similar to the one described in Section 4.4, is built on a multilayer perceptron neural network. It features an input layer with seven features: ACFL, MFDR, OQ, SQ, H1-H2, f_0, and SPL, with the outputs being

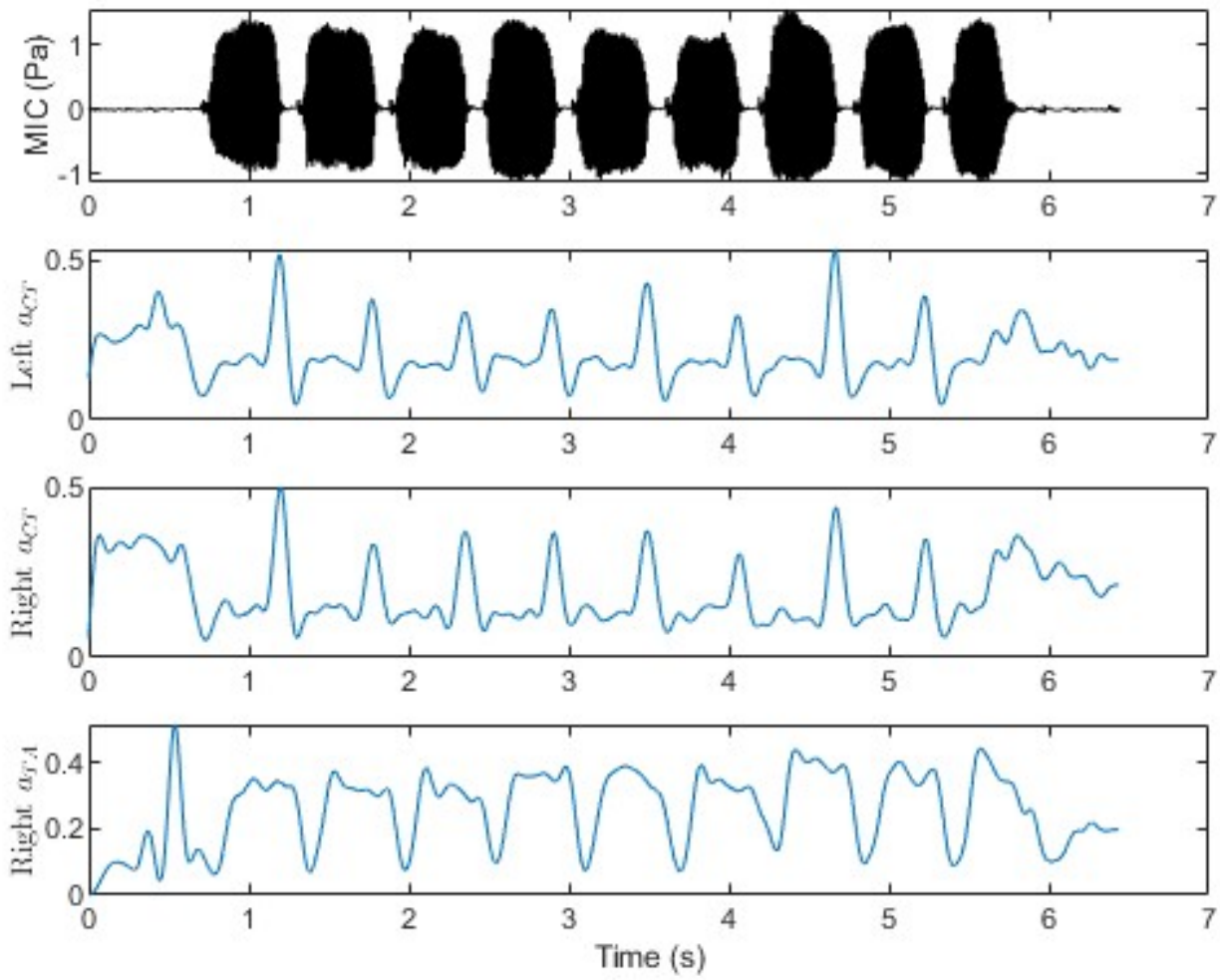


Figure 5.5: Microphone and normalized EMG signals during comfortable /pæ/.

P_s, a_{TA}, and a_{CT}. Each interconnected hidden layer incorporates a rectified linear unit activation function and is followed by a dropout layer to prevent overfitting. For these experiments, the hyperparameters of the baseline model were fine-tuned using a 5-fold cross-validation strategy applied to a dataset modeled after voice production, employing Talos [135] for optimization. The search space for the

hyperparameters is detailed in Table 5.2. The model that exhibited the best performance on the validation set across the five folds was adopted as the baseline model for all further experiments. The synthetic data utilized is as described in Section 4.2, with the exception that bias corrections for SPL and P_s were not applied in these experiments.

Table 5.2: Hyperparameters search space for 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

The baseline regression model was retrained using TL through layer freezing on laboratory datasets. Initially, the NN outputs were reduced to only P_s. The initial experiments aimed to develop a generalized model capable of estimating P_s across multiple subjects, utilizing datasets 3 and 4. Subsequently, the focus shifted to subject-specific models. In this second scenario, an NN with a single output, P_s, was utilized using Dataset 4; with dual outputs, a_{TA} and a_{CT}, from pitch glide samples in Dataset 5; and with three outputs, P_s, a_{TA}, and a_{CT}, from plosive /pæ/ gesture samples in Dataset 5. The optimal performance of the TL strategy was achieved by sequentially freezing the hidden layers. This method

enabled the system to retain high-level features from the baseline model while adapting to the specific requirements of the target models through the unfrozen layers.

The baseline model and the TL training used the Adam optimization algorithm with the MSE as the loss function. A learning rate schedule was used, initialized at 0.001. For all experiments, synthetic and laboratory data were min-max normalized. The neural networks in this study were implemented on a Google Colaboratory virtual machine, powered by two Intel(R) Xeon(R) CPUs @ 2.00GHz, using Python 3.6.9 and PyTorch library version 2.1.0.

5.5 Results

The regression performance of NN estimations is evaluated using several metrics, such as RMSE, MAE, mean absolute percentage error (MAPE), and R^2. These metrics were selected for their ability to provide a comprehensive assessment of model accuracy and to facilitate a quantitative comparison with previous studies in the context of P_s estimation. Initially, results from both general and subject-specific adapted baseline models for estimating P_s are presented. Subsequently, preliminary results for estimating muscle activation using the subject-specific adapted baseline model for a case study in Dataset 5 are discussed.

5.5.1 General neural network with transfer learning for subglottal pressure estimation

The initial set of experiments demonstrates the interest in applying a TL strategy for the refinement of the NN to adapt the estimate of the subglottal pressure to a population of speakers. Initially, the baseline NN was fine-tuned using laboratory Dataset 3, which includes data from 76 subjects without voice disorders. The evaluation of this dataset employed a stratified subject-independent 10-fold cross-validation strategy, ensuring that speakers did not overlap across different folds. This method was strategically chosen to robustly test the ability of the model to generalize across a diverse subject group. Subsequently, the baseline NN was adapted for each specific group within laboratory Dataset 4. Although this dataset has a larger number of samples, it involves a smaller number of subjects. Therefore, to assess the performance of the model for each group in this dataset, a leave-one-subject-out cross-validation approach was used. 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 available data. In these experiments, model performance with and without TL was compared to investigate whether this new approach leverages pre-trained knowledge to enhance accuracy and efficiency compared to models trained without pre-existing knowledge.

Estimating P_s from laboratory Dataset 3

The results in Table 5.3 show the error metrics for estimating the subglottal pressure when the model is trained from scratch and when fine-tuning the baseline model —originally trained using a physiological voice synthesizer— via TL with sequential freezing of its hidden layers. 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, specifically synthetic signals and laboratory recordings. Furthermore, the improvements achieved through TL, in contrast to training the model from scratch, suggest that maintaining the parameters of the first hidden layer in the baseline model contributes significantly to the robustness of the non-linear regression estimation.

The optimal performance result (achieved when the first hidden layer was frozen) demonstrates an improvement over the results shown in the previous chapter, utilizing an NN with 2 hidden layers and 4 neurons per layer, trained solely on a synthetic dataset. In that study, the subglottal pressure estimation metrics for the same laboratory data yielded an RMSE of 2.48 cmH_20, an MAE of 1.84 cmH_20, a MAPE of 24.9%, and an R^2 of 0.65. Figure 5.6, a scatter plot, contrasts the NN-estimated subglottal pressure against the reference subglottal pressure,

Table 5.3: Subglottal pressure estimation errors metrics for a neural network training using random initialization and transfer learning (TL) strategy with sequential frozen layers (FL) in Dataset 3.

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

comparing the previous estimations shown in Chapter 4, Figure 4.7 (represented by blue dots), with the current approach (represented by red dots). To ensure a comprehensive comparison, the plot illustrating the TL-based estimation compiles the validation set results from the 10-fold cross-validation. Generally, the TL estimation more closely aligns with the dashed line, representing a 1:1 correspondence between the estimated and reference subglottal pressure, compared to the previous results. Importantly, the TL refinement yields P_s estimations in ranges where the previous NN model was less effective, especially for P_s values below 5 cm H_20. This enhancement is quantitatively manifested as an increase of 0.04 absolute points in the coefficient of determination when applying TL.

