PHYSIOLOGICALLY BASED FEATURES RELATED TO VOCAL HYPERFUNCTION: FROM LABORATORY TO AMBULATORY DATA

BY

JUAN PABLO CORTÉS SOTOMAYOR, B.S., M.S.

A dissertation submitted

in partial fulfillment of the requirements

for the degree

Doctor of Philosophy

Major Subject: Electronic Engineering

Universidad Técnica Federico Santa María

Valparaíso, Chile

January 2020

"Physiologically Based Features Related to Vocal Hyperfunction: From Laboratory to Ambulatory Data," a dissertation prepared by Juan Pablo Cortés Sotomayor in partial fulfillment of the requirements for the degree, Doctor of Philosophy, has been approved and accepted by the following:

Matías Zañartu
Thesis Advisor

Juan I. Yuz
Chair of the Examining Committee

Date

Committee in charge:

Dr. Matías Zañartu

Dr. Juan I. Yuz, Chair

Dr. Alejandro Weinstein

Dr. Robert E. Hillman

DEDICATION

I dedicate this work to my parents Angélica and Juan for their unconditional support on this journey, as well as my siblings, Carolina and Claudio, for being sources of inspiration to continue higher on my education.

ACKNOWLEDGMENTS

There are many people who I am grateful to have contributed to this thesis work. First of all, I would like to thank my advisor, Professor Matías Zañartu, for his encouragement, interest, and patience to help me pursue this research work. Specifically, I would like to thank him for sharing his vast knowledge which has enriched my studies in speech science, acoustic modeling, and signal processing.

I would also like to thank the members of my thesis committee: Professor Robert E. Hillman, whose clinical work and advice were highly helpful for the goal of this thesis. Professors Juan I. Yuz and Alejandro Weinstein had significant impact on this work due to their useful advice and long-term commitment.

I am grateful for the continuing support and fruitful discussions with members from the Center for Laryngeal Surgery & Voice rehabilitation: Daryush Mehta and Jarrad Van Stan, who were cooperative with the management and processing of data for this thesis. I want to thank Marzyeh Gahssemi and John Guttag from MIT for sharing their vast knowledge on machine learning. I would like thank members from the VPLab at UTFSM: Víctor Espinoza, Gabriel Almazendi, and Christian Castro, for their great support and advice for this thesis work.

Finally, I would like to acknowledge the invaluable help from Andrea during difficult times in this journey, as well as the guidance from Mónica to close this cycle. For them I am truly grateful.

Special acknowledgement to BASAL award number FB0008.

VITA

October 24, 1984	Born at Santiago, Chile
2007-2010	B.S.E.E., University of California Los Angeles,
	Los Angeles, California, USA,
2011-2013	M.S.E.E., University of California Los Angeles,
	Los Angeles, California, USA

PROFESSIONAL AND HONORARY SOCIETIES

Student Member, Institute of Electrical and Electronics Engineers (IEEE)

Student Member, Eta Kappa Nu Honor Society (HKN)

PUBLICATIONS [or Papers Presented]

1. J. P. Cortés, V. M. Espinoza, M. Ghassemi, D. D. Mehta, J. H. Van Stan, R. E. Hillman, J. V. Guttag, and M. Zañartu. Ambulatory assessment of phonotraumatic vocal hyperfunction using glottal airflow measures estimated from neck-surface acceleration. *PLOS ONE*, 13(12):1–22, 12 2018

2. D. D. Mehta, J. H. Van Stan, M Zañartu, M Ghassemi, J. V. Guttag, V. M. Espinoza, J. P. Cortés, H. A. II Cheyne, and R. E. Hillman. Using ambulatory voice monitoring to investigate common voice disorders: research update. *Front. Bioeng. Biotechnol. 3:155. doi: 10.3389/fbioe.2015.00155*, 2015

3. J. P. Cortés, V. M. Espinoza, C. Castro, R. Manríquez, A. Testart, and M. Zañartu. Classification performance of paired subjects with vocal hyperfunction in the presence of subglottal inverse filtering uncertainties: Pilot study under laboratory conditions. Philadelphia, Pennsylvania, May 2019. 48th Voice Foundation Annual Symposium: Care of the Professional Voice (Best Poster Award)

4. J. P. Cortés, G. Alzamendi, A. Weinstein, J. Yuz, V. Espinoza, D. Mehta, J. Van Stan, R. Hillman, and M. Zañartu. Uncertainty of ambulatory airflow estimates and its effect on the classification of phonotraumatic vocal hyperfunction. Quebec, Canada, June 2019. The 13th International Conference on Advances in Quantitative Laryngology, Voice and Speech Research

5. J. P. Cortés, V. M. Espinoza, D. D. Mehta, J. H. Van Stan, R. E. Hillman, and M. Zañartu. Estimación de medidas aerodinámicas ambulatorias con un filtro de kalman para la evaluación de la hiperfunción vocal. Santa Cruz, Chile, Nov. 2018. LXXV Congreso Chileno de Otorrinolaringología

6. V. M. Espinoza, J. P. Cortés, and M. Zañartu. Métodos de evaluación clínica de la voz basados en medidas aerodinámicas y vibroacústicas. Santa Cruz, Chile, Nov. 2018. LXXV Congreso Chileno de Otorrinolaringología

7. J. P. Cortés, V. M. Espinoza, M. Ghassemi, J. V. Guttag, D. D. Mehta, J. H. Van Stan, R. E. Hillman, and M. Zañartu. Aerodynamic ambulatory assessment for phonotraumatic vocal hyperfunction. East Lansing, Michigan, August 2018. 11th International Conference on Voice Physiology and Biomechanics

8. J. P. Cortés, V. M. Espinoza, M. Ghassemi, D. D. Mehta, J. H. Van Stan, R. E. Hillman, J. V. Guttag, and M. Zañartu. Using aerodynamic features and their uncertainty for the ambulatory assessment of phonotraumatic vocal hyperfunction. Las Vegas, Nevada, March 2018. IEEE International Conference on Biomedical and Health Informatics

9. J. P. Cortés and M. Zañartu. Ambulatory classification of patients with muscle tension dysphonia vs. control group. Hong Kong, Oct. 2017. The 12th International Conference on Advances in Quantitative Laryngology, Voice and Speech Research

10. J. P. Cortés, V. M. Espinoza, M. Zañartu, M. Ghassemi, J. V. Guttag, D. D. Mehta, J. H. Van Stan, and R. E. Hillman. Discriminating patients with vocal fold nodules from matched controls using acoustic and aerodynamic features from ambulatory voice monitoring data. pages 95–96, Viña del Mar, Chile, 2016. 10th International Conference on Voice Physiology and Biomechanics

ABSTRACT

PHYSIOLOGICALLY BASED FEATURES RELATED
TO VOCAL HYPERFUNCTION:
FROM LABORATORY TO AMBULATORY DATA

JUAN PABLO CORTÉS SOTOMAYOR, B.S., M.S.

Doctor of Philosophy

Universidad Técnica Federico Santa María

Valparaíso, Chile, 2020

Dr. Matías Zañartu Salas, Advisor

The following thesis proposal describes a framework for the analysis of signal-based features related to vocal pathologies, namely phonotraumatic vocal hyperfunction (PVH) and non-phonotraumatic vocal hyperfunction (NPVH), using an accelerometer attached to the neck-skin in an ambulatory setting. The first stage consists of extracting physiologically relevant features that are associated with PVH on a daily basis. A clinical set-up (In Lab) that captures key components of vocal function, such as acoustics (microphone) and aerodynamics (oral airflow) from a reading passage, provides a set of model parameters to characterize vocal

function. An impedance-based inverse filtering (IBIF) technique is used to estimate glottal airflow and related features from the accelerometer signal and to obtain the same features for the ambulatory data (In Field). An in-depth analysis of IBIF aerodynamic measures is done in the context of machine learning classifiers. Subsequently, an adaptive version of the IBIF filter (i.e., Kalman smoother) is proposed in order to estimate the airflow signal, incorporating modeling and observation noise. The Kalman smoother is compared to the original IBIF filter with In Lab and In Field data within a classification task to determine the efficiency and relevance of both approaches. Additional efforts are presented to provide insights on the capabilities of machine learning tools to be used on PVH and NPVH patients when compared to their matched-controls. First, a case study with 4 pairs of PVH and controls is used to determine how the variability of the IBIF parameters can affect the classification performance with In Lab data. Later, classical machine learning algorithms are used to investigate the nuances in the classification of NPVH subjects vs. controls, while a final effort explores the use of wavelets with deep learning to separate Pre vs Post therapy in NPVH patients. The main contributions of this thesis are: 1) to develop a machine learning framework for the analysis and classification of neck-surface acceleration signals using aerodynamic features; 2) to propose an alternative filtering scheme to IBIF based on adaptive filtering; and 3) to support the first two contributions by exploring pilot studies on salient features for NPVH, IBIF parameter uncertainty, and ther-

apy effects on NPVH. Discussions and conclusions are included in each chapter to interconnect the ambulatory analysis of glottal flow with machine learning, to establish the potential benefits and limitations of these approaches in clinical settings.

Contents

List of Tables

List of Figures

ABBREVIATIONS

VH	Vocal Hyperfunction
PVH	Phonotraumatic Vocal Hyperfunction
NPVH	Non-Phonotraumatic Vocal Hyperfunction
ACC	Neck-skin accelerometer signal
VHM	Voice Health Monitor
GVV	Glottal volume velocity
ACFL	Peak to peak amplitude of the unsteady glottal airflow
MFDR	Maximum flow declination rate
OQ	Open quotient
SQ	Speed Quotient
H1-H2	Difference between first harmonic and second harmonic
HRF	Harmonic Richness Factor
NAQ	Normalized Amplitude Quotient
SPL	Sound pressure level
$f0$	Fundamental frequency
LPC	Linear Predictive Coding
IBIF	Impedance-based Inverse filtering
L1LR	L1-Logistic regression
SVM	Support vector machine

dB Decibels

RMS Root-Mean-Square

KF Kalman Filter

AR Auto Regressive

MA Moving Average

ARMA Auto Regressive Moving Average

pdf Probability Density Function

CWT Continuous Wavelet Transform

CNN Convolutional Neural Network

Chapter 1

Introduction

1.1 Motivation

The overall goal of this work is to improve the contribution of ambulatory voice data to clinical assessment of vocal hyperfunction (VH) [2]. It is hypothesized that some voice disorders such as the formation of vocal fold polyps, nodules, and muscle tension dysphonia are a result of the misuse and inappropriate compensatory mechanisms [12]. Symptoms of VH can go from vocal fatigue [13] to permanent tissue damage [14], and they could be developed during long periods of time through accumulation of vocal loading. However, voice therapists typically access only a snapshot of time using clinical recording techniques such as acoustic and aerodynamic measurements and laryngeal imaging [15, 16, 17, 18] to obtain objective voice data. Standardized subjective clinical assessment is very relevant for this purpose, but it is also limited to specific instances of time where the evaluation takes place [19]. Most information regarding daily voice use is subjective to the perception of the patient [20], which has been found to be er-

ratic and less useful. Therefore, a key step for the objective voice assessment is the incorporation of ambulatory sensors that can measure relevant voice function features over long periods of time. Assessment of voice function through ambulatory monitoring of relevant physiological parameters could have great benefits for clinicians and patients, as it could provide complementary information for clinical voice assessment, which imply improvements in voice therapy, real-time biofeedback, and, hopefully, some progress on the understanding of the etiology for some vocal pathologies. In the last decade there have been some efforts to monitor and quantify voice function using different tools that measure neck-skin vibration with an accelerometer attached just below the thyroid notch [21, 22, 23]. Most of these devices provide estimations of Sound Pressure Level (SPL), fundamental frequency (f_0), and three vocal doses: Phonation time, cycle dose, and distance dose [24]. The purpose of these measures is to quantify vocal loading, which is related to the stress affecting the vocal folds during long periods of time. At evaluation, these measures might provide indirect evidence of the severity of a voice problem, even if they cannot pinpoint the specific etiologies or pathologies [18]. In contrast, aerodynamic measures of voice production, such as glottal airflow and subglottal pressure, provide physiologically insights and can differentiate VH related pathologies from healthy vocal subjects [12, 25, 26, 27]. To improve the assessment of vocal function, a smartphone-based ambulatory monitoring framework was proposed to analyze neck-skin accelerometer signals from daily voice

use. The smartphone application, referred as the Voice Health Monitor (VHM) system [23], has the advantage to be user friendly and easily programmable using the Android operating system. In this thesis, we aim to use the VHM system to include features that are based on physiological events from voice production, for example, the peak to peak amplitude of the unsteady glottal airflow (ACFL or AC-flow) and the maximum flow declination rate (MFDR), using estimations of the glottal airflow [28], and to use an adaptive filter (e.g, Kalman filter) to quantify the uncertainty on those measurements. The goal of this effort is to make the ambulatory assessment of vocal function more aligned with previous methods and findings on VH [12]. In addition, the large amount of data provided by subjects with VH and healthy controls opens up the possibility to explore Machine learning algorithms [29] for supervised and unsupervised classification tasks. Even though there has been a great advance on complex Machine Learning algorithms (e.g., deep learning [30]) for several applications, they usually behave as a "black box" where the input is raw data and little is known about the learning process. In clinical applications such as voice assessment, it is important to consider physiological relevant features as input to learning algorithms, such that clinicians can better interpret the results of such algorithms. This thesis presents the first attempt to incorporate estimations of aerodynamic features from VH patients wearing a VHM during a week with the purpose of applying machine learning algorithms for an analysis of the differences between VH patients and healthy control subjects.

1.2 Goals

1.2.1 General aim

To develop a signal processing and classification framework to study vocal function using ambulatory aerodynamic features.

1.2.2 Specific aims

1. **Specific Aim 1 (SA1)**: To develop a signal processing and machine learning framework using ambulatory aerodynamic measures to distinguish statistical features from both subjects with PVH and NPVH and healthy controls.

2. **Specific Aim 2 (SA2)**: To develop a Bayesian framework to quantify and to explore the role of uncertainty in the estimation of the ambulatory aerodynamic features in the signal structure and classification tasks.

1.3 Hypotheses

Hypothesis 1 (H1) Events of phonotraumatic vocal hyperfunction will be correlated with amplitude-based glottal features (AC-flow,MFDR), while non-phonotraumatic vocal hyperfunction will be correlated with frequency-based glottal features (H1-H2, open quotient). Specifically, ambulatory measures of AC-flow and MFDR will be positively correlated to higher vocal fold collision forces on patients with PVH than healthy-matched controls, while ambulatory measures of

H1-H2 (or other frequency-dependent measure) will have higher values on patients with NPVH than their healthy-matched controls. The results will be reflected on the classification tasks when using the measures alone and normalized with SPL (for AC-flow and MFDR) as well.

Hypothesis 2 (H2) Variability and errors in the ambulatory aerodynamic data will be reduced by incorporating a Bayesian framework in the estimation of glottal features by discarding those with high variability. The incorporation of physiologically model-based features with low variance error in machine learning algorithms will improve classification of PVH subjects vs. healthy-matched controls.

Hypothesis 3 (H3) Classification of NPVH patients vs. matched controls is a more difficult task than classification of PVH subjects vs. matched controls. A deep learning approach to classification, for example the classification of pre vs. post therapy for NPVH subjects, will improve accuracy, at the cost of feature interpretation.

1.4 Contributions

This thesis introduces the first attempt to make a connection between glottal aerodynamic signals taken in clinical settings and related measures from ambulatory settings. By using an impedance-based inverse filtering scheme to estimate the unsteady glottal airflow component from a neck-surface accelerometer, it is

obtained and quantified, for the first time, aerodynamic features in an ambulatory assessment and a comprehensive framework. These features have been shown to be physiologically relevant for vocal hyperfunction in laboratory settings and computational studies. Prior efforts to obtain aerodynamic features from neck surface acceleration were limited to sustained vowels and simple proof of concept examples. The incorporation of subject-specific model parameters from clinical recordings and their application to the same subjects in ambulatory data within a stochastic framework is introduced for the first time in this work. Machine learning algorithms are used to for the first time with highly relevant clinical features (i.e., glottal airflow) to find model parameters on vocal use that are unique on subjects with VH.

There is an opportunity to enhance clinical medicine based on machine learning models. In the specific case of voice assessment, tracking data from patients could provide valuable information regarding the state of a specific voice disorder. The results of machine learning algorithms provide insights on what voice related features have more impact on a daily basis for patients with PVH and NPVH. As a result, clinicians can make better decisions regarding what procedure or therapy to follow, and how the patient is responding to such approach. The ultimate goal would be to optimize the health care system in the voice area by lowering costs of unnecessary treatments.

Chapter 2

Background

2.1 Voice pathologies

Vocal pathologies are a health problem of growing concern in our society. In the US these pathologies affect about 6.6 % of the working population [31] and in Chile there is more than 50 % prevalence of vocal pathologies in teachers [32]. One of the reasons that cause many voice pathologies is the misuse of the voice from a functional point of view, i.e., normal and balanced functioning of the voice is altered causing efforts mainly affecting the vocal folds. These appear when the quality, pitch, or loudness of an individual differ from voice characteristics that are normal for the gender, age, and population of the individual. Fig 2.1 shows healthy vocal folds when they are closed (adducted). Overall, voice pathologies might arise by three distinct factors: Maladaptive or inappropriate voice use, structural, medical, and neurologic alterations of the respiratory, laryngeal, and vocal tract mechanisms, while others originate in direct response to psychological factors. There is also a considerable overlap between these three groupings [18]. A major

grouping of laryngeal pathologies and voice disorders was organized by the Special Interest Division 3 of the American Speech-Language-Hearing Association in a volume called the *Classification Manual for Voice Disorders -I* [33], where clinical pathologies of the laryngeal mechanism where divided broadly into 8 groups:

1. Structural Pathologies of the Vocal Fold
    (a) Malignant Epithelial Dysplasia of the Larynx
    (b) Bening Epithelial and Lamina Propia Abnormalities of the Vocal Fold
    (c) Congenital and Maturational Changes Affecting Voice
2. Inflammatory Conditions of the Larynx
    (a) Cricoarytenoid and Crycothyroid Arthritis
    (b) Acute Laryngitis
    (c) Laryngopharyngeal Reflux
    (d) Chemical Sensitive/ Irritable Larynx Syndrome
3. Trauma or Injury of the Larynx
    (a) Internal Laryngeal Trauma
    (b) External Trauma and Arytenoid Dislocation
4. Systematic Conditions Affecting Voice
    (a) Endocrine Disorders
    (b) Immunologic Disorders
5. Nonlaryngeal Aerodigestive Disorders Affecting Voice
    (a) Respiratory Diseases

(b) Gastroesophageal Reflux Disease

(c) Infectious Diseases of the Aerodigestive Tract

6. Psychiatric and Psychological Disorders Affecting Voice

(a) Psychogenic Conversion Aphonia and Dysphonia

(b) Factitious Disorders or Malingering

(c) Gender Dysphoria or Gender Reassignment

7. Neurologic Disorders Affecting Voice

(a) Peripheral Nervous System Pathology

(b) Movements Disorders Affecting the Larynx

(c) Central Neurologic Disorders Affecting Voice

8. Other Disorders of Voice Use

(a) Vocal Abuse, Misuse, and Phonotrauma

(b) Vocal Fatigue

(c) Muscle Tension Dysphonia (Primary and Secondary)

(d) Ventricular Phonation (Plica Ventricularis)

(e) Paradoxical Vocal Fold Motion (Vocal Fold Dysfunction) or Episodic Dyspnea

The main body of this work will be concentrated on the analysis of vocal fold nodules and polyps (structural pathologies that are benign in the epithelial and lamina propia). Nodules represent an inflammatory degeneration on the superficial layer of the lamina propia. They are usually bilateral and tend to appear

in the location of greatest amplitude of the vocal folds. Figure 2.2 shows vocal folds with bilateral nodules in the close phase. Even in adducted state, there is an opening in the posterior and anterior part of the glottis due to the blocking of the nodules for a complete closure.

Vocal fold nodules/polyps are hypothesized to be developed through conditions of abuse and/or misuse of the vocal mechanism due to excessive and/or "imbalanced" muscular forces [12]. Due to the lack of loudness, patients try to compensate with excessive muscle tension causing a sub-optimal position of the vocal cords [34]. This alteration to the status of the voice, is known as *hyperfunctional voice* (VH) [12, 35] and is associated with a number of recurring pathologies in patients with positive clinical diagnosis.

Figure 2.1: Normal vocal folds (adducted).

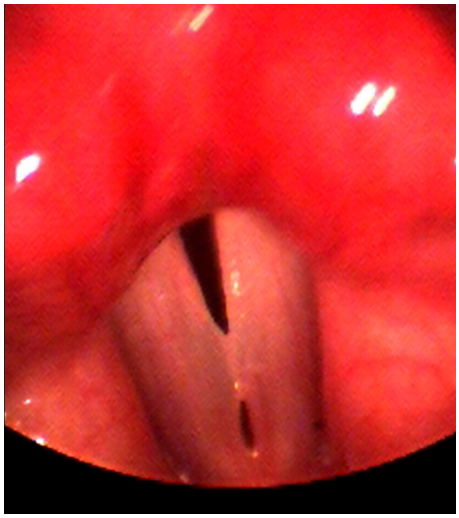

Figure 2.2: Vocal folds with bilateral nodules (adducted).

Vocal hyperfunction can manifest in two conditions: Phonotraumatic (PVH) and non-phonotraumatic (NPVH) [2]. PVH is a type of pathology associated to the increase of muscular tension on the vocal folds, producing high stiffness and higher collision forces on these. Subglottal pressure and higher glottal airflow could also increase above normal levels [12, 36, 37]. There are different degrees PVH, from vocal fatigue, inflammation of vocal fold tissue, to the formation of inflammatory lesions the superficial layer of the lamina propia, such as polyps or nodules. During any of these stages, if the vocal fold trauma diminishes due to vocal therapy, it is possible to recover to a normal voice state. However, if the trauma persists due to higher contact forces from lesions on the vocal folds, these lesions would become permanent and surgery should be the next step for recovery [12]. On the other hand, NPVH is a type of pathology that is associated with

dysphonia due to the incomplete closure and high stiffness of the vocal folds. The etiology of NPVH is less well understood because there is no vocal fold trauma. Fig 2.3 shows the vocal folds of a subject with primary muscle tension dysphonia (MTD, one manifestation of NPVH) during the close phase of the vibratory cycle, where is incomplete closure due to the pathology.

Figure 2.3: Muscle tension dysphonia (adducted).

In order to understand the causes of VH, it is necessary to have objective measures that can describe the mechanisms of laryngeal function. Objective clinical assessment of vocal function includes the use of instruments that measure components, such as acoustic output, aerodynamic function, laryngeal imaging, among others. The main purpose of acoustic analysis is to obtain estimations of f_0, intensity (i.e., sound pressure level, SPL), signal/harmonics-to-noise-ratio, per-

turbation measures, and spectral/cepstral features, most of them related to voice quality. The aerodynamic assessment includes measures of airflow rate and lung volume, subglottal (intraoral) pressure, phonation threshold pressure, and laryngeal resistance. Laryngeal imaging provides information of the gross structure and movements of the vocal folds, vibratory characteristics, and vibratory onset/offset (only with high-speed video). It is recommended to record acoustic signals with a professional grade, condenser type of microphone with unidirectional, or cardioid filtering characteristics [38, 39]. Ideally, the microphone must be positioned from the subject's mouth an approximate distance between 3 to 8 cm for sustained vowels, even though longer distances are acceptable for connected speech, and ambient noise should be minimal (42 dB signal-to-noise ratio recommended) in order to obtain reliable measurements [40]. Aerodynamic measurements are usually taken with a pneumotachohaph device that uses differential pressure across a known resistance to estimate flow rate. To measure oral airflow and pressure, a mask developed by Rothenberg [41] is typically used. The mask contain wire mesh screen vents that serve as the resistance and the pressure drop is calculated from inside the mask relative to atmospheric pressure. Fig 2.4 shows the procedure of taking aerodynamic measurements with the Rothenberg mask. Laryngeal imaging provides more information about the severity and possible etiology of a voice disorder than other measurements [17]. However, it is more invasive to the patient and there is a learning curve for the interpretation of a given video/image

due to visual perception judgment. Typical imaging techniques to assess voice disorders, using a rigid or flexible endoscope, include stroboscopy, kymography, and high-speed video. Each one has its advantages and limitations, and they can be used synchronously with devices that record other measures (i.e., acoustic) to better cross-validate perceptual judgment of vocal fold movement [42].

Figure 2.4: Extraction of aerodynamic measurements with a Rothenberg mask. Voice Production Laboratory (VPLAB), Universidad Técnica Federico Santa María, Valparaíso, Chile

The main purpose of objective measurements for voice functions is to provide clinicians reliable information on speech behavior. All objective tools men-

tioned before provide indirect measurements of vocal function, which are limited by equipment errors, calibration, and subjective interpretation of results [18]. Nevertheless, objective measurements have provided useful knowledge on the predictive power of vocal measures with respect to perceptual or self-rating measures [43, 44]. Using multivariate analysis, Wuyts et al. [45] created the Dysphonia Severity Index (DSI) to assess voice quality based on perceptual analysis and four objective measures: f_0, SPL, maximum phonation time, and jitter (measure of frequency instability). The DSI has been used to discriminate between organic and non-organic etiologies in a large population [46]. Due to the large variability of acoustic measurements between subjects, there is not a single indicator that could predict a voice disorder. Multiple measures are used instead, and subjective analysis is combined to assess a vocal disorder. In the case of dysphonia, there is one feature that has been used to assess it: the Cepstral Peak Prominence (CPP). The cepstral domain, or cepstrum, is the inverse discrete-time Fourier Transform (IDTFT) of the logarithm of the magnitude of the discrete-time Fourier Transform (DTFT) of a signal [47]. This transformation alters the spectrum plot of all harmonic peaks to a cepstrum plot, which emphasizes the peaks of the strongest harmonics, including the fundamental frequency f_0. The CPP compares the magnitude of the cepstral peak to the linear regression of all frequencies in the voice signal [18]. A number of variations on cepstral analysis have been shown to correlate with perceptual judgements of breathy and rough voices [48, 49, 50, 51]. In a

meta-analysis of acoustic measures, CPP resulted as a robust measure correlated to listener judgments of dysphonia severity [52]. The main issue with acoustic measures is that they do not provide an understanding of the etiology of the voice disorder. They provide relationships with symptoms of a voice pathology. For example, vocal effort has been associated with SPL, LH ratio and HNR for both speaker and listener ratings [53]. Voice quality, such as breathy voice, has strong correlation with spectral features such as the difference in magnitude between the first and second harmonic (H1-H2) [54, 55]. A measure called relative fundamental frequency (RFF) has been associated with vocal effort for subjects with VH [56] and spasmodic dysphonia [57]. Objective assessment using aerodynamic measures show promising results on the etiology of VH that includes nodules and polyps [12]. By using inverse filtering of the oral flow signal, reliable measurements of the glottal flow can be achieved in different voice conditions [25]. Subjects with VH have higher values of AC-flow, MFDR and subglottal pressure compared to normal groups [12, 26, 58, 27, 59]. Numerical models of the vocal folds can explain the experimental results of the resulting compensation of VH subjects due to increasing aerodynamic efforts to keep an adequate acoustic level [60]. The present thesis continues the effort of providing tools and evidence to calculate glottal flow and investigating how the signal compares between subjects with VH and controls.

2.2 Ambulatory voice monitoring

Most vocal pathologies develop during a period of time where subjects might develop stress on the vocal folds due to high vocal demand, which usually is described as vocal fatigue [13]. As a consequence, it is beneficial for the speech therapist to have objective information of the voice use during long periods of time for a subject who suffers from vocal fatigue and might develop an organic pathology. Therefore, researchers have developed monitor devices whose purpose is to monitor voice use and/or obtain real-time biofeedback to the user. Holbrook et al. [61] developed one of the first devices for ambulatory bio-feedback with a contact mic over the traquea. Zicker et al. [62] developed a similar device but instead used a throat microphone for biofeedback. Both devices only use sound level as vocal parameters. Ryu et al. [63] and Szabo et al. [64] designed ambulatory monitors with a contact mic (over trachea) to monitor duration of voicing while at the same time sound level and fundamental frequency, respectively. Other devices include a monitor with a piezoelectric microphone attached to the thyroid that measures duration and f_0 [65], a contact microphone that measures duration and sound level [66], and free-air microphone (head mount) that measures sound level and duration [67].

Currently, there have been three commercially available ambulatory voicing devices for clinical and research purposes: The Ambulatory Phonation Monitor

(APM) (KayPENTAX, Lincoln Park, NJ, USA), the VoxLog (Sonvox AB, Umeå, Sweden), and the VocaLog (Griffin Laboratories, Temecula, CA). While the three of them can monitor SPL and phonation time with biofeedback, in addition, the APM and VoxLog provide f_0 measurements and additional vocal doses related to vocal loading [24]. The APM employs medical-grade adhesive to hold the accelerometer against the neck, while the VoxLog employs a collar for the same purpose, and the latter includes an air microphone to measure environmental noise. The VocaLog employs a collar as well but uses a contact microphone instead of an accelerometer. The devices only store the vocal measurements, i.e., no data is recorded due to privacy issues. Calibration for SPL is different for all three devices, and there is report of some differences in performances, mostly due to the analysis window employed by each device and by the differences in the calibration procedure [68].

Different efforts have been reported to monitor daily voice use in order to obtain insights on the pathophysiological process leading to vocal pathologies. A Digital Signal Processor (DSP) prototype for a portable device captures features related to voice quality, such as f_0, jitter, and relative average perturbation, using a small, contact-less, microphone signal which only captures vowels for offline processing [69]. An ambulatory system based on surface electromyography for detecting muscle activation is used to classify simulated characteristics of MTD from normal voice in a laboratory setting, with potential for ambulatory mon-

itoring [70]. A portable vocal accumulator (PVA) designed to record neck-skin vibration signals using an accelerometer (similar to the APM) was tested on 99 participants with normal or dysphonic voices,and ambulatory testing was made with 4 participants for an average of a day of recordings [21]. Using a similar accelerometer setup, a voice dosimetry developed by The National Center for Voice and Speech (NCVS) [22] was used to measure silence and voice accumulation from teachers during two weeks in which subjects voiced for about 23% of the total time at work [71]. The same dosimeter was used for an objective study of vocal fatigue where seven professional singers used the device for a two-week period [72]. In a study of silence and voicing accumulation, the APM was used to determine voicing duration for 26 primary school teachers during an average of 4 hours per workday, that were divided in three groups: those with organic voice disorders, those with functional voice symptoms, and normal voice quality [73]. Another study estimated the voice quality of healthy college singing students using the VoxLog device plus an acoustic microphone to obtain parameters such as long-term average spectrum slope, alpha ratio, and harmonic-to-noise ratio [74]. In addition to those devices, the Voice Health Monitor (VHM) system improves on the previous ones by incorporating a smartphone application [23]. The cellphone is connected to a special designed circuit that conditions the acceleration signal coming from an accelerometer attached to the neck-skin, in the same design as the APM. Similarly, the VHM extracts SPL estimations (previous a calibration

procedure), f_0, and vocal doses every 50 ms. Studies using the VHM on vocal hyperfunction have been reported on ambulatory data from subjects wearing the device from 1 to 4 weeks [11, 75].

Perceptual assessments of voice use have also been reported in ambulatory monitoring. In a study, school teachers with self-estimated voice problems were recorded along match control subjects using the APM device, and self-subjective voice evaluations were correlated with the APM results [76]. A four-day follow-up study on self-reported voice condition on 27 teachers was performed using a Voice-Care device [77], where variation in vocal SPL was significantly associated with self-reported voice conditions [78]. Music theory teachers were monitored during one week with a voice dosimeter and correlations were calculated between vocal load index (VLI) and self-assessed voice quality, vocal fatigue, and amount of singing and speaking voice produced [79]. Six patients with voice disorders were matched with healthy controls and six low voice users for one week of phonation monitoring using the APM and self-reports were correlated with phonation time [80]. In a large study of ambulatory monitoring of 84 subjects with VH (74 matched controls) during one week concluded the reliability of ratings of difficulty producing soft, high-pitched phonation, discomfort, and fatigue using the VHM device [81].

The estimated features obtained from ambulatory data usually are f_0, SPL, and a set of vocal doses that include phonation time, cycle dose, and distance

dose [24], derived from the same estimations of SPL and f_0. These features are related to pitch, intensity, and vocal loading, respectively. However, these are not features related to aerodynamic process of the vocal folds that could contribute information on the etiology of a particular pathology. For example, it has been shown that SPL measures do not have significant differences between a group of PVH and normal subjects [75]. Changes in pitch or vocal loading might be a secondary effect due to a primary lesion, for example, the change in mass of the vocal folds due to nodules or polyps. In the specific case of PVH, features based on aerodynamic measures (i.e., AC-flow, MFDR, collision forces) are in direct connection to the mechanical change that might produce a primary organic vocal pathology. Estimation of features based on empirical equations for estimating collision forces of the vocal folds have been investigated in [82]. The estimation of another aerodynamic measure, subglottal pressure, is currently under research for ambulatory purposes through its estimation from neck-surface acceleration [83, 84]. Most common vocal pathologies have origin in biomechanical disorders. These disorders are intrinsically related to changes in aerodynamic forces compared to normal voice production. Measures of those changes can be critical to determine the origin of a specific pathology, such as PVH [85]. However, estimation of aerodynamic features is a difficult task, even in clinical cases. It usually involves a process of inverse filtering to remove resonances of the vocal tract [25] or estimation of the neck-skin impedance if the filtering is done through

an accelerometer [85]. Moreover, the inverse filtering is limited to vowels or short phrases, with a few examples on ambulatory settings with a small pool of subjects [10].

2.3 Classification and clustering with machine learning

Since early 1980s there has been studies for the detection and classification of vocal fold pathologies using basic techniques of pattern recognition, when machine learning where not as popular as it is nowadays [86]. Most of the work related to classification of vocal fold pathologies use sustained vowels for their experiments, which its limited to the estimation of acoustics of the vocal tract to estimate the glottal flow itself. Therefore, most of them use an acoustic microphone. Typical datasets are recordings of normal and pathological subjects with sustained vowels or continuous speech. Open available standard datasets include the Massachusetts Eye and Infirmary (MEEI) dataset, Saarbruecken Voice Database (SVD), and the Arabic Voice Pathology Database (AVPD). Details of these databases can be found in [87]. Some work related to the classification of phonotraumatic vocal hyperfunction (vocal nodules and /or polyps) can be found in Krishna et al. [88], where they utilized energy spectrum features and k-nearest neighbor (KNN) to classify several vocal disorders, including vocal fold polyps and nodules, using sustained vowels from MEEI. Behroozmand et al. [89] used

Wavelet packet sub-band and Mel-Frequency Cepstrum Coefficients (MFCCs) to detect vocal fold nodules and polyps with artificial neural networks (ANN) and Support Vector Machines (SVM) from a local dataset. Fonseca et al. [90] utilized Daubechies' discrete wavelet transforms with SVM in order to detect vocal fold nodules from voice signals with sustained Brazilian Portuguese phonemes. Nayak et al. [91] used Wavelet transformation patterns with ANNs to detect hyperfunctioning of vocal folds from the MEEI database. Turkmen et al. [92] used videolaryngostroboscopy to perform nodule-cyst classification with SVM, Random Forest and KNN. Aguiar et al. [93] utilized a vector-quantizing-trained distance classifier with LPC and Mel cepstral features to classify different vocal pathologies, including nodules, using the MEEI database. Arias-Londoño et al. [94] used Hidden Markov Models (HMM) with MFCC and short-term noise parameters to classify a variaty of vocal pathologies, including nodules and polyps. They used the MEEI and an internal database from the Universidad Politécnica de Madrid. Carvalho et al. [95] used ANN with Wavelet features to detect vocal fold nodules from sustained /a/ Portuguese vowels. Saeedi et al. [96] tested the MEEI database with Wavelet Filter Banks and SVM to detect a variety of pathologies, including nodules and polyps, where they reach 100% accuracy. Similarly, Muhammad et al. [97] developed a multidirectional regression-based features using Gaussian Mixture Models (GMM) to classify several vocal pathologies, including nodules, with a 99% accuracy on a database composed of arabic digits. Al-Nasheri et al.

[98] utilized three databases: the MEEI, SVD, and AVPD to classify several vocal pathologies using autocorrelation and entropy features.

Regarding NPVH, Schlotthauer et al. [99] used acoustical measurements with an ANN; Hemmerling et al. [100] used several acoustic features with their statistics to classify hyperfunctional dysphonia (among other pathologies) using Random Forest; and Markaki et al. [101] investigated the combination of modulation spectral features and MFCCs to classify dysphonia, nodules, and polyps using an SVM with the MEEI database. Recently, Sklanny et al. [102] developed a genetic classification algorithm to assess glottal insufficiency of vocal fold nodules in children using acoustic and electroglottographic signals. There are plenty of work related to classify different voice pathologies, however, they all share similar databases: sustained vowels from healthy and pathological subjects with microphones under laboratory conditions. For example, a common database, the MEEI, has been used for several studies in vocal pathologies. Usually, classification scores for this database are very good, above 90% in accuracy. However, a study from Daoudi et al. [103] revealed some issues with MEEI. Specifically, they found that the dataset is perfectly separable and they found a single scalar parameter capable of perfect classification accuracy. Therefore, MIIE could be considered as a "toy example" [103] not suitable for voice research where high classification is not the most important characteristic. Having only specific vowels on these databases could contribute to a high bias for separability, but the etiology and development

of a voice pathology is complex and it requires the analysis of a richer set of data to better understand the problem.

Classification/detection of other vocal pathologies using running speech have been reported in [104] for Parkinson's disease, [105] for physiological and neuromuscular larynx pathologies, and [106] for a variety of organic, neurological, traumatic, and psychogenic factors. Nonetheless, these studies with continuous speech have been done under laboratory conditions as well, which limits all the dynamics that can occur in spontaneous speech in an ambulatory setting. There are very few studies of ambulatory data, mainly on the use of a portable dosimeter, as mentioned in the previous subsection, [22] to measure SPL, f_0, and vocal doses [24]. Relevant work using this device has been reported on the variations of these measures between occupational versus non-occupational settings for teachers [107], the silence and voicing accumulations between teachers with and without voice disorders [73], and the comparison of vocal-dose measures using formulas for estimating collision stress on the vocal folds from schoolteachers [82]. However, those studies do not use advance statistical learning (i.e., machine learning) for classification. On ambulatory voice monitoring, Ghassemi et al. used the VHM system with supervised machine learning to classify PVH subjects vs. controls using high-order statistical features of SPL, f_0 and vocal doses [11]. The same group used an unsupervised machine learning technique called symbolic mismatch to identify clusters within the ambulatory data from NPVH subjects and controls

[108]. No prior studies have included aerodynamic features in an ambulatory setting.

Most of the work related to classification of vocal pathologies use supervised learning: it consists on the use of features to construct a parametric model to predict a label (or response) [109]. The process for finding the best parameters that fits the label is called *training*. The process of applying the trained model to unseen data is called *testing*.Supervised learning works well when the label assigned to an instance of the data corresponds to the true label of classification. However, there are cases in which it is difficult to assign a label to a data point. For example, some subjects with voice pathologies do not always exhibit the pathology during daily voice use, just as well as some normal subjects might show characteristics of a pathology in their voice use. People experience different degrees of vocal effort depending on how they are using their voice during the day. For example, a person could have more difficult voicing after work than at the beginning of the day. Individuals who have vocal pathologies might not always exhibit features related to the pathology, in the same way as individuals with normal vocal folds sometimes might exhibit voicing that is related to a particular pathology. Therefore, it is necessary to identify segments of time for which phonation becomes more problematic. In the case of lacking a reliable set of labels, unsupervised learning [109] can be useful to find patterns or clusters in the data without requiring a ground truth label set for every data point. Un-

supervised learning involves clustering data without having ground truth labels. The task of finding common patterns becomes difficult when voicing has a great variability within days and across subjects.The challenge becomes difficult as we narrow down specific voice behavior patterns and cluster them in categories to aid diagnosis and therapy.

2.4 Bayesian tracking

One of the limitations of previous models is the lack of a mechanism to quantify measurement uncertainty for clinical applications. This is important since different combination of parameters in a model could give the same output, but there is no way to explain the most likely solution within a credibility interval for a particular combination. In this case, a probabilistic state-pace approach to modeling might be useful for tracking uncertainty in measurements. In dynamic systems, the state-space approach focuses on the state vector, which contains all relevant information of the system under investigation. For example, in vocal tract estimation problems, the state vector could represent the air-volume of traveling waves from a model with concatenated tubes [110]. The measurement vector represents observations of the system that are related to the state vector. In order to analyze and make inferences about this system, it is necessary to have a state model describing the evolution of the system and an observation model

relating the measurements to the state. Assuming a probabilistic framework, both models should incorporate noisy distributions as well. In this context, Bayesian estimation is the process of constructing the posterior probability density function (pdf) of the state based on all available information, including the received measurements [111]. Recursive filtering is used when the problem requires a new estimate from past and current measurements, where the posterior pdf of the state is predicted with past measurements (prediction stage) and then updated with the current measurement (update stage). Bayes's theorem is the mechanism to update information about the state based on new data information. The general framework for nonlinear Bayesian tracking is the following: Consider a state sequence $\boldsymbol{x}_n, n \in \mathbb{N}$ of a target given by:

$$\boldsymbol{x}_n = \boldsymbol{f}_n(\boldsymbol{x}_{n-1}, \boldsymbol{v}_{n-1}) \tag{2.1}$$

where $\boldsymbol{f}_n : \Re^{m_x} \times \Re^{m_v} \longrightarrow \Re^{m_x}$ is a possible nonlinear function of the state $\boldsymbol{x}_{n-1}$, $\boldsymbol{v}_n, n \in \mathbb{N}$ is an i.i.d. process noise sequence, m_x, m_v are the dimensions of the state and process noise, respectively. The objective of tracking is to recursively estimate $\boldsymbol{x}_n$ from measurements:

$$\boldsymbol{z}_n = \boldsymbol{h}_n(\boldsymbol{x}_n, \boldsymbol{w}_n) \tag{2.2}$$

where $\boldsymbol{h}_n : \Re^{m_x} \times \Re^{m_w} \longrightarrow \Re^{m_z}$ is a possible nonlinear function of the state

$\boldsymbol{x}_n$ and $\boldsymbol{w}_n, n \in \mathbb{N}$ is an i.i.d. measurement noise sequence. The goal of Bayesian estimation is to find filtered estimations of $\boldsymbol{x}_n$ at time n given the measurements $\boldsymbol{z}_{1:n}$ up to time n. Therefore, it is necessary to construct the pdf $p(\boldsymbol{x}_n|\boldsymbol{z}_{1:n})$. Assuming that $p(\boldsymbol{x}_{n-1}|\boldsymbol{z}_{1:n-1})$ at time $n-1$ is available, we can obtain the prior pdf of the state at time n with the Chapman-Kolmogorov equation:

$$p(\boldsymbol{x_n}|\boldsymbol{z}_{1:n-1}) = \int p(\boldsymbol{x}_n|\boldsymbol{x}_{n-1}) p(\boldsymbol{x_{n-1}}|\boldsymbol{z}_{1:n-1}) d\boldsymbol{x_{n-1}} \tag{2.3}$$

At time n, when a new measurement $\boldsymbol{z}_n$ becomes available, the prior can be updated using Bayes's rule:

$$p(\boldsymbol{x_n}|\boldsymbol{z}_{1:n}) = \frac{p(\boldsymbol{z}_n|\boldsymbol{x}_n) p(\boldsymbol{x_n}|\boldsymbol{z}_{1:n-1})}{p(\boldsymbol{z_n}|\boldsymbol{z}_{1:n-1})} \tag{2.4}$$

where the normalizing constant $p(\boldsymbol{z_n}|\boldsymbol{z}_{1:n-1}) = \int p(\boldsymbol{z}_n|\boldsymbol{x}_n) p(\boldsymbol{x_n}|\boldsymbol{z}_{1:n-1})$ depends on the likelihood function $p(\boldsymbol{z}_n|\boldsymbol{x}_n)$ defined by the measurement model in eq. 2.2 and the known statistics of $\boldsymbol{w}_n$. The recurrence relations of prediction (eq. 2.3) and update (eq 2.4) form the basis of the optimal Bayesian solution. In general, these equations cannot be determined analytically. There are restrictive cases when the solution is traceable, as well as sub-optimal solutions, which will be explained in chapter 4. Bayesian methods for estimating underlying physical properties have been applied successfully in dynamical models of the vocal folds, such as in [112] and [113], where they use particle filtering and an extended

Kalman filter, respectively.

