UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTMENT OF ELECTRONIC ENGINEERING

Stationary and dynamic aerodynamic assessment of vocal hyperfunction using enhanced supraglottal and subglottal inverse filtering methods

A dissertation submitted by

Víctor M. Espinoza

in partial fulfillment of the requirement for the degree of

Doctor of Philosophy
in Electronic Engineering

Thesis Advisor
Matías Zañartu, Ph.D.

Valparaiso, 2018.

UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA
DEPARTAMENTO DE ELECTRÓNICA

Evaluación aerodinámica estacionaria y dinámica de la hiperfunción vocal utilizando métodos de filtrado inverso supraglotal y subglotal

Documento entregado por

Víctor M. Espinoza

como requerimiento parcial para la obtención del grado de

Doctor en Ingeniería Electrónica

Profesor Guía
Matías Zañartu, Ph.D.

Valparaiso, 2018.

TITLE:

Stationary and dynamic aerodynamic assessment of vocal hyperfunction using enhanced supraglottal and subglottal inverse filtering methods

AUTHOR:

Víctor M. Espinoza

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Electronic Engineering at Universidad Técnica Federico Santa María.

Documento enviado como requerimiento parcial para la obtención del grado de Doctor en Ingeniería Electrónica de la Universidad Técnica Federico Santa María.

Matías Zañartu, Ph.D. _____

Juan I. Yuz, Ph.D. _____

Cara E. Stepp, Ph.D. _____

Daryush D. Mehta, Ph.D. _____

Stationary and dynamic aerodynamic assessment of vocal hyperfunction using enhanced supraglottal and subglottal inverse filtering methods

Author

Víctor M. Espinoza

Thesis Advisor

Matías Zañartu, Ph.D.

Abstract

This thesis describes the guidelines, experimental design, and initial results for stationary and dynamic aerodynamic assessments of vocal hyperfunction. This work aims to improve the understanding and clinical assessment of vocal hyperfunction by advancing current methods for the inverse filtering of both oral airflow and neck skin acceleration signals and by incorporating statistical analysis tools in this framework. New algorithms to perform inverse filtering including a frame-based approach are explored and applied in an automatic framework to estimate multiple aerodynamic measures of vocal function, and later utilized with enhanced clinical methods to differentiate vocal hyperfunction from normal vocal behavior. To achieve this goal, a clinical assessment is performed through aerodynamic, vibroacoustic, and acoustic recordings of vocal function in laboratory conditions. Various methods to improve the estimation of aerodynamic measures using different vocal gestures are explored, including sustained vowels and continuous speech. Selected inverse filtering techniques were enhanced to estimate glottal airflow in normal and pathological voices, wherein many of them constitute the most challenging conditions for inverse filtering, namely female voices during running speech. The underlying hypothesis considers that supraglottal inverse filtering methods can be enhanced under challenging conditions by limiting the signal bandwidth down to the first formant. For estimating inverse filtering quality, several error metrics that are based on the signal behavior and the contrast with glottal waveform simulations, are proposed and studied. In addition, the subglottal impedance-based inverse filtering (IBIF) scheme is explored in the context of running speech. The

automated supraglottal inverse filtering technique was implemented in a frame-based framework to evaluate the uncertainties of the IBIF model parameters and the derived aerodynamic measures from a neck skin acceleration signal. Robust statistical analysis was used to reduce the influence of sporadic events and outliers for the objective parameter estimation. Accounting for the uncertainties of glottal airflow from the neck skin acceleration signal would allow for improving the subglottal inverse filtering, and future directions for this approach are discussed to advance the aerodynamic ambulatory monitoring of vocal function. Finally, an updated dataset of aerodynamic measures was derived with the proposed supraglottal and subglottal methods along with a robust approach to differentiate hyperfunctional patients with paired matched controls, using multivariate statistical models based on oral airflow and neck acceleration recordings.

Evaluación aerodinámica estacionaria y dinámica de la hiperfunción vocal utilizando métodos de filtrado inverso supraglotal y subglotal

Autor

Víctor M. Espinoza

Profesor Guía

Matías Zañartu, Ph.D.

Resumen

Esta tesis describe los lineamientos, el diseño experimental y los resultados iniciales de evaluaciones aerodinámicas estacionarias y dinámicas de la hiperfunción vocal. Este trabajo tiene como objetivo avanzar en la comprensión y evaluación clínica de la hiperfunción vocal, mediante el mejoramiento de los métodos actuales de filtrado inverso para las señales de flujo oral y de aceleración de la piel del cuello, incorporando herramientas de análisis estadístico en esta estructura. Nuevos algoritmos de filtrado inverso, incluyendo uno basado en segmentos de señal, son explorados y aplicados de manera automática para estimar múltiples medidas aerodinámicas de la función vocal que luego son utilizadas con métodos clínicos con el fin de diferenciar la hiperfunción vocal respecto a un comportamiento normal de la voz. Para lograr este objetivo, se realiza una evaluación clínica de la función vocal a través de grabaciones aerodinámicas, vibroacústicas y acústicas, en condiciones de laboratorio. Varios métodos son explorados para mejorar la estimación de medidas aerodinámicas, incluyendo vocales sostenidas y habla continua. Una selección de técnicas de filtrado inverso fueron refinadas para estimar el flujo de aire glotal en voces normales y patológicas, donde muchas de ellas constituyen las condiciones más desafiantes para el filtrado inverso, como son las voces femeninas en habla continua. La hipótesis latente considera que los métodos de filtrado inverso supraglotal pueden ser perfeccionados en condiciones difíciles al limitar el ancho de banda de la señal al primer formante. En este contexto, múltiples métricas de error son propuestas y estudiadas con el fin de estimar la calidad del filtrado inverso, las cuales son fundamentadas en el comportamiento esperado de la señal