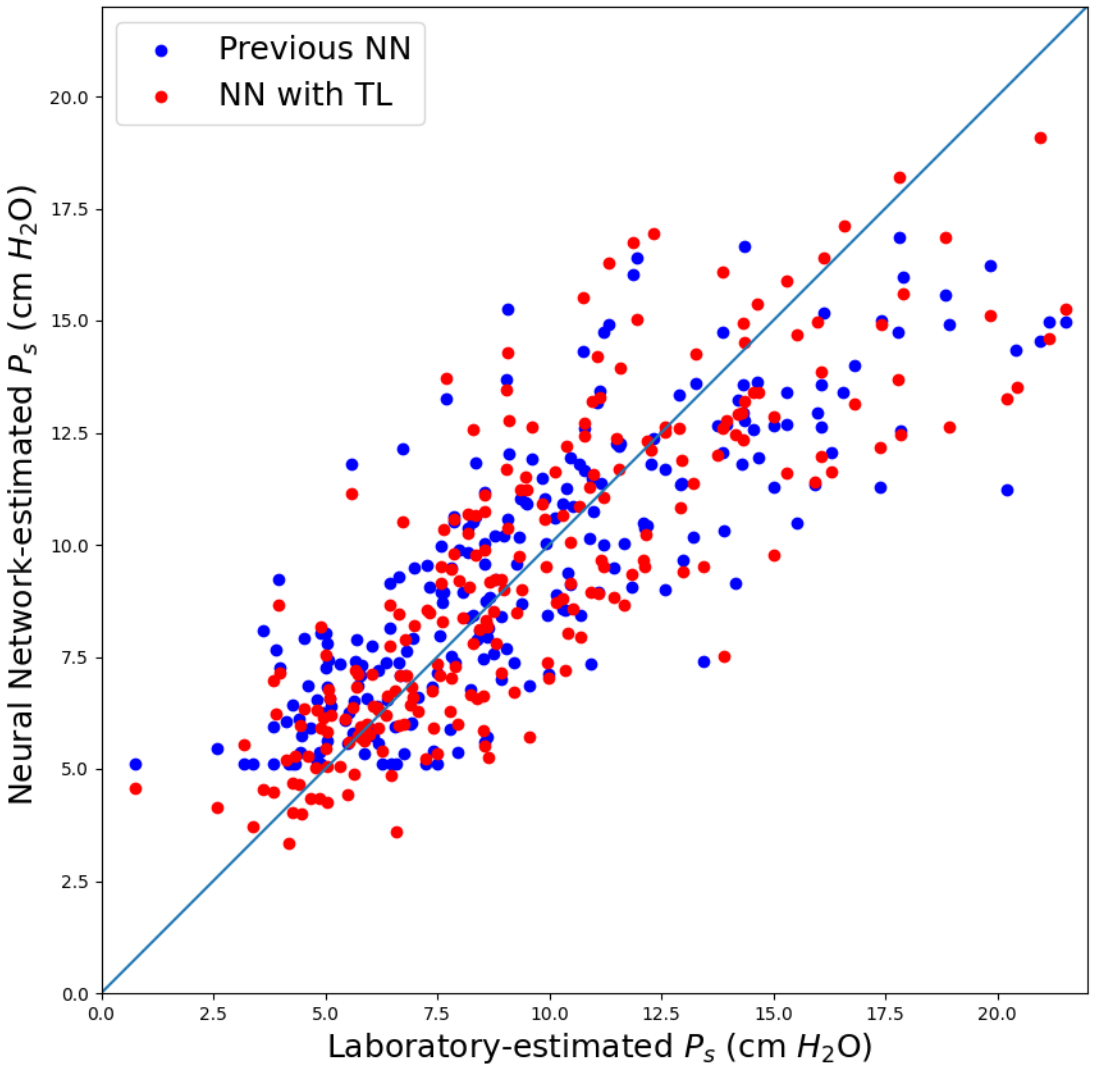


Figure 5.6: Comparison between laboratory-estimated subglottal pressure and the corresponding estimates from the trained neural network, for previous results and using transfer learning. The dashed line represents the theoretical 1:1 perfect matching.

For all folds, the NNs were trained using 100 epochs, similar to the baseline model training. This number of epochs ensures model convergence, as demonstrated by the graph in Figure 5.7, which compares MSE across epochs for both the re-training and validation phases for the 10 trained folds of an NN with one layer frozen. Additionally, as observed for the baseline model (Chapter 4, Figure 4.6), there is no evidence of overfitting. This absence of overfitting is indicated by the stable trend in the MSE for both training and validation observed after completing the 100 epochs. Similar behavior was observed across the remaining networks.

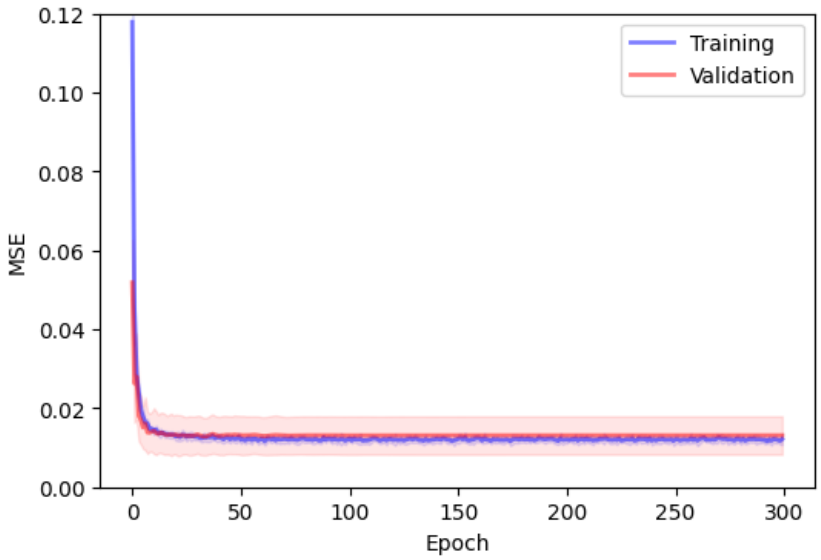


Figure 5.7: Mean squared error (MSE) versus epochs for the re-training and validation of a neural network with one layer frozen, across 10 folds. The solid lines represent the mean values, and the shaded regions indicate the standard deviation.

Estimating P_s from laboratory Dataset 4

Table 5.4 presents the results of the analysis for control and pathological cases in the context of subglottal pressure estimation. When TL is applied to the model, notable improvements in error metrics are observed. For instance, in the case of average RMSE, a 4.2% improvement was achieved for the control group, while for pathological cases, 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 estimation are consistent with the observations made in previous work [116] and can be attributed to the high variability in the intrapathological group. Specifically, patients with UVFP exhibited the highest RMSE, which may be due to the various manifestations of this pathology. For example, during the laboratory recording measurements, patients with UVFP experienced more difficulty maintaining a steady pitch throughout production, changing pitch, and managing breath control for phonatory tasks [35]. Additionally, a recent study employing high-speed video analysis revealed that a patient with UVFP exhibited chaotic behavior in vocal fold vibration dynamics [128].

Figure 5.8 contrasts the RMSE of the estimation of P_s using TL with two

Table 5.4: Error metrics for general subglottal pressure estimation using neural network training with and without transfer learning (TL) across four participant groups: Control, Phonotraumatic Vocal Hyperfunction (PVH), Nonphonotraumatic Vocal Hyperfunction (NPVH), and Unilateral Vocal Fold Paralysis (UVFP) in Dataset 4.

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

methods reported in the literature. Method 1 refers to the empirically derived formula proposed by Titze et al. in [31]. Method 2 is the NN (2 hidden layers with 4 neurons) trained using only synthetic data presented in Chapter 4. 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 [136] identified statistical differences between the TL-based method and Method 1 (p=0.0066), as well as between the TL-based method and Method 2 (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 the pathological groups, ANOVA analysis did not reveal significant differences between methods, with all p-values above the conventional significance threshold of 0.05.

5.5.2 Subject-specific neural network for subglottal pressure estimation

For subject-specific NNs, each model was developed following a 5-fold cross-validation as in [34]. The mean results for all subjects by groups are shown in Table 5.5. The results show that for subject-specific NN, the TL also improves es-

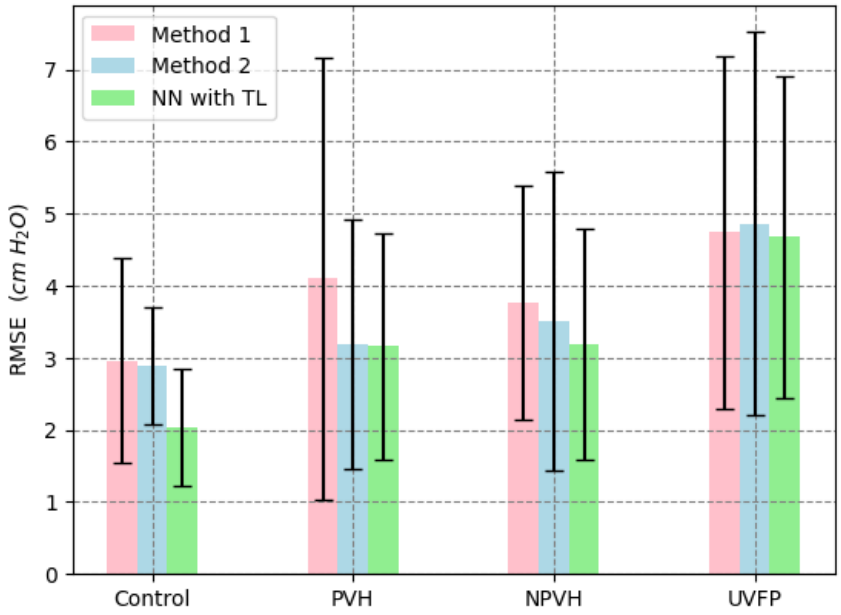


Figure 5.8: Comparison of RMSE in general P_s estimation among three methods: Proposed neural network (NN) with transfer learning (TL), empirical equation (Method 1), and neural network trained exclusively with synthetic data (Method 2).

timation in contrast to being trained from scratch. For all four groups, it is evident that the error metrics are lower in subject-specific NN models compared to the general NN estimation. This is directly associated with the fact that constrained data to unique subjects discards the inter-subject variability.

In Figure 5.9, the RMSE results of the subject-specific NN with TL are contrasted against other subject-specific methodologies from the literature, detailed in Chapter 2, Subsection 2.1.4. Method 3, a linear regression model that estimates P_s from the RMS magnitude of the ACC signal [36]. Method 4, which uses a multilinear regression function that combines RMS with additional ACC-based

Table 5.5: Error metrics for subglottal pressure estimation using subject-specific neural network training with and without transfer learning (TL) across four participant groups: Control, Phonotraumatic Vocal Hyperfunction (PVH), Nonphonotraumatic Vocal Hyperfunction (NPVH), and Unilateral Vocal Fold Paralysis (UVFP) in Dataset 4.

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**

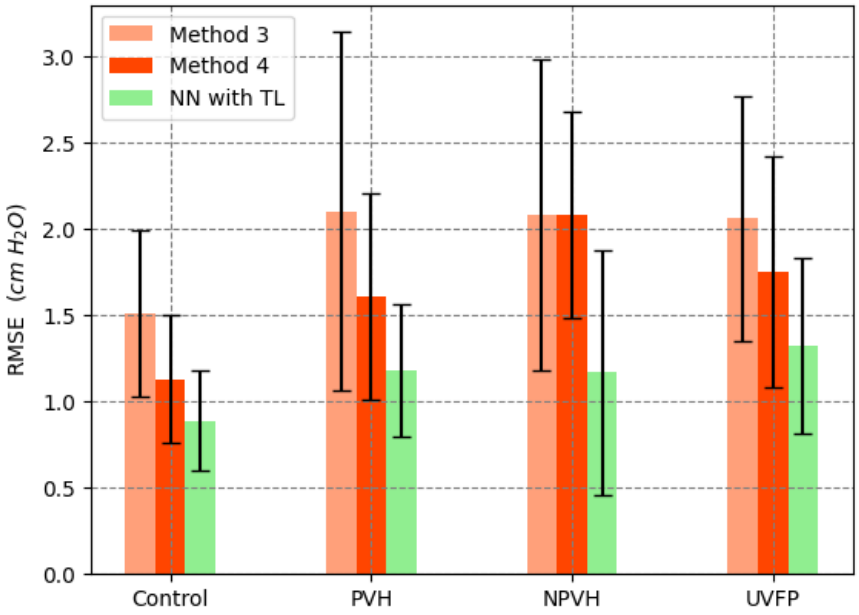


Figure 5.9: Comparison of the mean RMSE for the best fold of subject-specific neural network estimation among three methods: Proposed neural network with transfer learning, linear regression model (Method 3), and multi-linear regression model (Method 3).

features [34]. The bar plots in the figure illustrate that the subject-specific NN with TL provides a lower average RMSE in the P_s estimation. It is important to highlight that, in Figure 5.9, the bars indicate the mean of the best fold estimation for each method to facilitate a direct comparison with the results reported in [116].