2.5 Summary

This section provided an overview on the topics related to this thesis work. It began with a description of vocal fold pathologies, specifically PVH and NPVH, which are common among women and are associated to occupations with high vocal demand (e.g., teachers, singers) [72, 107, 78, 114, 115]. Bad vocal behavior, including talking loudly, using inappropriate pitch, and inefficient phonation might produce VH. Estimating aerodynamic measures, e.g., glottal airflow, can provide insights on the pathophysiology of VH, specially in the case of PVH, where organic lesions occur. Although aerodynamic measures can be extracted in laboratory, vocal behavior is meaningful in the long term, as patients develop VH through time in daily activities. Ambulatory monitoring can provide a useful tool for detecting and/or summarizing specific vocal behavior. Using an accelerometer attached to the neck-skin of the patient has become common in recent research of ambulatory monitoring. This thesis combines for the first time efforts on obtaining estimation of aerodynamic measures on ambulatory settings for the objective assessment of VH. In addition, a bayesian framework is included to improve the estimations of the glottal airflow signal by incorporating variability to the estimation. Additional analysis is provided using machine learning for difficult tasks such as classifying

NPVH with matched-controls and pre-therapy vs. post therapy. The conclusions explain how the hypotheses and aims relate to each experiment.

Chapter 3

Assessment of vocal hyperfunction using ambulatory data

The following chapter provides methods and results on the use of ambulatory voice data from neck-skin acceleration signal by estimating features from the glottal flow. Following the main objective of aim 1 (SA1), the analysis will be based on statistical features from the ambulatory data using IBIF and compare PVH subjects (pre-therapy) with healthy-matched controls. Machine learning algorithms will allow to classify subjects in a weekly basis, while statistical analysis will determine the features that are salient for PVH subjects. A final discussion on how these results relate to hypothesis 1 (H1) finishes the chapter.

3.1 Subglottal impedance based inverse filtering for ambulatory monitoring of voice

In this section, the IBIF algorithm [85] is summarized but also extended and

optimized for ambulatory voice monitoring. The IBIF is a model-based scheme to estimate the glottal airflow from neck-surface acceleration [85]. The method uses a mechano-acoustic transmission line model to account for the acoustic propagation in the subglottal system and neck skin characteristics. The scheme is illustrated in Fig 3.1, where the electrical equivalent circuit shows the interconnection between the subglottal tracts above and below the location of the accelerometer (sub1 and sub2, respectively) and load impedance of the skin Z_{skin}, that also includes the radiation load of the accelerometer sensor Z_{rad}. The glottal airflow signal estimate $\hat{u}_g(t)$ to be obtained from the accelerometer signal $\dot{u}_{skin}(t)$ is calculated using Eq (3.1):

$$\hat{u}_g(t) = \mathcal{F}^{-1}\left(-\frac{\dot{U}_{skin}(\omega)\cdot A_{acc}}{T_{skin}(\omega)}\right), \tag{3.1}$$

with

$$T_{skin}(\omega) = \frac{H_{sub1}(\omega)\cdot Z_{sub2}(\omega)\cdot j\omega}{Z_{sub2}(\omega)+Z_{skin}(\omega)}, \tag{3.2}$$

$$Z_{skin}(\omega) = \frac{1}{A_{acc}}\left\{R_m + j\omega M_m - \frac{j}{\omega}K_m + Z_{rad}(\omega)\right\}, \tag{3.3}$$

$$Z_{rad}(\omega) = \frac{j\omega\cdot M_{acc}}{A_{acc}}, \tag{3.4}$$

where $\mathcal{F}^{-1}(\cdot)$ is the inverse Fourier transform, $H_{sub1}(\omega) = U_{sub1}(\omega)/U_{sub}(\omega)$ is

the transfer function of subglottal section *sub1* (see Fig 3.1), A_{acc} the accelerometer area (cm^2), M_{acc} the accelerometer mass (gr), and $\dot{U}_{skin}(\omega)$ is the acceleration signal in frequency domain. Z_{sub2} and H_{sub1} are calculated using an anatomically based, acoustic model of the subglottal system [116, 117, 85]. Z_{rad} corresponds to the radiation impedance from the accelerometer. All frequency and time-domain expressions are sampled and processed appropriately [118].

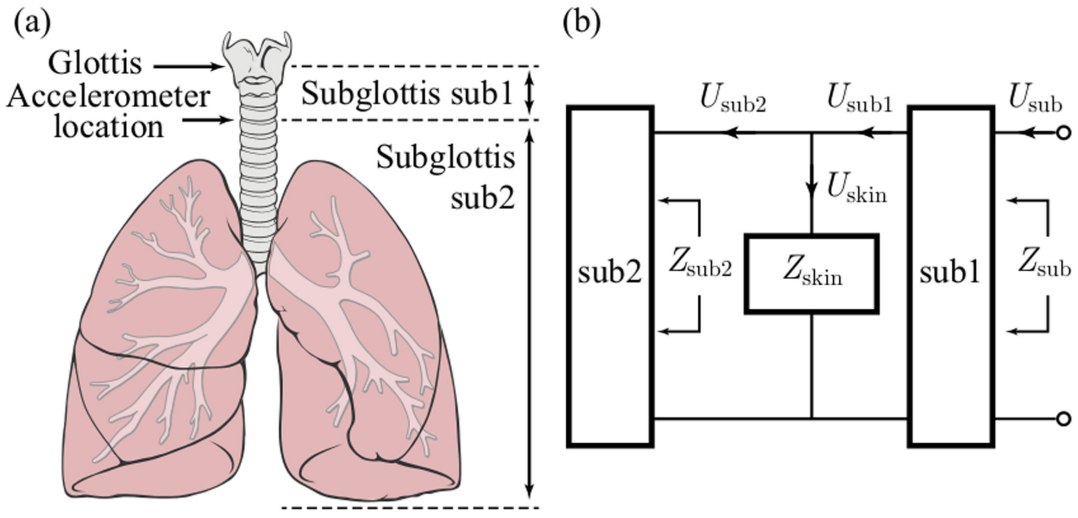

Figure 3.1: Representation of the subglottal system. (a) Accelerometer position and *sub1* and *sub2* system parts. (b) A mechano-acoustic analogy of the subglottal system including load impedance from skin. Reproduced with permission.

In order to use IBIF as a signal processing tool, subject-specific parameters need to be estimated. These IBIF parameters are scaling factors that adjust default values of the mechanical impedance model of neck skin surface, length of

the trachea, and accelerometer location. The parameters are represented in a set $\mathbf{Q} = \{Q_i\}_{i=1,\ldots,5}$ for neck skin resistance R_m, mass M_m, and stiffness K_m, as well as length of the trachea $L_{trachea}$ and accelerometer placement L_{sub1}. Each of these Q parameters is bounded to maintain physiological plausibility [85]. The magnitude terms in Eq (3.3) are the default values for each parameter [119], which are scaled for normalized Q factor as, $R_m = 2320 \cdot Q_1$ in $(\mathrm{g} \cdot \mathrm{s}^{-1} \cdot \mathrm{cm}^{-2})$, $M_m = 2.4 \cdot Q_2$ in $(\mathrm{g} \cdot \mathrm{cm}^{-2})$, $K_m = 491000 \cdot Q_3$ in $(\mathrm{dyn} \cdot \mathrm{cm}^{-3})$, and for $Z_{sub2}(\omega)$, $L_{trachea} = 10 \cdot Q_4$, and $L_{sub1} = 5 \cdot Q_5$ are in (cm). Note that default model parameter are obtained for $\mathbf{Q} = [1, 1, 1, 1, 1]$ [85].Using these subject-specific factors will allow to filter out neck-skin and subglottal resonances, making the estimated glottal airflow signals comparable between subjects.

To obtain subject-specific IBIF parameters, we compare the IBIF-derived glottal airflow waveform estimates with that from the current gold standard, namely an inverse filtered glottal airflow signal obtained from recordings using a CV pneumotachograph mask [120]. Inverse filtering in this case is a challenging task given the reduced bandwidth of the CV mask due to the airflow transducers (PT-2E, Glottal Enterprises) and the type of voices that will be analyzed (high-pitched female voices exhibiting pathology). Inverse filtering of the oral airflow was performed using a semi-automatic approach, as recently described in[27]. This approach was particularly designed to inverse filter normal and pathological high-pitched voices from a CV mask signal.

Once we obtain an estimate of the glottal airflow from the CV mask, we run a Particle Swarm Optimization (PSO) scheme [121], which consists in the optimization of a non-linear continuous fitness function thorough the search of optimal "particles" (parameters) by searching its best set. For this case, PSO searches the optimal Q parameters that represent the subject's anatomical features. The fitness function in this optimization process needs to yield robust and consistent solutions. We minimize the following normalized weighted absolute error (NWAE) function, such that

$$\mathrm{NWAE}(\mathbf{Q}) = \sum_{i=1}^{3} w_i \cdot e_i(\mathbf{Q}), \tag{3.5}$$

with

$$\sum_{i=1}^{3} w_i = 1, \quad 0 \leq w_i \leq 1, \tag{3.6}$$

and

$$e_i(\mathbf{Q}) = \frac{\sum_{n=0}^{N-1} \left|\Delta^{(i-1)}\tilde{u}_g - \Delta^{(i-1)}\hat{u}_g\right|}{\sum_{n=0}^{N-1} \left|\Delta^{(i-1)}\tilde{u}_g\right|}, \tag{3.7}$$

where $\tilde{u}_g$ is the CV mask-based inverse-filtered glottal airflow signal, $\hat{u}_g$ is a time-aligned IBIF-based glottal airflow signal, $\Delta^{(i-1)}$ the time-derivative operator of order $(i-1)$, and i represents the index of the corresponding error function e_i and its weight w_i. Each weighting w_i was set to $0.\bar{3}$. The increased order of the time-derivative operator is used to balance the energy of higher harmonics in NWAE to avoid over-fitting in the low frequency range. Therefore, the optimization problem

is stated as:

$$\hat{\mathbf{Q}} = \arg\min_{\mathbf{Q}} \text{NWAE}(\mathbf{Q}), \text{ subject to } \mathbf{Q} \in \mathbf{D} \quad , \tag{3.8}$$

where $\mathbf{D} = \{D_i\}_{i=1,\ldots,5}$, is a set of restrictions for each parameter within the $\mathbf{Q}$ set that is designed to maintain physiological plausibility [85]. To reduce the computational load of PSO, several configurations of subglottal systems were pre-calculated (i.e., before the PSO algorithm started) for a set of equally spaced values of tracheal length and accelerometer position. Each pre-calculated (Z_{sub} and H_{sub1}) transfer function was indexed and retrieved inside the PSO algorithm. This approach substantially reduces the computational time of the optimization process.

The time-alignment of the oral airflow and acceleration signals is as follows. A first approximation is to align using the sample cross-correlation function [118] and find the maximum peak shifted in the neighborhood of mid-lag position [122]. To improve this initial approximation, a delay parameter d is added in the PSO algorithm by shifting the indices of signal vectors (oral airflow and neck acceleration). Since the shifted signal (oral airflow) is delayed for only a few samples, the search space is limited to $d \in D_0 = [-d_0, d_0]$ where d_0 is a small number $\in \mathbf{Z}^+$. Then, given $N(\gg d_0)$ samples of data, $\tilde{u}_g$ and $\hat{u}_g$ are replaced in (3.7) by

$$\hat{u}_{gt}(nT) \quad ; \quad n \in [d_0, N-1-d_0], \text{ and} \tag{3.9}$$

$$\tilde{u}_{gtd}(nT) \quad ; \quad n \in [d_0+d, N-1-d_0+d]. \tag{3.10}$$

Note that $\hat{u}_{gt}(nT)$ is a trimmed version of $\hat{u}_g(nT)$ and $\tilde{u}_{gtd}(nT)$ is a trimmed, delayed version of $\tilde{u}_g(nT)$ both with $N - 2d_0$ samples, where T is the sampling period. An initial value for d_0 was half the average glottal cycle duration.

In the case of incomplete glottal closure, coupling between the subglottal tract and vocal tract is embedded in the resulting dipole source [123]. Therefore, the glottal flow with all the source-filter interactions can be estimated without the need to model glottal coupling.

3.2 Experimental setup and participants

The human studies protocol used to collect the data for this study (Ambulatory monitoring of vocal function to improve voice disorder assessment: #2011P002376) was approved by the Institutional Review of the Partners Healthcare System - the Massachusetts General Hospital is a founding member of this organization. Study participants were 48 pairs of adult females (total of 96 subjects) with each pair comprised of one patient with PVH (diagnosed with vocal nodules) and one normal control subject matched to the patient by age and occupation (see Table 5.1 for more details). Diagnoses were based on a complete team evaluation by laryngologists and speech-language pathologists at the Massachusetts General Hospital Voice Center that included (a) a complete case history, (b) endoscopic imaging of the larynx, (c) aerodynamic and acoustic assessment of vocal function based on Mehta et. al. [34], (d) a patient-reported Voice-Related Quality of Life question-

naire, and (e) a clinician-administered Consensus Auditory-Perceptual Evaluation of Voice assessment (CAPE-V). All patients were enrolled prior to the administration of any voice treatment. Written informed consent was obtained from all subjects. All subjects were 18 years of age or older. Due to the higher incidence of female patients with PVH than men in the overall population [37], only women were subjects for this study. Zhukhovitskaya et. al.[124] have shown significant differences ($p < 0.0001$) in the number of bilateral midfold lesions between males and women. Moreover, the inclusion of men would create confounding variables due to sex-specific characteristics. The matching is done to normalize for general vocal behavior differences. For example, males and females have anatomical differences, there are voice changes with age (for example, presbyphonia usually occurs when people gets older), and the type of occupation is related to how much voicing is used during a typical day at work. On the other hand, the subject-specific parameters from IBIF are normalized for each individual, so signals can be comparable, due to differences in neck-skin and subglottal anatomy. Therefore, these are not matched on healthy-patient pairs.

Each subject was recorded as they engaged in normal daily activities during one week using the smartphone-based ambulatory voice monitor [2, 23]. The system employs an accelerometer attached to the front of the neck below the larynx as the phonation sensor (see Fig 3.2). The sampling frequency was 11,025 Hz and the average total recording time for a subject was approximately 80 hours,

as in [2, 75].

Figure 3.2: Example of VHM system. Illustration of the smartphone-based ambulatory voice monitor that uses a neck-surface accelerometer attached to the skin halfway between the thyroid prominence and the suprasternal notch of a female subject.

Each subject underwent a session in the laboratory to obtain a subject-specific calibration for the IBIF algorithm. The session involved simultaneous and synchronous recordings of CV mask-based oral airflow and neck skin acceleration in an acoustically treated room. Each subject performed a series of sustained vowels gestures (/a/ and /i/) with a constant pitch using comfortable and loud (approximately 6 dB increase) voice. For each gesture, a bandpass filter ($60 - 1100$ Hz) oral airflow vowel segment was used to perform inverse filtering with a single notch filter constrained to unitary gain at DC [125].

Once a glottal airflow approximation is obtained from the CV mask, Q pa-

rameters are estimated using the optimization scheme described in the *Subglottal Impedance Based Inverse Filtering for Ambulatory Monitoring of Voice* section. The whole process, from estimation of parameters to classification and statistical analysis was done with MATLAB (The MathWorks, Inc.).

Table 3.1: Occupations and mean age of adult females with PVH and matched-control participants analyzed (48 pairs)

Occupation	**Pairs**	**Age** [a]	**Diagnosis**	**CAPE-V overall** [b]
Singer	34	21.3 (3.7)	Nodules (31)	21.2 (12.6)
			Polyp (3)	–
Teacher	5	38.9 (12.1)	Nodules	33.8 (18.8)
Consultant	2	23 (1.4)	Nodules (1)	22.0 (5.7)
			Polyp (1)	–
Psychologist	1	34(P) 30 (C)	Nodules	–
Recruiter	2	23.5 (0.8)	Nodules	40.5 (13.4)
Marketer	1	22 (P) 25 (C)	Nodules	25
Media relations	1	32 (P) 31 (C)	Nodules	30
Registered nurse	1	57 (P) 58 (C)	Polyp	40

[a]Mean age and (standard deviation) are shown for pairs ≥ 2. Otherwise, the age is shown for the phonotraumatic (P) and control (C) subject.

[b]Mean overall severity score (0-100) and (standard deviation) are shown patients from pairs ≥ 2. Otherwise, the patient's score is shown.

3.3 Ambulatory glottal airflow assessment

Estimates of individual Q parameters, which were assumed to be time-invariant for each subject, were applied in Eq (3.1). The assumption of time-invariance is due to the properties of the neck skin, which should not change over time. Preliminary studies of the use of IBIF calibrated for a sustained vowel [126] and on the variability of these calibrated parameters [127], have shown that using a sustained vowel works well on running speech (e.g., the rainbow passage). Current research aims to explain in more detail the estimation and variability of these parameters under different speech conditions. $\dot{u}_{skin}(k)$ the discrete time-domain equivalent of the acceleration signal $\dot{U}_{skin}(\omega)$, is convolved with $t_{skin}(k)$ the inverse transfer function of the skin in time domain, where its frequency domain expression is represented by Eq 3.2.

By taking the inverse fast Fourier transform (IFFT) with 1102 coefficients, we obtain $t_{skin}(k)$, a FIR filter. We take every consecutive hour of the acceleration signal $\dot{u}_{skin}(k)$ and convolved it with $t_{skin}(k)$ to obtain the estimated glottal flow signal $\hat{u}_g(k)$. This signal was segmented into 50 ms non-overlapping windows. Voiced frames from the ACC signal were identified based on the same voice activity detection algorithm used in [2], where a combination of periodicity and spectral metrics determines whether a frame is voiced or unvoiced. In addition, we discarded frames in which the absolute ratio of the RMS values of the first

half divided by the second half of the frame was greater than a threshold (1.5 for these experiments); thus, frames exhibiting onsets or offsets were removed since they typically result in incorrect inverse filtering estimates due to cycle-by-cycle variations in the signal. As with many inverse filtering methods [128], IBIF has difficulty analyzing signal with high f_0 values due to the short closed phase during which vocal tract information must be estimated (females and singers, especially, produce high-pitched phonation). Performance of traditional glottal inverse methods could be accurate up to a f_0 of 400 Hz [129]. By visual inspection, the estimation of IBIF voiced frames deteriorated around a f_0 of 500 Hz. Thus, voiced frames with f_0 higher than 500 Hz were not processed by IBIF. Future research will analyze sensitivity tests to find the range of frequencies for which the IBIF method fails.

Table 3.2 lists the 11 glottal airflow measures computed within each analyzed frame. Fig 3.3 shows an example of the estimated glottal airflow signal and its derivative for a single frame. Since the accelerometer is an AC signal, the glottal airflow does not have a DC component. As in previous studies [12, 25, 27], ACFL was obtained as the difference between the maximum and minimum amplitude (peak-to-peak) within each glottal cycle. MFDR was the minimum value of the derivative of one glottal cycle. For open and speed quotient, the closed phase in ambulatory settings often exhibits more fluctuations than in laboratory conditions

using sustained vowels. For robust estimations of open and speed quotient, two lines are fit from the glottal cycle peak to median values left and right. The lines are extended until 80% of ACFL is passed. The points of the slopes in the x-axis are the beginning and end of the open phase (see Fig 3.3 (A)). Then open quotient is defined as the open phase divided by the period T0 = $\frac{1}{f_0}$ (OQ = $\frac{\text{t1+t2}}{\text{T0}}$), speed quotient as SQ = $\frac{\text{t1}}{\text{t2}}$, and the normalized amplitude quotient (NAQ) as NAQ = $\frac{\text{ACFL}}{\text{MFDR}\cdot\text{T0}}$.

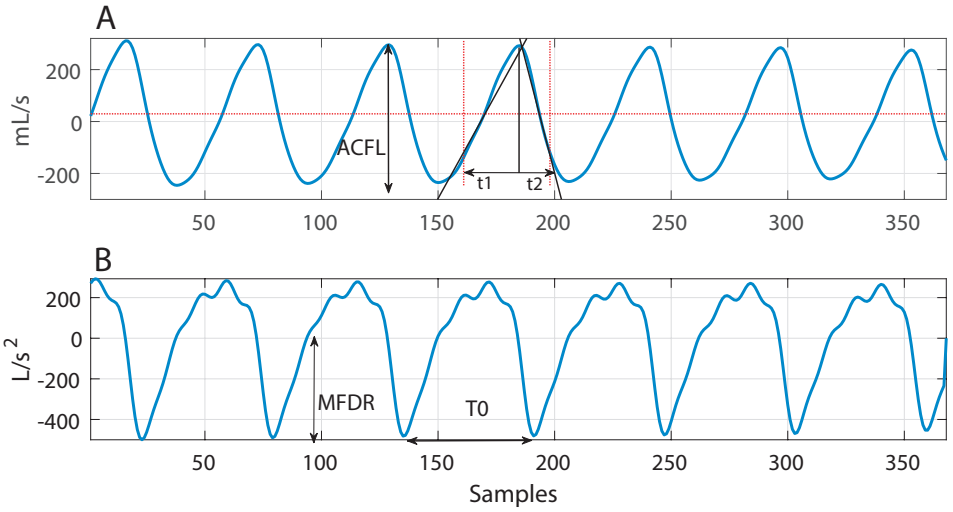

Figure 3.3: Example of ambulatory IBIF analysis. (A) Estimated glottal airflow waveform and (B) its derivative, showing how time-domain measures were derived per glottal cycle. Measures were then averaged over all cycles to yield a single value per frame for each time-domain measure.

We also included 4 additional measures derived from the time-domain measures:

Table 3.2: Frame-based glottal airflow measures estimated from the ambulatory neck-surface accelerometer signal using impedance-based inverse filtering.

Glottal airflow measures	**Description**	**Units**
ACFL	Peak-to-peak glottal airflow.	mL/s
MFDR	Negative peak of the first derivative of the glottal waveform.	L/s^2
Open Quotient (OQ)	Percentage of the open time of the glottal vibratory cycle to the corresponding cycle period.	%
Speed Quotient (SQ)	Percentage of the opening time of the glottis to the closing time.	%
H1-H2	Difference between the magnitude of the first two harmonics.	dB
Harmonic Richness Factor (HRF)	Ratio of the sum of the amplitudes of the first 8 harmonics to the amplitude of the first harmonic.	dB
Normalized Amplitude Quotient (NAQ)	Ratio of ACFL to MFDR divided by the glottal period.	–
logMFDR	$10log_{10}\|\text{MFDR}\|^2$.	dB
logACFL	$10log_{10}\|\text{ACFL}\|^2$	dB
MFDR'	Ratio of estimated SPL (dB SPL) to logMFDR.	–
ACFL'	Ratio of estimated SPL (dB SPL) to logACFL.	–

- Logarithmic versions of ACFL and MFDR squared: $10\log_{10}|\text{ACFL}|^2$ (dB) and $10\log_{10}|\text{MFDR}|^2$ (dB).

- SPL normalized by ACFL (dB) and MFDR (dB): $\text{SPL}/(10\log_{10}|\text{ACFL}|^2)$ and $\text{SPL}(/10\log_{10}|\text{MFDR}|^2)$. Estimates for SPL are calculated using a linear regression equation: $y = mx + b$, where m and b are the coefficients from the subject obtained from accelerometer amplitude (x) and corresponding acoustic SPL (y). The calibration is done daily in the morning with a handheld microphone yielding the reference SPL [23, 130]. These ratios have shown to be significantly different between PVH and control subjects [27].

Given that many of the glottal airflow features applied for vocal hyperfunction analysis are cycle-based [12, 27, 25] and multiple glottal cycles occur within each 50 ms frame, we computed average features across all glottal cycles in each frame. The idea was to provide a more consistent estimate of each measure, especially given the inherent fluctuations from continuous speech in the ambulatory signal. Fig 3.4 shows the spectrum of the estimated glottal airflow, from which spectral measures H1-H2 and harmonic richness factor (HRF) were computed. These measures have been correlated with voice quality [131, 55].

Histograms of features through weekly data provide insights on how the data is distributed and how separable are controls with respect to PVH subjects. Overall, there is a large overlap between groups for most of the features, but there are more

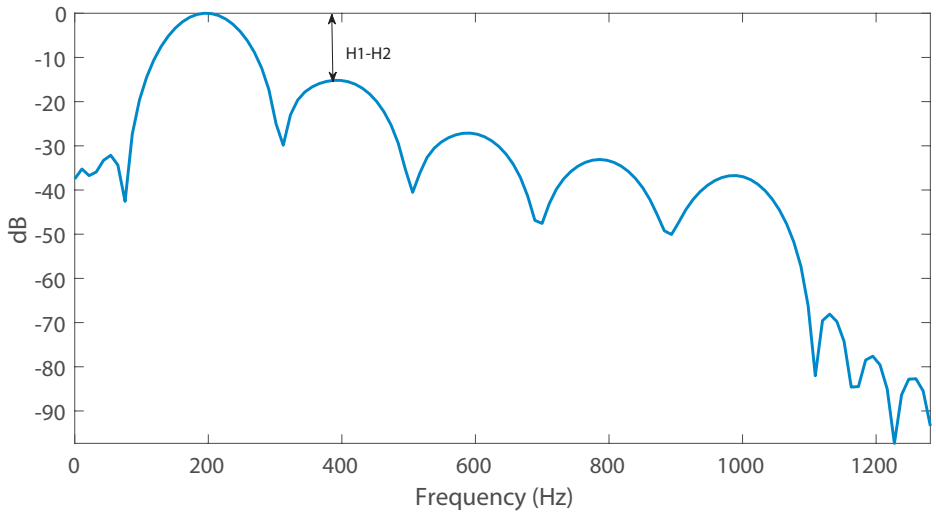

Figure 3.4: Spectrum of the frame in Fig 3.3 (A)

.

differences between individual subject pairs. There is also an issue of variability among subject pairs that contributes to the overlap of all group distributions. For example, Fig 3.5 shows an histogram of MFDR' with all subjects (controls vs. PVH) and one particular example of the distribution of the pair F031. Even though histograms for the general population, by visual inspection, overlap largely, the histograms of the F031 pair show separation, where the distribution of the PVH subject lags behind the distribution of the control. This indicate that the PVH subject has a higher MFDR per SPL ratio overall, supporting the evidence encountered in [12, 27]. However, this is not always the case for all subjects. For example, Fig 3.6 shows another pair where the distribution is the other way around: The PVH subject has a lower MFDR per SPL ratio than its control.

Therefore, some features overlap for the general population.

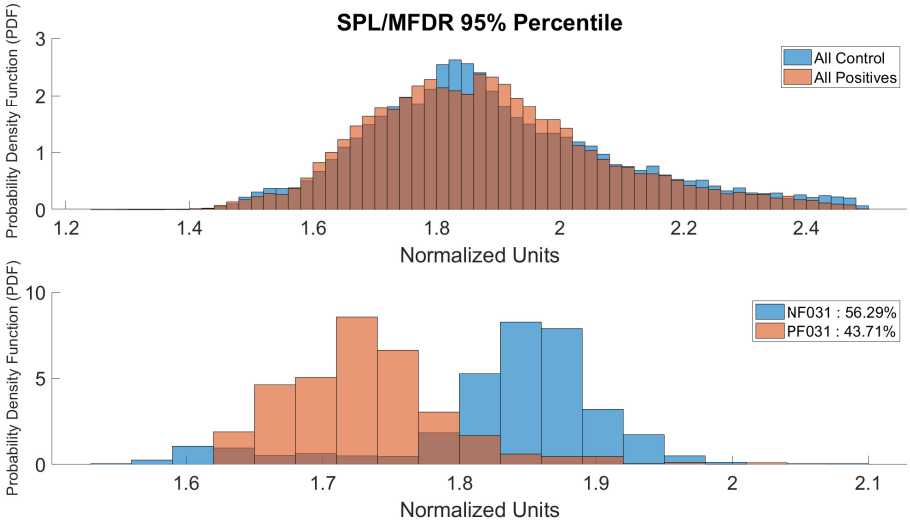

Figure 3.5: Histogram of MFDR' (SPL/MFDR) for 48 subject pairs (top) and single pair F031 (bottom)

.

There are other features where the differences in distribution are more visible, such as the standard deviation of H1-H2 (Fig 3.7). In this case, PVH subjects have a tendency to have lower deviation of H1-H2 values compared to controls.

3.4 Week-long univariate statistics for paired hypothesis testing

The purpose of the following series of tests is to find the most differentiating statistics between the PVH group and controls. Within-subject univariate statistics were calculated for each week-long time series data from each subject: mean,

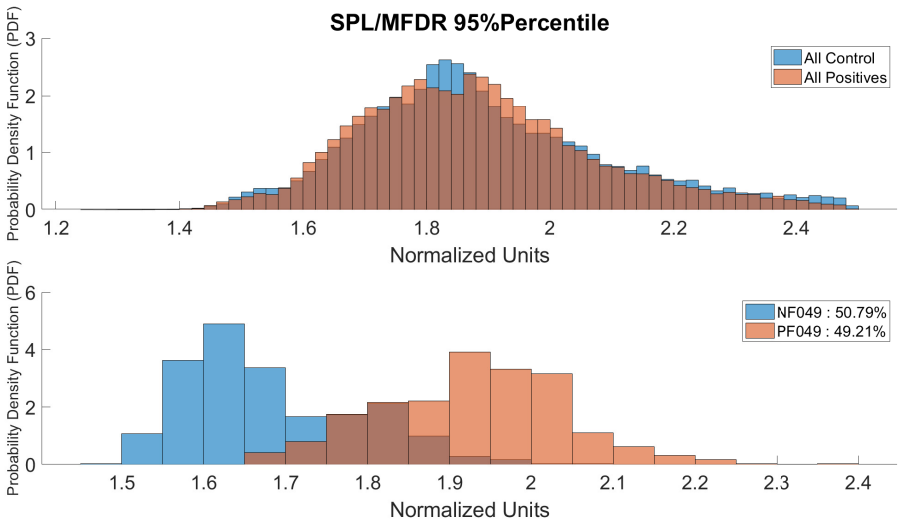

Figure 3.6: Histogram of MFDR' (SPL/MFDR) for 48 subject pairs (top) and single pair F049 (bottom)

.

median, 5th percentile (trimmed minimum), 95th percentile (trimmed maximum), standard deviation, skewness, and kurtosis. These statistics were used for paired t-tests with 48 data points (number of subject pairs). Normality was tested with a Chi-square goodness-of-fit test, and each statistic was not significantly different from a normal distribution with $p < 0.05$. The false discovery rate is described by Eq 3.11, where V is the percentage of false positives (type I error) and S is the percentage of true positives. Since the false discovery rate is an expectation, we have m possible outcomes from the hypothesis tests.

$$\text{False discovery rate} = \left(\frac{V}{V+S}\right) \tag{3.11}$$

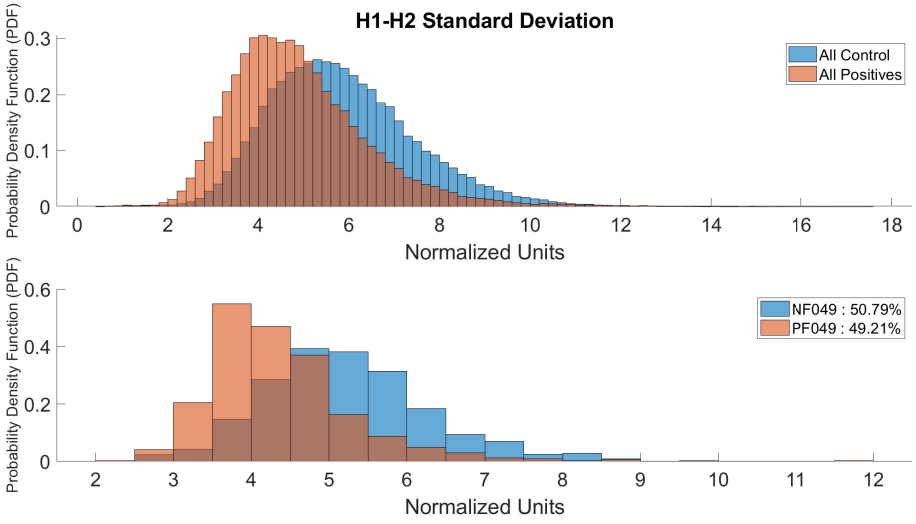

Figure 3.7: Histogram of standard deviation of H1-H2 for 48 subject pairs (top) and single pair F049 (bottom)

.

If we have $H_1, H_2, \ldots H_m$ independent hypotheses, Benjamini-Hochberg (BH) [132] showed that regardless of how many null hypotheses are true and regardless of the distribution of the p-values, when the null hypothesis is false, we have the following property [109]:

$$\text{False discovery rate} \leq \frac{U+V}{m}\alpha \leq \alpha \tag{3.12}$$

where U is the proportion of true negatives. By setting $\alpha = 0.1$, the procedure sorts the m p-values and defines a threshold L:

$$L = \max\left\{k : P_k \leq \frac{k}{m}\alpha\right\} \tag{3.13}$$

We reject all hypotheses H_k for which $p_k \leq p_{(L)}$, the BH rejection threshold. This procedure will find those statistics with at most an α false discovery rate between PVH subjects and controls. It is important to remember that the false discovery rate is not the same as the type I error, but is the expected proportion of false positive features among the list of features that are significant according to the test. An example in reference [109] (page 687) uses a false discovery rate of 0.15, which is typical for analyses that are exploratory in nature [133]. In this case, we find the most differentiating statistics using this test, in contrast with a Bonferroni-corrected t-test, which yields a conservative comparison for which there is no statistically significant difference between any statistic.

3.5 Classification methods

Following the same procedure as Ghassemi et al. [11], each subject's week-long ambulatory recording was subdivided into 5-minute windows (6000 frames, non-overlapping). Only windows exhibiting voicing were only included in the classification task; voiced windows were defined as containing at least 0.5% voicing (30 voiced frames). We then calculated the following univariate statistics over the voiced frames within each window for each measure in Table 3.2: mean, median, 5th percentile, 95th percentile, standard deviation, skewness, and kurtosis. Windows with less than 0.5% voicing were discarded due to data sparsity. Each

window-based statistic was z-normalized (subtracting by the mean and dividing the result by the standard deviation) in two ways: a) by week, across voiced windows from all subjects (PVH and controls) and b) by day, across voiced windows within their respective days.

The full feature vector is composed of 154 features: 77 weekly and 77 daily z-score normalization the features derived from the 7 window-based univariate statistics for each of the 11 frame-based glottal airflow measures in Table 3.2. Since we only have a small amount of training data, we reduce feature dimensionality before training. As a first pass, forward feature selection (FFS) [134] is applied to the full feature matrix. The procedure is a greedy search algorithm that starts with an empty set I and iteratively selects a new feature x from the set of features not in I that minimizes a cost function J (a quadratic discriminant analysis classifier). The feature x is added to I, and the procedure is repeated until a threshold (10^{-6} in this case) of consecutive results is achieved. E is the quadratic discriminant analysis classification error using 5-fold cross-validation. The final reduced feature vector is composed of 55 features. It is worth mentioning that this subset is suboptimal since further reduction can be achieved through LASSO selection, which is applied later on. We use these features to build both logistic regression and support vector machine (SVM) supervised classifiers.

Logistic regression is a type of discriminative classifier that models the class-

conditional probability as:

$$P(y = 1|x) = \frac{1}{1 + e^{-x^T\beta}} \tag{3.14}$$

where $x \in R^n$ is the feature vector, $y = 1 \in R^l$ is the class labeled as $y_i = 1$ (PVH) or $y_i = 0$ (control), and β is the vector of coefficient weights. In order to find the coefficients β, we maximize the following penalized log-likelihood using N data points of the training set with p feature vectors:

$$\max_{\beta \in R^{p+1}} \frac{1}{N} \sum_{i=1}^{N} \{y_i \log p(x_i) + (1 - y_i) \log(1 - p(x_i))\} + \lambda \|\beta\|_1, \tag{3.15}$$

where x_i is the data point for instance i and λ is the regularization parameter for the LASSO constraint. The L_1 penalty reduces the number of features used in the model. The regularization parameter λ is selected from 50 values ranging from 1 to 10^{-5}, through a 5-fold cross-validation procedure with the training data, as shown in Fig 3.8. A lower λ will tend to lower the deviance (error) of the model. However, a small value of λ will tend to overfit the model to the training data, which is not the purpose of regularization. Therefore, instead of choosing the λ that outputs the minimum deviance, we select λ to be 1 standard deviation error (blue line in Fig 3.8) from the value with minimum deviance error (green line in Fig 3.8). The optimization method to maximize Eq 3.15 is a gradient coordinate

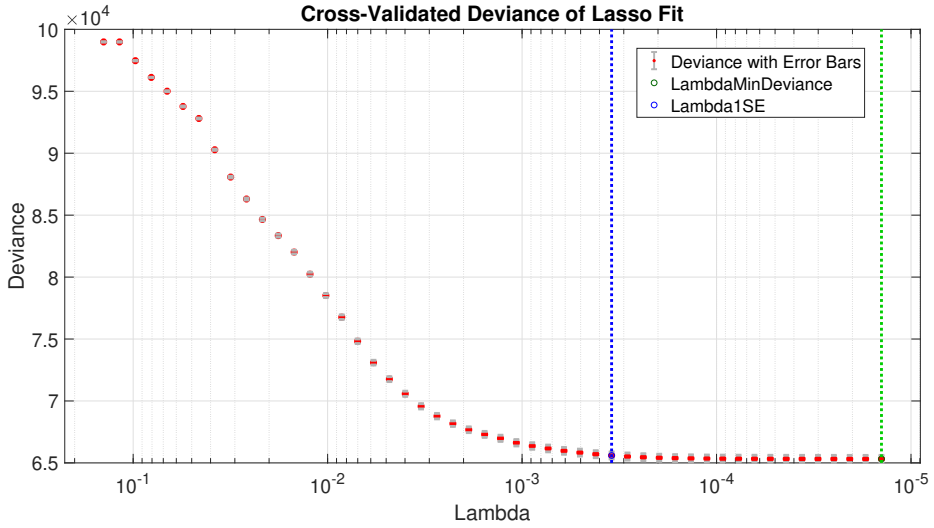

Figure 3.8: Deviance of the model for a given training data set with 50 different λ values using 5-fold cross-validation.

descent [135]. The procedure is repeated for each of the 48 trained models.

SVMs are commonly used machine learning tools for classification [136]. The weight vectors $\boldsymbol{w} \in R^n$ are optimized to create a linear L_1 SVM classifier:

$$\min_{\boldsymbol{w}} \quad \sum_{i=1}^{N} (\max(0, 1 - y_i \boldsymbol{w}^T \boldsymbol{x}_i))^2 + C\|\boldsymbol{w}\|_1 \tag{3.16}$$

where C is a regularization parameter similar to λ for logistic regression. The goal is to create a sparse $\boldsymbol{w}$ that solves the L_2-loss support vector classifier [137].

Fig 3.9 shows a flowchart of the feature extraction and classification process. We first divided data using leave-one-out cross-validation to generate 48 datasets, each consisting of 47 training pairs and one test pair. All windows from the 47 training pairs (94 subjects total) were then subdivided using 5 cross-validation

(1/5th validation and 4/5ths training in each fold). The validation sets are used to find the best set of parameters with respect to the area under the ROC curve (AUC) and these are selected for the model to be used in the test set. The following metrics are used to check the performance of the logistic model on the test pair: AUC, F-score, accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV). From this procedure, we test two scenarios: Classification with all the features after selection and using subsets of those features. The latter is done by sorting the absolute Beta values and running L_1 logistic regression again by starting with all selected features. Then we took out the feature with lowest Beta value in magnitude and ran the classification again, and so on. The positive Beta weights are associated with subjects with PVH, whereas the negative weights are associated with control subjects.