The two-way ANOVA performed on RMSE for the estimation of the P_s, reported a group (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 groups. 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.

5.5.3 Subject-specific neural network for muscle activations estimation

For this study, the subject-specific NN was trained using a leave-one-task-out cross-validation approach, involving tasks with ascending and descending pitch glides for vowels /a/ and /i/. This method systematically excludes data from one task at a time during training and uses it for validation. This approach effectively prevents overfitting and enhances the generalizability of the model across various tasks for the subject. Figure 5.10 illustrates the MSE across epochs for both

re-training and validation. The behavior, similar to previous experiments that estimated only P_s, demonstrates that the mean MSE across the different tasks decreases and stabilizes after more than 100 epochs, indicating the absence of overfitting.

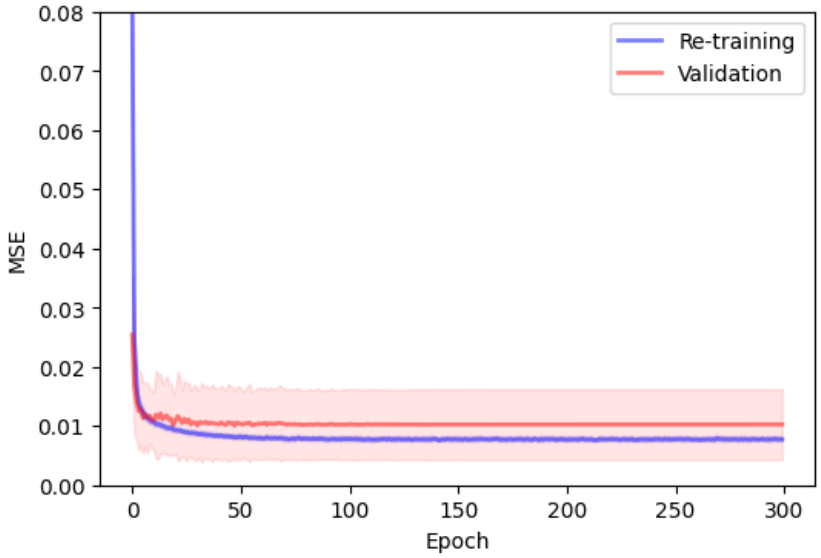


Figure 5.10: Mean squared error (MSE) versus epochs for the re-training and validation of a neural network with one layer frozen, across tasks. The solid lines represent the mean values, and the shaded regions indicate the standard deviation.

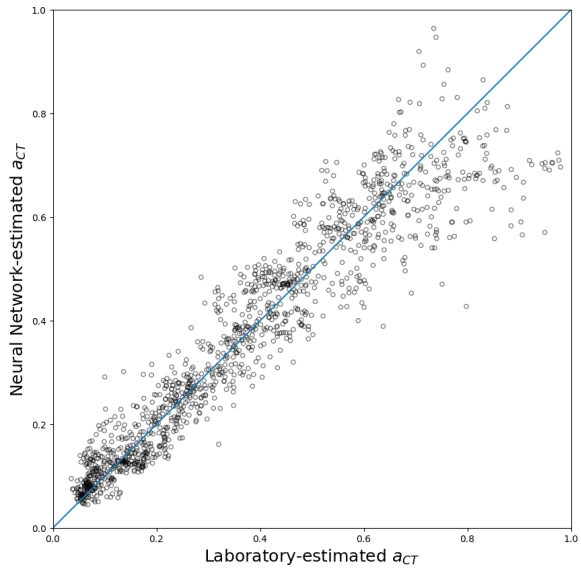
The mean error metrics for estimating muscle activation from Dataset 5, utilizing pitch glides in vowels /a/ and /i/, are detailed in Table 5.6. The error metrics for MAE, RMSE, and MAPE are similar for both types of muscle activations (a_{TA} and a_{CT}), with differences being less than 0.03 for RMSE and MAE, and 1.1% for MAPE, where the lowest error is observed for a_{TA}. However, as shown with syn-

Table 5.6: Error metrics for muscle activation estimation from a subject-specific neural network for pitch glides of vowels /a/ and /i/ in Dataset 5.

Output	RMSE ($cm\ H_2O$)	MAE ($cm\ H_2O$)	MAPE (%)	R^2
a_{CT}	0.08	0.05	16.70	0.87
a_{TA}	0.05	0.04	15.60	0.61

thetic data (Table 4.7 in Chapter 4), the R^2 value is higher for a_{CT}. The scatter plot in Figure 5.11, contrasting laboratory measurements with NN estimations, illustrates that the deviations of estimated values from the ideal match (indicated by the blue line) are comparable for both activations. Nonetheless, the variation range for a_{CT} is broader than that for a_{TA}. Consequently, for all data, NN predictions for a_{TA} are nearer to the mean observed values, resulting in a lower coefficient of determination. In this context, an R^2 value of 0.61 is considered very good, especially when compared to the 0.54 obtained from synthetic data. The observed limited variation in a_{TA} is attributable to the phonation gestures studied (pitch glides), which predominantly require a_{CT} to generate a wide frequency spectrum during phonation. This strong correlation between CT muscle activation and fundamental frequency was further evidenced by the simulations using the CEKF approach, as discussed in Experiment 2 of Chapter 3.

Figure 5.12 offers a detailed visualization of subject-specific NN estimations



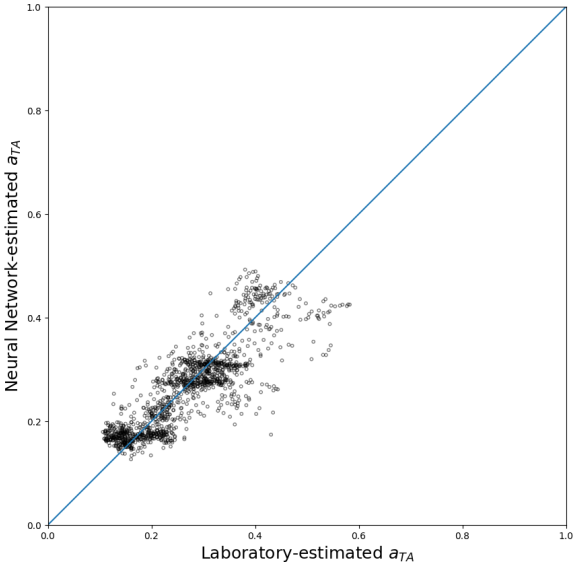
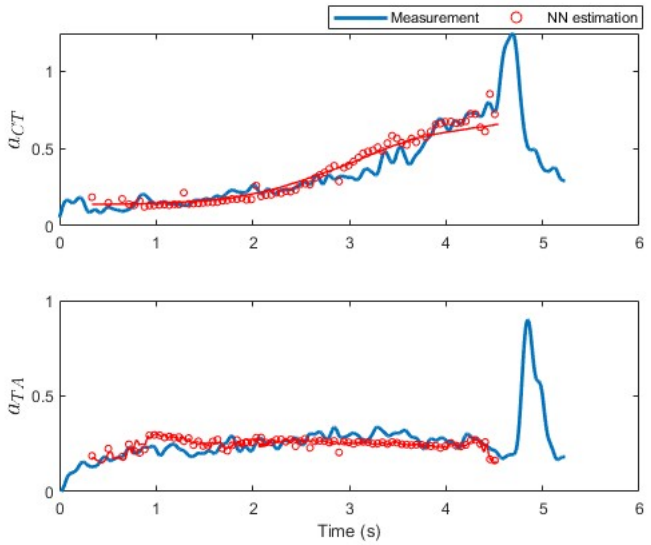


Figure 5.11: Comparison between laboratory-measured normalized muscle activations and the corresponding estimates from subject-specific NN.

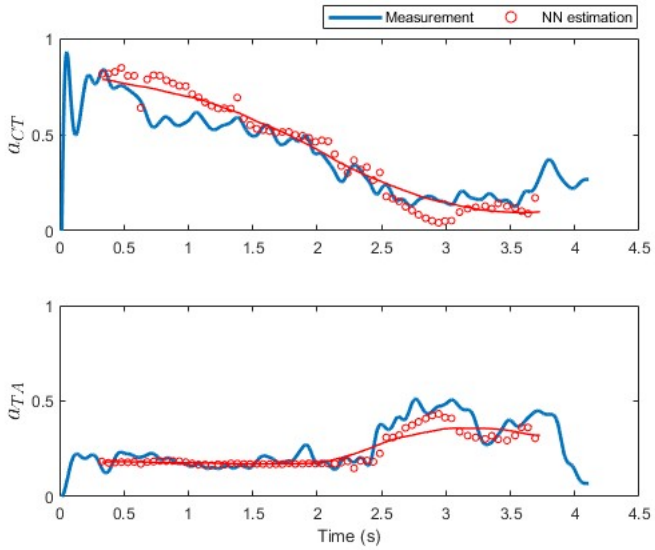
for muscle activations during both ascending and descending pitch glides of the vowel /a/. This analysis contrasts the normalized laryngeal EMG signal with NN estimations, captured within 50 ms windows without overlap, and depicted as red circles. The red line represents the interpolation and smoothing of these estimation points. The estimations for both muscle activations closely align with laboratory measurements, underscoring the precision of the NN model. Furthermore, the regressor accurately captures the trends observed in the derived EMG signals, including variations in a_{CT} across both pitch glides and the significant increase in a_{TA} activation at 2.5 s during the descending pitch glide. It is noted that the absence of estimations at the beginning and end of the EMG signals marks a limitation of this method. The approach is designed to focus on the stable portions of the vowel sound, thus excluding transient periods of activation at the onset and offset, due to the instability in speech-derived signals during these times.

5.5.4 Subject-specific neural network for subglottal pressure and muscle activations estimation

From the plosive /pæ/ phonatory task in Dataset 5, the base model was fine-tuned to estimate subglottal pressure and muscle activations. In this experiment, leave-one-task-out cross-validation was used, and the MSE across epochs exhibited behavior similar to that observed in Figure 5.10. The error metrics for these



(a) Pitch glide ascending



(b) Pitch glide descending

Figure 5.12: Normalized muscle activation obtained from laboratory measurements and estimated from subject-specific NN, from phonatory task pitch glides vowel /a/. The solid red line represents the interpolation and smoothing of NN estimations.

Table 5.7: Error metrics for subglottal pressure and muscle activation estimation from a subject-specific neural network for plosives /pæ/ phonatory tasks in Dataset 5.

Output	RMSE ($cm\ H_2O$)	MAE ($cm\ H_2O$)	MAPE (%)	R^2
P_s	1.24	0.89	8.07	0.86
a_{CT}	0.04	0.03	12.13	0.95
a_{TA}	0.07	0.05	14.56	0.16

estimations are presented in Table 5.7. The metrics for P_S are consistent with the mean values shown in Table 5.5 for the control group. Additionally, the RMSE, MAE, and MAPE metrics for muscle activation are in line with those presented in Table 5.6. These results indicate that incorporating additional vocal function estimations into the subject-specific NN does not compromise its accuracy. The coefficient of determination for a_{CT} is slightly higher, whereas for a_{TA}, it is lower compared to those obtained from pitch glides analysis. This decrease in the coefficient for a_{TA} suggests that the estimations for this muscle are closer to the mean value, indicating that the correlation between the input features and a_{TA} was diminished for /pæ/ phonatory tasks.

Figures 5.13 and 5.14 provide a direct comparison between measurements and subject-specific NN estimations. In these figures, the red circles represent the NN estimations within a 50 ms window with 50% overlap, while the red line

indicates the interpolation and smoothing of these estimation points. Across both figures, the NN estimations closely follow the measurements, with more significant discrepancies observed in a_{TA}. These findings highlight the effectiveness of the proposed approach in estimating vocal functions using an ACC sensor. Notably, for a_{TA}, the estimation discrepancies become more evident at the peaks of the signals. This observation accounts for the reduced R^2 value in comparison to pitch glide analysis, where the segment for vowel production is longer and more stable.

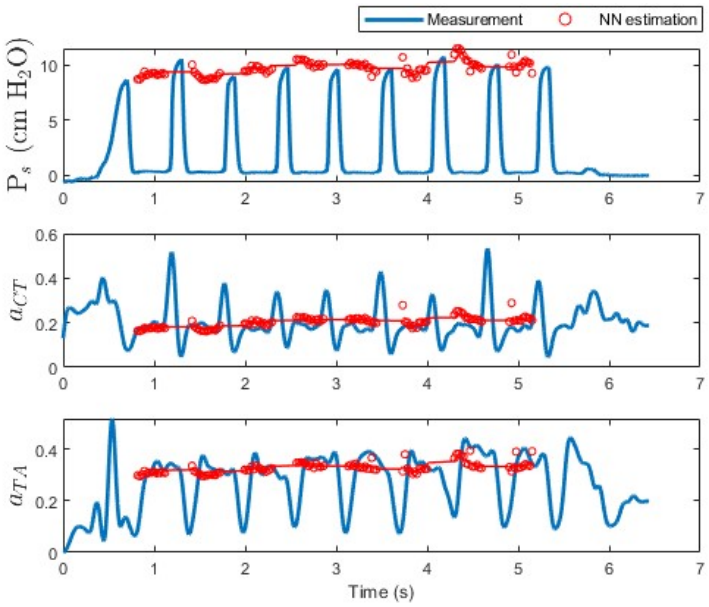


Figure 5.13: P_s and normalized muscle activation obtained from laboratory measurements and estimated from subject-specific NN, for plosive /pæ/ task in comfortable loud and normal pitch. The solid red line represents the interpolation and smoothing of NN estimations.

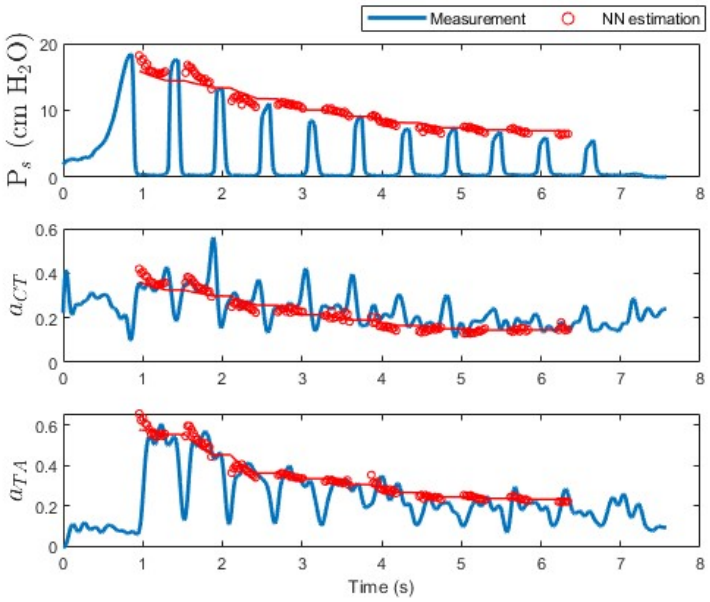


Figure 5.14: P_s and normalized muscle activation obtained from laboratory measurements and estimated from subject-specific NN, for plosive /pæ/ task descending loudness with normal pitch. The solid red line represents the interpolation and smoothing of NN estimations.

5.6 Chapter conclusions

The study described in this chapter demonstrates notable advancements in the estimation of subglottal pressure from neck-surface vibrations. It was found that utilizing neural networks initially trained on a numerical voice production model and subsequently refined with transfer learning significantly enhances subglottal pressure estimation. This method outperformed existing approaches in both con-

trol subjects and those with voice pathologies, marking a substantial improvement over previous work and previous methods in this thesis. These findings are particularly promising for developing advanced, non-invasive P_s assessment tools for clinical and ambulatory applications, that could offer new pathways for diagnostic and therapeutic strategies of voice disorders. Additionally, preliminary results demonstrate the feasibility of expanding subject-specific NN estimation to other vocal function measures such as muscle activation.

The effectiveness of methods integrating NN with a numerical voice production model is directly correlated with the ability of numerical models to accurately replicate laboratory data distributions [58]. In the previous chapter, bias corrections were applied to the synthetic features P_s and SPL to reduce the range and distribution discrepancies between clinical and synthetic datasets. This procedure forced $\mathcal{D}_S = \mathcal{D}_T$ to treat the problem as a traditional machine learning approach. In this current study, it was found that TL techniques address these discrepancies by fine-tuning the final layers of the NN. The results suggest that in this new proposal, the relevant knowledge obtained from synthetic data is conserved across the first hidden layers, and the refinement of subsequent parameters allows for optimal domain adaptation. Consequently, a 7% reduction in the RMSE of P_s estimation from laboratory Dataset 3 was observed, in contrast to the preliminary proposal. This shift in methodology not only demonstrates the efficacy of TL but also more effectively harnesses the physiological relevance of low-order lumped

models of voice production and clinical recordings for vocal function analysis.

According to the results, it is evident that the subject-specific approach significantly improves P_s estimation compared to the general model (see Tables 5.4 and 5.5), with an improvement of over 38% in the RMSE value across all groups. In this sense, it is maintained that while the simulation of the voice production model provides a good general representation of typical sustained phonation, the simplicity of the numerical model behind it does not allow for a representation of the variability of a universal population, thus failing to mimic the expected inter and intra-subject variability of real data. Under these circumstances, TL benefits the subject-specific representation more than the general model. Additionally, capturing the inter-subject variability will require clinical recordings that include a large number of subjects. The improvement obtained with the subject-specific NN regressor, compared to previous techniques based on subject-specific calibration [116], was over 21% in RMSE values for the four groups, as demonstrated in Figure 5.9. This highlights the contribution of this new proposal to the state-of-the-art.

It is worth noting that although the study was extensively validated for P_s estimation, the results for the estimated activation of the TA and CT muscles were tested using the only case study available for analysis. Nevertheless, this preliminary study demonstrates that the accuracy of P_s estimation is maintained even when new outputs are added to the NN. Additionally, the results indicate

that the selected input features are more strongly correlated with a_{CT} than with a_{TA}, a finding supported by the high correlation between f_0 and a_{CT}. While these initial results in muscle activation estimation are promising, extensive validation is required to substantiate these preliminary findings. However, broad validation of TL-based approaches for estimating muscle activation, or the inclusion of new features such as P_c, is limited by the scarcity of clinical recordings that encompass these measures.

Chapter 6

Conclusions

This thesis investigated the feasibility of employing a low-order voice production model to assess vocal function parameters that are challenging to measure in clinical settings. The primary aim was to develop a non-invasive method for estimating relevant vocal function measures, including subglottal pressure, muscle activation, and vocal fold collision pressures, utilizing an ACC sensor, suitable in both clinical and ambulatory settings.

Three research components were undertaken to achieve this goal: firstly, studying the capabilities of numerical simulations of a low-order voice production model through a Bayesian Framework; secondly, combining a low-order voice production model with machine learning tools to develop an inverse mapping that estimates targeted vocal function parameters from neck-surface vibrations in both normal and disordered voices; and thirdly, conducting a domain adaptation study to transfer network parameters learned from voice model simulations to assess vocal function using an accelerometer sensor.

Generally, the observations from the Bayesian inference framework, nonlinear

mapping, and domain adaptation provided evidence supporting previous studies while also contributing new insights. These include the potential of linking the voice production model to laboratory recordings to facilitate access to vocal function parameters that are challenging to measure in clinical environments, comparing the numerical low-order voice production model with clinical data to inform further model refinements and the integration of a machine learning tool to transfer synthetic knowledge from the numerical model to clinical or ambulatory settings. The primary clinical contribution of this research lies in its compelling demonstration of the effectiveness of the proposed noninvasive technique, which leverages neck-surface acceleration for the assessment of vocal function, particularly in the subglottal pressure estimation.