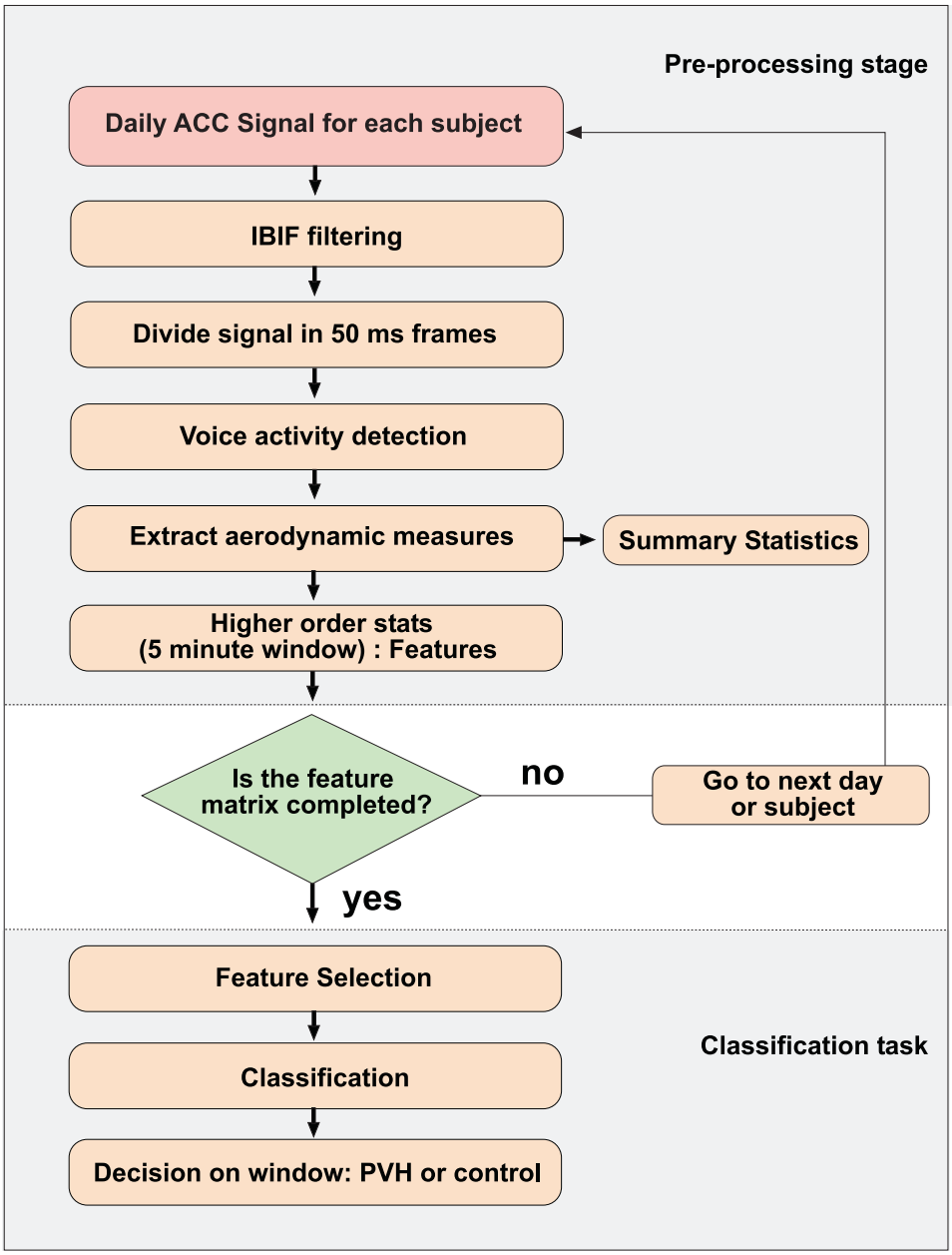

Figure 3.9: Flowchart: Feature extraction and classification process for 96 subjects

3.6 Results

3.6.1 Week-long univariate statistics for paired hypothesis testing

Table 3.3 shows the first 11 features sorted from lowest to highest p-value from the paired t-tests $H_1 \dots H_{11}$. The BH test rejects the first 8 null hypotheses $H_1 \dots H_8$, i.e., they are significantly different at the 95% confidence level. Minimum and median ACFL were the most discriminative statistics, with medium effect sizes (Cohen's d [138]) of 0.59 and 0.55, respectively. In general, statistics of the ACFL measure had the best differentiating power among all the week-long paired t-tests. In contrast, average values for estimated SPL for subjects from the same database were not significantly different between subjects with PVH and control subjects [75, 2]. This result suggests that high ACFL values are potentially good indicators of subjects with PVH, if the SPL distributions of both groups are statistically similar.

3.6.2 Supervised classification task

Most classifiers output a decision vector with probabilities that a data window is positive (labeled 1) or negative (labeled 0). Decision labels are created with a threshold number between 0 and 1, where the data window is labeled as positive if its probability is above the threshold, or labeled negative otherwise. The threshold is selected as the one that provides maximum accuracy in the test set. An example

Table 3.3: Top 11 week-long summary statistics (from a total of 77) sorted by p-value from the 48 paired t-tests. Statistically significant differences (*) were found by applying the Benjamini-Hochberg method using a false discovery rate of 0.1.

Voice Use Summary Statistic	Patient Group	Matched-Control Group	p-value	Effect Size
logACFL minimum	38.5 ± 3.3	36.6 ± 2.9	0.0011^*	0.59
logACFL median	49.2 ± 3.5	47.3 ± 3.7	0.0015^*	0.55
ACFL minimum	90.8 ± 40.6	71.6 ± 23.5	0.0016^*	0.58
logACFL mean	48.7 ± 3.5	46.9 ± 3.5	0.0025^*	0.52
ACFL median	315 ± 140	251 ± 99.0	0.0030^*	0.53
H1-H2 kurtosis	10.9 ± 4.30	8.8 ± 2.6	0.0061^*	0.59
logACFL kurtosis	3.17 ± 0.60	2.93 ± 0.4	0.0076^*	0.50
ACFL mean	359 ± 163	296 ± 117	0.0091^*	0.45
H1-H2 minimum	2.39 ± 4.20	0.38 ± 4.4	0.0120	0.48
HRF kurtosis	11.6 ± 4.7	10.0 ± 2.9	0.0230	0.42
MFDR median	365.4 ± 171.8	310.6 ± 127.7	0.0270	0.37

of the selection of a threshold for a test set is shown Fig 3.10, where the top panel displays the accuracy score with respect to threshold, and the bottom panel is the receiving operating characteristic (ROC) curve. The threshold selected is shown in red.

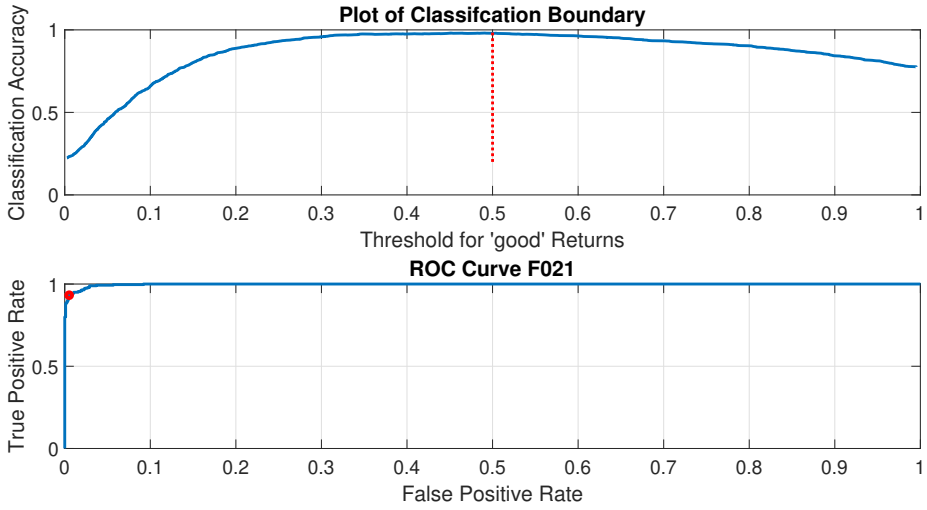

Figure 3.10: Top: Accuracy of model vs. threshold. Bottom: ROC curve of true positives vs. false positives

Table 3.4 shows a summary of the classification results for both implemented classifiers using the multiple performance metrics. Fig 3.11 displays performance of the L_1 logistic regression classifier for each of the 48 pairs for a subset of the performance metrics. There is a large spread of AUC scores across the subjects with an average of 0.82. AUC scores less than 0.5 indicate that the model places weight on positive examples versus negative ones and vice versa. The large AUC variance, including values less than 0.5, could be explained from the labels; e.g.,

Table 3.4: Classification performance of L1 logistic regression (L1-LR) and support vector machine (SVM) approaches for 96 subjects using IBIF features. Mean (standard deviation) is reported for the performance metrics. Previous results using 51 pairs [2] and 20 pairs [11] are also shown. It is worth noting that the distribution of metrics such as AUC, across all models, may be non-normal and may benefit from other summary statistics such as median (IQR).

Method	AUC	Accuracy	F-score	Sensitivity	Specificity	PPV	NPV	Threshold
L1-LR (IBIF)	0.82 (0.25)	0.83 (0.14)	0.77 (0.27)	0.78 (0.29)	0.85 (0.22)	0.81 (0.21)	0.82 (0.18)	0.54 (0.25)
SVM (IBIF)	0.82 (0.26)	0.84 (0.14)	0.78 (0.27)	0.79 (0.28)	0.84 (0.24)	0.83 (0.22)	0.82 (0.21)	0.02 (0.67)
Mehta et al.[2]	0.74 (0.27)	-	0.77 (0.20)	0.74 (0.30)	0.77 (0.29)	-	-	-
Ghassemi et al.[11]	0.71 (-)	0.66 (-)	0.63 (-)	0.50 (-)	0.81 (-)	0.72 (-)	0.62 (-)	-

subjects with PVH do not always exhibit vocal behavior typical for the pathology, whereas control subjects might exhibit some vocal behavior that differs substantially from healthy vocal behavior.

Logistic regression and SVM have similar good results on all performance metrics. Since L1 regularization was used in both cases, it could be that the removal of redundant features in every training case helped the performance. The mean (standard deviation) of the performance metrics for both classifiers improved when compared with previous results on 51 matched-paired subjects: 0.74 (0.27) for AUC, 0.77 (0.20) for F-score, 0.74 (0.30) for sensitivity, and 0.77 (0.29) for

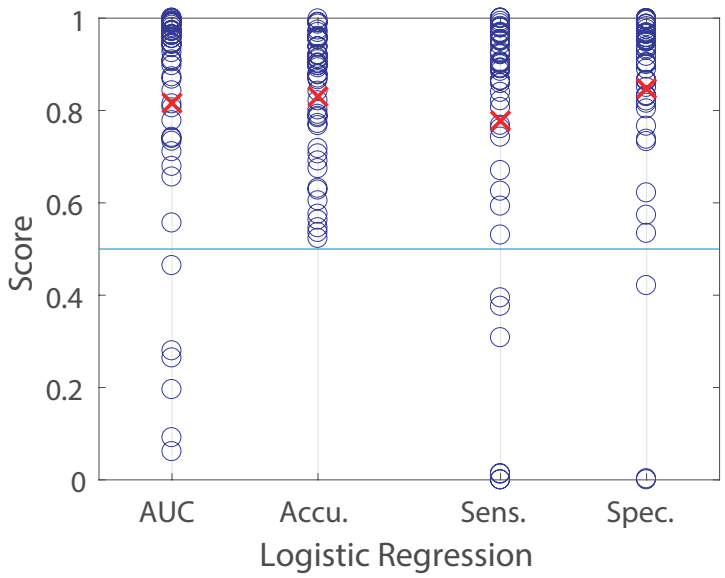

Figure 3.11: Performance results across subject pairs with L1-logistic regression: Area Under the ROC Curve (AUC), Accuracy, Sensitivity, Specificity. The red crosses indicates the average value for each performance metric

specificity [2]. Fig 3.12 shows the proportion of labels classified as positive (VH) for all subjects. 79 subjects from 96 were classified correctly by using a threshold of 0.57. This corresponds to 82.3% of accuracy.

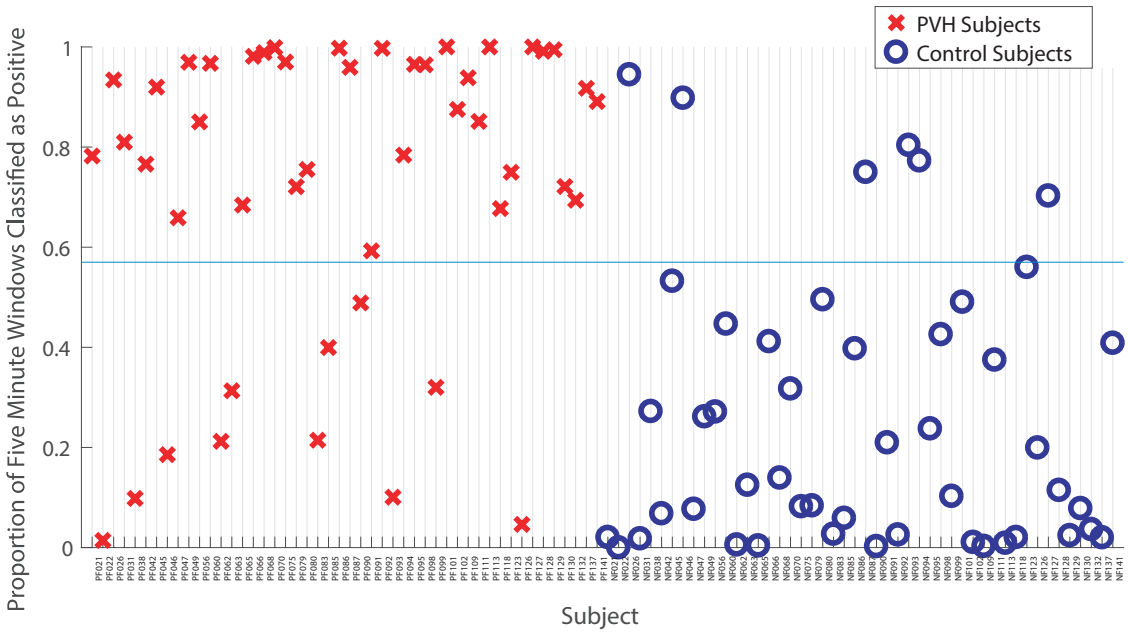

Figure 3.12: Classification results from L1-logistic regression. The threshold (blue line) at 0.57 classifies correctly 79 from 96 subjects (82.3%)

Feature selection is important for identifying the most relevant features that can help to further understand the underlying process, as well as reducing the complexity for future biofeedback applications. Table 3.5 shows the total number of features (26) that were present in all 48 models after using LASSO with the resulting 55 features after FFS. Table 3.6 shows the results for all 26 models and

the subset of features by sorting beta values. The mean F-score is stable in the 0.7 region until the number of features is 9. After that, the performance degrades moderately, where the AUC is 0.68 and the accuracy is 0.71 with only 7 features. Fig 3.13 shows boxplots of the same models versus F-score, where we can see the same trend: classification performance is more or less similar if we left in 9 features or more in the classifier.

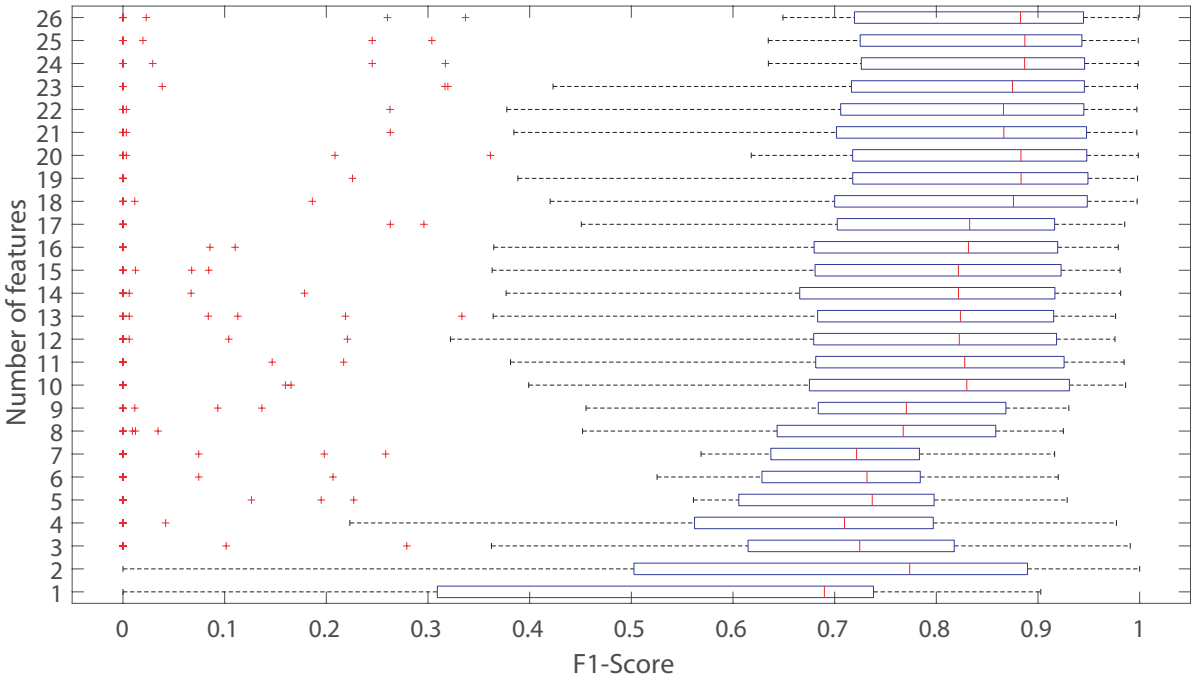

Figure 3.13: F-score distributions from Table 3.5. From all 26 features (rightmost box plot) to only one feature (H1-H2 95th%, leftmost box plot)

.

Fig 3.14 shows the association counts of features with PVH subjects as odds ra-

Table 3.5: Association count of Beta (weight) variables that were included in all 48 models. These 26 features were present in each logistic regression model.

Associated Feature	Phonotraumatic	Control	Beta Weight Mean	Standard Deviation
H1-H2 95th% (Daily Normalized)	48	0	2.50	0.16
NAQ mean	48	0	1.42	0.11
HRF skewness	48	0	1.38	0.09
logACFL standard deviation	48	0	1.30	0.10
HRF 5th% (daily normalized)	48	0	1.21	0.15
logACFL skewness (daily normalized)	48	0	1.17	0.05
SQ 5th%	48	0	1.16	0.05
SQ standard deviation	48	0	1.12	0.06
MFDR' 95th%	48	0	1.01	0.13
OQ 5th%	48	0	0.94	0.12
H1-H2 standard deviation (daily normalized)	48	0	0.71	0.09
HRF standard deviation (daily normalized)	48	0	0.43	0.06
logMFDR 5th% (daily normalized)	48	0	0.32	0.06
ACFL' standard deviation (daily normalized)	48	0	0.18	0.02
SQ skewness (daily normalized)	0	48	-0.12	0.03
SQ standard deviation (daily normalized)	0	48	-0.21	0.02
OQ 5th% (daily normalized)	0	48	-0.27	0.04
SQ 5th% (daily normalized)	0	48	-0.41	0.02
NAQ mean (daily normalized)	0	48	-0.47	0.05
HRF skewness (daily normalized)	0	48	-0.89	0.06
logACFL standard deviation (daily normalized)	0	48	-0.97	0.07
OQ mean	0	48	-1.00	0.13
H1-H2 standard deviation	0	48	-1.28	0.13
logACFL skewness	0	48	-1.58	0.07
HRF 5th%	0	48	-1.86	0.29
H1-H2 95th%	0	48	-4.47	0.31

Table 3.6: Mean and (standard deviation) performance metrics from L1-logistic regression for different group of features from Table 3.5, starting with the whole set of 26 features. Iteratively, the following group is obtained by taking out the feature with the smallest absolute Beta value.

Added feature	Number	AUC	F-score	Accuracy	Sensitivity	Specificity	PPV	NPV
Daily norm. Log ACFL Skew	26	0.82 (0.25)	0.76 (0.30)	0.84 (0.14)	0.77 (0.32)	0.87 (0.19)	0.79 (0.27)	0.83 (0.19)
Daily norm. ACFL' stand. dev.	25	0.82 (0.25)	0.76 (0.30)	0.83 (0.14)	0.77 (0.32)	0.87 (0.19)	0.79 (0.27)	0.83 (0.19)
Daily norm. SQ stand. dev.	24	0.82 (0.25)	0.76 (0.30)	0.84 (0.14)	0.77 (0.32)	0.87 (0.19)	0.79 (0.27)	0.83 (0.19)
Daily norm. OQ 5th%	23	0.82 (0.25)	0.77 (0.28)	0.83 (0.14)	0.77 (0.30)	0.86 (0.19)	0.80 (0.24)	0.83 (0.19)
Daily norm Log MFDR 5th%	22	0.82 (0.25)	0.77 (0.28)	0.83 (0.15)	0.78 (0.30)	0.85 (0.23)	0.81 (0.25)	0.82 (0.22)
Daily norm. SQ 5th%	21	0.82 (0.25)	0.77 (0.28)	0.83 (0.15)	0.78 (0.30)	0.85 (0.22)	0.81 (0.25)	0.81 (0.22)
Daily norm. HRF stand. dev.	20	0.82 (0.27)	0.78 (0.28)	0.83 (0.15)	0.79 (0.30)	0.85 (0.23)	0.81 (0.25)	0.82 (0.22)
Daily norm. NAQ mean	19	0.82 (0.27)	0.78 (0.28)	0.84 (0.15)	0.79 (0.30)	0.85 (0.22)	0.79 (0.27)	0.82 (0.22)
Daily norm. H1-H2 stand. dev.	18	0.82 (0.26)	0.77 (0.28)	0.83 (0.15)	0.78 (0.30)	0.85 (0.22)	0.80 (0.25)	0.81 (0.22)
Daily norm. HRF skew	17	0.79 (0.24)	0.74 (0.28)	0.80 (0.14)	0.75 (0.30)	0.81 (0.23)	0.76 (0.26)	0.78 (0.21)
OQ 5th%	16	0.77 (0.24)	0.71 (0.30)	0.79 (0.15)	0.73 (0.33)	0.80 (0.26)	0.74 (0.26)	0.77 (0.21)
Daily norm. Log ACFL stand. dev.	15	0.77 (0.24)	0.71 (0.30)	0.79 (0.14)	0.73 (0.32)	0.80 (0.25)	0.77 (0.24)	0.77 (0.21)
OQ mean	14	0.78 (0.24)	0.72 (0.29)	0.79 (0.14)	0.73 (0.32)	0.80 (0.26)	0.77 (0.24)	0.78 (0.21)
MFDR' 95th%	13	0.78 (0.24)	0.71 (0.29)	0.79 (0.14)	0.72 (0.32)	0.81 (0.25)	0.78 (0.21)	0.77 (0.21)
SQ stand. dev.	12	0.78 (0.24)	0.71 (0.30)	0.79 (0.14)	0.72 (0.33)	0.82 (0.25)	0.77 (0.24)	0.77 (0.21)
SQ 5th%	11	0.78 (0.24)	0.73 (0.27)	0.79 (0.14)	0.75 (0.30)	0.79 (0.26)	0.76 (0.24)	0.78 (0.21)
Daily norm. Log ACFL skew	10	0.78 (0.25)	0.73 (0.28)	0.79 (0.15)	0.75 (0.31)	0.79 (0.27)	0.76 (0.24)	0.78 (0.21)
Daily norm. HRF 5th%	9	0.75 (0.23)	0.71 (0.25)	0.75 (0.13)	0.74 (0.28)	0.72 (0.30)	0.73 (0.20)	0.75 (0.19)
H1-H2 stand. dev.	8	0.74 (0.22)	0.69 (0.26)	0.75 (0.13)	0.72 (0.29)	0.72 (0.29)	0.73 (0.20)	0.73 (0.19)
Log ACFL stand. dev.	7	0.68 (0.22)	0.62 (0.29)	0.71 (0.12)	0.66 (0.33)	0.69 (0.28)	0.63 (0.26)	0.66 (0.23)
HRF skew	6	0.69 (0.22)	0.63 (0.29)	0.71 (0.12)	0.67 (0.32)	0.68 (0.29)	0.63 (0.26)	0.67 (0.23)
NAQ mean	5	0.68 (0.23)	0.63 (0.29)	0.71 (0.12)	0.68 (0.34)	0.68 (0.29)	0.63 (0.26)	0.68 (0.24)
Log ACFL skew	4	0.66 (0.24)	0.63 (0.27)	0.70 (0.13)	0.67 (0.32)	0.67 (0.33)	0.64 (0.25)	0.66 (0.23)
HRF 5th%	3	0.63 (0.30)	0.64 (0.29)	0.71 (0.15)	0.69 (0.34)	0.67 (0.36)	0.66 (0.37)	0.74 (0.22)
Daily norm. H1-H2 95th%	2	0.63 (0.33)	0.64 (0.33)	0.74 (0.16)	0.67 (0.38)	0.76 (0.34)	0.69 (0.32)	0.70 (0.29)
H1-H2 95th%	1	0.58 (0.22)	0.53 (0.30)	0.65 (0.10)	0.58 (0.37)	0.64 (0.35)	0.58 (0.26)	0.64 (0.16)

tios. Odds ratios represent the association with a one-unit increase in the features. These features represent a combination of time and frequency-domain features that were consistently present in all 48 logistic regression models with $p < 0.05$ [2]. The 95th percentile of H1-H2 (daily normalized) had a large association with PVH labels, which is a voice measure correlated with voice quality [131]. However, the large confidence interval for this feature represents low level of precision of the odds ratio. The 95th percentile ratio of SPL and MFDR (MFDR' 95%ile in Fig 3.14) has a moderate association compared to the rest of the features with a small confidence interval, representing a higher precision on the odds ratio.

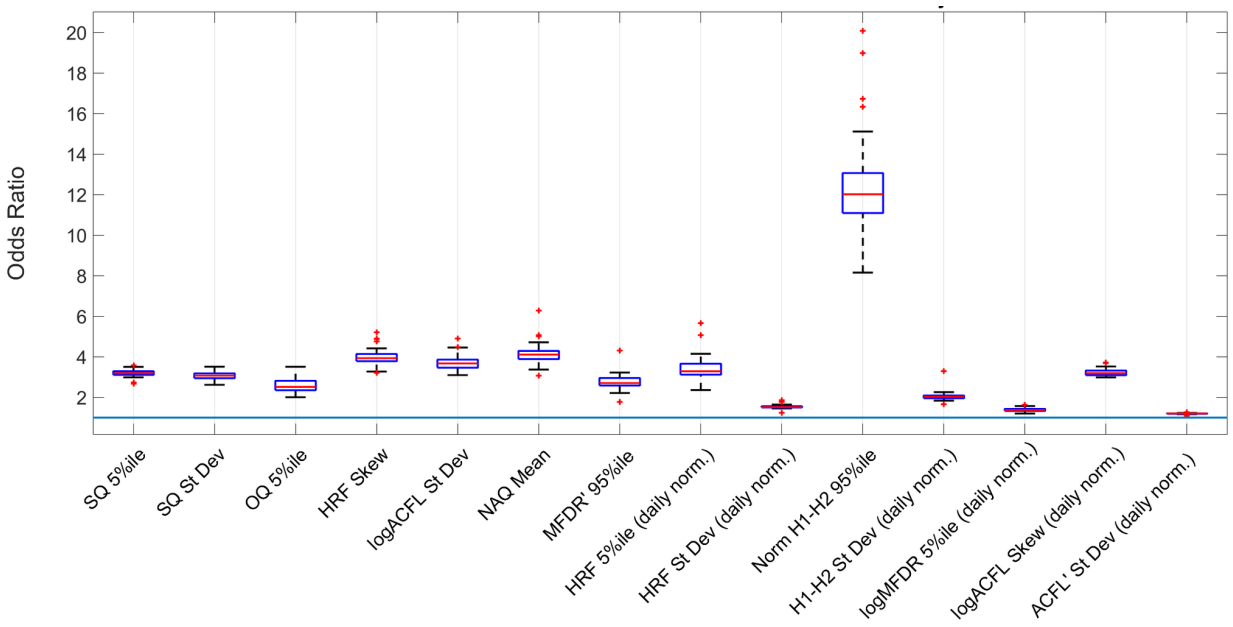

Figure 3.14: Odds Ratio association with phonotraumatic subjects.

3.7 Discussion

The efforts described in this chapter sought to determine whether optimized IBIF-based estimates of glottal airflow measures extracted from ambulatory voice (accelerometer-based) recordings can be used to differentiate between normal vocal function and pathophysiological mechanisms associated with PVH. Results showed that this approach can be quite successful in classifying subjects as being normal or having PVH. Within-subject univariate analyses identified eight aerodynamic features that were statistically different between the patient and matched control groups. ACFL was the most significant measure with medium effect sizes exhibited. These findings are in agreement with previous laboratory studies that used measures extracted from the inverse filtered oral airflow [12, 27] and with computer modeling that suggests that increases in ACFL may reflect the type of increased compensatory effort (i.e., increased vocal hyperfunction) that is necessary for PVH patients to maintain adequate phonation in the presence of vocal fold fold trauma/lesions [139, 60]. Such increases in vocal effort are believed to reflect the "vicious cycle" of progressive concomitant increases in PVH and vocal fold trauma that contribute to perpetuating these disorders.

From Table 3.4, use of the IBIF-based glottal airflow measures in the supervised classification task produced results that outperformed previous reports that used acoustic-based features extracted from ambulatory recordings of the accelera-

tion signal to differentiate between subjects with PVH and normal controls [11, 2]. The improvement in performance using IBIF-based features, in combination with the capability of such features to provide better insights into pathophysiological mechanisms, supports the potential that this approach has to improve the clinical assessment of hyperfunctional voice disorders. Future research could explore the performance of IBIF-based features with other pathologies, such as unilateral vocal fold paralysis [140].

There are several limitations to the current study which may serve to constrain any addition improvement in classification performance. First, even though the use of univariate statistics over 5-minute windows showed good performance, such an approach could smooth out fast variations in some features that may provide important information related to pathophysiology. Moreover, discarding silence periods from the analysis windows might be eliminating information that could further differentiate normal and pathological vocal function by indicating relative differences in non-vocal (non-phonatory) recovery times.

In addition, determination of the IBIF Q parameters is based on accurate estimates of the glottal volume velocity waveform obtained by inverse filtering the oral airflow recorded in the laboratory during sustained vowel production. However, the process of inverse filtering to estimate the glottal flow is still a topic of research and any method will have a degree of error (see [141, 142] for general discussions). The inverse filtering process is particularly challenging when applied

to pathological female voices, as was done in this study. The process was made even more demanding by that fact that many subjects in this study were singers who regularly reached very high pitches (above 400 Hz) daily during practice that tend to cause the inverse filtering and IBIF methods to fail. In addition, every feature has an associated uncertainty from the accelerometer measurements, and the task becomes difficult when we combine multiple estimated features (e.g., for the SPL-normalized measures of ACFL' and MFDR') since those errors may propagate and increase the total uncertainty in an ambulatory setting.

Finally, the task of differentiating normal and pathological subjects was made more difficult because the patients with PVH in this study were classified as having only mild-to-moderate voice disorders. We know from clinical experience that such patients can display periods of seemingly normal vocal function, and, conversely, normal speakers can display transient episodes of VH that do not develop into chronic conditions. Future studies could attempt to address these issues by developing estimates of uncertainty for the extracted IBIF parameters and using other analysis methods such as unsupervised learning to better pinpoint specific segments of abnormal vocal function, as has been initially demonstrated in [108]. In addition, efforts to incorporate aerodynamic features in the framework of ambulatory biofeedback to improve voice therapy are currently underway [143].

3.8 Conclusion

In this chapter, we further develop prior ambulatory efforts, by improving the ability to discriminate pathological voices from healthy ones. Using an impedance-based inverse filtering scheme to estimate the unsteady glottal airflow component from a neck-surface accelerometer and a smartphone platform, we obtain and quantify, for the first time in an ambulatory assessment and a comprehensive framework, aerodynamic features that have been shown to be physiologically relevant for vocal hyperfunction in recent laboratory settings and computational studies. Prior efforts to obtain aerodynamic features from neck surface acceleration were limited to sustained vowels [85] and simple proof of concept examples [2]. Therefore, the work addresses the specific aim 1 (SA1) for this thesis, in which ambulatory features from glottal airflow help to discriminate PVH subjects from controls.

Regarding Hypothesis 1 (H1), higher-order statistics of glottal features are able to discriminate PVH from controls. Specifically, the amplitude-based features (i.e., ACFL and MFDR) with statistics such as daily-normalized skewness and standard deviation (ACFL) and daily-normalized 5th and 95th percentiles (MFDR) contribute to the classification models, confirming H1. However, spectral features from the glottal flow contribute strongly to the classification as well. The standard deviation and 95th percentile of H1-H2, and the skewness and 5th

percentile of HRF are indicators, to some extent, of PVH. It is a surprising result, since H1-H2 would predict that those features would be more associated to NPVH subjects. H1-H2 and HRF are related to the spectral tilt of the glottal flow. Since those measures are related to perceptual voice quality (e.g, breathiness) [144, 48, 145, 146], it can be inferred that PVH subjects with distinct voice quality than controls tend to have salient glottal spectral features. For example, the low standard deviation in H1-H2 for PVH subjects could be related to a tendency in keeping vocal quality around a high mean value, which would correspond to breathy vocal quality. It is known that he resultant dysphonia from vocal fold nodules is perceived as breathy with various degrees of turbulent noise, strained vocal quality, roughness, instability and vocal fry/creak, with a tendency toward a low pitch [147, 148, 149]. Control subjects tend to have a larger variability of H1-H2, for which there is some evidence that the phonatory characteristics of H1-H2 seems to be speaker dependent [55]. Therefore, it seems that healthy subjects have greater control in varying voice quality, which is reflected by the greater variance in H1-H2. However, further studies need to investigate the role of H1-H2 in healthy subjects compared to PVH subjects.

The result of this comprehensive quantitative analysis show that these ambulatory glottal airflow measures can be successfully used to differentiate between normal vocal function and pathophysiological mechanisms associated with phonotraumatic vocal hyperfunction, and outperform state-of-the-art reports us-

ing sound pressure level, fundamental frequency, and related vocal doses. Due to its physiological relevance, the proposed aerodynamic ambulatory approach has already potential to improve the clinical assessment of hyperfunctional voice disorders, including the evaluation of treatment outcomes. One important aspect on the analysis of estimated glottal flow features is to determine if the signal estimated is robust against external factors, such as deviations from the ground truth flow signal and uncertainty on the variation of Q parameters. The former is investigated in chapter 4 by using a different implementation of the IBIF filter, while the latter is tested in section 5.1 with a pilot study using reading passages in lab conditions.

Chapter 4

Calculating deviations from IBIF measures using a state-space model for Bayesian processing

Subglottal Impedance-Based Inverse Filtering (IBIF) allows for the continuous, non-invasive estimation of glottal airflow from a surface accelerometer over the anterior neck skin below the larynx, which has been shown to be advantageous for the ambulatory monitoring of vocal function, as shown in chapter 3. In spite of these advances, there is a need to quantify the magnitude of the deviation of the parameters in the estimation process of the IBIF filter, and to potentially improve the estimation of the aerodynamic signals.

During the long-term ($>$ 1 week) ambulatory recordings, conditions may drift from the laboratory environment where the IBIF parameters are initially estimated. There are unquantified uncertainties in the estimation of the glottal airflow signal with the IBIF scheme due to a number of factors. First, the determination of the IBIF model parameters uses inverse filtering of the oral airflow

from sustained vowels, which can lead to IBIF parameter variations for different vowels and pitch conditions [150]. The latter becomes especially challenging for high-pitch female pathological voices, which is common in ambulatory studies. In addition, there is measurement uncertainty from the accelerometer due to sensor positioning, skin attachment, temperature, etc. Furthermore, there is no direct reference that can be used to quantify these combined effects in ambulatory scenarios. Observation and model uncertainties may result in significant deviations in the glottal airflow estimates, but are very difficult to quantify in ambulatory conditions. An example of the drifting conditions of the IBIF filter with constant parameters can be seen in Fig 4.1 and Fig 4.2 Fitting the Q parameters to an /a/ vowel results in a good match between the accelerometer signal and the GVV waveform, including its derivative, as seen in Fig 4.1. If we switch to an /i/ vowel using the Q parameters from vowel /a/, there is a deviation of the accelerometer signal from the GVV waveform, specially in the close phase as seen in Fig 4.2.

During ambulatory recordings, there is no access to the estimated glottal airflow from oral flow but only the accelerometer signal is available. By applying IBIF we can obtain an estimated GVV signal. Since constant Q parameters obtained from an /a/ vowel with comfortable SPL are used in the IBIF filter, we expect deviations from the ground-truth GVV. Taking frames from the ambulatory data, Fig 4.3 shows an expected GVV waveform, possible an /a/ vowel.

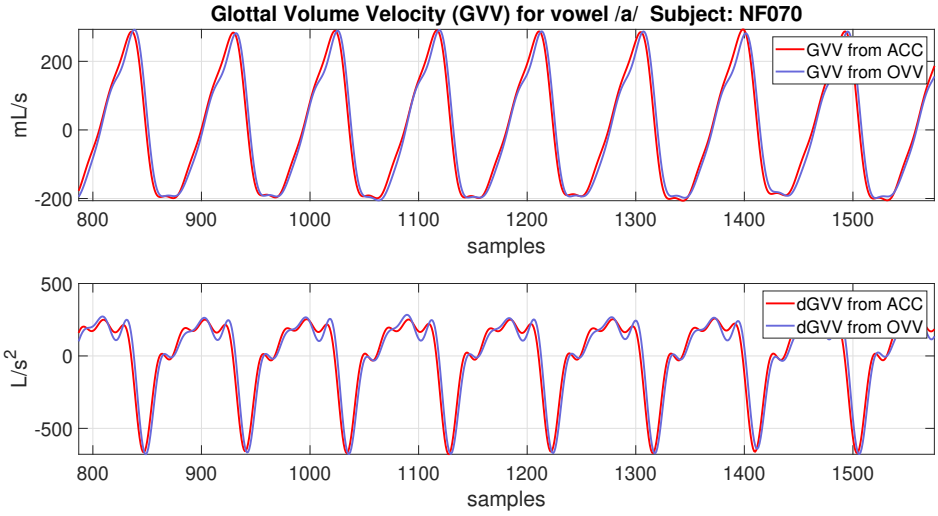

Figure 4.1: Glottal volume velocity (GVV) using inverse filtering from oral flow (blue waveform, top) and IBIF filtering from accelerometer (red waveform, top) for the vowel /a/, and their respective derivatives (bottom).

Fig 4.4 shows a waveform that is unlikely to be a good filtered signal due to the morphology of the waveform.

To address this issue, we propose a discrete-time space model implementation of the IBIF filter, which allows for both estimating the deviations of the IBIF model and adapting the airflow estimates to correct for potential deviations in the airflow signal estimates. In addition, the state-space model can be used as a data pruning tool to identify segments of the glottal airflow estimates with high adaptation to the IBIF model. Thus, the proposed approach is tested in the context of a classification task to discriminate between vocal fold nodules patients and healthy control subjects using ambulatory accelerometer data.

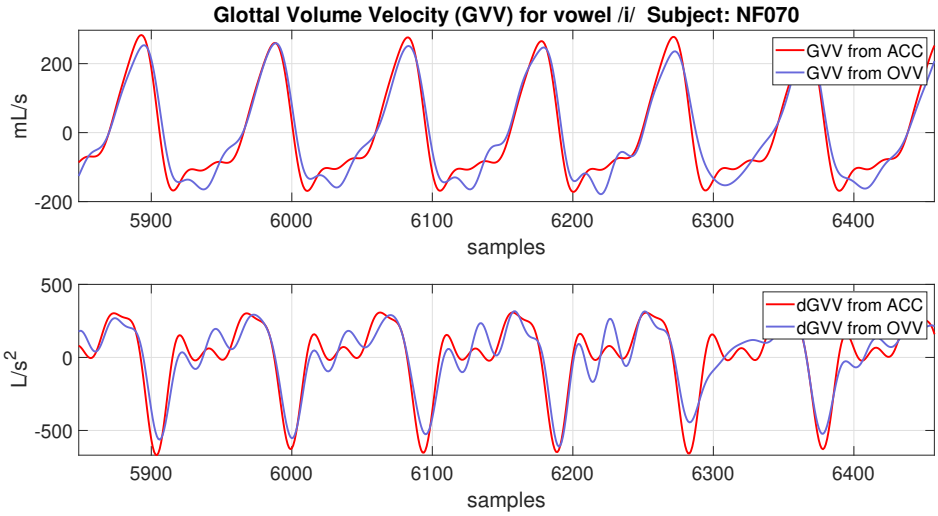

Figure 4.2: Glottal volume velocity (GVV) using inverse filtering from oral flow (blue waveform, top) and IBIF filtering from accelerometer (red waveform, top) for the vowel /i/ (using the Q parameters for an /a/ vowel), and their respective derivatives (bottom).

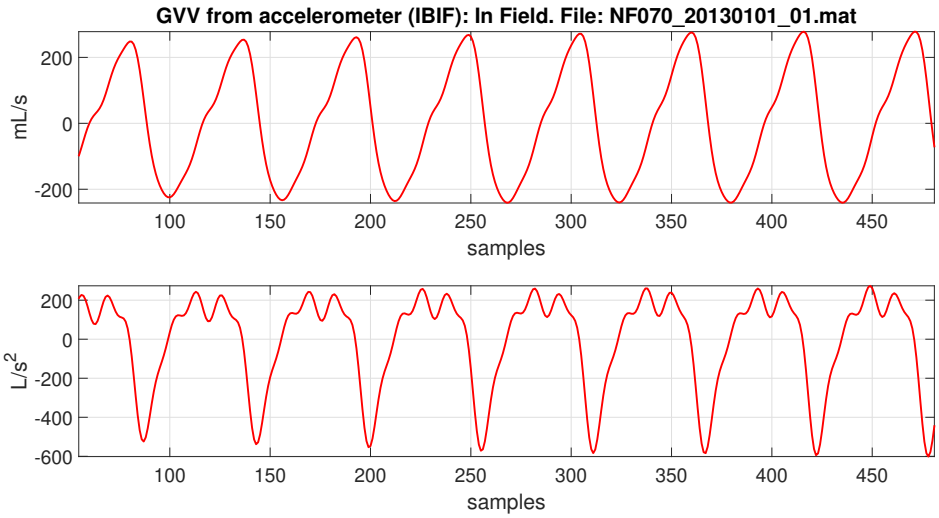

Figure 4.3: Estimated GVV signal (top figure) from IBIF during ambulatory recordings and its derivative (bottom figure).

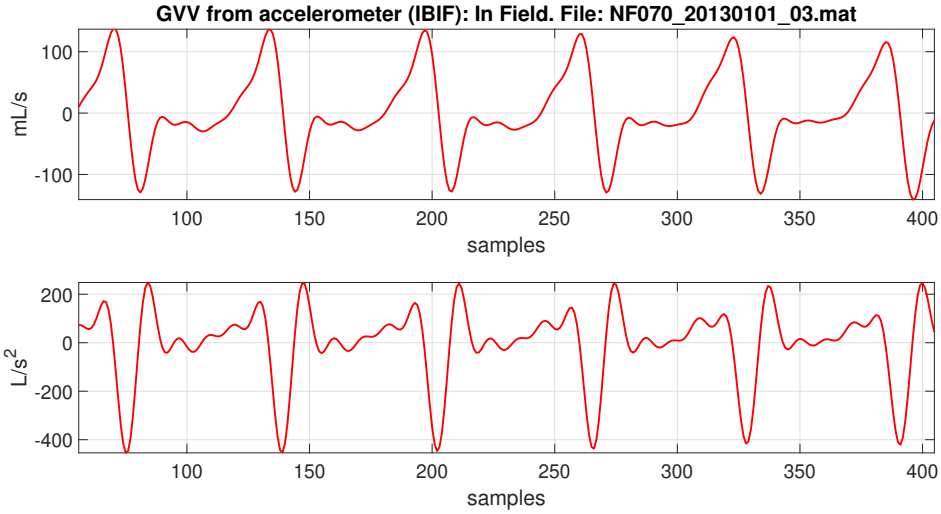

Figure 4.4: Estimated GVV signal (top figure) from IBIF during ambulatory recordings and its derivative (bottom figure).

Results show that the deviation between the proposed state-space model and direct IBIF implementation is not sufficiently strong to alter the classification performance, thus indicating that the effects of modeling variations with respect to the deterministic IBIF filter are not significant for these kinds of tasks. As a consequence, it is not necessary to calibrate the parameters of the IBIF filter for the collection of large amount of data. Other applications may take more advantage from the adaptation offered by using a Bayesian state-space implementation.