The literature review presented in Chapter 2 explored research proposing methods based on numerical models of voice production to estimate clinically relevant parameters. These methods encompass optimization-based voice inversion techniques [44, 45, 46, 47, 48, 49, 50], multi-parameter estimation techniques grounded in Bayesian estimation [96, 51, 52, 53, 54, 55, 56, 57], and the integration of machine learning tools with voice production models [58, 59]. Building on this review, it was concluded that optimization-based voice inversion methods are impractical due to the significant computational cost involved in processing large volumes of data during inversion [47, 58]. In contrast, Bayesian estimation, particularly those approaches utilizing the EKF, was identified as the most

promising technique for clinical vocal function assessment [51]. However, challenges associated with this algorithm limit its direct application in ambulatory settings. Conversely, machine learning methods for voice inversion, once trained, can accurately estimate the targeted vocal function parameters without the need for ongoing model simulations [58], offering a computationally efficient alternative suitable for both clinical and ambulatory settings.

Chapter 3 introduced a CEKF scheme, which simplifies the experimental requirements of the EKF approach presented in [51]. Unlike the preliminary method, which requires multi-sensor or simultaneous recordings as observation states, the CEKF can estimate model states from a single measurement (e.g., glottal area waveform or glottal airflow) by constraining model states with prior information on the physiological phonation process. The results from experiments 1 and 3 demonstrated that this new approach enables the estimation of the glottal airflow state from the GAW, and vice versa, with RMSE values similar to those obtained by [51], which utilized both measurements as observations. These results confirm that Hypothesis 1 was valid, as incorporating constraints derived from prior physiological knowledge of the phonation process into the Bayesian framework of a low-order voice production model allowed this to estimate vocal function using only single observation measurements with accuracy comparable to that of recent Bayesian inference models.

The applicability of the Bayesian inference method has been broadened by

correlating the low-order voice production model with subject-specific laboratory recordings, utilizing only measurements from HSV or Rothenberg masks, as evidenced in experiments 2 and 4 of Chapter 3. The results demonstrated that the proposed Bayesian framework could correlate a_{CT} with the fundamental frequency as in [111, 83], the magnitude of peak vocal fold collision pressure with subglottal pressure as in [112, 21], and reproduce the tendency for a_{CT} and a_{TA} to increase with rising sound pressure levels as suggested in [113]. These experiments achieve Aim 1 of this work, which involved investigating the capabilities of a low-order voice production model to mimic the behavior of vocal function as observed in laboratory measurements within a Bayesian framework.

The results of the CEKF-based approach represent one of the contributions of this thesis, enhancing the efficiency of Bayesian inference to yield reliable estimates of vocal function measures. Moreover, these findings highlight the capabilities of the numerical low-order voice production model in simulating a broad spectrum of pitch and loudness conditions for sustained vowels through the modulation of control variables such as muscle activation. The promising outcomes of the CEKF method pave the way for the application of the numerical lumped-element voice production model in clinical settings, particularly for data fusion across different recording sessions. Despite these advances in Bayesian inference, further experimentation is necessary to expand the number of subjects studied and to include pathological cases. Additionally, there are challenges associated with model pa-

rameter adjustments, such as the selection of covariance matrix values, which could affect the effectiveness of the method.

In Chapter 4, a more direct solution for estimating target vocal functions using ACC was introduced. This chapter outlines the second specific aim of this thesis: to evaluate an NN for inverse mapping accelerometer-based features to vocal function measures, such as subglottal pressure, vocal fold collision pressure, and intrinsic laryngeal muscle activation. Validation with synthetic data demonstrated that the NN could effectively correlate accelerometer-based features with vocal function measures, achieving a determination coefficient over 0.8 for P_s, P_c, and a_{CT}, and over 0.5 for a_{TA}. Testing with clinical recordings from 79 vocally healthy female participants revealed that the MAE and RMSE for subglottal pressure were 1.95 cm H_2O and 2.48 cm H_2O, respectively. These outcomes are on par with previous studies [125, 58, 34], but offer the advantage of a universal mapping that applies to all patients, providing simultaneous estimates of collision pressure and muscle activation. However, the clinical validation of these latter features proved to be challenging, leading to the exclusive use of synthetic data for validation purposes.

In contrast, when the NN was evaluated using laboratory dataset 4, the RMSE values increased for patient groups, highlighting the limitations of the selected voice production model in accurately simulating pathological cases. This increase in error rates suggests that the relationship between accelerometer-based features

and subglottal pressure is significantly impacted by non-modal phonations and the presence of voice disorders, as highlighted in [36, 35]. These results indicate that the second hypothesis of this thesis is invalid. Quantitatively, the estimation of subglottal pressure using a nonlinear regressor, trained solely with data from a numerical low-order voice production model, yielded higher RMSE values compared to those obtained by a subject-specific linear regression model for both normal and pathological voices.

For this stage of the thesis, it is concluded that integrating machine learning with a physiologically relevant voice synthesizer presents several advantages. Notably, this approach facilitates access to vocal features that are clinically challenging to measure, such as subglottal pressure, muscle activation, and vocal fold contact pressure [115]. The training process includes thousands of simulations, covering a wide range of sustained vowel phonations. However, while numerical voice production models offer a reliable representation of the phonatory process, the signals derived from these models approximate the complex interactions between human vocal fold physiology and voice production. Therefore, it is critical to acknowledge that regressor models trained with synthetic data operate under the assumption that the training domain (synthetic data from the numerical voice production model) and the target domain (laboratory data) exist within the same feature space and share identical distributions. Nonetheless, domain shifts are anticipated, not only due to differences between control and pathological groups but

also owing to individual subject variabilities. Consequently, evidence indicates that domain adaptation is essential for enhancing model performance.

In Chapter 5, domain adaptation through TL was proposed, wherein the NN regressor, initially trained with a dataset from the synthetic model of voice production, was refined using *in vivo* laboratory data. The methodology described in this chapter focused on achieving the third specific aim: to determine if the domain adaptation method improves the performance of the NN. In general, the results illustrated that domain adaptation significantly improves the estimation of vocal function parameters in frameworks combining machine learning methods with numerical models of voice production. Although this method allowed for the development of a universal model to estimate P_s, the scarcity of laboratory recordings covering a broad population hinders the creation of a robust model capable of accounting for the complex variability among subjects. Therefore, subject-specific NN regressors, based on domain adaptation, represent the best alternative for improving the estimation of P_s with a reduced number of recordings. This finding underscores how subject-specific regression models effectively estimate P_s for both normophonic subjects and individuals with voice disorders. The results demonstrated improvements in the P_s estimates compared to previous techniques, achieving over a 21% reduction in RMSE. In this sense, these results are the best estimates of P_s from the ACC signals reported in the literature. Additionally, the ANOVA study confirmed Hypothesis 3: fine-tuning

the NN (trained with synthetic voice production data) with individual laboratory recordings significantly enhances the accuracy of estimating vocal function parameters from neck-surface vibration recordings.

Furthermore, the potential to expand the subject-specific NN for estimating other vocal functions, such as muscle activation, was demonstrated. The preliminary results from the case study in Chapter 5 revealed that muscle activation could also be estimated with a MAPE lower than 20%, while maintaining accuracy in P_s estimation. These results further endorse that TL is an effective strategy for domain adaptation from synthetic to *in vivo* data. The subject-specific NN approach represents a significant contribution of this thesis, proposing a non-invasive method for estimating vocal function from ACC signals.

In an effort to assess physiologically relevant metrics, the versatility of ACC sensors enables extending the application of this method to ambulatory settings. Subject-specific fine-tuning enhances the ability of the NN to estimate subglottal pressure by relying solely on ACC-based features within brief 50 ms windows. This approach not only bolsters confidence in the non-invasive estimation of vocal functions, particularly P_s, but also broadens its potential use in clinical, laboratory, and ambulatory monitoring of vocal function during natural voice production. The long-term goal is to develop algorithms for analyzing neck-surface vibrations monitored via smartphone devices, aiming to improve the diagnosis, prevention, and treatment of voice disorders by deepening our understanding of their under-

lying mechanisms. Future efforts will focus on applying this method to measure P_s during spontaneous speech as part of ambulatory monitoring and biofeedback, as in [116]. This application will take place as individuals go about their daily activities in various settings, such as home, work, and social environments.

The results showed that the proposed domain adaptation (for both general and subject-specific approaches) was more efficient for the healthy than for the pathological groups. 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 future works the present findings could be significantly enhanced by further exploring numerical voice production models that more accurately mimic pathophysiological behavior [137, 138, 139, 140], which could facilitate the transfer of knowledge in cases involving subjects with voice disorders.

The domain adaptation approach presented in this study was extensively validated with respect to subglottal pressure and, in a case study, for two muscle activations. By utilizing the synthetic voice production model, access was gained to a broader set of phonatory measurements, such as vocal fold collision pressure and additional muscle activations, including the LCA, IA, and PCA. However, the

capability of this approach to estimate outputs is limited by the scarcity of clinical recordings that encompass such measurements. Future efforts will be directed towards exploring transductive TL techniques to enhance domain adaptation in scenarios where labeled data are abundant in the source domain but scarce or nonexistent in the target domain [131].

Bibliography

[1] Elaine Smith, Steven Gray, Katherine Verdolini, and Jon Lemke. Effects of voice disorders on quality of life. *Otolaryngology–Head and Neck Surgery*, 113(2):P121–P121, 1995.

[2] Neil Bhattacharyya. The prevalence of voice problems among adults in the united states. *The Laryngoscope*, 124(10):2359–2362, 2014.

[3] N. Roy, R. M. Merrill, S. D. Gray, and E. M. Smith. Voice disorders in the general population: Prevalence, risk factors, and occupational impact. *Laryngoscope*, 115:1988–1995, 2005.

[4] Adrián Castillo, César Casanova, Daniel Valenzuela, and Sebastian Castañón. Prevalencia de disfonía en profesores de colegios de la comuna de Santiago y factores de riesgo asociados. *Ciencia trabajo*, 17:15 – 21, 04 2015.

[5] Robert E. Hillman, Eva B. Holmberg, Joseph S. Perkell, Michael Walsh, and Charles Vaughan. Objective assessment of vocal hyperfunction. *Journal of Speech, Language, and Hearing Research*, 32(2):373–392, 1989.

[6] Daryush D. Mehta, Jarrad H. Van Stan, Matías Zañartu, Marzyeh Ghassemi, John V. Guttag, Víctor M. Espinoza, Juan P. Cortés, Harold A. Cheyne, and Robert E. Hillman. Using ambulatory voice monitoring to investigate common voice disorders: Research update. *Frontiers in Bioengineering and Biotechnology*, 3:155, 2015.

[7] Robert E. Hillman, Cara E. Stepp, Jarrad H. Van Stan, Matías Zañartu, and Daryush D. Mehta. An updated theoretical framework for vocal hyperfunction. *American Journal of Speech-Language Pathology*, 29(4):2254–2260, 2020.

[8] Víctor M. Espinoza, Matías Zañartu, Jarrad H. Van Stan, Daryush D. Mehta, and Robert E. Hillman. Glottal aerodynamic measures in women with phonotraumatic and nonphonotraumatic vocal hyperfunction. *Journal of Speech, Language, and Hearing Research*, 60(8):2159–2169, 2017.