4.1 Discrete state-space methods

The state-space form the relationship between input, output, and noise signal in a convenient way of a linear dynamical system of a first-order of difference

equations. It has been used in numerous cases of physical phenomena [151] due to its simplicity and efficacy. The general framework in matrix form is the following:

$$\mathbf{x}(k+1) = \mathbf{A}(\theta)\mathbf{x}(k) + \mathbf{B}(\theta)\mathbf{w}(k) + \mathbf{E}(\theta)\mathbf{x}_o(k) \tag{4.1}$$

$$\mathbf{z}(k) = \mathbf{C}(\theta)\mathbf{x}(k) + \mathbf{D}(\theta)\mathbf{v}(k) \tag{4.2}$$

where $\mathbf{x}(k)$ is the vector state of the process at time k, $\mathbf{z}(k)$ is the vector observation model at time k, $\mathbf{x}_o(k)$ is a deterministic vector input signal, $\mathbf{w}(k)$ and $\mathbf{v}(k)$ are vectors with random noise of the process and observation, respectively, and $\mathbf{A}(\theta)$, $\mathbf{B}(\theta)$, $\mathbf{E}(\theta)$, $\mathbf{C}(\theta)$, and $\mathbf{D}(\theta)$ are matrices with parameters θ that control the state equations. Assuming the following conditions:

1. $\mathbf{w}(k)$ and $\mathbf{v}(k)$ are drawn from Gaussian distributions of known parameters, where $\mathbf{B}(\theta)$ and $\mathbf{D}(\theta)$ are identity matrices.

2. $\mathbf{A}(\theta)$ is known and is a linear function of $\mathbf{x}(k)$ and $\mathbf{w}(k)$.

3. $\mathbf{C}(\theta)$ is known and is a linear function of $\mathbf{x}(k)$ and $\mathbf{v}(k)$

it can be proved, when $p(\mathbf{x}(k)|\mathbf{z}_{1:k})$ is Gaussian, that the Kalman filter, a recursive algorithm to solve the state-space equations, is the optimal linear estimator of $\mathbf{x}(k)$ that minimizes the trace of the covariance of the estimation error [111, 152, 153]. The wide use of Kalman filters for real-time tracking of hidden states when observations are available is an attractive tool for the problem of glottal flow

estimation. Even though more advance tools could be applied to the problem of flow estimation, such as deep neural networks, the simplicity and tractability of the Kalman filter makes it ideal for a first approach to the specific problem of estimating glottal flow from neck-skin acceleration. Moreover, if the estimation of $\mathbf{x}(k)$ has available a delay of N time units, during which $\mathbf{z}(K+1), \ldots, \mathbf{z}(k+N)$ appear, the estimation of $\mathbf{x}(k)$ becomes:

$$\hat{\mathbf{x}}(k) = \mathbf{E}[\mathbf{x}(k)|\mathbf{z}(0), \mathbf{z}(1), \ldots, \mathbf{z}(k+N)] \tag{4.3}$$

which is a smoothed estimate. Even though there is a delay of N samples, the smoothed estimate produced by a smoother is expected to perform better than a filter [154]. This smoothed estimate will be used in the construction of the Kalman filter in the following section.

4.2 Formulation of IBIF model based on Kalman filter

Even though the IBIF algorithm works well in laboratory settings where the calibration procedure is done with a Rothenberg mask, there are uncertainties related to the application of the IBIF filter in ambulatory settings. First, the position and arrangement of the sensor during in field monitoring might not match the laboratory specifications for inverse filtering, so the subject-specific parameters could change slightly. In this case, one approach for tracking an unknown

signal (i.e., GVV) of a given process (i.e., IBIF) based on the observation of a related noisy/perturbed signal (i.e., neck-skin acceleration) is the use of a Bayesian approach to estimate the unknown signal and its uncertainty [111]. Under the assumption of linearity and Gaussian distributions for the posterior distributions of the unknown states, a Kalman Filter is the optimal estimator. In this work, we propose an alternative formulation of IBIF combining the state-space framework with the moving average (MA) canonical form [155]:

$$\mathbf{x}(n+1) = \mathbf{A}\mathbf{x}(n) + \mathbf{w}(n), \tag{4.4}$$

$$z(n) = \mathbf{C}\mathbf{x}(n) + v(n). \tag{4.5}$$

What follows is the instantiation of the general Kalman filter described by Eq 4.4 and Eq 4.5, to our particular problem, where $\boldsymbol{x}(n)$ is the state vector that will contain the GVV signal: $\mathbf{x}(n) = [x(n-N+1) \quad x(n-N+2) \quad x(n-N+3) \quad \cdots \quad x(n)]^T$ where N is the length of the skin-impulse response. $\mathbf{A}$ is a transition matrix that in this case is a simple time-shift matrix:

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & 0 & 0 \end{bmatrix} \in \mathbb{R}^{N \times N}$$

and $\mathbf{w}(n)$ is a stochastic driving noise with zero mean and covariance:

$$\mathbf{R}_{\mathbf{w}} = \begin{bmatrix} 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & \dots & 0 & \sigma_w^2 \end{bmatrix} \in \mathbb{R}^{N \times N}$$

The initial conditions for this set of equations are the mean $\mathbf{m}_0 = \mathbf{E}(\mathbf{x}_0)$ and covariance $\mathbf{P}_0 = \mathbf{E}((\mathbf{x}_0 - \mathbf{m}_0)(\mathbf{x}_0 - \mathbf{m}_0)^T)$ of the initial state $\mathbf{x}_0$.

The observation equation Eq 4.5 relates the accelerometer signal $z(n)$ as a filtering process between the unobserved state $\mathbf{x}$ and the neck-skin impulse response $h(n)$ with coefficients:

$$\mathbf{C} = \begin{bmatrix} h(0) & h(1) & h(2) & \cdots & h(N-1) \end{bmatrix}^T \in \mathbb{R}^{1 \times N}$$

According to Eq 4.5, Gaussian noise $v(n)$ with mean zero and variance σ_v^2 is assumed to perturb the observed signal. The Kalman filter tracks point estimates of state up to time n : $\mathbf{x}(n|n)$ and its covariance matrix $\mathbf{P}(n|n)$. It is important to notice that, even though $\mathbf{P}(n|n)$ is usually used as a measure of uncertainty for the point estimates $\mathbf{x}(n|n)$, it is not the measurement for uncertainty, or deviation,

that is utilized in this work. Since the matrices A and C are constant, $\mathbf{P}(n|n)$ converges to a constant value (see [156], chapter 6), which is not useful for the current task. Instead, a comparison is made between the point estimate $\mathbf{x}(n|n)$ and the output of the IBIF filter. The algorithm for the standard MA Kalman filter as follows:

Algorithm 1 Kalman Filter Algorithm

procedure KALMAN($\mathbf{A}, \mathbf{C}, \mathbf{R_w}, \mathbf{R}_v, \mu_0, \mathbf{\Sigma}_0, z(n)$)

Initialization: Set $\mathbf{x}(0|0) = \mu_0$ and $\mathbf{P}(0|0) = \mathbf{\Sigma}_0$

Filtering: For $n = 1, 2 \ldots T$

Prediction equations:

$$\mathbf{x}(n|n-1) = \mathbf{A}\mathbf{x}(n-1|n-1)$$

$$\mathbf{P}(n|n-1) = \mathbf{A}\mathbf{P}(n-1|n-1)\mathbf{A}^T + \mathbf{R_w}$$

Update equations:

$$\mathbf{K}(n) = \mathbf{P}(n|n-1)\mathbf{C}^T(\mathbf{C}\mathbf{P}(n|n-1)\mathbf{C}^T + \mathbf{R}_v)^{-1}$$

$$\mathbf{x}(n|n) = \mathbf{x}(n|n-1) + \mathbf{K}(n)(z(n) - \mathbf{C}^T\mathbf{x}(n|n-1))$$

$$\mathbf{P}(n|n) = \mathbf{P}(n|n-1) - \mathbf{K}(n)\mathbf{C}\mathbf{P}(n|n-1)$$

Due to the structure as a time sequence of the state $\mathbf{x} = [x(n-N+1) \quad x(n-N+2) \quad x(n-N+3) \quad \cdots \quad x(n)]^T$, $x(n-N+1)$ is a smooth estimate of the GVV waveform. This corresponds to a fixed-lag smoother, where $x(n-N+1) = \hat{x}(n-N|n) = \mathbf{E}[x(n)|z(0), \quad z(1), \quad \ldots, z(n+N)]$. In this framework, the application

of the traditional Kalman filter allows us to estimate a state that depends on N samples ahead, for which is a smooth estimate that improves the estimation.

The implementation of the IBIF method in a Kalman filter framework has two important additions: The adaptive tracking of the GVV signal using the accelerometer and the modeling of state and observation noise. In the first case, the adaptive tracking is done through the sample by sample correction of the expected accelerometer signal by the Kalman gain $\mathbf{K}(n)$. In our hypothesis, the correction allows to track a GVV signal that minimizes the deviations from the GVV signal obtained with IBIF. There are different methods to select the state process noise variance σ_w^2 (mL^2/sec^2) and the observation noise variance $\mathbf{R}_v = \sigma_v^2$ (cm^2/sec^4). These include the use of the Expectation-Maximization (EM) algorithm to estimate those variables [157, 158], through iterative adaptation of the state [159] and observation model [160], and on heuristics based on previous knowledge of the state and noise models [110, 161]. For simplicity, we assume constant noise statistics that are obtained with data through a grid-search process in which we compare the root-mean-square error (RMSE, mL/sec) between the Kalman state $x(n - N + 1)$ and a reference GVV signal obtained from the OVV signal [27]. Fig 4.5 shows different values of σ_w^2 and σ_v^2 where multiple minimums ($RMSE = 17.268$) are found within a range for one subject saying the vowel /a/. Most blue RMSE values in Fig 4.5 corresponds to $RMSE = 17.273$ which are very close to the minimum. Similar trends were found for other subjects and

vowels. We selected $\sigma_w^2 = 100$ and $\sigma_v^2 = 1$ in this work, which are plausible values for the state and measurement noises due to the assumption of higher process noise due to glottal flow variance with low observation noise, while they produce a minimum RMSE value.

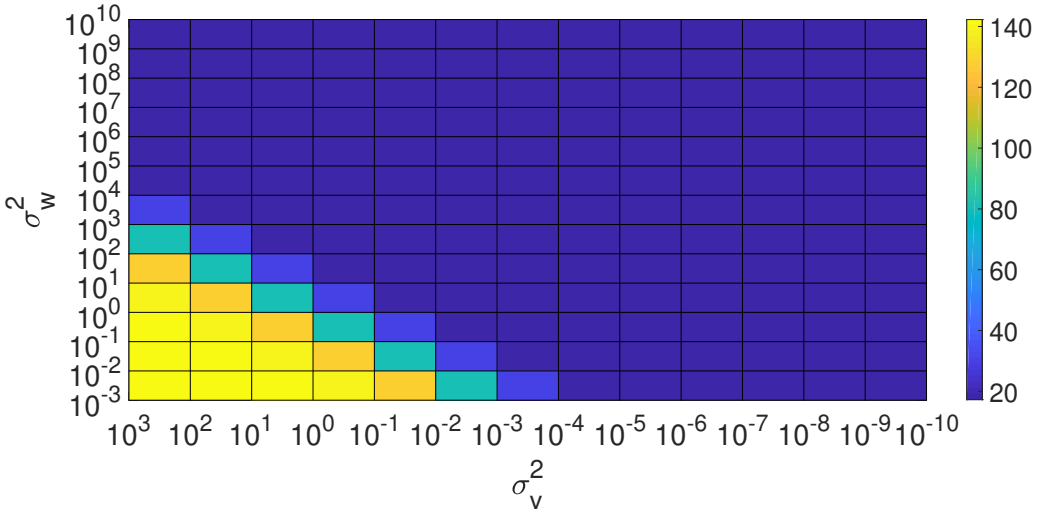

Figure 4.5: RMSE values for different combinations of σ_w^2 and σ_v^2.

4.3 Glottal flow model as input to Kalman filter

According to Fant's theory of the source-filter model of speech production [162], the laryngeal excitation can be consider independent of the vocal tract. Even though there is evidence for certain cases of non-linear coupling between the glottal source and the vocal tract [71, 123], the source-filter theory has served well for the development of glottal source modeling and estimation. In terms of modeling the glottal source, there is parametric modeling in the time domain, for

example the Rosenberg model of glottal pulse [163] and the Lijecrants-Fant (LF) model of the derivative of the glottal pulse [164] are the most known parametric models in the time domain, from which other models have been derived.

4.3.1 Glottal source as a low-pass filter

The source-filter theory models the glottal source as a low-pass system in the spectral domain. Therefore, the glottal flow is the output of this low-pass filter to an impulse train. Originally, Fant [162] used four poles on the negative real axis to model the glottal source:

$$U_g(s) = \frac{U_{g0}}{\prod_{i=1}^{4}\left(1 - \frac{s}{s_{ri}}\right)} \tag{4.6}$$

where $|s_{r1}| \simeq |s_{r2}| = 2\pi 100\ Hz$ and $s_{r3} = 2\pi 2000\ Hz$, $s_{r4} = 2\pi 4000\ Hz$, and U_{g0} is a gain factor. According to Fant, s_{r3} and s_{r4} are fixed, $|s_{r1}|$ and $|s_{r2}|$ account for variability to regard speaker and stress, where the fundamental frequency f_0 it's included. Therefore, this is a 6 parameters spectral model (f_0, U_{g0}, and four poles) [165].

This model of glottal flow has been used in numerous applications, including for the derivation of linear prediction equations for speech [166], where only two poles are used due to the linearity of the acoustic model that only holds for frequencies below 4000 Hz. In this case, the model simplifies to three parameters: The

constant gain U_{g0}, f_0 and a frequency parameter $s_{r1} \simeq s_{r2}$. Therefore, the spectral characteristics of the glottal pulse are those of a second-order filter frequency response, with a spectral tilt of about $-12dB/oct$.

A low-pass Butterworth filter [118] is designed to model the glottal source. The magnitude-squared frequency response of a second-order Butterworth filter is:

$$|H(\Omega)|^2 = \frac{G_0^2}{1 + (\Omega/\Omega_c)^4}, \tag{4.7}$$

where Ω_c is the cut-off frequency approximately at $-3dB$ and G_0 is the DC gain. In the s domain ($s = j\Omega$), Eq 4.7 can be rewritten as:

$$H(s)H(-s) = \frac{G_0^2}{1 + (-s^2/\Omega_c^2)^2}. \tag{4.8}$$

The poles s_k of the filter correspond to:

$$s_k = \Omega_c e^{j\pi 2/2} e^{j(2k+1)\pi/4}, \quad k = 0, 1 \tag{4.9}$$

The transfer function corresponds to:

$$H(s) = \frac{\Omega_c G_0}{(s - s_0)(s - s_1)} \tag{4.10}$$

The impulse invariance method [118, 167] is used to convert the filter to digital form, so the conversion:

$$z = e^{sT} \tag{4.11}$$

is used, where $T = 1/f_s$ is the sampling period and f_s the sampling frequency. Therefore, the poles corresponding to the z domain correspond to $z_0 = e^{s_0 T}$ and $z_1 = e^{s_1 T}$. By substituting the poles from equation 4.9 into the transfer function 4.10, we obtain the following:

$$H(z) = \frac{\Omega_c^2}{\left(z - e^{-\frac{\sqrt{2}}{2}\Omega_c(1+j)}\right)\left(z - e^{-\frac{\sqrt{2}}{2}\Omega_c(1-j)}\right)} \tag{4.12}$$

By expanding and doing some algebraic work, the following transfer function is obtained:

$$H(z) = \frac{\Omega_c^2 z^{-2}}{1 - 2\alpha_1 z^{-1} + \alpha_2 z^{-2}} \tag{4.13}$$

where $\alpha_1 = e^{\frac{-\sqrt{2}}{2}\Omega_c T}\cos\frac{\sqrt{2}}{2}\Omega_c T$, $\alpha_1 = e^{-\sqrt{2}\Omega_c T}$, and $\Omega_c = 2\pi f_c$, where f_c is the cut-off frequency of the low-pass filter.

From Eq 4.13 we can modify matrix A from the general Kalman filter to an

AR model, so the white noise entering the system is "colored" by the low-pass filter representing the glottal source. The new A_{lp} matrix is:

$$\mathbf{A_{lp}} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \\ 0 & 0 & 0 & \dots & 0 & 1 \\ 0 & 0 & 0 & \dots & -2\alpha_1 & \alpha_2 \end{bmatrix} \in \mathbb{R}^{N \times N}$$

An example of the estimated GVV using matrix $\mathbf{A_{lp}}$ is shown in Fig 4.6 and compared to the estimated GVV using the original matrix $\mathbf{A}$. The upper plot shows the tracking of the first state, which corresponds to the smoothed GVV. It is visible that there are no differences between the original proposal of Kalman and the one incorporating a low-pass filter as input noise. There is a slight difference in the tracking of the last state of the GVV, which corresponds to the state considering all past samples up to the current sample n. The original Kalman outputs a zero-mean signal, while the Kalman with a low-pass input tracks a very-small amplitude signal similar to an expected GVV.

4.3.2 Rosenberg model for glottal pulse

A parametric model of the glottal pulse can be obtained from the Rosenberg model [163]. Following an example from Rabiner et al. [47], the Rosenberg model

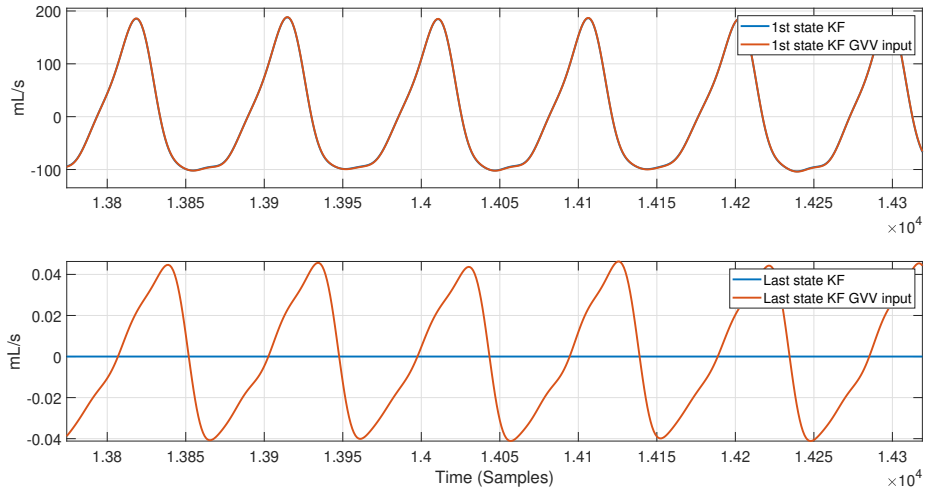

Figure 4.6: Top panel: GVV output (first state) using $\mathbf{A}$ (blue) and $\mathbf{A_{lp}}$ (red). Bottom panel: GVV output (last state) using $\mathbf{A}$ (blue) and $\mathbf{A_{lp}}$ (red)

can be written as:

$$g[n] = \begin{cases} 0.5[1 - cos(\pi(n+1))/N_1] & 0 \leq n \leq N_1 - 1 \\ cos(0.5\pi(n+1-N_1)/N_2) & N_1 \leq n \leq N_1 + N_2 - 2 \\ 0 & \text{otherwise} \end{cases}$$

where N_1 is the number of samples of the open phase and N_2 is the number of samples of the closing phase. For a sequence of 96 samples, with $N_1 = 25$ and $N_2 = 10$, the z-transform $G(z)$ has the form:

$$G(z) = z^{-95} \prod_{k=1}^{95} (-b_k^{-1}) \prod_{k=1}^{95} (1 - b_k z) \tag{4.14}$$

where b_k corresponds to the zeros of $G(z)$, which can also be written in the

following form:

$$G(z) = g[0] + g[1]z^{-1} + g[2]z^{-2} + \cdots + g[N-1]z^{-(N-1)} \tag{4.15}$$

The glottal pulse $g[n]$ and its spectrum are plotted in Fig 4.7

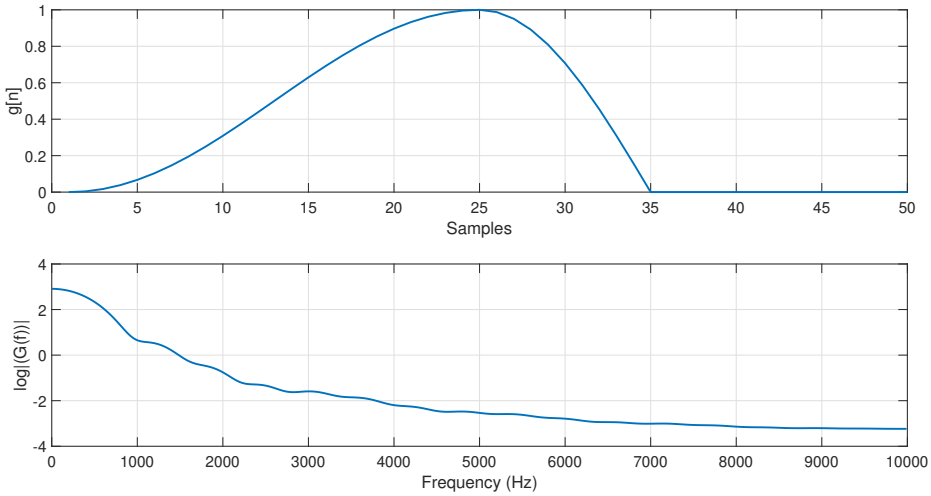

Figure 4.7: Rosenberg model in time domain (only first 50 samples shown, top panel) and its frequency response (bottom panel).

The periodic excitation $p[n]$ is modeled as one-sided quasi-periodic impulse train:

$$p[n] = \sum_{k=0}^{\infty} \gamma^k \delta[n - kN_p] \tag{4.16}$$

which has z-transform:

$$P(z) = \sum_{k=0}^{\infty} z^{-kN_p} = \frac{1}{1 - \gamma z^{-N_p}} \tag{4.17}$$

where $N_p = f_s/f_0$ (fundamental period in samples) and γ is a number close to 1 (e.g, 0.999) to make the filter stable. Fig 4.8 shows the spectrum of the periodic input $P(z)$ with a fundamental frequency of $f_0 = 210Hz$ ($N_p = 96$).

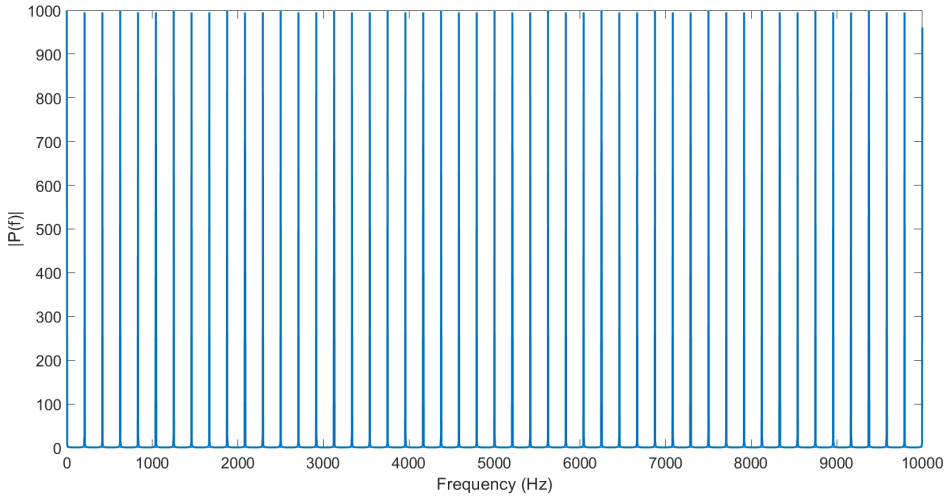

Figure 4.8: Frequency response of periodic input for voicing modeling ($f_0 = 210Hz$) before multiplication with Rosenberg source spectrum.

Therefore, $P(z)G(z)$ is the z-transform of the glottal flow model (spectrum shown in Fig 4.9). This is an ARMA model that can be constructed as a shaping filter input to the canonical MA model (Eq 4.18) [168, 169].

$$x_{sf}(n) = -\sum_{k=1}^{p} \alpha_k x_{sf}(n-k) + \sum_{k=0}^{q} \beta_k w_2(n-k) \tag{4.18}$$

where $x_{sf}(n)$ is the state of the shaping filter, α_k and $beta_k$ are the kth coefficient of the AR and MA model, respectively, and $w_2(n)$ is Gaussian noise with mean 0 and variance $\sigma^2_{w_2}$. The state-space equations for this model is:

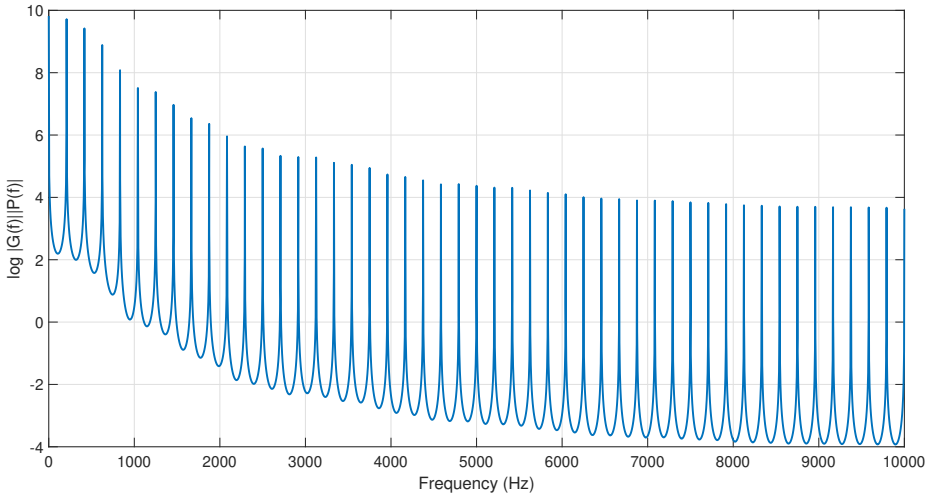

Figure 4.9: Periodic input P(z) multiplied by Rosenberg model G(z), which corresponds to an ARMA model ($f_0 = 210Hz$).

$$\mathbf{x_{SF}}(n+1) = \mathbf{A_{SF}}\mathbf{x_{SF}}(n) + \mathbf{B}_{SF}w_2(n), \tag{4.19}$$

$$w_1(n) = \mathbf{C_{SF}}\mathbf{x_{SF}}(n). \tag{4.20}$$

where $\mathbf{x_{sf}}(n) = (x_{SF}(n-p+1) \quad x_{SF}(n-p+2) \quad \dots \quad x_{SF}(n))^T$ is the state vector, where p is the order of the AR model. Since the periodic input has N_p poles, the order of the AR model is $p = N_p$. $\mathbf{A_{SF}}$ is the transition matrix $p \times p$:

$$\mathbf{A_{SF}} = \begin{bmatrix} 0 & 1 & 0 & \dots & 0 & 0 \\ 0 & 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & 1 \\ -\alpha_p & -\alpha_{p-1} & -\alpha_{p-2} & \dots & -\alpha_2 & -\alpha_1 \end{bmatrix} \in \mathbb{R}^{p\times p}$$

$\mathbf{B}_{SF} = [0 \quad 0 \ldots 1]^T \in \mathbb{R}^{p\times 1}$ and $w_2(n)$ is a stochastic driving noise with zero mean and variance $\sigma^2_{w_2}$. The MA equation (Eq 4.20) contains $\mathbf{C_{sf}} = [\beta_q \quad \beta_{q-2} \ldots \beta_1 \quad \beta_0] \in \mathbb{R}^{1\times(q+1)}$ and the colored noise $w_1(n) \in \mathbb{R}$ is the dot product of $\mathbf{C_{sf}}$ and $\mathbf{x_{sf}}(n)$. $w_1(n)$ is the input the MA state model in Eq 4.4. A continuous time diagram of this augmented system is shown in Fig 4.10. The white noise $w_2(t)$ is input to the shaping filter, which is the Rosenberg model convolved with the periodic input (Fig 4.9). the output of this filter is the colored white noise $w_1(t)$, which is the input noise to the canonical MA system (physical system in Fig 4.10), whose output $z(t)$ is the observed signal, i.e., the neck-skin acceleration. The new state-space equations in discrete-time are:

$$\mathbf{X}_T(n+1) = \mathbf{A}_T\mathbf{X}_T(n) + \mathbf{B}_T\mathbf{w}(n), \tag{4.21}$$

$$z(n) = \mathbf{C}_T\mathbf{X}_T(n) + v(n). \tag{4.22}$$

where

$$\mathbf{A}_T = \begin{bmatrix} \mathbf{A} & \mathbf{B}\mathbf{H}_{SF} \\ \mathbf{0} & \mathbf{A}_{SF} \end{bmatrix} \tag{4.23}$$

$$\mathbf{B}_T = \begin{bmatrix} \mathbf{0} \\ \mathbf{B}_{SF} \end{bmatrix} \tag{4.24}$$

$$\mathbf{C}_T = \begin{bmatrix} \mathbf{C} & \mathbf{0} \end{bmatrix} \tag{4.25}$$

$$\mathbf{X}_T = \begin{bmatrix} \mathbf{x} \\ \mathbf{x}_{SF} \end{bmatrix} \tag{4.26}$$

An example of the Kalman filter with a Rosenberg (ARMA) noise input is shown in Fig 4.11. The GVV from a sustained vowel /a/ from a healthy female with an average frequency of $f_0 = 210$ Hz using the MA Kalman filter and the MA Kalman filter with ARMA noise as input. The top panel shows the last state $x(n - N - P + 1)$ which is the smooth estimate. Again, N is the length of the impulse response of the neck-skin (350 pts.) while p is the order of the AR model (96 pts.). Both signals are very similar, therefore, the smoothing in both cases converges to the same type of signal. For the case of the bottom panel, the MA Kalman filter tracks the current state $x(n)$, while the MA Kalman filter with ARMA input tracks the state $x(n - p)$, which is the state after the ARMA filter is applied (corresponding to $x(n)$ in the original MA KF). The state $x(n - p)$ from the MA KF with ARMA input is able to track a GVV signal similar to the smoothed state from the top panel with the same amplitude.

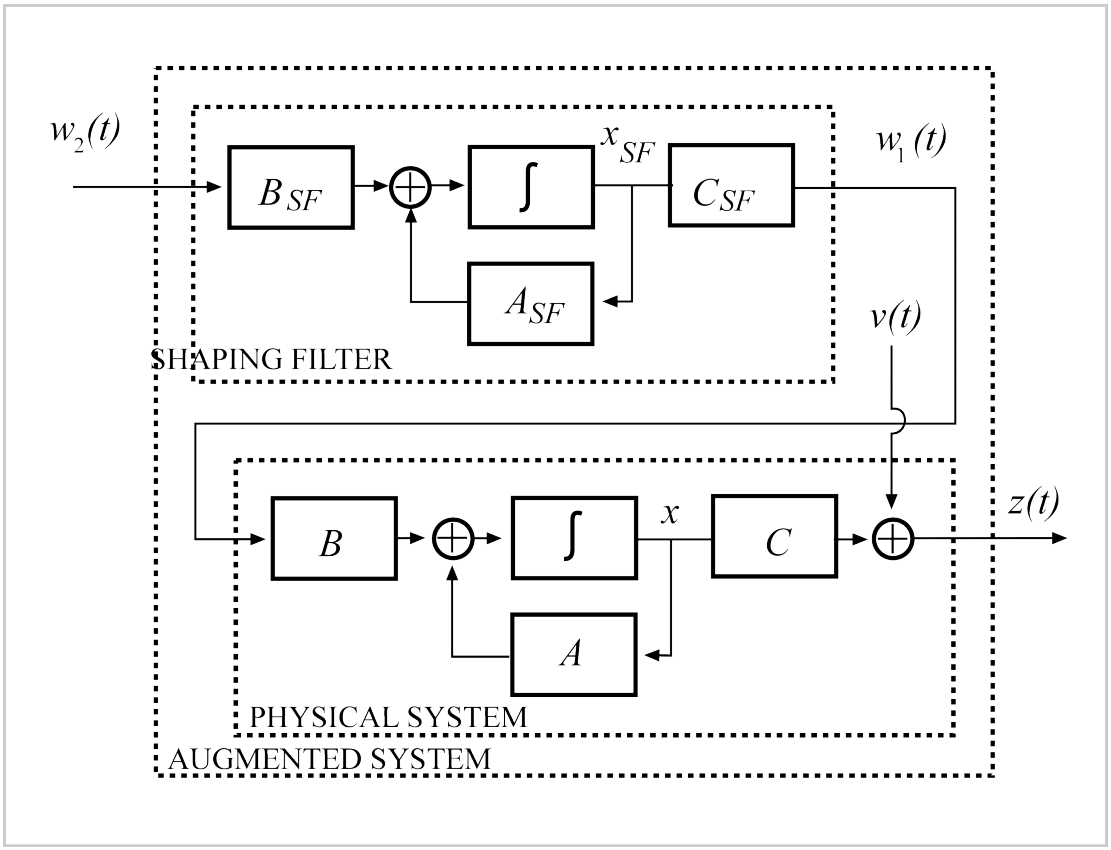

Figure 4.10: Diagram of augmented Kalman Filter in a continuous time domain. The Physical system corresponds to the MA Kalman Filter, the shaping filter is the colored noise system composed of an ARMA Kalman Filter.

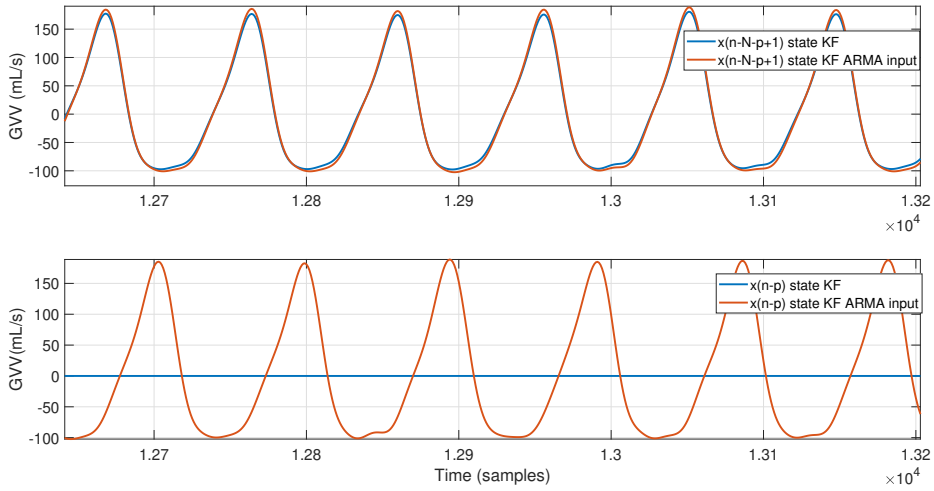

Figure 4.11: Top panel: GVV output (first state) using $\mathbf{A}$ (blue) and $\mathbf{A_{lp}}$ (red). Bottom panel: GVV output (last state) using $\mathbf{A}$ (blue) and $\mathbf{A_{lp}}$ (red)

The augmented Kalman Filter has a physiological meaning in the sense that a stochastic input, a parametric model of the GVV signal, is used within the Kalman framework to estimate the GVV signal from an IBIF modeling. The issues with this framework is the necessity to know before hand N_p (i.e., f_0) in order to construct the ARMA filter, and this might change over time, so the matrices need also to adjust their sizes for these changes. Another issue is the simplicity of the model, which might no be applicable to pathological voices due to the lack of parameters in the Rosenberg model to modify the signal to pathological cases. Moreover, it is necessary to know opening and closing phases percentages to construct the model, which can only be assumed since there is no prior information on the shape of the GVV waveform. Since the smoothed state is very similar

to the smoothed state incorporating the stochastic ARMA input, only the MA Kalman Filter will be used on the following experiments.

4.4 Experimental setup

4.4.1 Participants

The human studies protocol used to collect the data for this study was approved by the Institutional Review of the Partners Healthcare System - the Massachusetts General Hospital (MGH) is a founding member of this organization. Study participants were 16 pairs of adult females (total of 32 subjects) with each pair comprised of one patient with PVH (diagnosed with vocal nodules) and one normal control subject matched to the patient by age and occupation. Due to the higher incidence of female patients with PVH than male in the overall population [37], only women were subjects for this study. Zhukhovitskaya et. al.[124] have shown significant differences ($p < 0.0001$) in the number of bilateral midfold lesions between males and women. Moreover, the inclusion of men would create confounding variables due to sex-specific characteristics. The patient matching is done to normalize for general vocal behavior differences. Clinical diagnoses were based on a complete team evaluation by laryngologists and speech-language pathologists at the MGH Voice Center that included (a) a complete case history, (b) endoscopic imaging of the larynx, (c) aerodynamic and acoustic assessment

of vocal function based on Mehta et. al. [34], (d) a patient-reported Voice-Related Quality of Life questionnaire, and (e) a clinician-administered Consensus Auditory-Perceptual Evaluation of Voice assessment (CAPE-V). All patients were enrolled prior to the administration of any voice treatment. Written informed consent was obtained from all subjects. All subjects were 18 years of age or older.

4.4.2 Ambulatory data

Each subject was recorded as they engaged in normal daily activities during one week using a smartphone-based ambulatory voice monitor [23, 2]. The system employs an accelerometer attached to the front of the neck below the larynx as the phonation sensor. The sampling frequency was 11,025 Hz and the average total recording time for a subject was approximately 80 hours [75] [2]. An example of the ambulatory setup for a patient is shown in Fig 3.2 from chapter 3. The Kalman filter is processed for every sample of the ambulatory data, except for those points where there is silence or unvoiced segments. Voiced segments are selected using a voice activity detector (VAD) every consecutive 50 ms. When the accelerometer data goes from an unvoiced to voiced segment, the filter continues with the state and covariance from the previous voiced segment.

4.4.3 Aerodynamic features

Table 4.1 is a subset from table 3.2 in chapter 3 and describes the features ob-

tained from the GVV waveform (and its derivative) from every 50 ms frame. Aerodynamic features have been shown to be good indicators of PVH within subjects [12, 25, 27]. Time-domain features (AC-flow, MFDR, OQ, SQ, and NAQ [170]) are obtained by taking the median value of all cycles within the frame. Frequency-domain features (H1-H2 and HRF) are computed using the whole frame. In contrast to [11], the median of the data frames are used as features, instead of their higher-order statistics. An example of data collection during one day (19,067 data frames) for the AC-flow feature using IBIF and Kalman is shown in Fig 4.12.

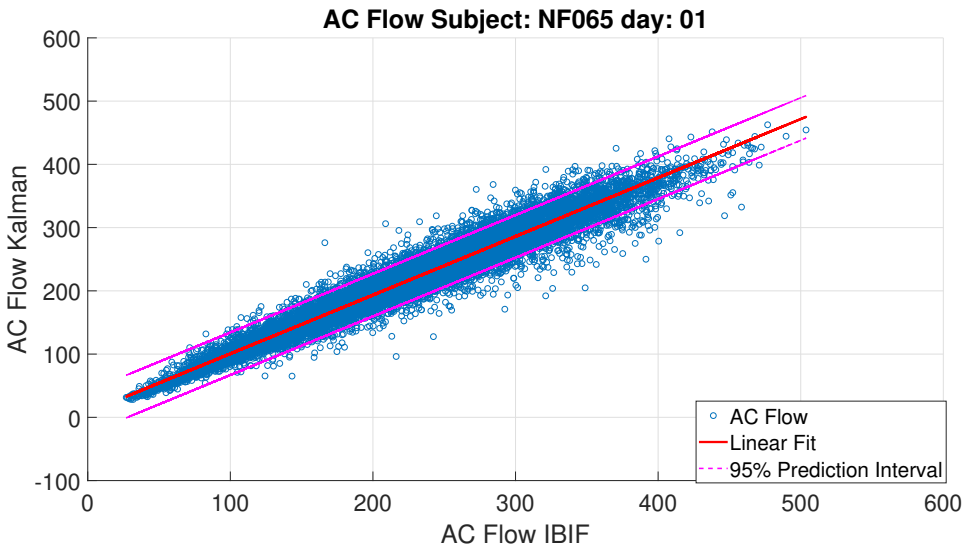

Figure 4.12: Day 1: AC-flow values using Kalman and IBIF for a control subject, including a linear fit and a 95% prediction interval.

Table 4.1: Frame-based derived glottal airflow measures to be estimated from the ambulatory neck-surface accelerometer signal using impedance-based inverse filtering and Kalman filter.

Glottal airflow measures	**Description**	**Units**
AC-flow	Peak-to-peak glottal airflow.	mL/s
MFDR	Negative peak of the first derivative of the glottal waveform.	L/s^2
Open Quotient (OQ)	Ratio of the open time of the glottal vibratory cycle to the corresponding cycle period.	–
Speed Quotient (SQ)	Ratio of the opening time of the glottis to the closing time.	–
H1-H2	Difference between the magnitude of the first two harmonics.	dB
Harmonic Richness Factor (HRF)	Ratio of the sum of the amplitudes of the first 8 harmonics to the amplitude of the first harmonic.	dB
Normalized Amplitude Quotient (NAQ)	Ratio of ACFL to MFDR divided by the glottal period.	–

4.4.4 IBIF calibration with laboratory data

Each subject underwent a session in the laboratory to obtain a subject-specific calibration for the IBIF algorithm. The session involved simultaneous and synchronous recordings of a circumferentially vented mask-based oral volume velocity (OVV) and neck skin acceleration in an acoustically treated room. Each subject performed a series of sustained vowel gestures (/a/ and /i/) with a constant pitch using comfortable and loud (approximately 6 dB increase) voice. For each gesture, a bandpass filtered ($60-1100$ Hz) oral airflow vowel segment was used to perform inverse filtering with a single notch filter constrained to unitary gain at DC [125].

Once a glottal airflow approximation is obtained from the mask, $\mathbf{Q}$ parameters are estimated using the optimization scheme described in [85]. These are the parameters describing the mechanical properties of the neck-skin, as well as the length of the trachea and the position of the accelerometer with respect to the glottis [85].

In order to reduce the complexity of the Kalman smoother, we need to reduce the size of the matrices $\mathbf{A}$ and $\mathbf{C}$. This is necessary due to the computational cost of Kalman filter in the multiplications of state-space matrices of size 550×550 when processing clinical signal with recording time of approximately 80 hours. Since $\mathbf{A}$ and $\mathbf{C}$ depend on the length of the neck-skin impulse response $h(n)$, the latter is truncated in the middle region and then windowed (Hann function) at a

length of 350 points. This procedure seeks to maintain the performance of IBIF filter because most of the energy of the impulse response is concentrated in the middle section, while the extremes are considerably low in energy. As an example, in Fig 4.13 we show a given $h(n)$ in blue and the resulting truncated version in red. The magnitude of the frequency response is shown in Figure 4.14.