[9] Steven M. Zeitels, Rosemary B. Desloge, Robert E. Hillman, and Glen A. Bunting. Cricothyroid subluxation: A new innovation for enhancing the voice with laryngoplastic phonosurgery. *Annals of Otology, Rhinology & Laryngology*, 108(12):1126–1131, 1999. PMID: 10605916.

[10] Steven M. Zeitels, Ramon A. Franco, Robert E. Hillman, and Glenn W. Bunting. Voice and treatment outcome from phonosurgical management

of early glottic cancer. *Annals of Otology, Rhinology & Laryngology*, 111(12_suppl):3–20, 2002.

[11] Dinesh K. Chhetri and Juergen Neubauer. Differential roles for the thyroarytenoid and lateral cricoarytenoid muscles in phonation. *The Laryngoscope*, 125(12):2772–2777, 2015.

[12] Nobuhiko Isshiki. Regulatory mechanism of voice intensity variation. *Journal of Speech and Hearing Research*, 7(1):17–29, 1964.

[13] Randall L. Plant and Allen D. Hillel. Direct measurement of subglottic pressure and laryngealresistance in normal subjects and in spasmodic dysphonia. *Journal of Voice*, 12(3):300–314, 1998.

[14] Johan Sundberg, Ronald C. Scherer, Markus Hess, Frank Müller, and Svante Granqvist. Subglottal pressure oscillations accompanying phonation. *Journal of Voice*, 27(4):411–421, 2013.

[15] Peter Ladefoged and Norris P. McKinney. Loudness, sound pressure, and subglottal pressure in speech. *The Journal of the Acoustical Society of America*, 35(4):454–460, 1963.

[16] J. Van den Berg. Direct and indirect determination of the mean subglottic pressure. *Folia Phoniatrica Et Logopaedica*, 8:1–24, 1956.

[17] Bert Cranen and Louis Boves. Pressure measurements during speech production using semiconductor miniature pressure transducers: Impact on models for speech production. *The Journal of the Acoustical Society of America*, 77(4):1543–1551, 04 1985.

[18] Daryush D. Mehta, James B. Kobler, Steven M. Zeitels, Matías Zañartu, Emiro J. Ibarra, Gabriel A. Alzamendi, Rodrigo Manriquez, Byron D. Erath, Sean D. Peterson, Robert H. Petrillo, and Robert E. Hillman. Direct measurement and modeling of intraglottal, subglottal, and vocal fold collision pressures during phonation in an individual with a hemilaryngectomy. *Applied Sciences*, 11(16), 2021.

[19] Philip Lieberman. Direct Comparison of Subglottal and Esophageal Pressure during Speech. *The Journal of the Acoustical Society of America*, 43(5):1157–1164, 07 2005.

[20] Jw. 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*, 8(1):1–24, 11 2009.

[21] Heather E. Gunter, Robert D. Howe, Steven M. Zeitels, James B. Kobler, and Robert E. Hillman. Measurement of vocal fold collision forces during

phonation. *Journal of Speech, Language, and Hearing Research*, 48(3):567–576, 2005.

[22] Varun Varadarajan and Jonathan M. Blumin, Joel H.and Bock. State of the Art of Laryngeal Electromyography. *Current Otorhinolaryngology Reports*, 1:171–177, 2013.

[23] Christopher J. Poletto, Laura P. Verdun, Robert Strominger, and Christy L. Ludlow. Correspondence between laryngeal vocal fold movement and muscle activity during speech and nonspeech gestures. *Journal of Applied Physiology*, 97(3):858–866, 2004. PMID: 15133000.

[24] Robert E. Hillman and Daryush D. Mehta. Ambulatory monitoring of daily voice use. *Perspectives on Voice and Voice Disorders*, 21(2):56–61, 2011.

[25] Daryush D. Mehta, Matías Zañartu, Shengran W. Feng, Harold A. Cheyne II, and Robert E. Hillman. Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform. *IEEE Transactions on Biomedical Engineering*, 59(11):3090–3096, 2012.

[26] Peter S. Popolo, Jan G. Švec, and Ingo R. Titze. Adaptation of a pocket pc for use as a wearable voice dosimeter. *Journal of Speech, Language, and Hearing Research*, 48(4):780–791, 2005.

[27] Jarrad H. Van Stan, Daryush D. Mehta, and Robert E. Hillman. Recent Innovations in Voice Assessment Expected to Impact the Clinical Management of Voice Disorders. *Perspectives of the ASHA Special Interest Groups*, 2(3):4–13, jan 2017.

[28] Meri L Andreassen, Juliana K Litts, and Derrick R Randall. Emerging techniques in assessment and treatment of muscle tension dysphonia. *Current opinion in otolaryngology & head and neck surgery*, 25(6):447—452, December 2017.

[29] Jan G. Švec, Ingo R. Titze, and Peter S. Popolo. Estimation of sound pressure levels of voiced speech from skin vibration of the neck. *The Journal of the Acoustical Society of America*, 117(3):1386–1394, 2005.

[30] Marzyeh Ghassemi, Jarrad H. Van Stan, Daryush D. Mehta, Matías Zañartu, Harold A. Cheyne II, Robert E. Hillman, and John V. Guttag. Learning to detect vocal hyperfunction from ambulatory neck-surface acceleration features: Initial results for vocal fold nodules. *IEEE Transactions on Biomedical Engineering*, 61(6):1668–1675, 2014.

[31] Ingo R. Titze, Jan G. Švec, and Peter S. Popolo. Vocal dose measures: Quantifying accumulated vibration exposure in vocal fold tissues. *Journal of Speech, Language, and Hearing Research*, 46(4):919–932, 2003.

[32] Ingo R. Titze and Eric J. Hunter. Comparison of vocal vibration-dose measures for potential-damage risk criteria. *Journal of Speech, Language, and Hearing Research*, 58(5):1425–1439, 2015.

[33] D. D. Mehta, V. M. Espinoza, J.H. Van Stan, M. Zañartu, and R.E. Hillman. The difference between first and second harmonic amplitudes correlates between glottal airflow and neck-surface accelerometer signals during phonation. *The Journal of the Acoustical Society of America*, 145(5):EL386–EL392, 2019.

[34] J. Z. Lin, V. M. Espinoza, K. L. Marks, M. Zañartu, and D. D. Mehta. Improved subglottal pressure estimation from neck-surface vibration in healthy speakers producing non-modal phonation. *IEEE Journal of Selected Topics in Signal Processing*, 14(2):449–460, 2020.

[35] Katherine L. Marks, Jonathan Z. Lin, James A. Burns, Tiffiny A. Hron, Robert E. Hillman, and Daryush D. Mehta. Estimation of subglottal pressure from neck surface vibration in patients with voice disorders. *Journal of Speech, Language, and Hearing Research*, 63(7):2202–2218, 2020.

[36] Katherine L. Marks, Jonathan Z. Lin, Annie B. Fox, Laura E. Toles, and Daryush D. Mehta. Impact of nonmodal phonation on estimates of subglottal pressure from neck-surface acceleration in healthy speakers. *Journal of Speech, Language, and Hearing Research*, 62(9):3339–3358, 2019.

[37] Victoria McKenna, Andres Llico, Daryush Mehta, Joseph Perkell, and Cara Stepp. Magnitude of neck-surface vibration as an estimate of subglottal pressure during modulations of vocal effort and intensity in healthy speakers. *Journal of Speech, Language, and Hearing Research*, 60:1–13, 12 2017.

[38] Amanda S. Fryd, Jarrad H. Van Stan, Robert E. Hillman, and Daryush D. Mehta. Estimating subglottal pressure from neck-surface acceleration during normal voice production. *Journal of Speech, Language, and Hearing Research*, 59(6):1335–1345, 2016.

[39] Andrés F. Llico, Matías Zañartu, Agustín J. González, George R. Wodicka, Daryush D. Mehta, Jarrad H. Van Stan, and Robert E. Hillman. Real-time estimation of aerodynamic features for ambulatory voice biofeedback. *The Journal of the Acoustical Society of America*, 138(1):EL14–EL19, 2015.

[40] Juan P. Cortés, Víctor M. Espinoza, Marzyeh Ghassemi, Daryush D. Mehta, Jarrad H. Van Stan, Robert E. Hillman, John V. Guttag, and Matías Zañartu. Ambulatory assessment of phonotraumatic vocal hyperfunction using glottal airflow measures estimated from neck-surface acceleration. *PLOS ONE*, 13(12):1–22, 12 2018.

[41] Jarrad H. Van Stan, Andrew J. Ortiz, Juan P. Cortes, Katherine L. Marks, Laura E. Toles, Daryush D. Mehta, James A. Burns, Tiffiny Hron, Tara Stadelman-Cohen, Carol Krusemark, Jason Muise, Annie B. Fox-Galalis,

Charles Nudelman, Steven Zeitels, and Robert E. Hillman. Differences in daily voice use measures between female patients with nonphonotraumatic vocal hyperfunction and matched controls. *Journal of Speech, Language, and Hearing Research*, 64(5):1457–1470, 2021.

[42] Jarrad H. Van Stan, Daryush D. Mehta, Dagmar Sternad, Robert Petit, and Robert E. Hillman. Ambulatory voice biofeedback: Relative frequency and summary feedback effects on performance and retention of reduced vocal intensity in the daily lives of participants with normal voices. *Journal of Speech, Language, and Hearing Research*, 60(4):853–864, 2017.

[43] Jarrad H. Van Stan, Daryush D. Mehta, Andrew J. Ortiz, James A. Burns, Katherine L. Marks, Laura E. Toles, Tara Stadelman-Cohen, Carol Krusemark, Jason Muise, Tiffiny Hron, Steven M. Zeitels, Annie B. Fox, and Robert E. Hillman. Changes in a daily phonotrauma index after laryngeal surgery and voice therapy: Implications for the role of daily voice use in the etiology and pathophysiology of phonotraumatic vocal hyperfunction. *Journal of Speech, Language, and Hearing Research*, 63(12):3934–3944, 2020.

[44] M. Dollinger, U. Hoppe, F. Hettlich, J. Lohscheller, S. Schuberth, and U. Eysholdt. Vibration parameter extraction from endoscopic image series of the vocal folds. *IEEE Transactions on Biomedical Engineering*, 49(8):773–781, 2002.

[45] Michael Döllinger, Thomas Braunschweig, Jörg Lohscheller, U. Eysholdt, and Ulrich Hoppe. Normal voice production: Computation of driving parameters from endoscopic digital high speed images. *Methods of Information in Medicine*, 42:271–276, 01 2003.