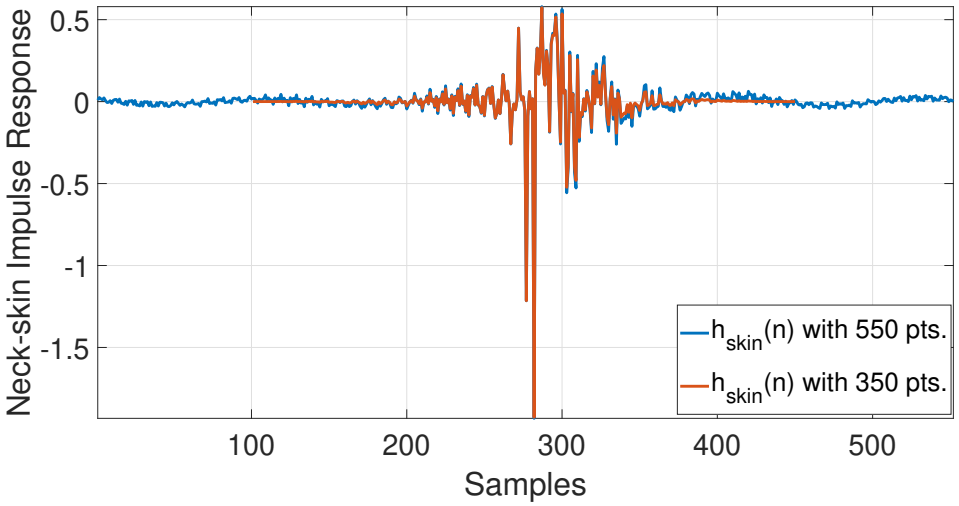

Figure 4.13: Neck-skin impulse response for a healthy female subject, full impulse (blue) and truncated version with a Hann window (red).

An example of a patient subject saying the vowel /a/ and /i/ is shown in Fig 4.15. The "ground truth" GVV is computed using a Single Notch Filter (SNF) on the OVV signal with an optimization procedure similar in [27] and [150]. The Kalman filter signal shown is the first state $x(n - N + 1)$, which is a smooth estimate computed considering future information in the acceleration signal $z(n)$ and

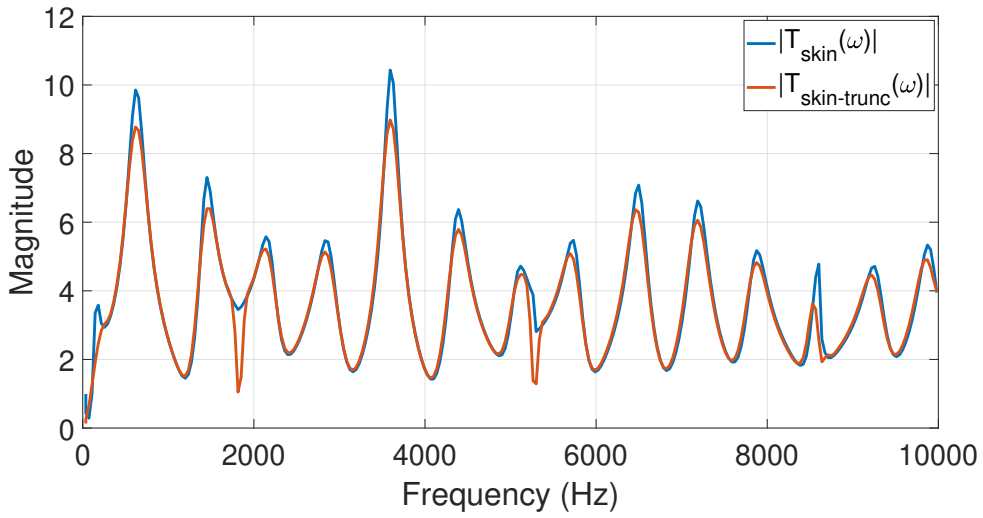

Figure 4.14: Neck-skin frequency response for a healthy female subject, full length (blue) and truncated version with a Hann window (red).

that is delayed due to the formulation of state transition equation (see Eqs 4.4 and 4.5). In this case, the "smooth" term does no refer to the classical approach of Kalman smoothing, which can be written in fixed-lag form, fixed-interval, or fixed-point form (for details, see [154, 153, 152]). The smoothing refers to the structure of the state vector x, which has N points, we are observing the estimate of the state $x(n-N+1|n)$, which is a delayed state of the current state at time n. This is the reason we do not need to do a backward pass to the filtering algorithm, as the state $x(n-N+1|n)$ already has information of future predictions.

As was previously stated the IBIF and, therefore, the model based on Kalman filter are both calibrated using a procedure based on the vowel /a/. The top plot of

Fig 4.15 shows a good fit between all the estimations of GVV. This is expected due to the calibration procedure of the **Q** parameters. The RMSE between Kalman and the reference signal is 16.79, while the RMSE between IBIF and the reference signal is 24.69. The bottom plot of Fig 4.15 shows the estimation results for the vowel /i/ where we can see some differences on the IBIF and Kalman estimations compared to the reference signal. Fig 4.16 shows a close zoom to one cycle of the vowel /i/. In this case, the method based on Kalman follows the reference signal a bit closer than IBIF. Even though the Kalman filter is an alternative implementation of the IBIF filter, the adaptive filtering nature of Kalman allows to track better the "ground truth" signal than IBIF. The RMSE between Kalman and the reference signal is 36.46, while between the IBIF and the reference signal is 46.01. Similar trends were found in different subjects and tokens.

For the selection of the window length to truncate the IBIF filter, we analyzed different window lengths of the impulse response and checked the RMSE (E_{abs}, Eq 4.27) obtained from 42 subjects saying the sustained vowel /a/ by comparing the original neck-skin impulse response and its shorter, windowed version,

$$E_{abs} = \sqrt{\frac{1}{N}\sum_{i=0}^{N-1} |h(n) - h_{trunc}(n)|^2}, \tag{4.27}$$

where N is 550 points (the original length), $h(n)$ is the original neck-skin impulse response, and $h_{trunc}(n)$ is the truncated, windowed impulse response of length L

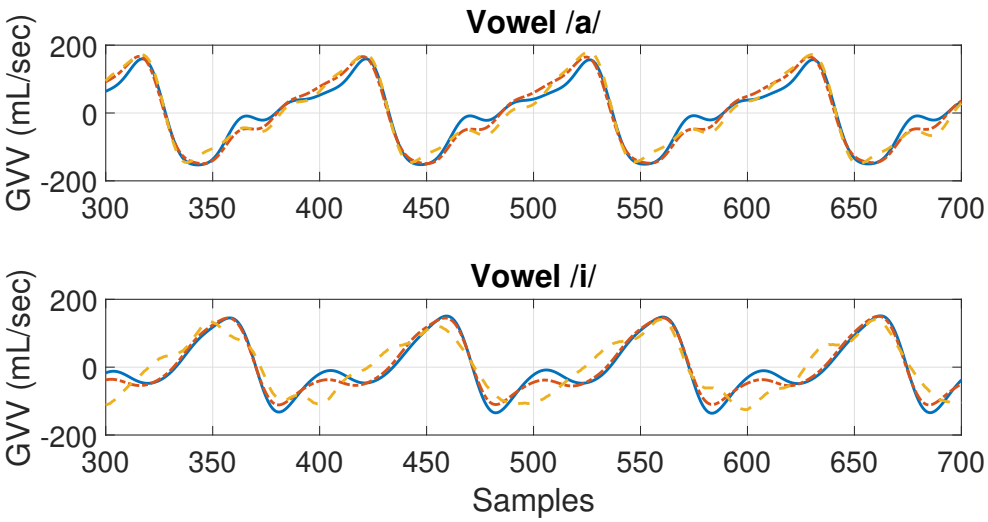

Figure 4.15: Sample of the Kalman filtered signal (red dashed), IBIF (blue solid) and a reference signal (SNF, yellow) for the vowels /a/ (top) and /i/ (bottom).

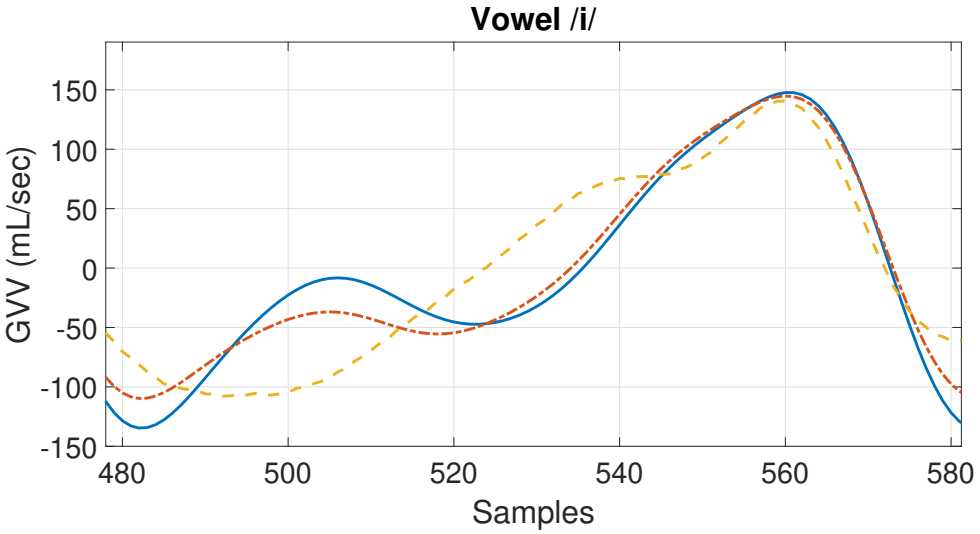

Figure 4.16: Zoom to one cycle of /i/ vowel signals corresponding to Fig 4.15 showing the KF signal (red dashed), IBIF (blue solid) and the SNF reference signal (yellow).

at the midpoint of $h(n)$ towards the beginning ($n = 0$), while length $L - 1$ goes from the midpoint towards the end ($n = 549$). From Fig 4.17 we can see that the median E_{abs} is below 0.1 after $L = 50$. By selecting $L = 175$, we have a conservative truncated window where the median gets closer to zero and all subjects are grouped together around that number. Therefore, we have a truncated window of length $2L = 350$.

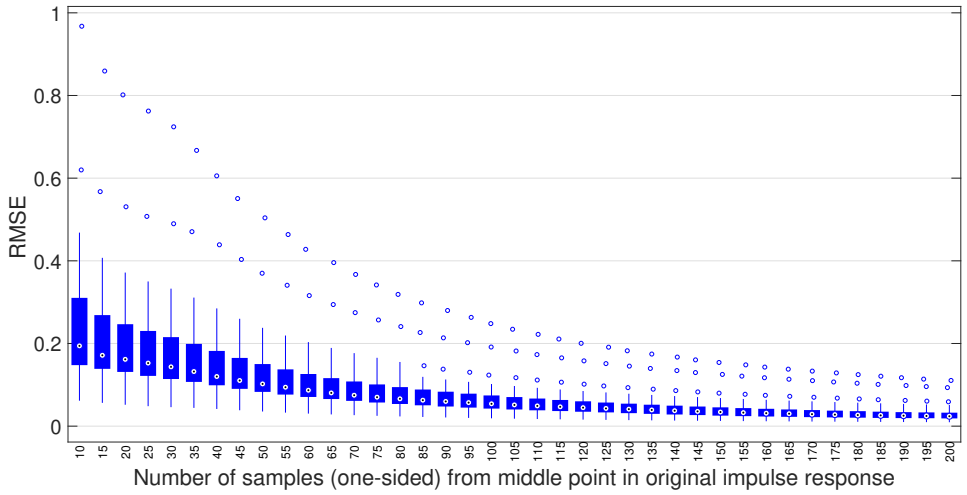

Figure 4.17: Number of samples L (one-sided) from middle point in original impulse response vs. E_{abs}.

4.5 Results

A subset of the features obtained in an ambulatory setting using the IBIF filter are compared with those obtained using Kalman Filter per subject in Table 4.2.

There is a large standard deviation for AC-flow and MFDR values for both the IBIF method and the Kalman filter. It is natural for large variations in vocal intensity to occur during the week for all subjects. We can see larger differences between KF and IBIF mean MFDR values, while mean values of H1-H2 are similar between KF and IBIF. This is due to the MFDR extraction, which is estimated from the derivative of both signals, and that might introduce extra differences between both the KF based method and the IBIF filter.

We compare classification results from Machine Learning simulations considering 4 different tasks/scenarios: 1) Using all the data frames available. 2) Randomly selecting 50% of the data available. 3) Selecting the best 10% data frames according to the lowest 10% RMSE between IBIF and KF waveforms, and 4) Selecting the worst 10% data frames according to the highest 10% RMSE between IBIF and KF waveforms. Examples of signals with error less than 10% of total RMSE and higher than 90% are shown in figures 4.18 and 4.19, respectively. The purpose of the comparison between different groups of frames is to identify any differences in classification performance if the Kalman filter is used to extract aerodynamic features. There are some differences in the morphology of the GVV signal when Kalman and IBIF is used. For example, by visual inspection, phase differences between peaks contribute to higher deviations from the IBIF output in Fig 4.19.

An ensemble of bagged decision trees (Random Forest [109]) was used as a

classifier. There are two hyperparameters: the number of trees and the maximal number of decision splits per node, which are optimized using Bayesian Optimization [171]. 25% of the total data was used (4 pair of subjects) as a test set. Table 4.3 shows the results of the area under the ROC curve (AUC), the accuracy, the f-score, the sensitivity, the specificity, the positive predicted value (PPV), and the negative predicted value (NPV). The similarity in between results might indicate that deviations of the estimated GVV signal using IBIF would not affect the detection performance for an ambulatory classification task. This implies that, in an ambulatory scenario, it is possible to have room for some error estimation in the GVV signal without compromising the task at hand.

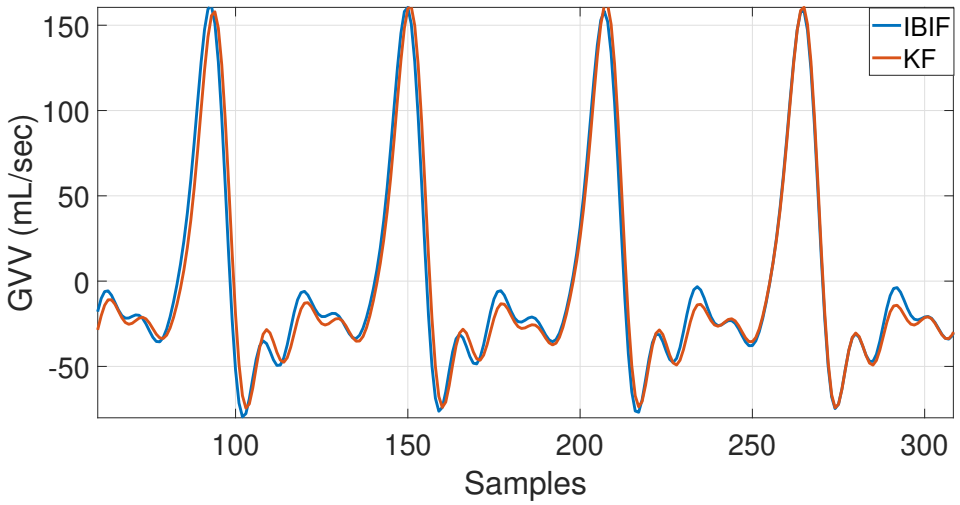

Figure 4.18: GVV with IBIF (blue) and Kalman (red) from In Field data. RMSE is less than 10% of the total day RMSE.

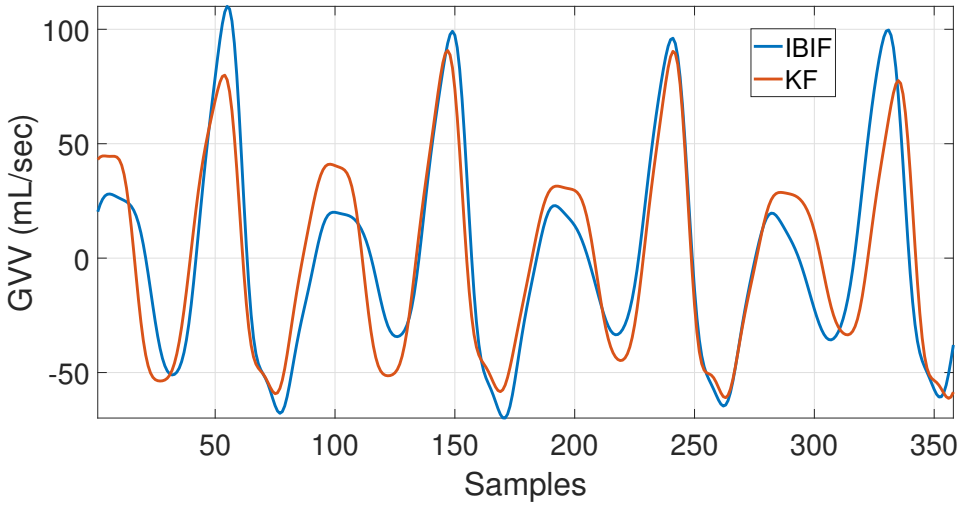

Figure 4.19: GVV with IBIF (blue) and Kalman (red) from In Field data. RMSE is more than 90% of the total day RMSE.

Subject	AC-flow (IBIF)	AC-flow (KF)	MFDR (IBIF)	MFDR (KF)	OQ (IBIF)	OQ (KF)	H1-H2 (IBIF)	H1-H2 (KF)
PF022	493.9 (359.6)	469.1 (315.1)	664.0 (488.8)	859.0 (696.0)	62.7 (19.7)	76.9 (28.1)	3.47 (7.73)	4.24 (7.14)
NF022	239.7 (231.4)	257.3 (237.8)	339.0 (385.6)	465.1 (555.9)	60.2 (20.0)	65.7 (22.0)	4.28 (7.04)	4.58 (6.79)
PF031	518.2 (331.2)	434.3 (269.8)	466.0 (387.2)	526.5 (520.3)	78.8 (18.5)	80.0 (20.9)	10.6 (6.41)	8.67 (5.99)
NF031	307.9 (236.6)	299.3 (247.2)	380.8 (416.3)	454.3 (567.4)	67.6 (16.5)	67.8 (17.8)	12.8 (6.16)	12.4 (6.14)
PF038	235.3 (115.3)	246.8 (127.3)	309.4 (190.0)	425.7 (306.3)	54.8 (11.7)	52.4 (12.4)	10.7 (4.90)	9.45 (4.89)
NF038	199.0 (142.1)	213.7 (153.9)	315.3 (271.6)	390.3 (366.6)	53.6 (20.2)	55.3 (21.3)	7.09 (5.23)	7.28 (5.07)
PF056	436.3 (259.2)	342.6 (202.2)	616.6 (391.9)	669.6 (502.1)	77.8 (12.7)	75.4 (14.4)	8.08 (4.19)	7.86 (3.87)
NF056	327.5 (165.7)	313.6 (154.9)	372.3 (228.5)	499.5 (336.6)	68.6 (12.6)	69.4 (13.9)	10.5 (4.87)	9.65 (4.72)
PF063	262.3 (130.8)	281.1 (147.3)	253.3 (154.4)	372.2 (301.5)	66.3 (13.1)	64.3 (16.9)	14.1 (5.81)	12.7 (5.91)
NF063	361.8 (172.5)	255.8 (147.4)	464.2 (260.9)	473.2 (342.5)	59.7 (11.6)	56.8 (14.2)	11.6 (5.17)	7.60 (5.90)
PF065	167.9 (91.9)	199.8 (117.1)	270.0 (184.3)	406.2 (319.0)	63.3 (14.2)	66.3 (14.8)	10.4 (4.33)	10.2 (4.25)
NF065	210.4 (128.7)	204.5 (121.2)	353.0 (222.3)	430.5 (274.1)	53.3 (17.2)	55.3 (17.6)	6.15 (6.28)	5.79 (5.95)
PF066	318.5 (188.7)	321.1 (195.0)	373.0 (355.6)	434.9 (451.5)	53.8 (15.2)	53.9 (18.8)	8.73 (3.82)	9.06 (3.68)
NF066	367.5 (236.2)	365.5 (228.3)	557.5 (398.9)	713.9 (521.4)	50.0 (17.6)	51.1 (19.2)	7.25 (5.54)	7.63 (5.15)
PF070	774.6 (460.9)	669.7 (415.9)	991.1 (654.9)	1164.3 (966.5)	75.4 (11.8)	75.5 (12.9)	10.8 (4.02)	9.64 (3.97)
NF070	323.8 (245.5)	321.8 (248.7)	476.8 (403.1)	529.7 (519.8)	74.5 (17.9)	76.2 (20.0)	8.08 (6.93)	7.82 (6.73)
PF079	280.8 (173.2)	314.2 (198.5)	415.8 (309.2)	510.1 (384.4)	48.2 (15.4)	48.2 (16.1)	7.82 (4.56)	7.26 (4.34)
NF079	150.5 (116.0)	159.4 (119.1)	239.1 (218.9)	309.5 (323.3)	64.0 (19.1)	66.7 (18.5)	8.11 (6.61)	8.69 (6.14)
PF080	492.3 (217.1)	473.3 (214.0)	744.6 (341.0)	1050.1 (599.4)	72.4 (14.6)	71.1 (15.5)	8.32 (4.73)	8.17 (4.53)
NF080	393.9 (248.2)	318.1 (208.6)	559.8 (385.7)	604.2 (467.1)	62.0 (15.6)	63.5 (17.0)	9.58 (4.77)	9.23 (4.66)
PF087	560.9 (224.0)	581.4 (246.5)	939.6 (483.7)	1174.5 (668.1)	76.3 (11.5)	85.6 (14.5)	9.81 (3.68)	9.18 (3.45)
NF087	403.8 (273.2)	435.1 (311.1)	697.5 (573.4)	914.1 (789.9)	50.0 (17.8)	60.9 (23.9)	0.00 (5.10)	0.00 (4.59)
PF090	175.7 (117.7)	142.9 (96.3)	241.2 (181.6)	239.3 (199.2)	72.7 (12.1)	72.5 (25.6)	12.8 (4.98)	12.5 (4.90)
NF090	115.2 (87.9)	100.4 (77.5)	124.7 (142.8)	137.8 (188.8)	61.5 (12.5)	59.8 (12.9)	12.2 (6.10)	11.5 (5.95)
PF091	249.7 (143.6)	242.8 (152.7)	261.6 (230.7)	370.8 (391.5)	70.8 (15.3)	70.3 (19.7)	8.44 (4.90)	7.88 (4.72)
NF091	291.1 (245.3)	274.2 (235.4)	368.1 (330.9)	400.7 (424.7)	62.7 (18.9)	62.5 (20.2)	9.18 (6.46)	8.75 (6.25)
PF094	206.5 (335.5)	143.4 (240.6)	291.5 (622.8)	230.4 (544.3)	80.0 (21.6)	87.5 (24.9)	7.59 (6.17)	6.93 (5.59)
NF094	563.7 (332.6)	469.0 (309.3)	560.1 (484.8)	609.9 (619.3)	59.4 (16.5)	55.8 (21.1)	12.1 (4.88)	10.5 (4.90)
PF098	203.8 (108.4)	201.7 (111.4)	246.3 (163.3)	263.2 (202.6)	70.1 (14.3)	71.6 (15.0)	12.2 (4.88)	12.0 (4.39)
NF098	243.6 (93.1)	235.7 (91.6)	279.5 (131.2)	331.8 (183.6)	60.0 (10.7)	60.0 (11.2)	12.4 (4.33)	12.2 (4.30)
PF126	286.4 (170.2)	243.6 (165.9)	338.3 (284.4)	395.4 (461.1)	63.6 (16.3)	62.2 (19.7)	12.2 (6.05)	10.9 (6.05)
NF126	262.2 (174.2)	252.9 (175.8)	367.6 (270.5)	420.7 (332.9)	71.8 (13.1)	71.8 (14.0)	12.7 (5.50)	12.3 (5.45)

Table 4.2: Weekly summary (mean and standard deviation) of features using IBIF and Kalman Filter (per subject).

Percentage of frames used	AUC	Accuracy	F-score	Sensitivity	Specificity	PPV	NPV
All frames	0.80	0.72	0.74	0.74	0.68	0.74	0.69
50 % random frames	0.80	0.71	0.74	0.74	0.68	0.74	0.68
10% lowest error	0.78	0.70	0.73	0.74	0.66	0.73	0.67
10 % highest error	0.78	0.70	0.73	0.73	0.67	0.73	0.67

Table 4.3: Classification performance of Ensemble bagged trees (Random Forest) using different portions of the data with aerodynamic features.

4.6 Discussion

The proposed method based on Kalman filter is an adaptive implementation of the IBIF scheme, and therefore has some differences with the original IBIF design, namely a forward prediction of the accelerometer signal (i.e., no filter is inverted) and a truncation of the finite impulse response. In spite of these differences, it is shown that the Kalman filter implementation allows for enhancing the glottal airflow estimates, as it adapts to better predict the accelerometer signal and to more closely resemble the glottal airflow estimates from a Rothenberg mask in benchmark experiments. It is important to note that there are still differences between the Kalman filter glottal airflow estimates and the reference signal from the Rothenberg mask, due to supraglottal inverse filtering errors and measurement uncertainty of the oral airflow signal, which are difficult to assess for the scenario of high-pitch female pathological voices.

The signal deviations between the Kalman filter and the original (time invariant) FIR IBIF glottal airflow estimates are relatively small, although the former reduces the RMSE in up to 40% in some cases. These differences can be relevant in some cases, depending on the application. When assessing the relevance of these differences in a classification task to discriminate between vocal fold nodules patients and control subjects using ambulatory accelerometer data, no significant variations in the classification were found, even when comparing frames with low and high error (or deviation). Thus, the classification task seems to be fairly insensitive to the uncertainty of the airflow estimates from IBIF model parameters, sensor positioning, and other effects. This supports the use of the original FIR version of the IBIF scheme for such classification tasks, which indicates that factors affecting the classification performance in chapter 3 were not degraded by the airflow estimates. However, other applications more sensitive to signal quality can further benefit from the enhancement offered by the proposed Kalman implementation to estimate more accurate glottal airflow in running speech and/or ambulatory scenarios.

The main current limitation of the proposed Kalman filter approach is its relatively high computational cost due to its FIR-inspired construction, which can become a problem when processing many hours of recordings (as in ambulatory monitoring) in numerous subjects. Future efforts can be devoted to optimize the approach via more efficient methods, using for example an autoregressive model

in the construction of the state space model. Other variations in the construction, e.g., addition of a random walk term or an extended Kalman filter could be explored as well to encompass non-linear implementations of the neck accelerometer to glottal airflow signal transformation.

4.7 Conclusion

A Kalman filter implementation of the subglottal impedance based inverse filtering scheme was introduced to enhance the estimates of glottal airflow from recordings of a neck surface acceleration signal and to assess the relevance of model uncertainty in such estimates. The approach is capable of adapting the signal estimates to correct for inverse filtering errors, as observed in benchmark experiments with sustained vowels. To explore the relevance of the signal deviation in the context of ambulatory study, accelerometer data from 32 subjects during 1 week-long recordings was utilized in the context of a classification task used in prior studies to discriminate normal from pathological voices. When comparing the performance using the standard and Kalman IBIF implementations, no strong differences in the classification performance were observed. Therefore, hypothesis H2 is rejected due to the insufficient evidence to show that reducing the variability of the IBIF filter would improve the classification performance between PVH and control subjects in an ambulatory setting. It is concluded that the IBIF filter has

good performance in the task of classification, especially with high-order statistics, even if there are error on the measurements. The high amount of data available from weekly recordings allows that the features obtained from the estimated GVV signal are robust enough to construct machine learning models with parameters that do not deviate enough to change performance scores drastically. With respect to the MA Kalman filter, although is more computationally expensive than the FIR IBIF, there is an opportunity to improve the design of an ARMA filter that estimates the GVV signal through a Bayesian framework. Further improvements on the design of a parametric glottal flow input could aid on the development of a physiological relevant filter that do not require as many states as the MA Kalman filter, with as good, or better performance on estimation of features relevant to the voice function.

Chapter 5

Additional tools on supervised classification and uncertainty analysis from accelerometer voice data

The following set of experiments are aimed to understand better the classification performance using different data sets from chapter 3 and to quantify uncertainty with a different approach from chapter 4. The goal is to answer aims SA1 and SA2 with pilot projects that could take a large scale (i.e., larger datasets) and at the same time, to answer hypothesis H1 and H2. Section 5.1 develops on the supervised classification task of ambulatory data from NPVH subjects. Section 5.2 quantifies the uncertainty of IBIF parameters using In Lab accelerometer data. Section 5.3 develops a framework for classification of NPVH therapy by using In Lab accelerometer data and deep learning. Results and conclusions, including their connections to the aims and hypothesis of this thesis, are included at the end of the chapter and each section.

5.1 Supervised classification of subjects with muscle tension dysphonia

Primary muscle tension dysphonia, or non-phonotraumatic vocal hyperfunction, refers to a speech disorder characterized by variable symptoms of voice disruption without the presence of a laryngeal pathology [18]. It is usually associated to high levels of steady (DC) flow [12] but not to high levels of unsteady AC-flow or MFDR [12, 27] which could explain the lack of organic lesions for this type of pathology. This type of dysphonia is referred as Primary MTD, which arises from apparent excessive user or misuse from hyperfunctional patterns in the absence of an organic vocal fold pathology, psychogenic, or neurological etiology [33].

The work described in this section of the chapter aims to classify healthy subjects from MTD subjects, similar to the task developed in chapter 3. In this case, in addition to the study of aerodynamic features, features related to speech processing are analyzed as well, to investigate further those features that are salient for this type of pathology. Due to the lack of enough components that could derive physiological features related to the pathology, there is no strong prior information on which voiced features could be important on detecting the characteristics of NPVH, especially in an ambulatory context.

5.1.1 Methods

Ten subjects with NPVH and 10 matched controls were recruited for a week of ambulatory monitoring. Occupation and CAPE-V scores of participants are shown in Table 5.1. Following the same methodology from chapter 3, ambulatory features were extracted for each 50 ms frames of non-overlapping accelerometer data. Statistical features were obtained for each subject during an average of 7 days of voicing data. IBIF features were extracted from the same subjects, with the exception of the pair whose occupation are teachers, due to difficulty of obtaining an ambulatory aerodynamic signal from the control subject. Therefore, the classification learning with IBIF features was applied to 9 pairs.

Accelerometer features were those features from which IBIF is not applied beforehand. These are taken directly from the calibrated accelerometer and are estimated from the time and spectral domains. Table 5.2 lists the accelerometer features used for ambulatory analysis. Sound pressure level is obtained using linear regression with respect to the skin acceleration signal from a calibration procedure [130]. The fundamental frequency of the skin-acceleration signal is directly correlated to the fundamental frequency of the vocal fold vibrating cycle. The cepstral peak prominence is the magnitude of the highest peak in the power cepstrum [172]. The zero crossing rate corresponds to the proportion of the number

Table 5.1: Occupations and mean age of adult females with PVH and matched-control participants analyzed (48 pairs)

Occupation	Age	Diagnosis	CAPE-V overall
Non-Classical Singer	19 (P) 18 (C)	MTD	11
	22 (P) 20 (C)	MTD	16
Teacher	48 (P) 51 (C)	MTD	60
Actress	59 (P) 60 (C)	MTD	5
Pilates Instructor	26 (P) 22 (C)	MTD	17
Nurse	59 (P) 51 (C)	MTD	28
Social Worker	54 (P) 50 (C)	MTD	19
Administrative Assistant	39 (P) 40 (C)	MTD	15
System Analyst	44 (P) 47 (C)	MTD	15
Home School Teacher	43 (P) 42 (C)	MTD	16

of times that the signal crosses its mean within a frame. H1-H2 is the magnitude difference between the first and second harmonic [173]. The harmonic spectral tilt is a linear regression of the first 8 spectral harmonics [174]. Similar to the harmonic spectral tilt, the low-to-high spectral ratio is the difference of spectral power below and above 2000 Hz [175]. The skin-acceleration level is the RMS of the accelerometer signal in dB per frame. The autocorrelation peak amplitude is the relative amplitude of the first non-zero peak in the normalized autocorrelation function [23].

Accelerometer Features	**Voicing criteria**	**Units**
Sound Pressure Level (SPL)	45-130	dB
Fundamental Frequency (f_0)	70-1000	Hz
Cepstral Peak Prominence (CPP)	10-35	dB
Zero Crossing Rate (ZCR)	0-1	–
H1-H2	-60-50	dB
Harmonic Spectral Tilt	-25-0	$dB/octave$
Low-High Spectral ratio	22-50	dB
Skin Acceleration Signal (SAL)	5-200	dB
Autocorrelation Peak Amplitude (NPeak)	0.60-1	dB

Table 5.2: Features estimated from the accelerometer.

Fig 5.1 and Fig 5.2 show the relationship between distributions of 5 IBIF and accelerometer features, respectively, from a given week between NPVH and

control subjects. The features were sorted by lowest p-value in a two-sample t-test with Bonferroni correction [176]. Therefore, the 95th percentile of MFDR (in dB) and the mean value of skin-acceleration Level (SAL) were the features that most separated NPVH subjects from controls, in the IBIF and accelerometer pool of features, respectively.

It is worth to notice that the top 5 IBIF features are derivations from ACFL and MFDR, which in turn are related to PVH behavior [1, 12, 27]. Previous results on aerodynamic measures using NPVH and controls showed significant differences only on subglottal pressure and open quotient [27]. One reason that OQ was not a salient feature for NPVH could be due to he estimation procedure, which is particularly difficult in ambulatory settings because it is necessary to know the exact closing time. Moreover, subglottal pressure is not estimated for this ambulatory task. Therefore, measures such as MFDR and ACFL seem to provide new insights on NPVH subjects during ambulatory monitoring. The fact that MLR features (SPL divided by MFDR) are included, indicate some type of compensation from NPVH subjects to keep a certain SPL levels, as the histograms of median and 95th percentile of MLR show an overall lower value for NPVH compared to controls, indicating that higher MFDR is needed to keep a certain SPL value [60].

The top 5 features from the accelerometer signal are mixed between intensity-related (SAL mean), spectral-related (H1-H2 and LH Ratio), and temporal-related

(zero-crossing rate).

In the task of machine learning, five different classifiers were used to identify 5-minute frames from MTD subjects versus controls. In addition of using a

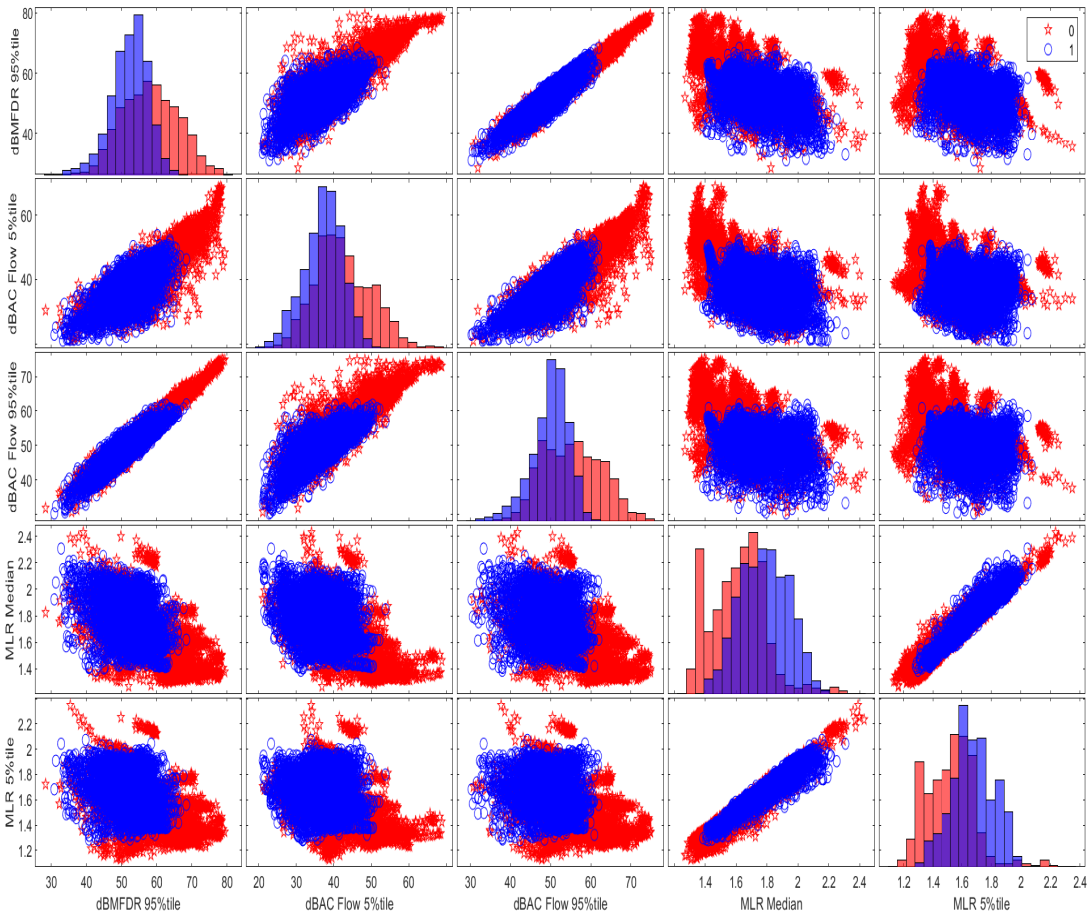

Figure 5.1: Distribution of 5 IBIF features for patients (red) and controls (blue). The features were selected based on a two-sample t-test procedure, where the features with lowest p-values, based on a Bonferroni correction, are displayed.

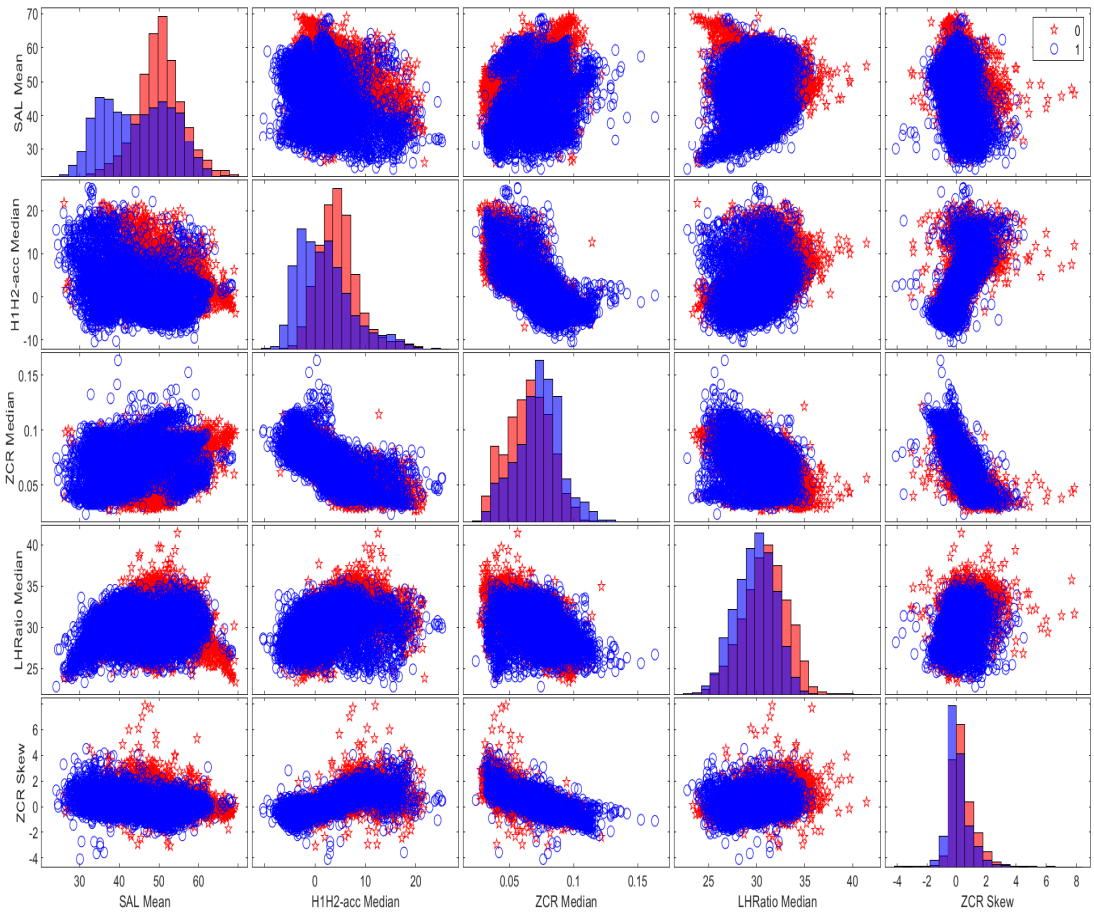

Figure 5.2: Distribution of 5 accelerometer features for patients (red) and controls (blue). The features are selected based on a two-sample t-test procedure where the features with lowest p-values, based on a Bonferroni correction, are displayed.

L1- Logistic Regression and Linear Kernel SVM, a SVM with Gaussian kernel, a Random Forest, and a Feedforward Neural Network were tested as well. Hyperparameters for the Gaussian SVM were found through Bayesian optimization [171]. Hyperparameters for Random Forest were found using 5 fold CV, while the number of nodes and layers of the neural network were chosen after trial of different configurations that would not overfit the data.

5.1.2 Results

Table 5.3 show results of the classifiers using IBIF features only. The mean performance is worse than the mean results for PVH subjects using the same IBIF features. The standard deviation is larger for NPVH subject pairs, indicating that some test subjects performed very well with the trained model, and others performed extremely poor with the trained model. These extreme cases explain the mean performance on all scores, as indicated in Fig 5.3.

The classification map in Fig 5.4 reflects the classification zones of both NPVH and control subjects, where those who are closer to 1 are more associated to NPVH, and those closer to 0 are more associated to the control group.

The odds ratio for features that have high association to NPVH (statistically significant effect of $p < 0.001$ for absolute odds ratio ≥ 1.10 in every pair) is shown in Fig 5.5. The skewness of the NAQ feature has on average a large odds ratio,

Method	AUC	Accuracy	F-score	Sensitivity	Specificity	PPV	NPV
LR-L1	0.55 (0.43)	0.54 (0.37)	0.51 (0.39)	0.50 (0.39)	0.59 (0.35)	0.53 (0.39)	0.56 (0.35)
SVM-Linear Kernel	0.53 (0.40)	0.69 (0.17)	0.59 (0.36)	0.68 (0.40)	0.69 (0.35)	0.55 (0.35)	0.68 (0.30)
SVM-RBF Kernel	0.61 (0.14)	0.61 (0.24)	0.62 (0.09)	0.70 (0.32)	0.47 (0.39)	0.56 (0.22)	0.46 (0.33)
Random Forest	0.62 (0.38)	0.72 (0.20)	0.57 (0.43)	0.62 (0.47)	0.78 (0.36)	0.65 (0.40)	0.68 (0.31)
Neural network	0.52 (0.34)	0.66 (0.12)	0.48 (0.37)	0.52 (0.42)	0.74 (0.33)	0.55 (0.33)	0.61 (0.27)

Table 5.3: Classification performance for 9 NPVH and 9 matched control pairs using IBIF features

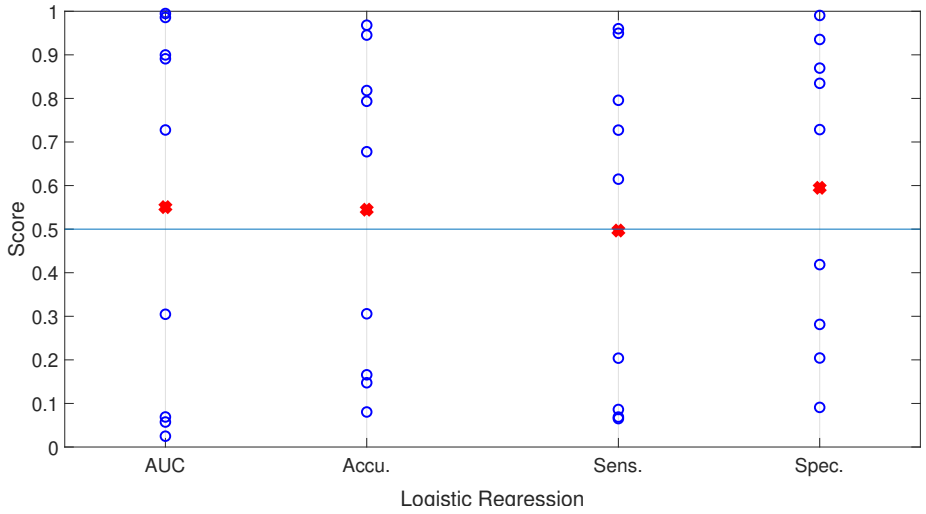

Figure 5.3: Logistic regression classification scores per pair (NPVH vs control) using IBIF features

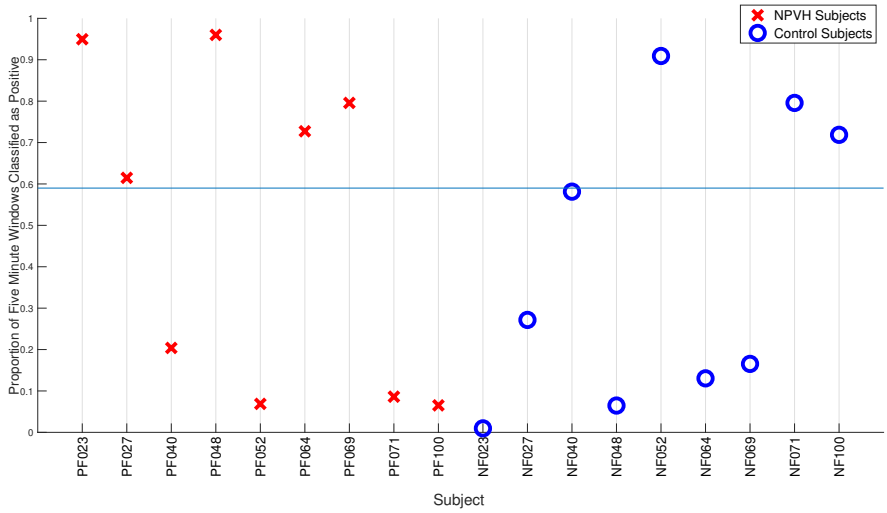

Figure 5.4: Classification map of 18 subjects (9 NPVH, 9 controls) during 1 week of ambulatory data using Logistic Regression.

however, it has a large deviation across subjects. The 5th percentile of MLR has lower deviation, indicating that MFDR has a strong association to NPVH subjects. H1-H2 has also a consistent association to NPVH, which is related to quality of voice [55].

The classification performance improves slightly when features from the accelerometer are used, as can be seen from Fig 5.6. However, it is difficult to interpret the features that are more relevant in NPVH with a physiological process. For example, in Fig 5.7 the features that are associated with NPVH could have multiple interpretations but since none are directly related to an amplitude-based aerodynamic process, it is a bit more difficult to interpret the results. The largest associated feature is the mean of the zero-crossing rates, which indicates

less periodicity in the signal. The skewness of CPP and the standard deviation of the spectral tilt are related to voice quality, which is at the same time related to spectral characteristics of the glottal flow. These last features can be comparable to features extracted to the ground-truth glottal flow, as there are positive correlation between the neck-skin acceleration and glottal flow signal for those features [177].

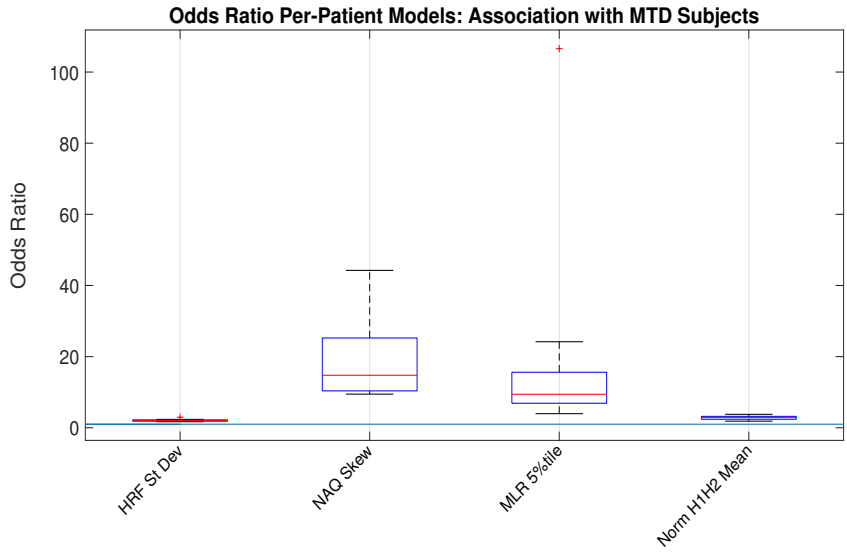

Figure 5.5: Features with odds ratio greater than 1 for all 9 pair of subjects.

Method	AUC	Accuracy	F-score	Sensitivity	Specificity	PPV	NPV
LR-L1	0.59 (0.27)	0.65 (0.25)	0.68 (0.12)	0.71 (0.30)	0.59 (0.35)	0.62 (0.24)	0.63 (0.25)
SVM-Linear Kernel	0.60 (0.26)	0.65 (0.25)	0.68 (0.12)	0.73 (0.31)	0.57 (0.35)	0.61 (0.24)	0.76 (0.15)
SVM-RBF Kernel	0.61 (0.14)	0.61 (0.24)	0.62 (0.09)	0.70 (0.32)	0.47 (0.39)	0.56 (0.22)	0.46 (0.33)
Random Forest	0.58 (0.17)	0.59 (0.23)	0.63 (0.09)	0.68 (0.35)	0.49 (0.41)	0.67 (0.13)	0.57 (0.32)
Neural network	0.64 (0.16)	0.66 (0.23)	0.66 (0.11)	0.77 (0.30)	0.49 (0.38)	0.68 (0.13)	0.71 (0.13)

Table 5.4: Classification performance for 10 NPVH and 10 matched control pairs using accelerometer features

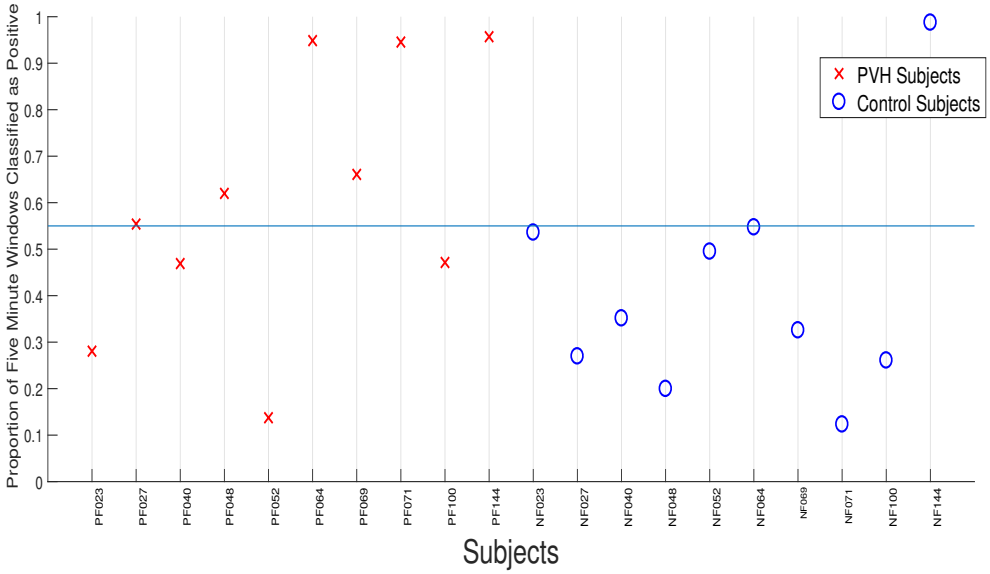

Figure 5.6: Classification map of 20 subjects (10 NPVH, 10 controls) during 1 week of ambulatory data using Logistic Regression.

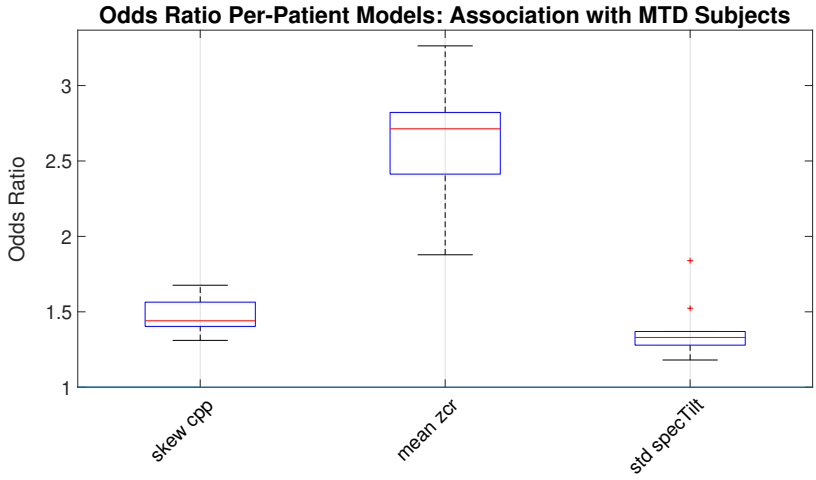

Figure 5.7: Features with odds ratio greater than 1 for all 10 pair of subjects.

5.1.3 Discussion

NPVH subjects suffer from excess laryngeal tension without other primary etiology that causes this dysfunction. Therefore, the causes of NPVH are difficult to assess without an organic or neurological symptom from which the pathology can be classified. The set of experiments shown in this section improves on the idea that aerodynamic forces might not be a strong indicator of the presence of NPVH. On average, classification scores between NPVH and control is almost random due to some of pair of subjects were correctly classified and others were not. An interpretation of these results indicate that aerodynamic features are highly indicative of NPVH, as indicative of the high AC-flow and MFDR values (normalized by SPL) compared to some controls. However, there are cases of pairs where the results are backwards: the NPVH subject behaves as the control (from the training data) and the control behaves as the NPVH. In the assumption that controls do not have aerodynamic alterations in their vocal behavior, this indicate that aerodynamic features are poor indicators of NPVH for these subjects.

5.2 Classification performance of paired subjects with vocal hyperfunction in the presence of inverse filtering uncertainties: Pilot Study

The impedance-based inverse filter in [85] is subject-specific due to the filter parameters that are related to default neck-skin mechanical properties such as resistance, inertance, and stiffness [119], as well as trachea length [178, 179] and sensor position. The main assumption is that these parameters are constant for each individual, even though there are anatomical differences that could affect parameter values, such as skin fat and changes in head position, or other components in the system such as tracheal diameter and losses in the subglottal system [85]. Even though IBIF filter parameters are calibrated for each subject through the Q scaling factors, these are assumed to be constant for ambulatory tasks, which might not be the case due to changes on the assumptions mentioned before. For instance, there is evidence that the Q parameters are not purely constant, but they behave as random variables when they are calculated from different vowels as reference [150].

The experiment in this section of the chapter aims to quantify the effect of uncertainty on the Q parameters for a simple classification task between Spanish speakers with healthy voice and PVH. The Q parameters will be assumed to be random variables from Gamma distributions [176] because Q parameters are pos-

itive (random variables from Gamma distributions are positive), the similarity to Gaussian distributions, and the possibility to parameterize long tails to the right for Q values that are much higher than its mode. Glottal airflow estimations will be obtained with different realizations of the Q parameters using Montecarlo simulation. Features from Table 3.2: ACFL, MFDR, OQ, H1-H2 and HRF will be used as input to a L1-logistic regression model. The classification scores will provide histograms from which distributions can be inferred. Uncertainty measures will be obtained in order to compare how the deviation of Q parameters affect the deviation of classification scores.

5.2.1 Methods

In this experiment, 8 female subjects participated in lab and ambulatory sessions, where 4 were diagnosed with PVH nodules and 4 were healthy-matched controls. Aerodynamic measurements were taken in the Voice Production Laboratory at Universidad Técnico Federico Santa María. Following a very similar procedure from chapter 3, the subjects phonated different token vowels with different intensity, as well as reading a Spanish passage called "El Abuelo" [180] that is phonetically balanced. Nominal Q parameters were obtained from a calibration procedure using the /a/ and /i/ vowels under different loudness conditions. Most of the subjects had similar Q values when the calibration was replicated under the mentioned conditions, therefore, they were kept fixed. Other subjects had different

Q values for different conditions, for which the mean and variance were estimated.

Following the procedure from [150], a Gamma distribution is used to extract random variables from the first 3 Q values:

$$Ga(T|shape = a, rate = b) \triangleq \frac{b^a}{\Gamma(a)} T^{a-1} e^{-Tb}, \tag{5.1}$$

where $\Gamma(a)$ is the gamma function:

$$\Gamma(x) \triangleq \int_0^\infty u^{x-1} e^{-u} du. \tag{5.2}$$

The parameters $a > 0$ and $b > 0$ are the shape and rate of the distribution, respectively, while T is the random variable. The mean of the distribution is $\frac{a}{b}$ and its variance is $\frac{a}{b^2}$. In order to fit a Gamma distribution for each Q parameter, we need to obtain the parameters a and b. These can be obtained using the mean and variance of all measures of Q's for each subject. For those subjects with constant Q's, the mean is determined by that same value, while the variance is fixed as 10% of the mean. Otherwise, the sample mean and variance is obtained for those subjects with different Q values.

Fig 5.8 shows an example of three Gamma distributions with mean 1 and variance 0.1 (blue), mean 2 and variance 0.2 (red), and mean 3, variance 0.3

(black). The distributions are very similar to Gaussian distributions, however, Gamma distributions range from 0 to infinity, which are the plausible values for Q parameters, being zero or very large values quite unlikely. The variance increases as the mean increases, indicating that the uncertainty is larger when the Q value is large as well.

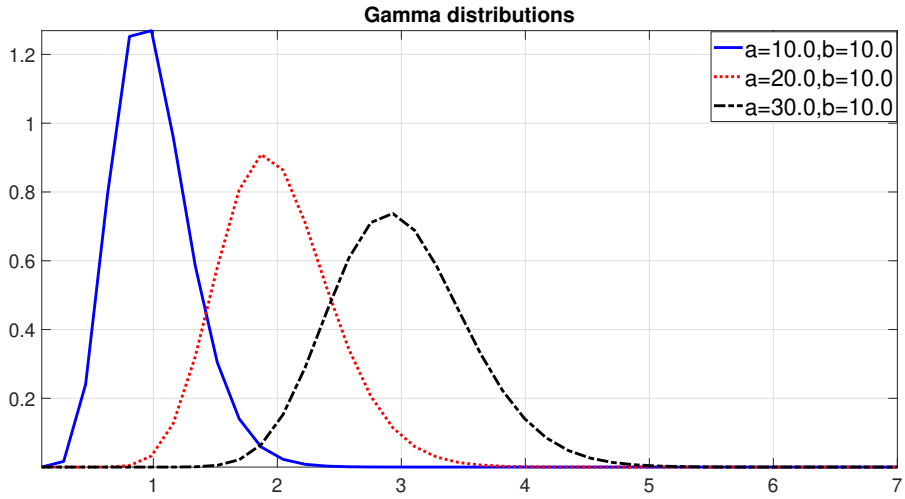

Figure 5.8: Example of Gamma distributions for different parameters a and b

The first three Q parameters (Q_1, related to skin resistance, Q_2, related to skin inertance, and Q_3, related to skin stiffness) are modeled with Gamma distribution. Q_4, related to trachea length, and Q_5, related to accelerometer position, are set to 1 (default values). Therefore, the trachea length and accelerometer position are fixed to 10 cm. and 5 cm., respectively [85]. Since deviations from the default values do not perturb the IBIF filter as much as the first three Q parameters [150], Q_4 and Q_5 are kept fixed for simplicity on the simulations.

Once a probability density function (pdf) is defined for each Q_1, Q_2, Q_3, for each subject, a Montecarlo simulation is performed to extract 1000 random values from each pdf. Therefore, 1000 IBIF filters are created randomly. Fig 5.9 shows an example of a subject pair with randomized values of Q_1 and Q_2. In this case, the mean values of Q_1 and Q_2 for the PVH subject (red) are lower than the mean values of the control subject (blue). Therefore, the variance is lower for the PVH subject, as the randomized Q values are more concentrated around their means, compared to the control subject.

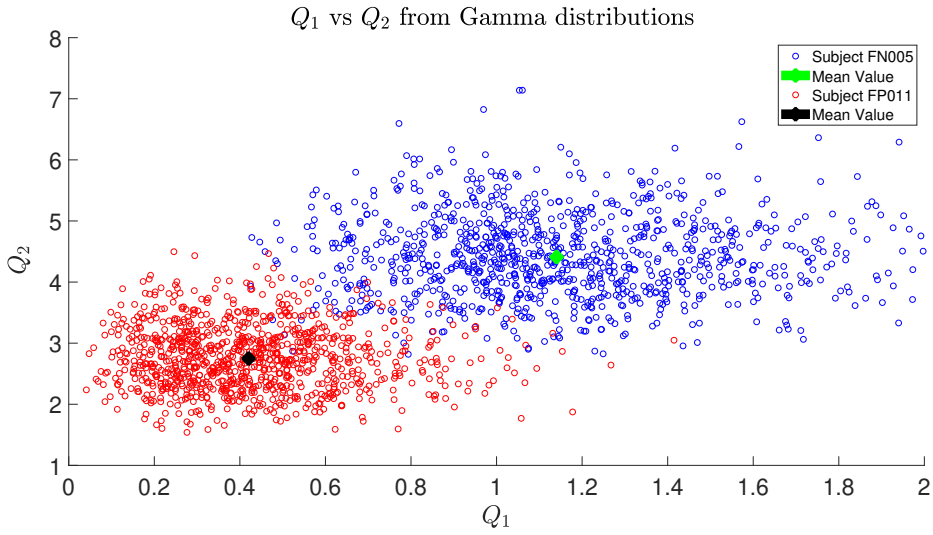

Figure 5.9: Example of the sample distribution of IBIF parameters Q_1 and Q_2 for a healthy subject (blue) and a PVH subject (red)

5.2.2 Results

A Logistic regression model with L1 regularization was applied with the same procedure as in chapter 3: Leave one pair out for testing, train on the rest of the data. Performance scores (median and standard deviations) are shown in Table 5.5 for Area under the curve (AUC), accuracy and F1-score.

Subjects	**AUC**	**Accuracy**	**F1-score**
Pair 1	0.67 (0.19)	0.64 (0.20)	0.61 (0.13)
Pair 2	0.83 (0.20)	0.75 (0.16)	0.73 (0.15)
Pair 3	0.72 (0.19)	0.50 (0.23)	0.57 (0.14)
Pair 4	0.69 (0.14)	0.40 (0.15)	0.58 (0.08)

Table 5.5: Classification performance for 4 pairs by median and (standard deviation) of 1000 simulations using LR-L1 and leave-one-pair-out method of training and testing.

AUC scores are fitted to a Beta distribution [181] because these have the support over the interval $[0, 1]$, exactly the same support of AUC scores. The Beta distribution is defined as follows:

$$Beta(x|a, b) \triangleq \frac{1}{B(a, b)} x^{a-1}(1 - x)^{b-1} \tag{5.3}$$

where $B(p, q)$ is the beta function:

$$B(a,b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{5.4}$$

Using the data from histograms, we can obtain values for the parameters a and b through maximum likelihood estimation (MLE). All parameters estimated for each subject pair data are above 1, which implies an unimodal distribution [181]. Moreover, we can obtain the standard deviation of the distribution, sd_{beta}, using MLE. The mode of the beta distribution is defined as:

$$mode_{beta} \triangleq \frac{a-1}{a+b-2} \tag{5.5}$$

We define the one-sided relative error (R.E.) of the distribution with respect to the mode as:

$$R.E._{\%} = \frac{sd_{beta}}{2 * mode_{beta}} * 100\% \tag{5.6}$$

A boxplot with AUC and F1-scores, with their respective R.E. for each pair is shown Figs 5.10 and 5.11:

There are a few observations for discussion from the results and boxplots. There are extreme cases where the AUC score is less than 0.5 (worse than random guessing), due to some random combinations of Q parameters that do not provide a correct GVV signal, therefore, the features obtained are not likely to

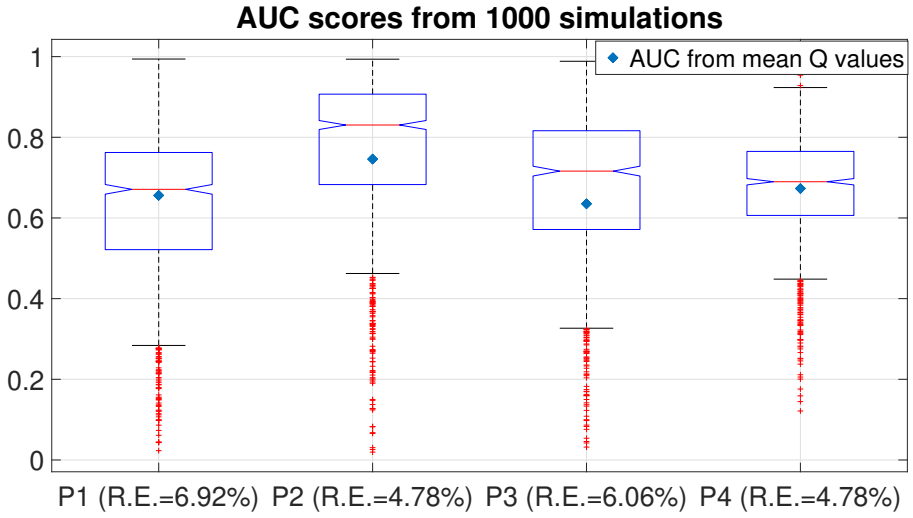

Figure 5.10: AUC boxplots for the 4 pairs analyzed using 1000 random classifications. Blue points indicate the original Q parameters values that correspond to the mean values on the parametric distributions.

be correct. Most of the concentration of AUC results is around 0.6-0.7 range, which is expected from previous work [11]. It is worth noting that, for every pair, the resulting AUC and F1-score from using the mean Q parameters are below the median of the simulations. Given that the mean Q parameters are expected to perform best due to a good estimation of the glottal flow, the differentiation between PVH and controls should be strong. However, the classification performance is not directly related to the estimation of the features using the mean Q parameters. This in an indication that aerodynamic features only estimated from glottal flow might not capture all differences between PVH and control subjects, even if their physical (Q) parameters are well calibrated. Fig 5.12 shows the AUC

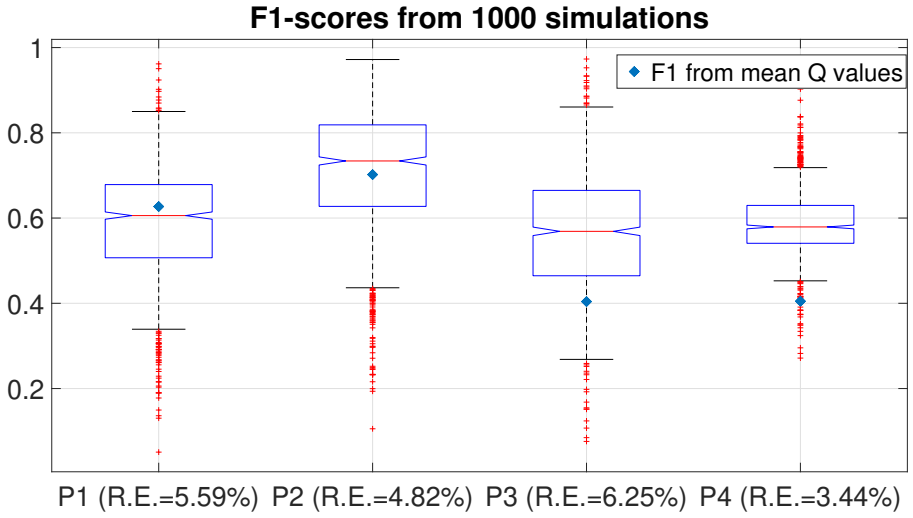

Figure 5.11: F1-scores boxplots for the 4 pairs analyzed using 1000 random classifications. Blue points indicate the original Q parameters values that correspond to the mean values on the parametric distributions.

histogram for a given pair of subjects. A kernel density estimated distribution (KDE) smooths the histogram [176], while the beta distribution is also plotted in the graph. Visually, KDE fits very well to the data, while the beta distribution is wider at the left of the histogram, but fits very close to the histogram and KDE, having almost the same peak. The vertical green dotted line indicates the AUC value using the mean Q parameters. Notice that this score is slightly less than the peak values of the beta and KDE distribution

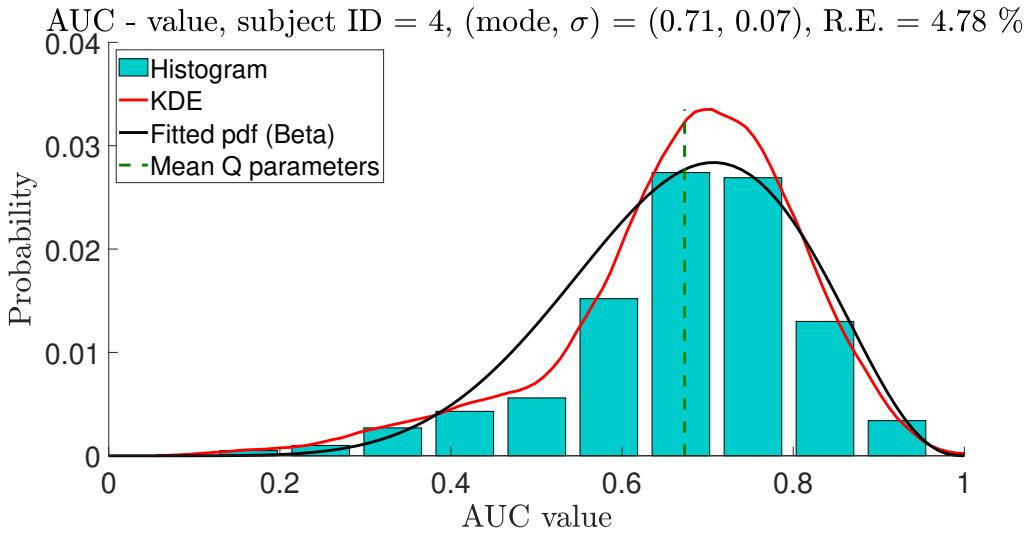

Figure 5.12: AUC histogram for pair number 2. The red curve is the kernel density estimated (KDE) distribution, while the black curve is the fitted beta distribution.

5.2.3 Conclusions

In conclusion, the purpose for the set of experiments shown in this section was to observe the variability of classification performance when the estimated glottal flow is drawn from a specific distribution. The Q parameters from the IBIF filter were drawn from a Gamma distribution using Montecarlo simulations from which glottal waveforms and features (e.g., ACFL, MFDR) were extracted and used in a supervised learning classifier. By using a Beta distribution for the classification scores, the relative error is calculated and compared to the error introduced in the Gamma distribution (10% of the mean). Results show that the relative error of classification are below 10%, therefore, they are within the margin established by the randomization of Q parameters. An observation made

with the simulations was that standard features from the glottal waveform might not capture all differences between subjects with PVH and controls. It is inferred that mean Q parameters extracted from PSO are the best estimation for a glottal waveform compared to the ground-truth signal (i.e., inverse filtering from oral airflow signal). However, the simulations indicate that the classification scores using those parameters are below the median of the classification distributions. One plausible explanation is that there are better estimations of the glottal flow using Q parameters that were not captured during the process of calibration and PSO calculation. One way to solve this issue is to incorporate more subjects to the classification task and to use a Montecarlo simulation to obtain Q parameters. It is worth noticing that a better classification score does not guarantee that the best estimation of Q parameters was found; it might just a random artifact where the features extracted, even though are not optimal, might perform better in a classification task.

5.3 Transfer Learning: Using wavelets and convolutional neural networks to classify pre and post therapy for NPVH subjects

The goal if this section is to address aim 1 (SA1) of this thesis, which looks to compare supervised classification tasks using physiological-relevant features and learned features from a deep learning algorithm. This section addresses the issue of differentiating pre-therapy against post-therapy for subjects with NPVH using learned features from the raw accelerometer signal. A supervised classification scheme is used for sustained vowels used in laboratory settings. Instead of using a classic machine learning algorithm, this experiment uses transfer learning, a technique used in deep learning to improve classification algorithms by using a pre-trained set of deep convolutional neural networks (CNN) from a large database of images [182].

Training a deep CNN from scratch is computationally expensive and requires a large amount of training data. In various applications, a sufficient amount of training data is not available, and synthesizing new realistic training examples are not feasible. In these cases, leveraging existing neural networks that have been trained on large data sets for conceptually similar tasks is desirable. This leveraging of existing neural networks is called transfer learning [183]. This technique has found success in different applications of data mining and machine learning [184, 185, 186] and it is especially useful when training data is limited.

The goal of this section is to assess the classification performance using advanced deep learning algorithms for the task of NPVH classification before and after therapy. It is hypothesized that classification performance will improve substantially due to the state-of-the-art algorithm, at the expenses of features that are harder to interpret in the context of vocal function.

5.3.1 Methods

GoogLeNet [187] is a deep CNN originally designed and pre-trained to classify images in 1000 categories. We reuse the network architecture of the CNN to classify accelerometer signals based on images from the time-frequency representation of the time series data. Due to the computational cost of training a CNN, the data is limited to vowels /a/ and /i/ from comfortable, loud, and soft conditions from each subject. There are 9 NPVH subjects and we compare pre-therapy vs. post-therapy using those tokens. Figures 5.13 to 5.18 show the example of sustained vowels /a/ and /i/ in three different conditions (soft, comfortable, and loud). All of these represent the same subject, where the top panel is the ACC signal pre-therapy and the bottom panel corresponds to post-therapy. The morphology of the ACC signals are different between vowels and conditions. Moreover, there are differences between pre and post therapy for the same condition. In the examples from figures 5.13-5.18, there are clear differences in amplitude for all conditions, in addition to changes in the shape of the waveforms.

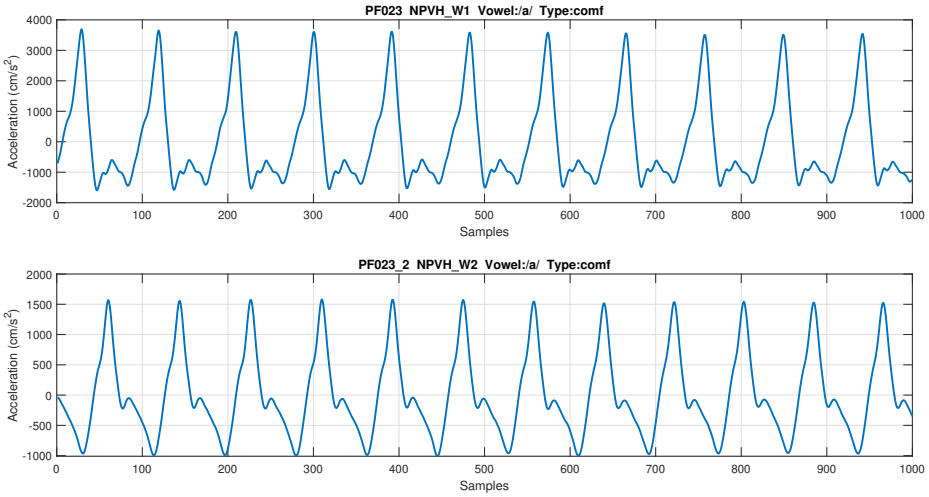

Figure 5.13: In Lab example of 50 ms ACC signal for subject PF023 in comfortable /a/ vowel for pre-therapy (top) and post-therapy (bottom).

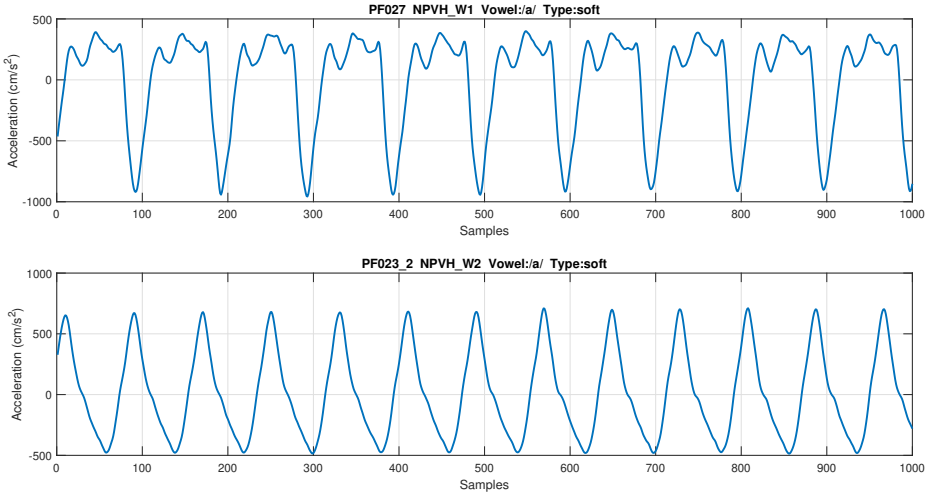

Figure 5.14: In Lab example of 50 ms ACC signal for subject PF023 in soft /a/ vowel for pre-therapy (top) and post-therapy (bottom).

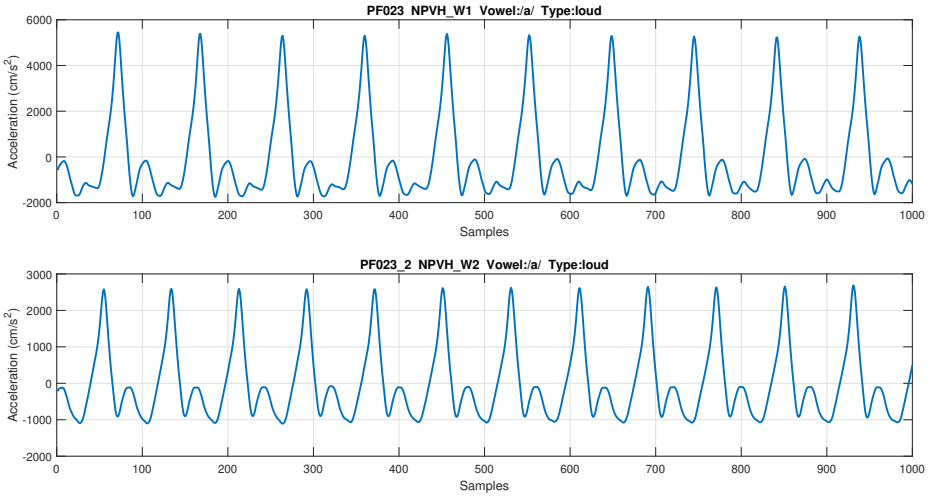

Figure 5.15: In Lab example of 50 ms ACC signal for subject PF023 in loud /a/ vowel for pre-therapy (top) and post-therapy (bottom).

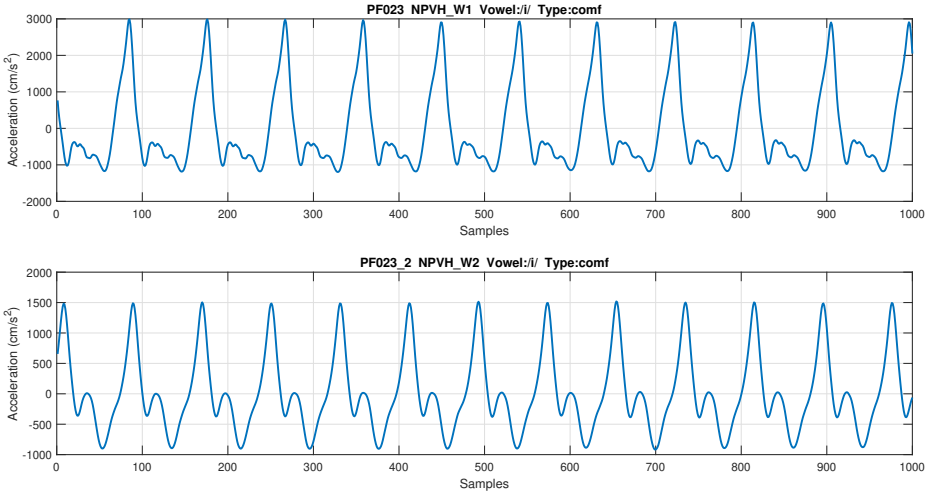

Figure 5.16: In Lab example of 50 ms ACC signal for subject PF023 in comfortable /i/ vowel for pre-therapy (top) and post-therapy (bottom).

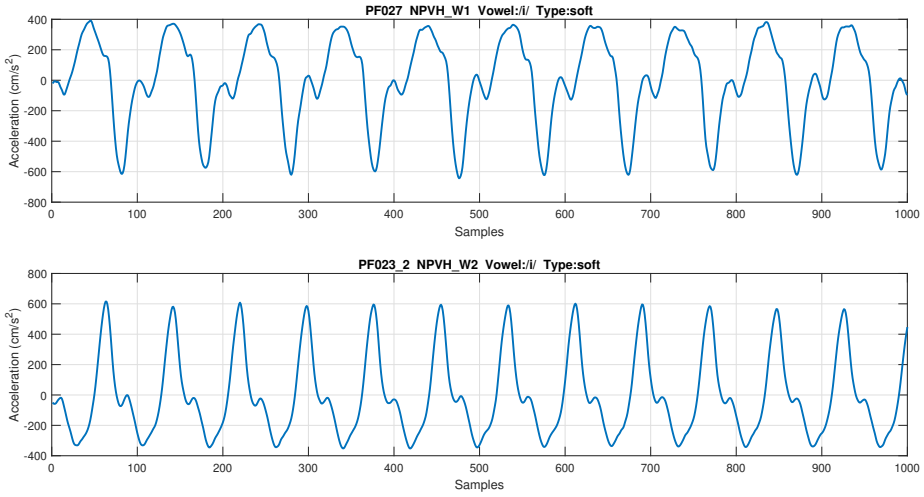

Figure 5.17: In Lab example of 50 ms ACC signal for subject PF023 in soft /i/ vowel for pre-therapy (top) and post-therapy (bottom).

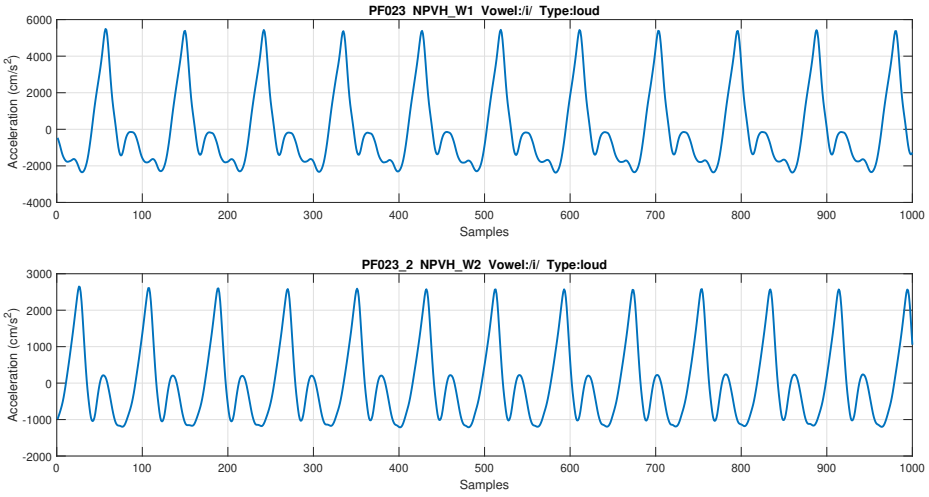

Figure 5.18: In Lab example of 50 ms ACC signal for subject PF023 in loud /i/ vowel for pre-therapy (top) and post-therapy (bottom).

Scalogram: Spectrogram of wavelets

The input to CNNs are 2-D images represented by 3-D matrices representing height, width, and RGB code (for color images). Since the ACC data is 1-dimensional, one approach is to transform the 50 ms signal in a time-frequency representation in order to obtain an image (2-dimensional). An option could be to use spectrograms from Short-Time Fourier Transforms (STFT), however, such approach fails to capture sudden frequency changes and transients that might occur in a non-stationary ACC signal. In contrast, the Wavelet Transform (WT) is of particular interest for non-stationary signals because it utilizes short windows at high frequencies and long windows at low frequencies (in contrast to STFT which uses a single analysis window) [188]. In this exercise we use a continuous wavelet transform (CWT) [189], which is defined as:

$$W_\psi(t,s) = \int_{-\infty}^{\infty} \frac{1}{s^n} \psi^* \left(\frac{\tau - t}{s} \right) x(\tau) d\tau \tag{5.7}$$

where $x(t)$ is the signal of interest and $\psi^*()$ is the conjugate of the wavelet (also called the "mother" wavelet). s is the scale parameter which is inversely proportional to the frequency, and $n = 1/2$ is a scale normalization as a set of projections. CWT supports good time-frequency resolution, specially for instantaneous changes in frequency and localizing transients in non-stationary signals [189]. A discretized version of the CWT is necessary to be implemented in a

computational environment. Specifically, the scale parameter s is discretized as a fractional power of 2, i.e., $s = 2^{j/v}$ where v is an integer greater than one and $j = 1, 2, 3, \cdots$. The parameter v is referred as "voices per octave" [189]. Therefore, the resulting discretized mother wavelets for the CWT are:

$$\frac{1}{2^{j/v}}\psi\left(\frac{n-m}{2^{j/v}}\right) \tag{5.8}$$

where m is a non-negative integer. The parameter v is referred as the number of voices per octave because increasing s by an octave (a doubling) requires v intermediate steps. Common values of v are 10,12,14,16, and 32. The larger the value of v, the finer the discretization of s, but it also requires higher computation. A special type of wavelets are desirable for CWT: Analytic wavelets, which are complex-valued wavelets whose Fourier transform vanish for negative frequencies.

Many analytical wavelets have been proposed, including the Morlet wavelet, the Cauchy-Klauder-Morse-Paul, Derivative of Gaussian, lognormal or log Gabor, Shannon, and Bessel wavelets [189, 190]. A wavelet that generalizes all wavelet types mentioned before is the Morse wavelet [191], which is defined in the frequency domain as:

$$\Psi_{\beta,\gamma}(\omega) = \int_{-\infty}^{\infty} \psi_{\beta,\gamma}(t)e^{-i\omega t}dt = U(\omega)a_{\beta,\gamma}\omega^{\beta}e^{-\omega^{\gamma}} \tag{5.9}$$

where $a_{\beta,\gamma}$ is a normalization constant, $U(\omega)$ is the unit step function, and β and γ are two parameters controlling the wavelet form. The parameter γ controls the symmetry of the wavelet in time through the demodulate skewness [192] and β, which is a decay or compactness parameter. A derived parameter $P^2 = \beta\gamma$ is the time-bandwidth product, whose square-root, P, is proportional to the wavelet duration in time. One advantage of generalized Morse wavelets is that different combinations of its parameters give rise to specific analytic wavelets. For example, Cauchy wavelets have $\gamma = 1$ and Bessel wavelets are approximated by $\beta = 8$ and $\gamma = 0.25$ [193]. Fig 5.19 shows Morse wavelets in the time domain for different combinations of γ and β. The solid red line corresponds to the real part of the wavelet, the dashed yellow line to the imaginary part, and the solid blue line to the modulus. Fig 5.20 corresponds to the frequency response of the wavelets in Fig 5.19.