[46] Michael Döllinger, Pablo Gómez, Rita R. Patel, Christoph Alexiou, Christopher Bohr, and Anne Schützenberger. Biomechanical simulation of vocal fold dynamics in adults based on laryngeal high-speed videoendoscopy. *PLOS ONE*, 12(11):1–26, 11 2017.

[47] Pablo Gómez, Anne Schützenberger, Stefan Kniesburges, Christopher Bohr, and Michael Döllinger. Physical parameter estimation from porcine ex vivo vocal fold dynamics in an inverse problem framework. *Biomech Model Mechanobiol*, 17(3):777–792, 2018.

[48] R. Schwarz, Ulrich Hoppe, Maria Schuster, Tobias Wurzbacher, Ulrich Eysholdt, and Jörg Lohscheller. Classification of unilateral vocal fold paralysis by endoscopic digital high-speed recordings and inversion of a biomechanical model. *IEEE transactions on bio-medical engineering*, 53:1099–108, 07 2006.

[49] Alan P. Pinheiro, David E. Stewart, Carlos D. Maciel, José C. Pereira, and Suely Oliveira. Analysis of nonlinear dynamics of vocal folds using

high-speed video observation and biomechanical modeling. *Digital Signal Processing*, 22(2):304–313, 2012.

[50] Chao Tao, Yu Zhang, and Jack Jiang. Extracting physiologically relevant parameters of vocal folds from high-speed video image series. *IEEE transactions on bio-medical engineering*, 54:794–801, 06 2007.

[51] Gabriel Alzamendi, Rodrigo Manríquez, Paul Hadwin, Jonathan Deng, Sean Peterson, Byron Erath, Daryush Mehta, Robert Hillman, and Matías Zañartu. Bayesian estimation of vocal function measures using laryngeal high-speed videoendoscopy and glottal airflow estimates: An in vivo case study. *The Journal of the Acoustical Society of America*, 147(5):EL434–EL439, 2020.

[52] Carlo Drioli and Gian Luca Foresti. Fitting a biomechanical model of the folds to high-speed video data through bayesian estimation. *Informatics in Medicine Unlocked*, 20:100373, jan 2020.

[53] Paul J. Hadwin, Byron D. Erath, and Sean D. Peterson. The influence of flow model selection on finite element model parameter estimation using bayesian inference. *JASA Express Letters*, 1(4):045204, 2021.

[54] Paul J. Hadwin, Mohsen Motie-Shirazi, Byron D. Erath, and Sean D. Peterson. Bayesian inference of vocal fold material properties from glottal area

waveforms using a 2d finite element model. *Applied Sciences*, 9(13), 2019.

[55] Paul J. Hadwin and Sean D. Peterson. An extended kalman filter approach to non-stationary bayesian estimation of reduced-order vocal fold model parameters. *The Journal of the Acoustical Society of America*, 141(4):2909–2920, 2017.

[56] Paul J. Hadwin, Gabriel E. Galindo, Kyle J. Daun, Matías Zañartu, Byron D. Erath, Edson Cataldo, and Sean D. Peterson. Non-stationary bayesian estimation of parameters from a body cover model of the vocal folds. *The Journal of the Acoustical Society of America*, 139(5):2683–2696, 2016.

[57] Manuel E. Díaz-Cádiz, Sean D. Peterson, Gabriel E. Galindo, Víctor M. Espinoza, Mohsen Motie-Shirazi, Byron D. Erath, and Matías Zañartu. Estimating vocal fold contact pressure from raw laryngeal high-speed videoendoscopy using a hertz contact model. *Applied Sciences*, 9(11), 2019.

[58] P. Gómez, A. Schützenberger, M. Semmler, and M. Döllinger. Laryngeal pressure estimation with a recurrent neural network. *IEEE Journal of Translational Engineering in Health and Medicine*, 7:1–11, 2019.

[59] Zhaoyan Zhang. Estimation of vocal fold physiology from voice acoustics using machine learning. *The Journal of the Acoustical Society of America*,

147(3):EL264–EL270, 2020.

[60] Gabriel A. Alzamendi, Sean D. Peterson, Byron D. Erath, Robert E. Hillman, and Matías Zañartu. Triangular body-cover model of the vocal folds with coordinated activation of the five intrinsic laryngeal muscles. *The Journal of the Acoustical Society of America*, 151(1):17–30, 01 2022.

[61] 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 Transactions on Audio, Speech, and Language Processing*, 21(9):1929–1939, 2013.

[62] A. Tarantola. *Inverse Problem Theory and Methods for Model Parameter Estimation*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, 2005.

[63] Jonathan J. Deng, Paul J. Hadwin, and Sean D. Peterson. The effect of high-speed videoendoscopy configuration on reduced-order model parameter estimates by bayesian inference. *The Journal of the Acoustical Society of America*, 146(2):1492–1502, 2019.

[64] Zhang Zhaoyan. Voice feature selection to improve performance of machine learning models for voice production inversion. *Journal of Voice*, 2021.

[65] Staffan Björklund and Johan Sundberg. Relationship between subglottal pressure and sound pressure level in untrained voices. *Journal of Voice*, 30(1):15–20, 2016.

[66] Jari Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*. Springer, Dordrecht, 2005.

[67] E. Cataldo, Christian Soize, R. Sampaio, and C. Desceliers. Probabilistic modeling of a nonlinear dynamical system used for producing voice. *Computational Mechanics*, 43:265–275, 2009.

[68] E. Cataldo, C. Soize, and R. Sampaio. Uncertainty quantification of voice signal production mechanical model and experimental updating. *Mechanical Systems and Signal Processing*, 40(2):718–726, 2013.

[69] Rita R. Patel, Johan Sundberg, Brian Gill, and Filipa M.B. Lã. Glottal airflow and glottal area waveform characteristics of flow phonation in untrained vocally healthy adults. *Journal of Voice*, 36(1):140.e1–140.e21, 2022.

[70] Zhaoyan Zhang. Cause-effect relationship between vocal fold physiology and voice production in a three-dimensional phonation model. *The Journal of the Acoustical Society of America*, 139(4):1493–1507, 2016.

[71] Martin Rothenberg. A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. *The Journal of the Acoustical Society of America*, 53(6):1632–1645, 1973.

[72] Judith R. Smitheran and Thomas J. Hixon. A clinical method for estimating laryngeal airway resistance during vowel production. *Journal of Speech and Hearing Disorders*, 46(2):138–146, 1981.

[73] Brad H. Story. An overview of the physiology, physics and modeling of the sound source for vowels. *Acoustical Science and Technology*, 23(4):195–206, 2002.

[74] Tomáš Vampola, Jaromír Horáček, and Ivo Klepáček. Computer simulation of mucosal waves on vibrating human vocal folds. *Biocybernetics and Biomedical Engineering*, 36(3):451–465, 2016.

[75] Michael Döllinger, Zhaoyan Zhang, Stefan Schoder, Petr Šidlof, Boğaç Tur, and Stefan Kniesburges. Overview on state-of-the-art numerical modeling of the phonation process. *Acta Acustica*, 7, 06 2023.

[76] Byron Erath, Matías Zañartu, Kelley Stewart, Michael Plesniak, David Sommer, and Sean Peterson. A review of lumped-element models of voiced speech. *Speech Communication*, 55:667690, 06 2013.

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

[78] Brad H. Story and Ingo R. Titze. Voice simulation with a body-cover model of the vocal folds. *The Journal of the Acoustical Society of America*, 97(2):1249–1260, 1995.

[79] Matías Zañartu, Luc Mongeau, and George R. Wodicka. Influence of acoustic loading on an effective single mass model of the vocal folds. *The Journal of the Acoustical Society of America*, 121(2):1119–1129, 2007.

[80] Takuya Koizumi, Shuji Taniguchi, and Seijiro Hiromitsu. Two-mass models of the vocal cords for natural sounding voice synthesis. *The Journal of the Acoustical Society of America*, 82(4):1179–1192, 10 1987.

[81] Peter Birkholz, Bernd Kröger, and Christiane Neuschaefer-Rube. Articulatory synthesis of words in six voice qualities using a modified two-mass model of the vocal folds. 01 2011.

[82] Matías Zañartu, Gabriel E. Galindo, Byron D. Erath, Sean D. Peterson, George R. Wodicka, and Robert E. Hillman. Modeling the effects of a posterior glottal opening on vocal fold dynamics with implications for vocal hy-

perfunction. *The Journal of the Acoustical Society of America*, 136(6):3262–3271, dec 2014.

[83] Ingo R Titze and Brad H Story. Rules for controlling low-dimensional vocal fold models with muscle activation. *The Journal of the Acoustical Society of America*, 112(3 Pt 1):1064, sep 2002.

[84] Jaromír Horáček, Anne Maria Laukkanen, and Petr Šidlof. Estimation of impact stress using an aeroelastic model of voice production. *Logopedics Phoniatrics Vocology*, 32(4):185–192, 2007.

[85] Gabriel E. Galindo, Sean D. Peterson, Byron D. Erath, Christian Castro, Robert E. Hillman, and Matías Zañartu. Modeling the pathophysiology of phonotraumatic vocal hyperfunction with a triangular glottal model of the vocal folds. *Journal of Speech, Language, and Hearing Research*, 60(9):2452–2471, 2017.

[86] Ingo R. Titze and Eric J. Hunter. A two-dimensional biomechanical model of vocal fold posturing. *The Journal of the Acoustical Society of America*, 121(4):2254–2260, apr 2007.

[87] Peter Birkholz, Bernd J Kröger, and Christiane Neuschaefer-Rube. Synthesis of breathy, normal, and pressed phonation using a two-mass model with a triangular glottis. In *Interspeech 2011: 12th Annual Conference of the*

International Speech Communi- cation Association, pages 2681–2684, 2011. Florence, Italy.

[88] Ingo R. Titze. *The Myoelastic Aerodynamic Theory of Phonation*. National Center for Voice and Speech, 1st edition edition, 2006.

[89] Dinesh K. Chhetri, Juergen Neubauer, Elazar Sofer, and David A. Berry. Influence and interactions of laryngeal adductors and cricothyroid muscles on fundamental frequency and glottal posture control. *The Journal of the Acoustical Society of America*, 135(4):2052–2064, apr 2014.

[90] Zhaoyan Zhang. Mechanics of human voice production and control. *The Journal of the Acoustical Society of America*, 140(4):2614–2635, oct 2016.

[91] Ingo R. Titze. Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model. *The Journal of the Acoustical Society of America*, 111(1):367–376, jan 2002.

[92] Jorge C. Lucero and Jean Schoentgen. Smoothness of an equation for the glottal flow rate versus the glottal area. *The Journal of the Acoustical Society of America*, 137(5):2970–2973, 05 2015.