A filter bank of wavelets can be constructed, as shown in Fig 5.21, from the Morse wavelet with $\gamma = 3$, $P = 60$ (2nd row, 2nd column from Fig 5.19), which is going to be the default wavelet used. The pre-computed filter bank allows for efficiently calculating the continuous wavelet transform (CWT). The ACC signal is filtered through this filter bank and the square absolute values of the CWT are used to construct the scalogram, which is the equivalent to a spectrogram, as the scale parameter s is inversely proportional to the frequency in a spectrogram.

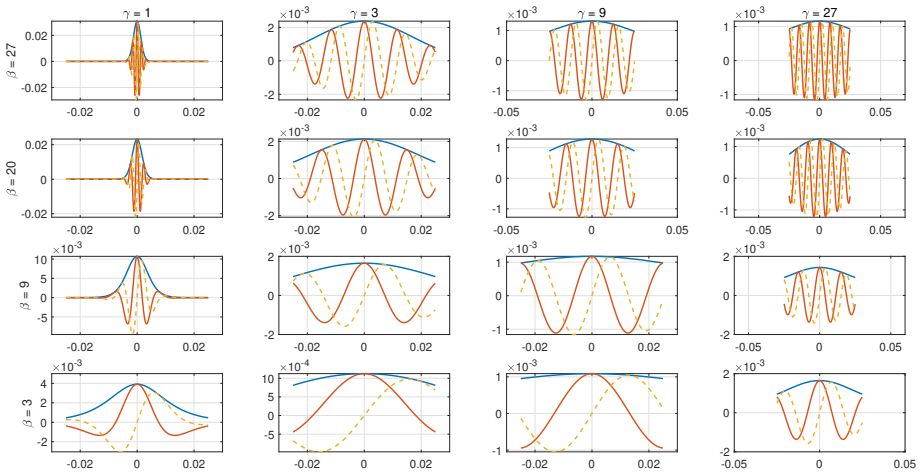

Figure 5.19: Morse wavelets with different γ and β values. Red line corresponds to real part, dashed yellow to imaginary part, and blue line to modulus of the wavelet. The sampling frequency is 20 kHz.

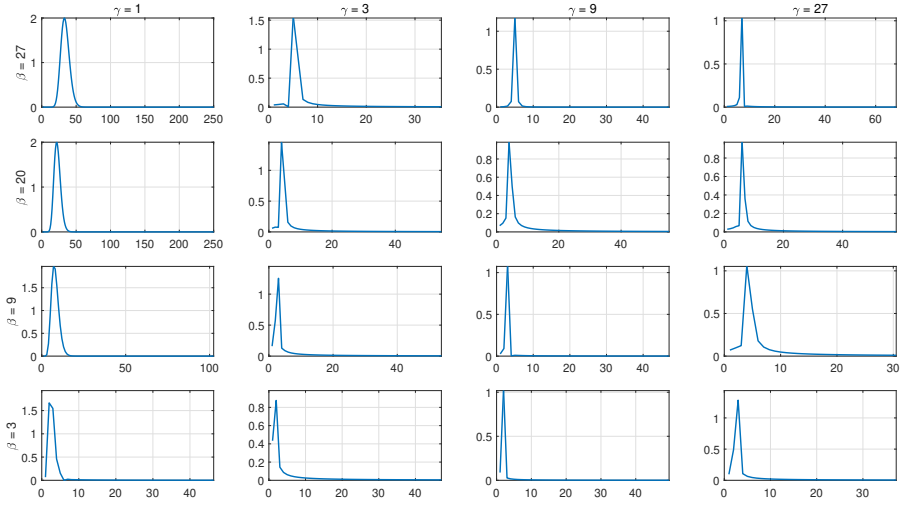

Figure 5.20: Spectrum corresponding to each mother wavelet in Fig 5.19

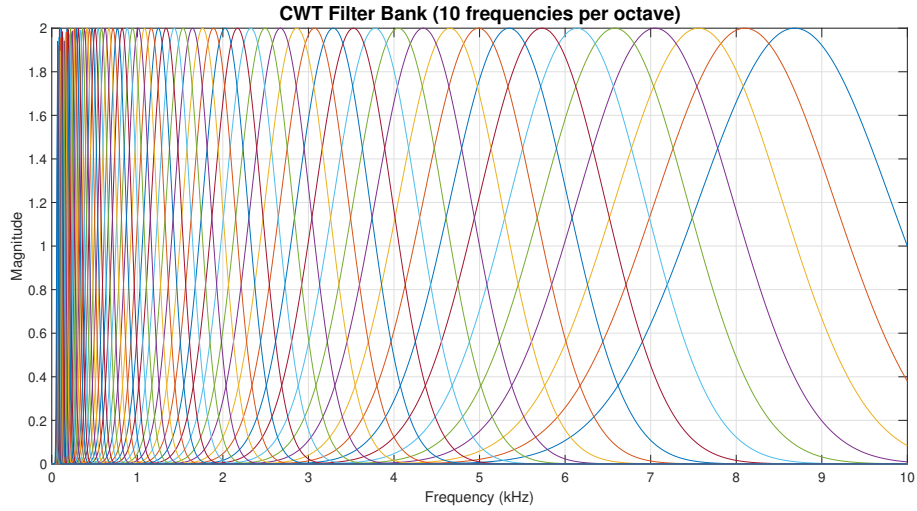

Figure 5.21: Filter bank of the Morse wavelet (3,60) for 10 frequencies per octave.

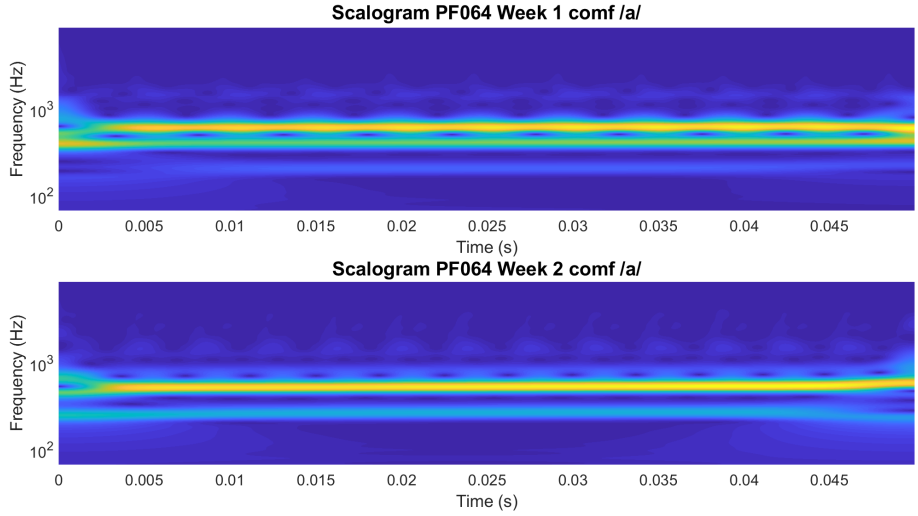

Figure 5.22: Scalogram of a 50 ms segment for the vowel /a/ (comfortable) from a NPVH subject pre-therapy (top) and post-therapy (bottom). Parameter values: $\gamma = 3$, $P^2 = 60$, VpO $= 26$

Transfer learning with CNNs

A common and highly effective approach to deep learning on relatively small datasets is to use a pre-trained network. A pre-trained network is a saved network that was previously trained on a large dataset, typically on a large-scale image classification task. If this original dataset is large and general enough, then the spatial hierarchy of features learned by the pre-trained network can effectively act as a generic model of the visual world, and hence its features can prove useful for many different computer vision problems, even if these new problems may involve completely different classes than those of the original task [194]. There are many available convolutional neural networks (convnet) architectures, usually trained on the ImageNet dataset (1.4 million labeled images and 1,000 classes [182]). One of these architectures is the GoogLeNet model [187], which is an state-of-the-art convnet with 144 layers that is widely used in the computer vision community .

The parameters of the GoogLeNet network are optimized to recognize 1000 categories of images, mostly animals, people, food, plants, among others. In this exercise, the three last layers are modified in order to adapt the task of binary classification:

1. The layer *pool5-drop_7x7_s1* is modified to have a dropout probability of 0.6. Dropout is a standard technique to avoid overfitting of networks [195].

2. The fully connected layer *loss3-classifier* replaces the layer with 1000 cate-

gories from the original GoogLeNet to the number of classes of the current task, which is two.

3. The classification layer specifies the output classes of the network. This one is replaced with a new one without classes. During training time, the network automatically sets the output classes of the layer.

The standard stochastic gradient descent (sgd) method [196] is an optimization algorithm that updates the network parameters (weights and biases) to minimize the loss function by taking small steps at each iteration in the direction of the negative gradient of the loss. Stochastic gradient descent with momentum (sgdm) [181] it's the same sgd algorithm but an extra momentum term is added to smooth the effect of zig-zagging when looking for the optimal solution. It is defined as:

$$\boldsymbol{\theta}_{l+1} = \boldsymbol{\theta}_l - \alpha \nabla \mathbb{E}(\boldsymbol{\theta}_l) + \eta(\boldsymbol{\theta}_l - \boldsymbol{\theta}_{l-1}) \tag{5.10}$$

where $\boldsymbol{\theta}_l$ represents the weights and biases of the network at iteration l, α is the learning rate, and $\mathbb{E}(\boldsymbol{\theta}_l)+\eta(\boldsymbol{\theta}_l-\boldsymbol{\theta}_{l-1})$ is the momentum term, where η is a number between $0 \leq \eta \leq 1$ that controls the importance of the momentum term. *eta* is set to 0.9 and α is set to 0.0001, a small number in order to avoid large deviations from the initial parameter values, that are already optimized to classify images.

There are hyperparameters in the network that affect the speed and accuracy

of the training and validation model, as well the results of the testing data. These include:

- **Dropout number**: The fraction of nodes that are randomly shut down in a fully connected layer during a pass of the training procedure. The number goes from 0 to 1, where 0 means the fully connected layer is active, while 1 leaves the layer inactive. Used to avoid overfitting, a higher number number indicates more protection against overfitting. The following results use a dropout number of 0.6.

- **Learning rate**: This parameter is used in the stochastic gradient descent algorithm, as indicates how fast the algorithm converges to its optimal parameter values. A large learning rate might find the optimal solution faster, however, there is a great possibility that instead it overshoots the search of optimal parameters and it never finds them. On the other hand, a small learning rate greater than zero will more likely find the optimal solution, at the expense of a slower training procedure. Since our objective is to find the optimal solutions with the assumption that GoogLeNet already has optimal parameters to detect images, from details such as edges and orientation, to higher image features, a low learning rate is expected. The following results use a learning rate of 10^{-4}.

- **Number of epochs**: Epochs are the number of training iterations over the

dataset [30]. The experiments run on 5 epochs, as the training and test are kept steady afterwards.

- **Number of mini-batches**: A mini-batch is a subset of the training set that is used to evaluate the gradient of the loss function and update the weights. The number of mini-batches is set to 10.

All training was done a graphical process unit (gpu) NVDIA card GeForce GTX 960M to optimize and speed up the processing task.

5.3.2 Results

Fig 5.24 shows the accuracy for validation and testing set using different voices per octave (VpO) in each filter bank of Morse wavelets. While the validation set is stable in accuracy across VpO, the test set accuracy fluctuates more across VpO, and the maximum accuracy for the test set is 81.5% using 26 VpO.

There could be an issue with overfitting (i.e., memorization of training data instead of generalizing to different data) due to the large gap between training/validation and test accuracy. Methods to avoid overfitting in neural networks include dropout (already used in the analysis) and L2-regularization [181, 29] applied to the loss function. It can be defined as:

$$\mathbb{E}_R(\boldsymbol{\theta}) = \mathbb{E}(\boldsymbol{\theta}) + \lambda\Omega(\boldsymbol{\omega}) \tag{5.11}$$

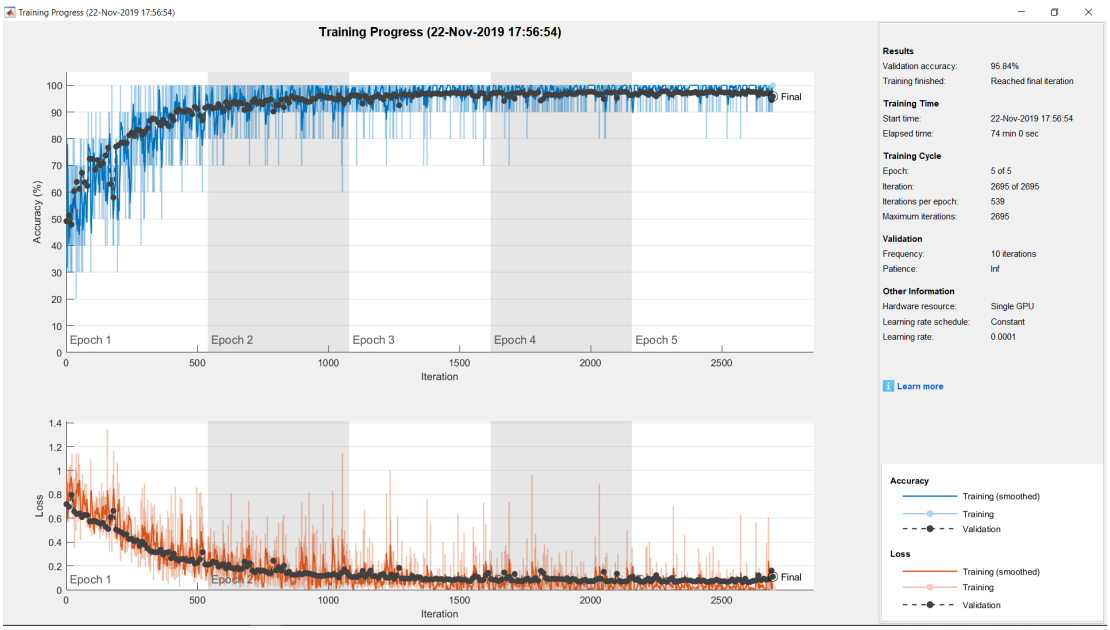

Figure 5.23: Training and validation schedule for 5 epochs using the GoogLeNet CNN network. The top plot indicates the accuracy of the training model (blue line) while the black line indicates the validation accuracy. The bottom plot indicates the error of the training (red line) and the validation error (black line).

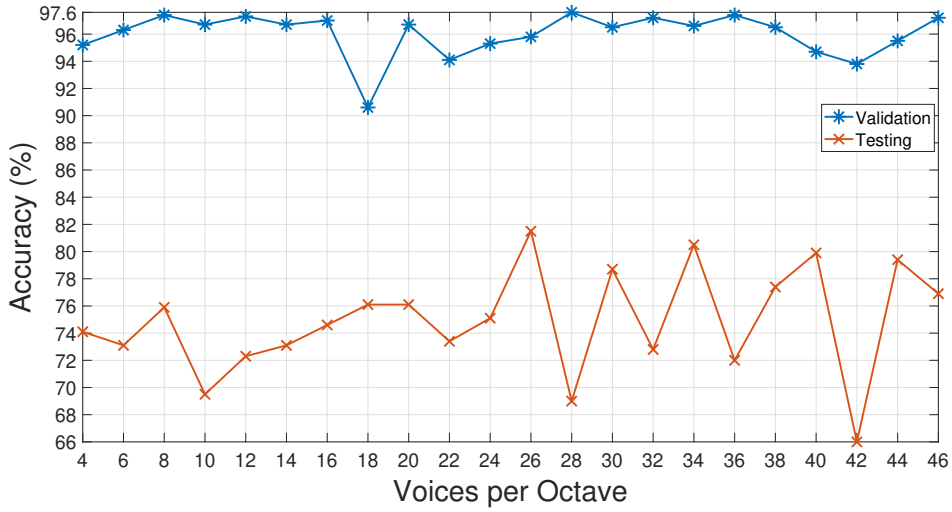

Figure 5.24: Accuracy percentage of classification for validation set (blue) and testing set (PF064, red) for different number of wavelet filters (voices per octave)

where $\mathbb{E}(\boldsymbol{\theta})$ is the loss function of the network with weight parameters $bm\theta$. The regularization term $\lambda\Omega(\boldsymbol{\omega})$ contains a regularization factor λ and a regularization function dependent on the weight parameters $\boldsymbol{\omega}$:

$$\Omega(\boldsymbol{\omega}) = \frac{1}{2}\omega^T\omega \tag{5.12}$$

Similar to the L1-regularization applied to logistic regression and SVM in chapter 3, the weight parameters $\boldsymbol{\omega}$ are penalized to be small values. The bias are not regularized and, for the following experiments, the factor λ is applied to all the weights in the layers of the CNN. Table 5.6 shows classification results using different λ values; all results include a dropout layer and the number of VpO for the CWT is 26.

Set	λ	AUC	Acc.	Sens.	Spec.	PPV	NPV	Fscore
Val.								
	0.0001	0.996	0.978	0.979	0.978	0.980	0.977	0.979
	0.001	0.993	0.955	0.969	0.940	0.946	0.965	0.957
	0.01	0.997	0.975	0.974	0.975	0.977	0.972	0.976
	0.1	0.997	0.976	0.970	0.983	0.984	0.968	0.977
Test								
	0.0001	0.831	0.688	0.341	0.986	0.954	0.635	0.502
	0.001	0.916	0.835	0.692	0.958	0.933	0.784	0.795
	0.01	0.760	0.640	0.264	0.962	0.857	0.604	0.403
	0.1	0.811	0.657	0.286	0.976	0.912	0.614	0.435

Table 5.6: Classification performance for subject PF064 using a dropout of 0.7 and L2-regularization with different values of the hyperparameter λ. An increasing value of λ results in higher regularization penalty.

As shown in Table 5.6, L2-improves slightly on top of the dropout layer, if using $\lambda = 0.001$. However, there is a dramatic drop in performance when higher lambdas are used. Surprisingly, the classification performance of the validation set remained almost the same, independently from the lambda value. The validation set is used to optimize the parameters of the CNN, which at the same time are

random images from the training set (not used for training). Therefore, even with high regularization, the CNN works very well with the training and validation set due to the large number of layers with optimized parameters for learning multiple type of images. However, as was the case using simpler learning classifiers, the test set seems to differ from the training set in terms of optimized parameters. One detail to notice is the high specificity value compared to the sensitivity for the testing case. This implies that the classifier is able to detect frames for the first week more accurate than frames for the second week. A larger analysis is needed to detect this pattern in more subject tests.

Each layer of a CNN produces a response, or activation, to an input image. However, there are only a few layers that are suitable for feature extraction. The layers at the beginning of the network capture basic image features, such as edges and blobs [30]. To see this, Fig 5.25 shows the network filter weights of the first layer, for which there are 64 individual sets of weights.

There is the possibility of extracting the activation of different layers of the network given an image. By examining the activations we can discover what features the network learns by comparing areas of activation with the original image. If we examine the first layer activations for Fig 5.26, which is a 50 ms vowel /a/ (comfortable) from pre-therapy, we obtain Fig 5.27:

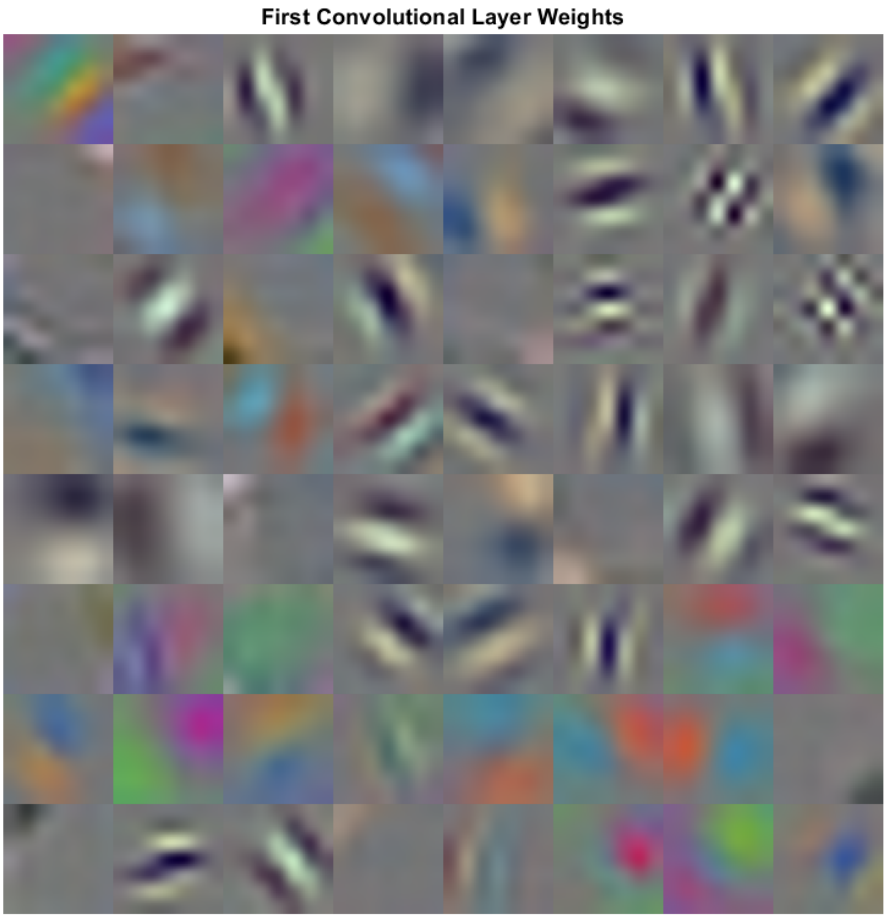

Figure 5.25: Weights for the first layer in CNN.

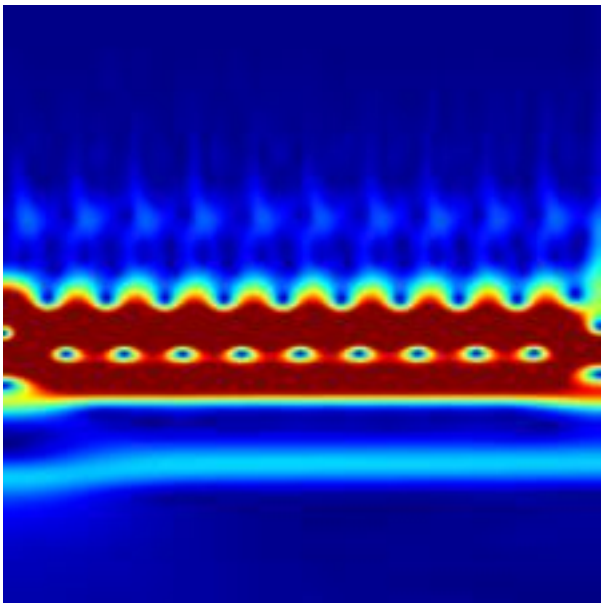

Figure 5.26: Scalogram for the vowel /a/ comfortable condition, pre-therapy, from subject PF064 (converted to 227x227x3 image as input to the network)

We then compare the scalogram from Fig 5.26 with the strongest activation channel in Fig 5.27, and find to what channel it corresponds, as shown in Fig 5.28.

Another image will create a different map of activation channels, as the input scalogram in Fig 5.29

Again, we compare the scalogram from Fig 5.29 with the strongest activation channel in Fig 5.30, and find to what channel it corresponds, as shown in Fig 5.31.

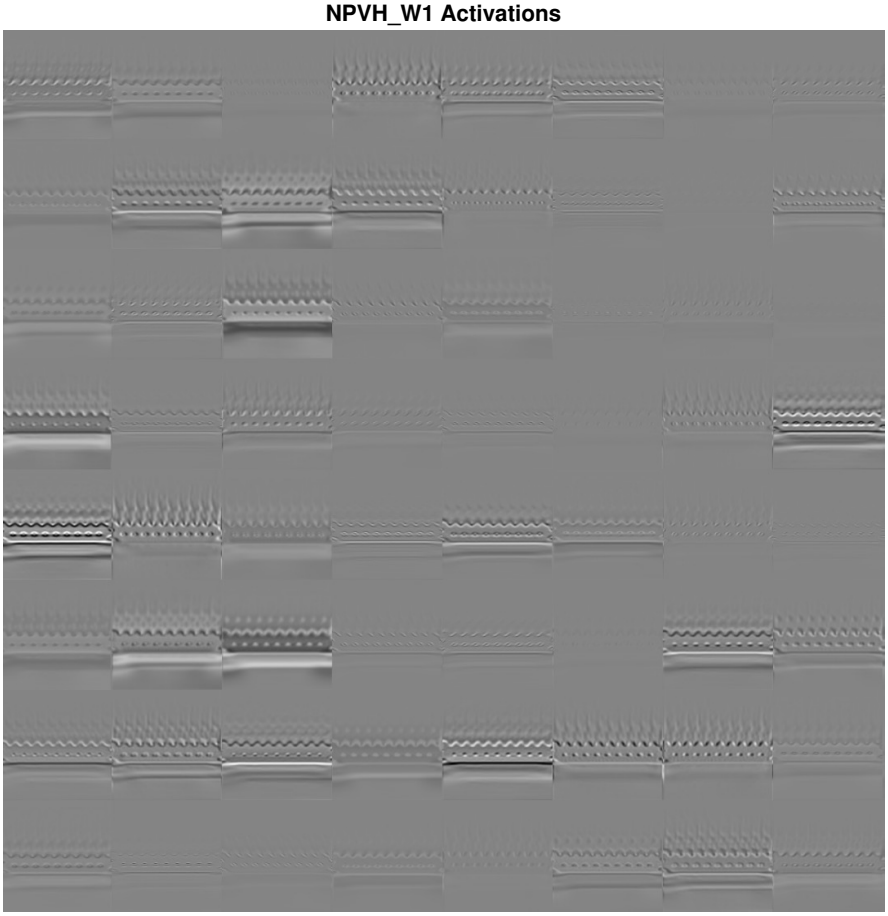

Figure 5.27: Channel activations from the first layer using the scalogram from Fig 5.26 as input

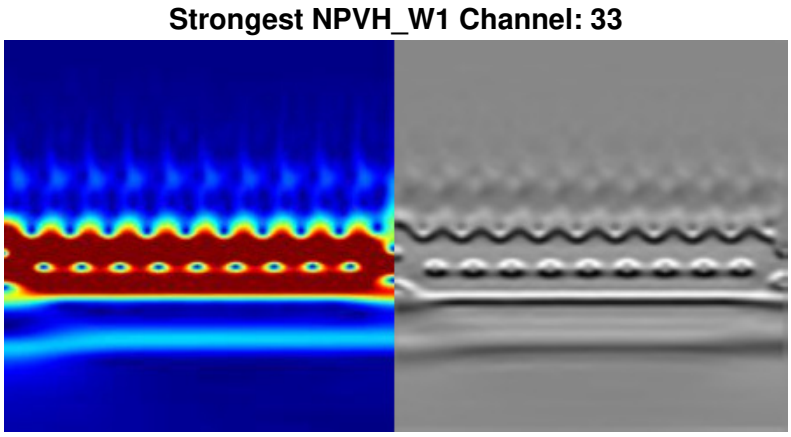

Figure 5.28: Input image (left) and strongest activation channel from Fig 5.27

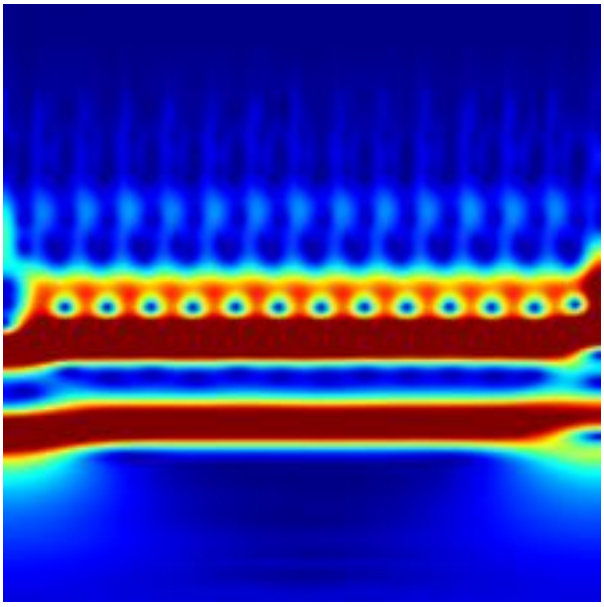

Figure 5.29: Scalogram for the vowel /a/ comfortable condition, post-therapy, from subject PF064 (converted to 227x227x3 image as input to the network)

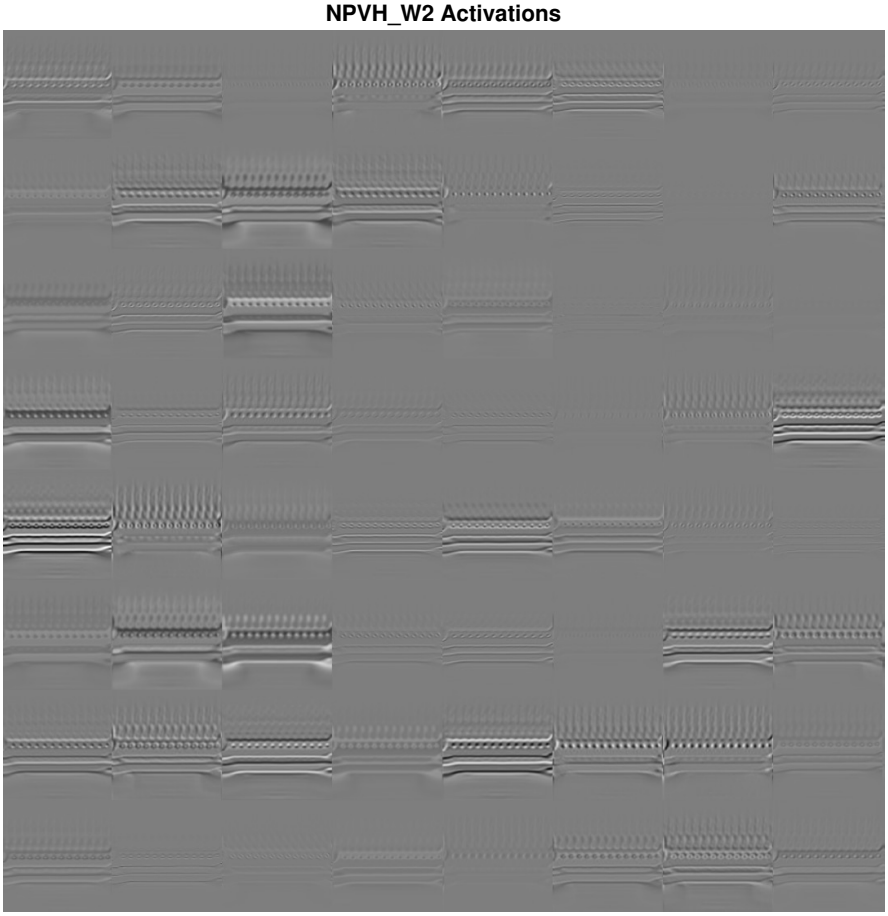

Figure 5.30: Channel activations from the first layer using the scalogram from Fig 5.29 as input

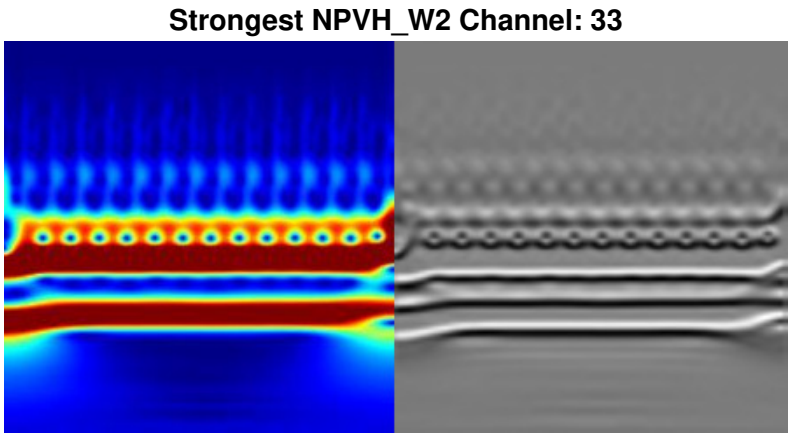

Figure 5.31: Input image (left) and strongest activation channel from Fig 5.27

5.3.3 Discussion

This section presented a supervised classification tool, Convolutional Neural Networks, to classify voiced frames from NPVH subjects with pre and post voice therapy. It is a first attempt to identify voiced frames in a pre-treatment vs. post-treatment using only neck-skin acceleration data within a deep learning approach. The main objective is to identify a network model for every subject that can classify pre-therapy vs. post-therapy. After a series of experiments using a single pair of test subjects, the following can be concluded:

1. From the classification tasks, as expected from methods and results from chapter 3, the variance of the accuracy is large across subjects, thus, only a subject (pre-therapy vs. post-therapy) is used as test data in the tuning of features/classifier parameters.

2. By keeping fixed γ and β values, a comparison of different voices per octave showed that the best performance is with wavelets having 26 voices per octave.

3. There is a tendency to produce overfitting, i.e., there is a large gap between the validation and test accuracy, even if traditional methods to avoid it (e.g., dropout and L2 regularization) are used.

4. Overall accuracy for different test sets might need fine tuning of hyperparameters.

It can be concluded that, similar to classical machine learning algorithms, deep learning could help classify "normal" voiced frames from pathological frames by tuning correctly training parameters, even though both approaches seem to have a problem with overfitting data, despite the use of the techniques for its reduction. Since the problem of this section is different to the problem in chapter 3, i.e., classification of control/PVH subjects vs. classification of NPVH subjects pre/post therapy, both approaches could be complementary to the assessment of voice problems. It is well known that NPVH is a difficult condition to assess due to the lack of knowledge on its etiology [18]. As it was reported in section 5.1 of this chapter, the classification of NPVH subjects vs. controls using ambulatory data does not perform as well as the classification of PVH subjects and controls. Due to the lack of organic lesion on NPVH subjects, there is no clear explanation to the causes of the problem, and therefore, there is no clear evidence that features from glottal aerodynamics are strongly different than controls. The main advantage of deep learning tools, such as CNN, is the capability to learn high-order complex features from an image. In this case, we used scalograms: time-frequency representations of filter banks passed through ACC signals of NPVH subjects. The purpose of this methodology is to extract characteristics of the ACC signal in different frequency bands with different scale/time dimension, in order to differentiate scalograms that are different from pre-therapy vs. post-therapy. Even though this procedure might require tuning many parameters, in the absence of a

strong prior for the etiology of NPVH, it is a good first approach to analyze the ACC signal and understand what are the aspects that are different from other type of ACC signals.

Future work might involve to optimize wavelet parameters, as well as CNN hyperparameters, using a robust optimization algorithm, such as PSO or Bayesian optimization. Even though there might be a significant improvement, the task of differentiating salient features from ACC signals in NPVH subjects is a difficult one, specially in ambulatory settings. As labeling is an issue with this type of signals, an unsupervised, or semi-supervised approach to identify NPVH patterns could be a solution this issue. Specifically, analyzing time-series data requires to look the relationship between neighboring frames and extract useful information about behavioral patterns. In addition, feedback from the user, such as periodically updating vocal status, could improve learning algorithms as these status are soft labels that might give indications to unsupervised algorithms how well the system is learning.

5.4 Conclusion

The purpose of this chapter was to provide different experiments that attended aim SA1 (section 5.1 and 5.3 with classification methods using aerodynamic and non-aerodynamic features) and aim SA2 (section 5.2 with the analysis of uncer-

tainty of the Q parameters from the IBIF filter). The common ground of these experiments was the sample size of data, which was small compared to the data used in chapters 3 and 4. The limited data provides an opportunity to try different methods and to quickly check if it is worth to continue using the same methods with larger data.

Section 5.1 aimed to classify ambulatory voice data from NPVH subjects and controls. It is the first time that aerodynamic measures from NPVH subjects are analysed in an ambulatory setting. Even though previous results on clinical aerodynamic voice data with NPVH showed statistical differences from controls, the accuracy of classification is not strong enough for ambulatory data. Using the leave-one-out method of classification, some subject pairs performed very well with the classification, while others performed very poorly. The only aerodynamic feature that was significant different based on a t-test and was associated to NPVH as an odd ratio is the 5th percentile of MFDR'. Therefore, the lowest values of MFDR (normalized by SPL) are statistically higher for NPVH than for healthy controls. Since MFDR is positively correlated to higher collision forces, NPVH subjects would have higher MFDR at lower glottal flow, probably at onset or offset of phonation. The mean of H1-H2 is significant different only when is obtained using the raw accelerometer signal (without IBIF). However, using IBIF, H1-H2 is a feature associated to NPVH from the odds ratio. Since H1-H2 from the accelerometer raw signal is positively correlated to H1-H2 obtained from the

glottal flow signal [173], the hypothesis from H1 is confirmed as H1-H2 is a spectral glottal feature associated to NPVH subjects.

Section 5.2 aimed to classify PVH subjects from controls using IBIF with random parameters Qs extracted from parametric distributions. The goal was to quantify the error in classification scores by setting the variance of the probability distributions from which the Q parameters were drawn. Even though the experiments were done with features extracted from reading passages (no ambulatory data), the phonetically balance of those passages provides assurance that the IBIF filtered a diverse set of voiced frames. This is important, as the hypothesis of H2 states that the error of Q parameters affects the filtering procedure. The production of different vowels might be a source of error for a set of Q parameters that were calibrated using a single vowel [127, 150]. After applying 10% of variance around the mode of nominal Q parameters, the classification results using random Q parameters were within an acceptable range of error. This implies that small variation of Q parameters do not affect strongly the classification results. Therefore, it would not satisfy hypothesis H2, since controlling the variance of those parameters do not necessarily improve the classification performance. The shortcomings of this experiment is the prior assumption of a parametric distribution for the Q parameters, which might not be the case in ambulatory scenarios, and the limited number of subjects for training and testing. Therefore, a larger experiment with ambulatory data could provide more evidence if the prior distribution

assumption is true.

Section 5.3 provided a different classification task: NPVH subjects with pre-therapy and post-therapy. This is a hard classification problem as the same subjects are tested for classification. Aerodynamic features classify almost randomly, therefore, the task in this section was to use advance machine learning algorithms (e.g., deep learning) to obtain high-level features from a time-frequency representation of the accelerometer signal. The experiment attends aim SA1 because it is a classification problem of NPVH subjects using non-aerodynamic features. The hypothesis H3 states that a deep learning approach will improve classification for complex tasks such as differentiating pre-therapy vs. post-therapy. Section 5.3 confirms H3 because a classification accuracy of 83% is achieved, similar to the one obtained for the task of PVH/controls in chapter 3. However, there are some issues: Overfitting (i.e., high accuracy training vs. a low accuracy test level) is a common problem in deep learning systems. Another issue is to have multiple test sets to generalize the average accuracy. These issues can be solved by implementing a system that optimizes parameters from the front-end (wavelet transform) to the back-end (CNN) in sequential steps. In conclusion, there is room for improvement to these technique, as it has been successful in other medical areas, such as electrocardiogram classification [197, 198, 199].

Chapter 6

Conclusion and future work

The work presented here summarizes a comprehensive analysis on ambulatory voice data using an accelerometer attached to the neck-skin of subjects. The analysis covered classification of subjects with Phonotraumatic vocal hyperfunction (PVH, e.g., nodules and polyps on the vocal folds) and Non-Phonotraumatic vocal hyperfunction (NPVH, e.g., non-organic functional disorder) during ambulatory settings and in laboratory conditions. The main purpose of chapter 3 was to differentiate ambulatory data from subjects with pathologies vs. healthy controls by using aerodynamic features from the glottal flow. The main hypothesis is that glottal features, such as AC-flow and MFDR, are more salient in subjects with PVH due to the obstruction of airflow through the glottis from the organic lesions, which results in higher subglottal pressure, and therefore, higher glottal airflow, in order to maintain a certain sound pressure level at the lips. The organic lesions also might induce to higher collision forces, which is reflected indirectly in higher MFDR. Previous work has shown significant differences between

PVH and healthy controls when glottal aerodynamic features are obtained under laboratory conditions. This thesis provides advances on the analysis of salient aerodynamic features for a large group of PVH and control subjects during ambulatory recording of voice function. These features improved on previously reported classification tasks using machine learning. Statistical significant features using summary statistics show that minimum and median measures of AC-flow have significant differences between PVH and control subjects. These are not normalized by SPL, which implies that PVH subjects had higher values in absolute AC-flow than controls in the minimum and median range. These results agree with previous results [6, 12], even though in the ambulatory case these are not normalized by SPL. However, there are not significant differences in SPL between PVH and controls in the ambulatory data [75], therefore normalization would not affect the results overall. Moreover, differences between the acoustic and ACC estimated SPL [130] might affected more robust estimations of normalized values of AC-flow and MFDR. Spectra-based features that were significant different in these summary statistics include the kurtosis and minimum value of H1-H2, and the kurtosis of HRF. These measures are related to voice quality, as a higher spectral slope represents the type of a breathy voice, while a lower spectral slope represents a pressed type of voice. Kurtosis is related to the tails of the distribution [200]: Higher values of kurtosis imply higher number of outliers within the distribution, and to some extent, how "peaky" the distribution is. Therefore, H1-H2 and HRF

have a higher number of outliers in the PVH population than the control, which is a sign that PVH subjects had more moments of extreme values for these measures that are related in voice quality. The analysis of odds ratio from logistic regression models associate a number of features to the PVH and control group. As expected, amplitude-based features such as AC-flow (standard deviation, skewness) and MFDR (5th and 95th percentile) are associated to the PVH group in 48 training. Another measure related to AC-flow, MFDR and f_0 (NAQ mean) was highly associated to PVH subjects. However, the most associated feature (highest odd ratio) with PVH was the 95th percentile of H1-H2, which, as mentioned before, is a strong indicator of voice quality. The variance of the odds ratio for this feature was the highest, which is an indication that subjects with PVH manifest quite differently H1-H2. It explains the statistically significant difference of kurtosis between PVH and controls for the analysis of summary statistics. In conclusion, chapter 3 confirmed hypothesis 1 (H1) with respect to the high prediction value of amplitude-based glottal features for the assessment of PVH. An unexpected result was the high prediction value of spectral features, specifically H1-H2, for PVH subjects. The hypothesis of H1 was that spectral features would be more associated to NPVH than PVH due to their relationship to voice quality. However, it seems that H1-H2 it is an important feature associated to dysphonia in general. Further studies should be directed on robust estimations of the glottal flow to verify the same results, as well as to study segments of the day to quantify

glottal features as a function of changes in voice quality.

Chapter 4 aimed to establish a different approach to the estimation of the glottal flow using neck-skin acceleration. The IBIF is a deterministic filter that doesn't take into account uncertainties and errors that might appear in the acceleration signal. The proposed method is to use the IBIF model as a forward filter, i.e., to use the transfer function of the neck-skin to calculate acceleration (output) from the GVV (input). This framework allows to use the inverse IBIF as an observation model in a Kalman filter, where the observation signal is the measured neck-skin acceleration, while the states of the filter are the present and past estimations of the GVV signal. In this case, the KF is a Moving Average filter. The goal is to estimate those states in time and select one state that best estimates the GVV signal. Beside white noise, other colored inputs to the KF were tested as well, all based on parametric models of the glottal waveform. Comparing waveforms from the IBIF and KF to the ground-truth GVV (obtained from an OVV signal), the KF signal it was similar to IBIF or closer to the GVV filtered from the OVV. This suggest that KF is able to follow better the ground-truth GVV than IBIF. In an ambulatory setting, the KF and IBIF were calculated for 48 pair of PVH-control subjects and the RMSE for each 50 ms frame was calculated. Frames with an error lower than 10% were selected for classification. However, there was no improvement over random frames selected for classification. Therefore, lowering the deviance between the IBIF and KF did not improve

classification results, and hypothesis H2 is rejected. Further improvements to the state-space model of the neck-skin might incorporate state uncertainty that can be useful for analysis of the error incorporated in the IBIF filter. Future work could be related to the exploration of other applications that can further benefit from the Kalman filter enhancement when estimating glottal airflow and to reduce its computational expense.

Chapter 5 is a collection of different experiments aimed to the assessment of VH. The study of section 5.1 has the same framework as chapter 3 and aim (SA1), except that NPVH subjects (with controls) are used for classification with aerodynamic and non-aerodynamic features. Different machine learning classifiers yielded similar results for both cases. Aerodynamic features provided excellent results for some test pairs, while others resulted in very poor performance. Non-aerodynamic features provided mixed results with almost random guessing. Therefore, aerodynamic features are promising indicators of NPVH for some subjects, but a larger number of test data is needed to establish a good confidence on the results. Since spectral features were associated to NPVH subjects, hypothesis H1 was partially confirmed.

The work on section 5.2 aimed to analyze the uncertainty of Q parameters from the IBIF filter and its relationship to classification performance on PVH/control subjects. The data was limited to 4 pair of subjects reading an Spanish passage, phonetically balanced similar to The Rainbow passage. Under prior assumption

of distributions for the Q parameters, a 10% of variance around the mode of the distribution resulted in different GVV waveforms, from which glottal features were extracted and used for classification. The error in the performance metrics were within an acceptable range for which the conclusion was that small perturbations to the Q parameters did not affect significantly the metric scores of classification. Therefore, hypothesis H2 is partially rejected as IBIF parameters are robust under controlled perturbations, so lowering the error on the IBIF estimation would not affect substantially the classification results.

Section 5.3 had the analysis of time-frequency signals (CWT) from the raw accelerometer signal from subjects with NPVH: pre-therapy and post-therapy. A deep learning framework (CNNs) provided a robust tool for classification using only vowels at different intensities. The system is able to learn training images in a short time with high accuracy, however, there is a gap of more than 15% of accuracy with testing sets, which is a clear evidence of overfitting. Standard tools to avoid overfitting, such as L2-regularization and Dropout, help to a certain limit. As a preliminary study, this section described a system that has good results on classification of pre-therapy vs. post-therapy, and it can be improved further by incorporating more data that shows the transients and dynamics of the neck-skin accelerometer signal. Moreover, an optimization procedure of the hyperparameters and incorporating CAPE-V scores to the classes (pre-post therapy) could be also an improvement to the classification task.

The results of this thesis provide a solid framework for future work related to ambulatory analysis, classification, and clustering of voice pathologies, specifically VH. There are some areas of improvement that can be developed next. For example, the analysis of 50 ms frames could be a limitation, since transients and dynamics of the accelerometer waveform could provide useful information on how VH develops through time. In that case, the continuous wavelet transform would be an excellent choice as a tool of feature extraction due to the time-frequency components that can capture. Therefore, variable length analysis windows can be used with the tools developed in this thesis. Another useful incorporation would be soft labels, i.e., subjective info such as self-perceptual scores from the patients during their ambulatory recordings. These scores could aid a semi-supervised classification/clustering algorithm to detect modes during the day on how the patient feels about his/her voice. These different modes can be connected to find patterns during the day/week a provide insights on how the cycle of VH develops through time. If numerical models of the vocal folds are incorporated in the ambulatory assessment of VH, through subject-specific modeling for prediction of modes with objective data from measurements of the accelerometer, the likelihood of better VH assessment and outcome treatment could improve significantly.

Bibliography

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

[2] D. D. Mehta, J. H. Van Stan, M Zañartu, M Ghassemi, J. V. Guttag, V. M. Espinoza, J. P. Cortés, H. A. II Cheyne, and R. E. Hillman. Using ambulatory voice monitoring to investigate common voice disorders: research update. *Front. Bioeng. Biotechnol. 3:155. doi: 10.3389/fbioe.2015.00155*, 2015.

[3] J. P. Cortés, V. M. Espinoza, C. Castro, R. Manríquez, A. Testart, and M. Zañartu. Classification performance of paired subjects with vocal hyperfunction in the presence of subglottal inverse filtering uncertainties: Pilot study under laboratory conditions. Philadelphia, Pennsylvania, May 2019. 48th Voice Foundation Annual Symposium: Care of the Professional Voice.

[4] J. P. Cortés, G. Alzamendi, A. Weinstein, J. Yuz, V. Espinoza, D. Mehta, J. Van Stan, R. Hillman, and M. Zañartu. Uncertainty of ambulatory airflow estimates and its effect on the classification of phonotraumatic vocal hyperfunction. Quebec, Canada, June 2019. The 13th International Conference on Advances in Quantitative Laryngology, Voice and Speech Research.

[5] J. P. Cortés, V. M. Espinoza, D. D. Mehta, J. H. Van Stan, R. E. Hillman, and M. Zañartu. Estimación de medidas aerodinámicas ambulatorias con un filtro de kalman para la evaluación de la hiperfunción vocal. Santa Cruz, Chile, Nov. 2018. LXXV Congreso Chileno de Otorrinolaringología.

[6] V. M. Espinoza, J. P. Cortés, and M. Zañartu. Métodos de evaluación clínica de la voz basados en medidas aerodinámicas y vibroacústicas. Santa Cruz, Chile, Nov. 2018. LXXV Congreso Chileno de Otorrinolaringología.

[7] J. P. Cortés, V. M. Espinoza, M. Ghassemi, J. V. Guttag, D. D. Mehta, J. H. Van Stan, R. E. Hillman, and M. Zañartu. Aerodynamic ambulatory assessment for phonotraumatic vocal hyperfunction. East Lansing, Michigan, August 2018. 11th International Conference on Voice Physiology and Biomechanics.

[8] J. P. Cortés, V. M. Espinoza, M. Ghassemi, D. D. Mehta, J. H. Van Stan, R. E. Hillman, J. V. Guttag, and M. Zañartu. Using aerodynamic features and their uncertainty for the ambulatory assessment of phonotraumatic vocal hyperfunction. Las Vegas, Nevada, March 2018. IEEE International Conference on Biomedical and Health Informatics.

[9] J. P. Cortés and M. Zañartu. Ambulatory classification of patients with muscle tension dysphonia vs. control group. Hong Kong, Oct. 2017. The 12th International Conference on Advances in Quantitative Laryngology, Voice and Speech Research.

[10] J. P. Cortés, V. M. Espinoza, M. Zañartu, M. Ghassemi, J. V. Guttag, D. D. Mehta, J. H. Van Stan, and R. E. Hillman. Discriminating patients with vocal fold nodules from matched controls using acoustic and aerodynamic features from ambulatory voice monitoring data. pages 95–96, Viña del Mar, Chile, 2016. 10th International Conference on Voice Physiology and Biomechanics.

[11] M. Ghassemi, J.H. Van Stan, D.D. Mehta, M. Zañartu, H.A. Cheyne, R.E. Hillman, and J.V. Guttag. Learning to detect vocal hyperfunction from ambulatory neck-surface acceleration features: Initial results for vocal fold nodules. *Biomedical Engineering, IEEE Transactions on*, 61(6):1668–1675, June 2014.

[12] R. E Hillman, E. B. Holmberg, J. S. Perkell, M. Walsh, and C. Vaughan. Objective assessment of vocal hyperfunction: An experimental framework and initial results. *Journal of Speech and Hearing Research*, 32:373–392, 1989.

[13] N. V. Welham and M. A. Maclagan. Vocal fatigue: Current knowledge and future directions. *Journal of Voice*, 17(1):21 – 30, 2003.

[14] G.P. Moore. *Organic voice disorders*. Prentice-Hall foundations of speech pathology series. Prentice-Hall, 1971.

[15] M. Hirano. *Clinical examination of voice*. Disorders of human communication. Springer London, Limited, 1981.

[16] R. T. Sataloff, J. R. Spiegel, L. M. Carroll, K. S. Darby, M. J. Hawkshaw, and R. K. Rulnick. The clinical voice laboratory: Practical design and clinical application. *Journal of Voice*, 4(3):264 – 279, 1990.

[17] M. Hirano and D.M. Bless. *Videostroboscopic Examination of the Larynx*. Singular Publishing Group, 1993.

[18] J.C. Stemple, N. Roy, and B.K. Klaben. *Clinical Voice Pathology: Theory and Management, Sixth Edition*. Plural Publishing, Incorporated, 2018.

[19] G. B. Kempster, B. R. Gerratt, K. Verdolini, J. Barkmeier-Kraemer, and R. E. Hillman. Consensus auditory-perceptual evaluation of voice: Development of a standardized clinical protocol. *American Journal of Speech-Languaje Pathology*, 18:124–132, 2009.

[20] B. H. Jacobson, A. Johnson, C. Grywalski, A. Silbergleit, G. Jacobson, M. S. Benninger, and C. W. Newman. The voice handicap index (VHI). *American Journal of Speech-Language Pathology*, 6(3):66–70, 1997.

[21] H. A Cheyne, H. M. Hanson, R. P. Genereux, K. N. Stevens, and E. H. Robert. Development and testing of a portable vocal accumulator. *Journal of Speech, Language, and Hearing Research*, 46:1457–1467, 2003.

[22] P. S. Popolo, J. G. Švec, and I. 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.

[23] D. D. Mehta, M. Zañartu, S. W. Feng, H. A. Cheyne, and R. E. Hillman. Mobile voice health monitoring using a wearable accelerometer sensor and a smartphone platform. *Biomedical Engineering, IEEE Transactions on*, 59(11):3090–3096, Nov 2012.

[24] I. R. Titze, J. G. Švec, and P. S. Popolo. Vocal dose measures: Quantifying accumulated vibration exposure in vocal fold tissues. *Journal of Speech, Language, and Hearing Research*, 46:919–932, 2003.

[25] E. B. Holmberg, R. E. Hillman, and J. S. Perkell. Glottal air-flow and transglottal air-pressure measurements for male and female speakers in soft, normal, and loud voice. *The Journal of the Acoustical Society of America*, 84:511–529, 1988.

[26] J. S. Perkell, R. E. Hillman, and E. 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.

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

[28] M. Zañartu, V. M. Espinoza, D. D. Mehta, J. H. Van Stan, H. A. Cheyne III, M. Ghassemi, J. V. Guttag, , and R. E. Hillman. Toward an objective aerodynamic assessment of vocalhyperfunction using a voice health monitor. In *8th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications, MAVEBA 2013, December 16 - 18 2013, Firenze, Italy.*, 2013.

[29] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[30] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. Adaptive Computation and Machine Learning series. MIT Press, 2016.

[31] N. Roy, R. M. Merrill, S. D. Gray, and E. M. Smith. Voice disorders in the general population: Prevalence, risk factors, and occupational impact. *The Laryngoscope*, 115(11):1988–1995, 2005.

[32] C. Morales. ¿De qué se enferman las trabajadoras chilenas? *Ciencia y Trabajo*, 23:20–24, 2007.

[33] K. Verdolini, C.A. Rosen, R.C. Branski, M.L. Andrews, Voice American Speech-Language-Hearing Association. Special Interest Division 3, and Voice Disorders. *Classification Manual for Voice Disorders-I.* Number v. 1. Lawrence Erlbaum, 2006.

[34] D. D. Mehta and R. E. Hillman. Voice assessment: Updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods. *Current Opinion in Otolaryngology Head and Neck Surgery*, 16:211–215, 2008.

[35] D. D. Mehta and R. E. Hillman. Use of aerodynamic measures in clinical voice assessment. *Voice and Voice Disorders*, 17(3):14–18, 2007.

[36] P. Karkos and M. McCormick. The etiology of vocal fold nodules in adults. *Current Opinion in Otolaryngology Head and Neck Surgery*, 17:420–3, 09 2009.

[37] M. Kunduk and A. Mcwhorter. True vocal fold nodules: The role of differential diagnosis. *Current Opinion in Otolaryngology Head and Neck Surgery*, 17:449–52, 09 2009.

[38] I.R. Titze, National Center for Voice, and Speech. *Workshop on Acoustic Voice Analysis: Summary Statement*. National Center for Voice and Speech, 1995.

[39] I. R. Titze, R. Baken, and H. Herzel. Evidence of chaos in vocal fold vibration. In I. R. Titze, editor, *Vocal Fold Physiology: New Frontiers in Basic Science*, pages 143–188. Singular Publishing Group, San Diego, CA, 1993.

[40] D. D. Deliyski, H. S. Shaw, and M. K. Evans. Adverse effects of environmental noise on acoustic voice quality measurements. *Journal of Voice*, 19(1):15–28, Mar 2005.

[41] M. Rothenberg. Measurement of airflow in speech. *Journal of Speech and Hearing Research*, 20(1):155–176, 1977.

[42] M. P. Karnell. Synchronized videostroboscopy and electroglottography. *Journal of Voice*, 3(1):68 – 75, 1989.

[43] K. M. Wheeler, S. P. Collins, and C. M. Sapienza. The relationship between vhi scores and specific acoustic measures of mildly disordered voice production. *Journal of Voice*, 20(2):308–317, Jun 2006.

[44] A. G. Askenfelt and B. Hammarberg. Speech waveform perturbation analysis. *Journal of Speech, Language, and Hearing Research*, 29(1):50–64, 1986.

[45] F. L. Wuyts, M. S. De Bodt, G. Molenberghs, M. Remacle, L. Heylen, B. Millet, K. Van Lierde, J. Raes, and P. H. Van de Heyning. The dysphonia severity index. *Journal of Speech, Language, and Hearing Research*, 43(3):796–809, 2000.

[46] M. M. Hakkesteegt, M. P. Brocaar, M. H. Wieringa, and L. Feenstra. The relationship between perceptual evaluation and objective multiparametric evaluation of dysphonia severity. *Journal of Voice*, 22(2):138–145, Mar 2008.

[47] L. R. Rabiner and R. W. Schafer. *Theory and Applications of Digital Speech Processing*. Prentice Hall, 2010.

[48] J. Hillenbrand and R. A. Houde. Acoustic correlates of breathy voice quality: Dysphonic voices and continous speech. *Journal of Speech and Hearing Research*, 39:311–321, 1996.

[49] Y. D. Heman-Ackah, D. D. Michael, and G. S. Jr Goding. The relationship between cepstral peak prominence and selected parameters of dysphonia. *Journal of Voice*, 16(1):20–27, Mar 2002.

[50] Y. D. Heman-Ackah. Reliability of calculating the cepstral peak without linear regression analysis. *Journal of Voice*, 18(2):203–208, Jun 2004.

[51] C. R. Watts and S. N. Awan. Use of spectral/cepstral analysis for differentiating normal from hypofunctional voices in sustained vowel and continuous speech contexts. *Journal of Speech, Language, and Hearing Research*, 54:1525–1537, 2010.

[52] Y. Maryn, N. Roy, M. De Bodt, P. Van Cauwenberge, and P. Corthals. Acoustic measurement of overall voice quality: A meta-analysis. *The Journal of the Acoustical Society of America*, 126(5):2619–2634, 2009.

[53] V. S. McKenna and C. E. Stepp. The relationship between acoustical and perceptual measures of vocal effort. *The Journal of the Acoustical Society of America*, 144(3):1643–1658, 2018.

[54] S. V. Narasimhan and K. Vishal. Spectral measures of hoarseness in persons with hyperfunctional voice disorder. *Journal of Voice*, 31(1):57 – 61, 2016.

[55] J. Kreiman, Y. Shue, G. Chen, M. Iseli, B. R. Gerratt, J. Neubauer, and A. Alwan. Variability in the relationships among voice quality, harmonic amplitudes, open quotient, and glottal area waveform shape in sustained phonation. *The Journal of the Acoustical Society of America*, 132(4):2625–2632, 2012.

[56] Y. S. Lien, C. R. Calabrese, C. M. Michener, E. H. Murray, J. H. Van Stan, D. D. Mehta, R. E. Hillman, J. P. Noordzij, and C. E. Stepp. Voice relative fundamental frequency via neck-skin acceleration in individuals with voice disorders. *Journal of Speech, Language, and Hearing Research*, 58(5):1482–1487, 2015.

[57] T. L. Eadie and C. E. Stepp. Acoustic correlate of vocal effort in spasmodic dysphonia. *Annals of Otology, Rhinology & Laryngology*, 122(3):169–176, 2013.

[58] E. B. Holmberg, P. Doyle, J. S. Perkell, B Hammarberg, and R. E. Hillman. Aerodynamic and acoustic voice measurements of patients with vocal nodules: Variation in baseline and changes across voice therapy. *Journal of Voice*, 17:269–282, 2003.

[59] C. M. Sapienza and E. T. Stathopoulos. Respiratory and laryngeal measures of children and women with bilateral vocal fold nodules. *Journal of Speech, Language, and Hearing Research*, 37(6):1229–1243, 1994.

[60] M. Zañartu, G. Galindo, B. D. Erath, S. D. Peterson, G. R. Wodicka, and R. E. Hillman. Modeling the effects of a posterior glottal opening on vocal fold dynamics with implications for vocal hyperfunction. *The Journal of the Acoustical Society of America*, 136:3262–3271, 2014.

[61] A. Holbrook, M. I. Rolnick, and C. W. Bailey. Treatment of vocal abuse disorders using a vocal intensity controller. *Journal of Speech and Hearing Disorders*, 39(3):298–303, 1974.

[62] J. E. Zicker, W. J. Tompkins, R. T. Rubow, and J. H. Abbs. A portable microprocessor-based biofeedback training device. *Biomedical Engineering, IEEE Transactions on*, BME-27(9):509–515, Sep. 1980.

[63] S. Ryu, S. Komiyama, S. Van Kannae, and H. Watanabe. A newly devised speech accumulator. *ORL*, 45:108–114, 1983.

[64] A. Szabo, B. Hammarberg, A Håkansson, and M. Södersten. A voice accumulator device: evaluation based on studio and field recordings. *Logopedics Phoniatrics Vocology*, 26(3):102–17, 2001.

[65] A. Ohlsson, O. Brink, and A. Lofqvist. A voice accumulator-validation and application. *Journal of Speech, Language, and Hearing Research*, 32(2):451–457, 1989.

[66] T. Masuda, Y. Ikeda, H. Manako, and S. Komiyama. Analysis of vocal abuse: Fluctuations in phonation time and intensity in 4 groups of speakers. *Acta Oto-Laryngologica*, 113(4):547–552, 1993.

[67] R. Buekers, E. Bierens, H. Kingma, and E. H. Marres. Vocal load as measured by the voice accumulator. *Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics*, 47 5:252–61, 1995.

[68] J. H. Van Stan, J. Gustafsson, E. Schalling, and R. E. Hillman. Direct comparison of three commercially available devices for voice ambulatory monitoring and biofeedback. *Perspectives on Voice and Voice Disorders*, 24:80, 07 2014.

[69] C. Manfredi, T. Bruschi, A. Dallai, A. Ferri, P. Tortoli, and M. Calisti. Voice quality monitoring: A portable device prototype. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 997–1000, Aug 2008.

[70] N. R. Smith, L. A. Rivera, M. Dietrich, C. R. Shyu, M. P. Page, and G. N. DeSouza. Detection of simulated vocal dysfunctions using complex semg patterns. *IEEE Journal of Biomedical and Health Informatics*, 20(3):787–801, May 2016.

[71] I. R. Titze, E. J. Hunter, and J. G. Švec. Voicing and silence periods in daily and weekly vocalizations of teachers. *The Journal of the Acoustical Society of America*, 121(1):469–478, 2007.

[72] T. Carroll, John Nix, E. Hunter, I. Titze, and M. Abaza. Objective measurement of vocal fatigue in classically trained singers: A pilot study of vocal dosimetry data. *Otolarynol. Head Neck Surg.*, 135(4):595–602, 2006.

[73] P. Bottalico, S. Graetzer, A. Astolfi, and E. Hunter. Silence and voicing accumulations in italian primary school teachers with and without voice disorders. *Journal of Voice*, 31, 2017.

[74] M. J. Schloneger and E. J. Hunter. Assessments of voice use and voice quality among college/university singing students ages 18-24 through ambulatory monitoring with a full accelerometer signal. *Journal of Voice*, 31(1):124.e21–124.e30, Jan 2017.

[75] J. H. Van Stan, D. D. Mehta, S. M. Zeitels, J. A. Burns, A. M. Barbu, and R. E. Hillman. Average ambulatory measures of sound pressure level, fundamental frequency, and vocal dose do not differ between adult females with phonotraumatic lesions and matched control subjects. *Annal. Otolog. Rhinol. Laryngol.*, 124(11):864–874, 2015.

[76] V. Lyberg Åhlander, D. Pelegrín García, S. Whitling, R. Rydell, and A. Löfqvist. Teachers' voice use in teaching environments: A field study using ambulatory phonation monitor. *Journal of Voice*, 28(6):841.e5–841.e15, Nov 2014.

[77] A. Carullo, A. Vallan, and A. Astolfi. Design issues for a portable vocal analyzer. *IEEE Transactions on Instrumentation and Measurement*, 62(5):1084–1093, May 2013.

[78] L. C. Cantor Cutiva, G. E. Puglisi, A. Astolfi, and A. Carullo. Four-day follow-up study on the self-reported voice condition and noise condition of teachers: Relationship between vocal parameters and classroom acoustics. *Journal of Voice*, 31(1):120.e1–120.e8, Jan 2017.

[79] I. S. Schiller, D. Morsomme, and A. Remacle. Voice use among music theory teachers: A voice dosimetry and self-assessment study. *Journal of Voice*, 32(5):578–584, Sep 2018.

[80] D. D. Mehta, H. A. Cheyne, A. Wehner, J. T. Heaton, and R. E. Hillman. Accuracy of self-reported estimates of daily voice use in adults with normal and disordered voices. *American Journal of Speech-Language Pathology*, 25(4):634–641, 2016.

[81] J. H. Van Stan, M. Maffei, M. L. Vaz Masson, D. D. Mehta, J. A. Burns, and R. E. Hillman. Self-ratings of vocal status in daily life: Reliability and validity for patients with vocal hyperfunction and a normative group. *American Journal of Speech-Language Pathology*, 26(4):1167–1177, 2017.

[82] I. R. Titze and E. J. Hunter. Comparison of vocal vibration-dose measures for potential-damage risk criteria. *Journal of Speech, Language, and Hearing Research*, 58:1425–1439, 2015.

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

[84] K. L. Marks, J. Z. Lin, A. Fox, L. E. Toles, and D. D. Mehta. Impact of non-modal 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.

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

[86] S. Hegde, S. Shetty, S. Rai, and T. Dodderi. A survey on machine learning approaches for automatic detection of voice disorders. *Journal of Voice*, In Press, 2018.

[87] A. Al-nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, K. H. Malki, and M. A. Bencherif. An investigation of multidimensional voice program parameters in three different databases for voice pathology detection and classification. *Journal of Voice*, 31(1):113.e9–113.e18, Jan 2017.

[88] A. Krishna, K. Shama, and U. Niranjan. K-means nearest neighbor classifier for voice pathology. In *Proceedings of the IEEE INDICON 2004 - 1st India Annual Conference*, pages 352– 354, 01 2005.

[89] R. Behroozmand and F. Almasganj. Comparison of neural networks and support vector machines applied to optimized features extracted from patients' speech signal for classification of vocal fold inflammation. In *Proceedings of the Fifth IEEE International Symposium on Signal Processing and Information Technology, 2005.*, pages 844–849, Dec 2005.

[90] E. S. Fonseca, R. C. Guido, A. C. Silvestre, and J. C. Pereira. Discrete wavelet transform and support vector machine applied to pathological voice signals identification. In *Seventh IEEE International Symposium on Multimedia (ISM'05)*, pages 5 pp.–, Dec 2005.

[91] J. Nayak, P.S. Bhatb, U. R. Acharya, and V. Aithal. Classification and analysis of speech abnormalities. *IRBM*, 26:319–327, 01 2005.

[92] H. I. Turkmen, M. E. Karsligil, and I. Kocak. Classification of vocal fold nodules and cysts based on vascular defects of vocal folds. In *2013 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, Sep. 2013.

[93] B. Aguiar, S. Costa, and J. Fechine. Lpc modelling and cepstral analysis applied to vocal fold pathology detection. *I. J. Functional Informatics and Personalised Medicine*, 1:156–170, 01 2008.

[94] J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, and G. Castellanos-Domínguez. An improved method for voice pathology detection by means of a hmm-based feature space transformation. *Pattern Recogn.*, 43(9):3100–3112, September 2010.

[95] R. T. S. Carvalho, C. C. Cavalcante, and P. C. Cortez. Wavelet transform and artificial neural networks applied to voice disorders identification. In *2011 Third World Congress on Nature and Biologically Inspired Computing*, pages 371–376, Oct 2011.

[96] N. Erfanian Saeedi, F. Almasganj, and F. Torabinezhad. Support vector wavelet adaptation for pathological voice assessment. *Computers in biology and medicine*, 41:822–8, 09 2011.

[97] G. Muhammad, T. A. Mesallam, K. H. Malki, M. Farahat, A. Mahmood, and M. Alsulaiman. Multidirectional regression (mdr)-based features for automatic voice disorder detection. *Journal of Voice*, 26(6):817.e19–817.e27, Nov 2012.

[98] A. Al-Nasheri, G. Muhammad, M. Alsulaiman, Z. Ali, K. H. Malki, T. A. Mesallam, and M. Farahat Ibrahim. Voice pathology detection and classification using auto-correlation and entropy features in different frequency regions. *IEEE Access*, 6:6961–6974, 2018.

[99] G. Schlotthauer, M. E. Torres, and M. Jackson-Menaldi. Automatic diagnosis of pathological voices. In *Proceedings of the 6th WSEAS International Conference on Signal, Speech and Image Processing*, 01 2006.

[100] D. Hemmerling, A. Skalski, and J. Gajda. Voice data mining for laryngeal pathology assessment. *Comput. Biol. Med.*, 69(C):270–276, February 2016.

[101] M. Markaki, Y. Stylianou, J. D. Arias-Londoño, and J. I. Godino-Llorente. Dysphonia detection based on modulation spectral features and cepstral coefficients. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5162–5165, March 2010.

[102] K. Szklanny and P. Wrzeciono. The application of a genetic algorithm in the noninvasive assessment of vocal nodules in children. *IEEE Access*, 7:44966–44976, 2019.

[103] K. Daoudi and B. Bertrac. On classification between normal and pathological voices using the meei-kaypentax database: issues and consequences. In Haizhou Li, Helen M. Meng, Bin Ma, Engsiong Chng, and Lei Xie, editors, *INTERSPEECH*, pages 198–202. ISCA, 2014.

[104] J. R. Orozco-Arroyave, F. Hönig, J. D. Arias-Londoño, J. F. Vargas-Bonilla, K. Daqrouq, S. Skodda, J. Rusz, and E. Nöth. Automatic detection of parkinson's disease in running speech spoken in three different languages. *The Journal of the Acoustical Society of America*, 139(1):481–500, 2016.

[105] H. Cordeiro and C. Meneses. Low band continuous speech system for voice pathologies identification. In *2018 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA)*, pages 315–320, Sep. 2018.

[106] K. Umapathy, S. Krishnan, V. Parsa, and D. G. Jamieson. Discrimination of pathological voices using a time-frequency approach. *Biomedical Engineering, IEEE Transactions on*, 52(3):421–430, March 2005.

[107] E. J. Hunter and I. R. Titze. Variations in intensity, fundamental frequency, and voicing for teachers in occupational versus nonoccupational settings. *Journal of Speech, Language, and Hearing Research*, 53(4):862–875, 2010.

[108] M. Ghassemi, Z. Syed, D. D. Mehta, J. H. Van Stan, R. E. Hillman, and J. V. Guttag. Uncovering voice misuse using symbolic mismatch. *Machine Learning for Healthcare Conference*, 56:239–252, 2016.

[109] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2009.

[110] S. Sahoo and A. Routray. A novel method of glottal inverse filtering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(7):1230–1241, July 2016.

[111] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing*, 50(2):174–188, Feb 2002.

[112] P. J. Hadwin, G. E. Galindo, K. J. Daun, M. Zañartu, B. D. Erath, E. Cataldo, and S. 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.

[113] P. J. Hadwin and S. 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.

[114] F. Lindstrom, K. P. Waye, M. Södersten, A. McAllister, and S. Ternström. Observations of the relationship between noise exposure and preschool teacher voice usage in day-care center environments. *Journal of Voice*, 25(2):166–172, Mar 2011.

[115] Sharon L. Morrow and Nadine P. Connor. Voice amplification as a means of reducing vocal load for elementary music teachers. *Journal of Voice*, 25(4):441–446, Jul 2011.

[116] S. S. Harper, P.and Kraman, H. Pasterkamp, and G. R. Wodicka. An acoustic model of the respiratory tract. *J. Appl. Physiol.*, 77:554–566, 2001.

[117] J. C. Ho, M. Zañartu, and G. R. Wodicka. An anatomically based, time-domain acoustic model of the subglottal system for speech production. *The Journal of the Acoustical Society of America*, 129(3):1531–1547, 2011.

[118] J. G. Proakis and D. G. Manolakis. *Digital Signal Processing: Principles, Algorithms and Applications*. Pearson Education Inc., 4th edition, 2007.

[119] K. Ishizaka, J.C. French, and J. L. Flanagan. Direct determination of vocal tract wall impedance. *IEEE Transaction on Acoustics, Speech and Signal Processing*, 23:370–373, 1975.

[120] M. 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.

[121] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Neural Networks, 1995. Proceedings., IEEE International Conference on*, volume 4, pages 1942–1948 vol.4, Nov 1995.

[122] L. R. Rabiner. *Digital Processing of Speech Signals*. Prentice Hall, 1978.

[123] M. Zañartu. *Acoustic Coupling in Phonation and its Effect on Inverse Filtering of Oral Airflow and Neck Surface Acceleration*. PhD thesis, Purdue University, West Lafayette, IN, 2010.

[124] Zhukhovitskaya A., Battaglia D., Khosla S. M., Murry T., and Sulica L. Gender and age in benign vocal fold lesions. *The Laryngoscope*, 125(1):191–196.

[125] J. S. Perkell, E. B. Holmberg, and R. 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.

[126] D. D. Mehta, M. Zañartu, J. H. Van Stan, S. W. Feng, H. A. Cheyne, and R. E. Hillman. Smartphone-based detection of voice disorders by long-term monitoring of neck acceleration features. In *IEEE 10th Annual Wearable and Implantable Body Sensor Networks Conference*, Cambridge, USA, 2013.

[127] V. M. Espinoza, D. D. Mehta, J. H. Van Stan, R. E. Hillman, and M. Zañartu. Uncertainty of glottal airflow estimation during continuous speech using impedance-based inverse filtering of the neck-surface acceleration signal. *Proceedings of the Acoustical Society of America*, 2017.

[128] P. Alku, J. Pohjalainen, M. Vainio, A. Laukkanen, and B. Story. Formant frequency estimation of high-pitched vowels using weighted linear prediction. *The Journal of the Acoustical Society of America*, 134(2):1295–1313, 2013.

[129] P. Alku, J. Horáček, M. Airas, F. Griffond-Boitier, and A. Laukkanen. Performance of glottal inverse filtering as tested by aeroelastic modelling of phonation and fe modelling of vocal tract. *Acta Acustica united with Acustica*, 92(5):717–724, 2006.

[130] J. G. Švec, I. R. Titze, and P. 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.

[131] D. H. Klatt and L. C. Klatt. Analysis, synthesis and perception of voice quality variations among male and female talkers. *The Journal of the Acoustical Society of America*, 87(2):820–856, 1990.

[132] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[133] K. E. Rieger, W. Hong, V. G. Tusher, J. Tang, R. Tibshirani, and G. Chu. Toxicity from radiation therapy associated with abnormal transcriptional responses to dna damage. *Proceedings of the National Academy of Sciences of the United States of America*, 101(17):6635–6640, 2004.

[134] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1):273 – 324, 1997.

[135] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software, Articles*, 33(1):1–22, 2010.

[136] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, 1992.

[137] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin. Liblinear: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, June 2008.

[138] J. Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale: Lawrence Erlbaum, 2nd ed. edition, 1988.

[139] G. E. Galindo, S. D. Peterson, B. D. Erath, C. Castro, R. E. Hillman, and M. Zañartu. Modeling the pathophysiology of phonotraumatic vocal hyperfunction with a triangular glottal model of the vocal folds. *Journal of Speech, Language, and Hearing Research*, 60(9):2452–2471, 2017.

[140] C. M. R. Pinho, L. M. T. Jesus, and A. Barney. Aerodynamic measures of speech in unilateral vocal fold paralysis (uvfp) patients. *Logopedics Phoniatrics Vocology*, 38(1):19–34, 2013.

[141] P. Alku. Glottal inverse filtering analysis of human voice production: A review of estimation and parameterization methods of the glottal excitation and their applications. *SADHANA - Academy Proceedings in Engineering Sciences*, 36:623–650, 2011.

[142] T. Drugman, P. Alku, A. Alwan, and Y. Yegnanarayana. Glottal source processing: From analysis to applications. *Computer Speech & Language*, 28:1117–1138, 2014.

[143] A. F. Llico, M. Zañartu, A. J. González, G. R. Wodicka, D. D. Mehta, J. H. Van Stan, and R. 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.

[144] J. Hillenbrand, R. A. Cleveland, and R. L. Erickson. Acoustic correlates of breathy vocal quality. *Journal of Speech, Language, and Hearing Research*, 37(4):769–778, 1994.

[145] J. Kreiman, B. Gerratt, M. Garellek, R. Samlan, and Z. Zhang. Toward a unified theory of voice production and perception. *Loquens*, 1(1), 2014.

[146] R. A. Samlan, B. H. Story, and K. Bunton. Relation of perceived breathiness to laryngeal kinematics and acoustic measures based on computational modeling. *Journal of Speech, Language, and Hearing Research*, 56(4):1209–1223, 2013.

[147] S. Fu, D. G. Theodoros, and E. C. Ward. Intensive versus traditional voice therapy for vocal nodules: Perceptual, physiological, acoustic and aerodynamic changes. *Journal of Voice*, 29(2):260.e31–260.e44, Mar 2015.

[148] C. Aronsson, M. Bohman, S. Ternström, and M. Södersten. Loud voice during environmental noise exposure in patients with vocal nodules. *Logopedics Phoniatrics Vocology*, 32(2):60–70, 2007.

[149] A. E. Aronson and D. M. Bless. *Clinical Voice Disorders*. Thieme Medical Pub, 2009.

[150] V. Espinoza. *Stationary And Dynamic Aerodynamic Assessment Of Vocal Hyperfunction Using Enhanced Supraglottal And Subglottal Inverse Filtering Methods*. PhD thesis, Universidad Técnica Federico Santa María, Valparaíso, Chile, 2018.

[151] L. Ljung. *System Identification: Theory for the User*. Prentice Hall, 1987.

[152] D. Simon and Y. S. Shmaliy. Unified forms for kalman and finite impulse response filtering and smoothing. *Automatica*, 49(6):1892 – 1899, 2013.

[153] D. Simon. *Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches*. Wiley-Interscience, New York, NY, USA, 2006.

[154] B.D.O. Anderson and J.B. Moore. *Optimal Filtering*. Dover Books on Electrical Engineering. Dover Publications, 2012.

[155] J. Benesty, M. M. Sondhi, and Y. Huang. *Springer Handbook of Speech Processing*. Springer-Verlag, Berlin, Heidelberg, 2007.

[156] C. K. Chui and G. Chen. *Kalman Filtering with Real-time Applications*. Springer-Verlag New York, Inc., New York, NY, USA, 1987.

[157] S. Gannot, D. Burshtein, and E. Weinstein. Iterative and sequential kalman filter-based speech enhancement algorithms. *IEEE Transactions on Speech and Audio Processing*, 6(4):373–385, July 1998.

[158] G. Alzamendi and G. Schlotthauer. Modeling and joint estimation of glottal source and vocal tract filter by state-space methods. *Biomedical Signal Processing and Control*, 37, 01 2017.

[159] K. Myers and B. Tapley. Adaptive sequential estimation with unknown noise statistics. *IEEE Transactions on Automatic Control*, 21(4):520–523, August 1976.

[160] L. Deng, L.J. Lee, H. Attias, and A. Acero. Adaptive kalman filtering and smoothing for tracking vocal tract resonances using a continuous-valued hidden dynamic model. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(1):13–23, 2007.

[161] D. D. Mehta, D. Rudoy, and P. J. Wolfe. Kalman-based autoregressive moving average modeling and inference for formant and antiformant tracking. *The Journal of the Acoustical Society of America*, 132(3):1732–1746, 2012.

[162] G. Fant. *The Acoustic Theory of Speech Production*. Mouton & Co. N.V. Publishers, 1960.

[163] A. E. Rosenberg. Effect of glottal pulse shape on the quality of natural vowels. *The Journal of the Acoustical Society of America*, 49(2B):583–590, 1971.

[164] G Fant, J Liljencrants, and Q. Lin. A four-parameter model of glottal flow. *STL-QPSR*, 4, 01 1985.

[165] B. Doval, C. D'Alessandro, and N. Henrich Bernardoni. The spectrum of glottal flow models. *Acta Acustica united with Acustica*, 92(6):1026–1046, 2006.

[166] J.D. Markel and A.H. Gray. *Linear Prediction of Speech*. Communication and cybernetics. Springer-Verlag, 1976.

[167] A. V. Oppenheim, R. Schafer, and J. Buck. *Discrete-Time Signal Processing*. Prentice Hall, Englewood Cliffs, N.J., 1999.

[168] M.S. Grewal and A.P. Andrews. *Kalman filtering: theory and practice using MATLAB*. [Wiley InterScience Online Books, Electronic and Electrical Engineering Collection]. Wiley, 2001.

[169] J. Benesty, M. Sondhi, and Y. Huang. *Springer Handbook of Speech Processing*. Springer, 2008.

[170] P. Alku, T. Bäckström, and E. Vilkman. Normalized amplitude quotient for parametrization of the glottal flow. *The Journal of the Acoustical Society of America*, 112(2):701–710, 2002.

[171] J. Snoek, H. Larochelle, and R. P. Adams. Practical bayesian optimization of machine learning algorithms. In *Advances in neural information processing systems*, pages 2951–2959, 2012.

[172] D. D. Mehta, S. M. Zeitels, J. A. Burns, A. D. Friedman, D. D. Deliyski, and R. E. Hillman. High-speed videoendoscopic analysis of relationships between cepstral-based acoustic measures and voice production mechanisms in patients undergoing phonomicrosurgery. *Ann. Otol. Rhinol. Laryngol.*, 121(5):341–347, 2012.

[173] 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*, 2019. accepted for publication.

[174] D. D. Mehta, M. Zañartu, T. F. Quatieri, D. D. Deliyski, and R. E. Hillman. Investigating acoustic correlates of human vocal fold phase asymmetry through mathematical modeling and laryngeal high-speed videoendoscopy. *The Journal of the Acoustical Society of America*, 130:3999–4009, 2011.

[175] S. N. Awan, N. Roy, M. E. Jetté, G. S. Meltzner, and R. E. Hillman. Quantifying dysphonia severity using a spectral/cepstral-based acoustic index:

Comparisons with auditory-perceptual judgements from the cape-v. *Clinical Linguistics & Phonetics*, 24(9):742–758, 2010.

[176] L. Wassermann. *All of Statistics: A concise Course in Statistical Inference*. Springer, 2010.

[177] D. D. Mehta, J. H. Van Stan, and R. E. Hillman. Relationships between vocal function measures derived from an acoustic microphone and a subglottal neck-surface accelerometer. *Audio Speech and Language Processing, IEEE/ACM Transactions on*, 24(4):659–668, April 2016.

[178] K. Ishizaka, M. Matsuidara, and T. Kaneko. Input acoustic-impedance measurement of subglottal system. *The Journal of the Acoustical Society of America*, 60:190–197, 1976.

[179] E. R. Weibel. *Morphometry of the Human Lung*. Springer, New York, 1963.

[180] F. Contreras, P. Prieto, and J. Valdés. Intrumentos de evaluación fonoaudiológica. una aproximación a las metodologías de evualuación. Technical Report Serie Creación número 15, Unidad Habla y Lenguaje Adultos, Carrera de Fonoaudiología. Facultad de Ciencias de la Salud. Centro de investigación en Educación Superior CIES- USS, 2017.

[181] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[182] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec 2015.

[183] L. Torrey and J. Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI Global, 2010.

[184] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[185] W. Dai, Q. Yang, G. Xue, and Y. Yu. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, pages 193–200. ACM, 2007.

[186] Y. Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36, 2012.

[187] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions, 2014.

[188] O. Rioul and M. Vetterli. Wavelets and signal processing. *Signal Processing Magazine, IEEE*, 8:14 – 38, 10 1991.

[189] S. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Elsevier Science, 2008.

[190] M. Holschneider. *Wavelets: an analysis tool*. Oxford mathematical monographs. Clarendon Press, 1995.

[191] J. M. Lilly and S. C. Olhede. On the analytic wavelet transform. *IEEE Trans. Inf. Theor.*, 56(8):4135–4156, August 2010.

[192] J. M. Lilly and S. C. Olhede. Higher-order properties of analytic wavelets. *IEEE Transactions on Signal Processing*, 57(1):146–160, Jan 2009.

[193] J. M. Lilly and S. Olhede. Generalized morse wavelets as a superfamily of analytic wavelets. *IEEE Transactions on Signal Processing*, 60, 03 2012.

[194] F. Chollet. *Deep Learning with Python*. Manning Publications Co., Greenwich, CT, USA, 1st edition, 2017.

[195] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958, 2014.

[196] A. Nemirovski and D. Yudin. On Cezari's convergence of the steepest descent method for approximating saddle point of convex-concave functions. In *Soviet Math. Dokl*, volume 19, pages 258–269, 1978.

[197] M. Engin. ECG beat classification using neuro-fuzzy network. *Pattern Recognition Letters*, 25(15):1715 – 1722, 2004.

[198] Zhao, Q. and Zhang, L. ECG feature extraction and classification using wavelet transform and support vector machines. In *2005 International Conference on Neural Networks and Brain*, volume 2, pages 1089–1092, Oct 2005.

[199] T. Li and M. Zhou. ECG classification using wavelet packet entropy and random forests. *Entropy*, 18(8):285, Aug 2016.

[200] K. Pearson. Das fehlergesetz und seine verallgemeiner-ungen durch fechner und pearson*. A rejoinder. *Biometrika*, 4(1-2):169–212, 06 1905.