[93] Matías 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, May 2006.

[94] Brad H. Story. Comparison of magnetic resonance imaging-based vocal tract area functions obtained from the same speaker in 1994 and 2002. *The Journal of the Acoustical Society of America*, 123(1):327–335, 2008.

[95] Brad H. Story, Ingo R. Titze, and Eric A. Hoffman. Vocal tract area functions for an adult female speaker based on volumetric imaging. *The Journal of the Acoustical Society of America*, 104(1):471–487, 1998.

[96] Emiro J. Ibarra, Gabriel A. Alzamendi, and Matíaz Zañartu. Constrained extended Kalman filter for improving Bayesian inference of vocal function from laryngeal high-speed videoendoscopy. In *18th International Symposium on Medical Information Processing and Analysis*, volume 12567, page 125671E. International Society for Optics and Photonics, SPIE, 2023.

[97] Gabriel Galindo. *Bayesian estimation of a subject-specific model of voice production for the clinical assessment of vocal function.* PhD thesis, Department of Electronic Engineering, Universidad Técnica Federico Santa Maria, Valparaiso, Chile, 2017.

[98] T.J. Durbin and S.J. Koopman. *Time Series Analysis by State Space Methods: Second Edition.* Oxford Statistical Science Series. OUP Oxford, 2012.

[99] Dan Simon. Kalman filtering with state constraints: A survey of linear and nonlinear algorithms. *Control Theory & Applications, IET*, 4:1303 – 1318, 09 2010.

[100] Víctor M. Espinoza, Daryush D. Mehta, Jarrad H. Van Stan, Robert E. Hillman, and Matías Zañartu. Glottal aerodynamics estimated from neck-surface vibration in women with phonotraumatic and nonphonotraumatic vocal hyperfunction. *Journal of Speech, Language, and Hearing Research*, 63(9):2861–2869, 2020.

[101] I.R. Titze. *Principles of Voice Production*. National Center for Voice and Speech, 2000.

[102] Andrew M. Vahabzadeh-Hagh, Zhaoyan Zhang, and Dinesh K. Chhetri. Hirano's cover–body model and its unique laryngeal postures revisited. *The Laryngoscope*, 128(6):1412–1418, 2018.

[103] Matías Zañartu, Daryush D. Mehta, Julio C. Ho, George R. Wodicka, and Robert E. Hillman. Observation and analysis of in vivo vocal fold tissue instabilities produced by nonlinear source-filter coupling: A case study. *The Journal of the Acoustical Society of America*, 129(1):326–339, 2011.

[104] Matías Zañartu. *Acoustic coupling in phonation and its effect on inverse filtering of oral airflow and neck surface acceleration*. PhD thesis, School of

Electrical and Computer Engineering, Purdue University, West Lafayette, IN, May 2010.

[105] James B. Kobler, Steven M. Zeitels, Robert E. Hillman, and Jeff Kuo. Assessment of vocal function using simultaneous aerodynamic and calibrated videostroboscopic measures. *Annals of Otology, Rhinology & Laryngology*, 107(6):477–485, 1998. PMID: 9635457.

[106] Joseph S. Perkell, Eva B. Holmberg, and Robert E. Hillman. A system for signal processing and data extraction from aerodynamic, acoustic, and electroglottographic signals in the study of voice production. *The Journal of the Acoustical Society of America*, 89(4):1777–1781, 1991.

[107] 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*, pages 118–121, 2006.

[108] Daryush D. Mehta, Dimitar D. Deliyski, Steven M. Zeitels, M. Zañartu, and Robert E. Hillman. *Integration of transnasal fiberoptic high-speed videoendoscopy with time-synchronized recordings of vocal function*, pages 105–114. Pacific Voice & Speech Foundation, San Francisco, 2015.

[109] Paavo Alku, Tiina Murtola, Jarmo Malinen, Juha Kuortti, Brad Story, Manu Airaksinen, Mika Salmi, Erkki Vilkman, and Ahmed Geneid. Openglot – an open environment for the evaluation of glottal inverse filtering. *Speech Communication*, 107:38–47, 2019.

[110] Peter Birkholz. Glottalimageexplorer – an open source tool for glottis segmentation in endoscopic high-speed videos of the vocal folds. In Oliver Jokisch, editor, *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2016*, pages 39–44. TUDpress, Dresden, 2016.

[111] Ingo R Titze, Jiaqi Jiang, and David G Drucker. Preliminaries to the body-cover theory of pitch control. *Journal of Voice*, 1(4):314–319, 1988.

[112] Zhaoyan Zhang. The physical aspects of vocal health. *Acoustics Today*, 17:60–68, 10 2021.

[113] Anil Palaparthi, Simeon Smith, and Ingo R. Titze. Mapping thyroarytenoid and cricothyroid activations to postural and acoustic features in a fiber-gel model of the vocal folds. *Applied Sciences*, 9(21), 2019.

[114] Georg Luegmair, Dinesh K. Chhetri, and Zhaoyan Zhang. The role of thyroarytenoid muscles in regulating glottal closure in an in vivo canine larynx model. *Proceedings of Meetings on Acoustics*, 22(1):060007, 11 2021.

[115] Emiro J. Ibarra, Jesús A. Parra, Gabriel A. Alzamendi, Juan P. Cortés, Víctor M. Espinoza, Daryush D. Mehta, Robert E. Hillman, and Matías 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. *Frontiers in Physiology*, 12, 2021.

[116] Juan P. Cortés, Jon Z. Lin, Katherine L. Marks, Víctor M. Espinoza, Emiro J. Ibarra, Matías Zañartu, Robert E. Hillman, and Daryush D. Mehta. Ambulatory monitoring of subglottal pressure estimated from neck-surface vibration in individuals with and without voice disorders. *Applied Sciences*, 12(21), 2022.

[117] Gail B. Kempster, Bruce R. Gerratt, Katherine Verdolini Abbott, Julie Barkmeier-Kraemer, and Robert E. Hillman. Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*, 18(2):124–132, 2009.

[118] Joseph S. Perkell, Robert E. Hillman, and Eva B. Holmberg. Group differences in measures of voice production and revised values of maximum airflow declination rate. *The Journal of the Acoustical Society of America*, 96(2):695–698, 1994.

[119] James Kennedy and Russell C. Eberhart. Particle swarm optimization. In *Proceedings of the IEEE International Conference on Neural Networks*, pages 1942–1948, 1995.

[120] Paul Boersma and David Weenink. Praat: Doing phonetics by computer.

[121] M.T. Hagan, H.B. Demuth, M.H. Beale, and O. De Jesús. *Neural Network Design*. Martin Hagan, 2014.

[122] Michael J. Bianco, Peter Gerstoft, James Traer, Emma Ozanich, Marie A. Roch, Sharon Gannot, and Charles-Alban Deledalle. Machine learning in acoustics: Theory and applications. *The Journal of the Acoustical Society of America*, 146(5):3590–3628, 2019.

[123] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[124] Jan G. Švec and Svante Granqvist. Tutorial and guidelines on measurement of sound pressure level in voice and speech. *Journal of Speech, Language, and Hearing Research*, 61(3):441–461, 2018.

[125] Zhaoyan Zhang. Estimating subglottal pressure and vocal fold adduction from the produced voice in a single-subject study (L). *The Journal of the Acoustical Society of America*, 151(2):1337–1340, 02 2022.

[126] Stellan Hertegård, Jan Gauffin, and Per Åke Lindestad. A comparison of subglottal and intraoral pressure measurements during phonation. *Journal of Voice*, 9(2):149–155, 1995.

[127] Martin Rothenberg. Rethinking the interpolation method for estimating subglottal pressure. In *Proceedings of the 10th International Conference on Advances in Quantitative Laryngology, Voice and Speech Research*, pages 111–112. AQL Press, Cincinnati, OH, 2013.

[128] Emiro J. Ibarra, Gabriel E. Galindo, Gabriel A. Alzamendi, Juan P. Cortes, Christian Castro, Rodrigo Manríquez, Alba Testart, and Matías Zañartu. Empirical distribution of glottal edges (edge): A statistical assessment of vocal fold kinematics using high-speed videoendoscopy. *IEEE Journal of Biomedical and Health Informatics*, 2024. In review.

[129] Emiro J Ibarra, Julián D Arias-Londoño, Juan Godino-Llorente, Daryush D Mehta, and Matías Zañartu. Subject-specific modelling of the subglottal pressure estimation from neck-surface vibration signals by domain adaptation. *Authorea Preprints*, 2024.

[130] Emiro J. Ibarra, Julián D. Arias-Londoño, Juan I. Godino-Llorente, Daryush D. Mehta, and Matías Zañartu. Improved subglottal pressure estimation from neck-surface vibrations using transfer learning of deep neural

networks trained from voice production model. In *13th International Conference on Voice Physiology and Biomechanics*, 2024. Accepted.

[131] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[132] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[133] Andreas Ebbehoj, Mette Østergaard Thunbo, Ole Emil Andersen, Michala Vilstrup Glindtvad, and Adam Hulman. Transfer learning for non-image data in clinical research: A scoping review. *PLOS Digital Health*, 1(2):1–22, 02 2022.

[134] Allen D. Hillel. The study of laryngeal muscle activity in normal human subjects and in patients with laryngeal dystonia using multiple fine-wire electromyography. *The Laryngoscope*, 111(S97):1–47, 2001.

[135] Autonomio. Talos computer software, 2020. Retrieved from http://github.com/autonomio/talos Accessed on= 26 July. 2023.

[136] John W. Tukey. *Comparing individual means in the analysis of variance.* Biometrics, 1949.

[137] Ina Steinecke and Hanspeter Herzel. Bifurcations in an asymmetric vocal-fold model. *The Journal of the Acoustical Society of America*, 97(3):1874–1884, 03 1995.

[138] Daryush D. Mehta, Matías Zañartu, Thomas F. Quatieri, Dimitar D. Deliyski, and Robert E. Hillman. Investigating acoustic correlates of human vocal fold vibratory phase asymmetry through modeling and laryngeal high-speed videoendoscopya). *The Journal of the Acoustical Society of America*, 130(6):3999–4009, 12 2011.

[139] Jack J. Jiang, Yu Zhang, and Jennifer Stern. Modeling of chaotic vibrations in symmetric vocal folds. *The Journal of the Acoustical Society of America*, 110(4):2120–2128, 10 2001.

[140] Jesús Parra, Carlos Calvache, Gabriel Alzamendi, Emiro Ibarra, Leonardo Soláque, Sean D. Peterson, and Matías Zañartu. Asymmetric triangular body-cover model of the vfs with bilateral intrinsic muscle activation. *bioRxiv*, 2024.