filtrada y el contraste con simulaciones numéricas del impulso glotal en estado estacionario. Agregado a lo anterior, el esquema de filtrado inverso basado en la impedancia subglotal (del inglés impedance-based inverse filtering, IBIF) es explorada en el contexto de habla continua. Las técnica de filtrado inverso supraglotal automática fue implementada en un enfoque basado en segmentos de señal con el fin de evaluar las incertezas de los parámetros del modelo IBIF y las medidas aerodinámicas derivadas de una señal de aceleración de la piel del cuello. Se utilizó un análisis estadístico robusto para reducir la influencia de eventos esporádicos y valores atípicos para la estimación objetiva de estos parámetros. Tomando en cuenta las incertezas de la señal del flujo de aire glotal y de aceleración de la piel del cuello, es posible mejorar el filtrado inverso subglotal, y en este sentido, se discuten las direcciones futuras de este enfoque para avanzar en el monitoreo ambulatorio aerodinámico de la función vocal. Finalmente, se obtuvo un conjunto de datos actualizado de medidas aerodinámicas con los métodos supraglotal y subglotal propuestos, junto con un enfoque robusto para diferenciar pacientes hiperfuncionales de controles pareados, utilizando modelos estadísticos multivariados basados en grabaciones del flujo de aire oral y de aceleración del cuello.

Acknowledgements

My deepest gratitude to my advisor, Prof. Matías Zañartu, for his guidance and encouragement during my graduate studies. His leadership and supervision contributed significantly to the final version of this manuscript. I appreciate the encouraging comments and feedback from my thesis committee, Prof. Cara E. Stepp, Prof. Juan I. Yuz and Prof. Daryush D. Metha, which largely improved the final version of this dissertation. A special thank to Prof. Robert E. Hillman for his support to my fruitful and enjoyable time as a visiting researcher at the Center for Laryngeal Surgery and Voice Rehabilitation at Massachusetts General Hospital, Boston, USA. I am grateful to my fellows members of the Voice Production Lab at UTFSM, Manuel Díaz, Andrés Llico, Gabriel Galindo, Juan Mucarquer, Rodrigo Manriquez, Christian Castro, and Juan P. Cortés, for their friendship during my time at UTFSM. I also appreciate the support of my colleagues at the Department of Sound (University of Chile), Sergio Floody, Luis Núñez, and Javier Jaimovich during this period. Funding for my doctoral studies was obtained through Comisión Nacional de Ciencia y Tecnología (CONICYT-Chile), Universidad Técnica Federico Santa María, MECESUP (UTFSM), FONDECYT (CONYCIT), PIIC programs (UTFSM), and Universidad de Chile.

On a personal level, me and my wife, appreciate the support and love of Julia and Ricardo during this period. I thank my parents Manuel and Noemí, and my brother Edgardo for their infinite love and support in my life. My beloved son and daughter, Tomás and Fernanda, were my inspiration for succes. Last but not least, my profound gratitude goes to my wife, Jocelyn, for her endless support and love. Without her, this journey would have never been possible.

Contents

List of Figures

List of Tables

Abbreviations

SNR	**S**ignal to **N**oise **R**atio
IBIF	**I**mpedance-**B**ased **I**nverse **F**iltering
IF	**I**nverse **F**iltering
CV	**C**ircumferentially **V**ented (mask)
CAPE-V	**C**onsensus **A**uditory-**P**erceptual **E**valuation of **V**oice
V-RQOL	**V**oice-**R**elated **Q**uality **o**f **L**ife
AR	**A**uto-**R**egressive (model)
std	**St**an**d**ard error or deviation
pdf	**P**robability **d**ensity **f**unction
ACFL	**A**lternating **C**urrent **fl**ow or AC-flow
MFDR	**M**aximum **F**low **D**eclination **R**ate
SGP	**S**ub**g**lottal **P**ressure
OQ	**O**pen **Q**uotient
CPP	**C**epstral **P**eak **P**rominence
H1H2	**H**armonic **1** to **H**armonic **2** ratio
SQ	**S**peed **Q**uotient
CQ	**C**losing **Q**uotient
ACFL'	SPL-Normalized **AC-fl**ow
MFDR'	SPL-Normalized **M**aximum **F**low **D**eclination **R**ate
SGP'	SPL-Normalized **S**ub**g**lottal **P**ressure
OQ'	SPL-Normalized **O**pen **Q**uotient
$\mathbf{ACFL_L}$	log-transformed **AC-fl**ow
$\mathbf{MFDR_L}$	log-transformed **M**aximum **F**low **D**eclination **R**ate
$\mathbf{SGP_L}$	log-transformed **S**ub**g**lottal **P**ressure
PVH	**P**honotraumatic **V**ocal **H**yperfunction

NPVH	**N**on **P**honotraumatic **V**ocal **H**yperfunction
VHM	**V**oice **H**ealth **M**onitor
OVV	**O**ral **V**olume **V**elocity
GVV	**G**lottal **V**olume **V**elocity
dGVV	time-**d**erivative of **G**lottal **V**olume **V**elocity
ACC	neck skin **ACC**eleration (sensor or signal)
SPL	**S**ound **P**ressure **L**evel
NAQ	**N**ormalized Amplitude **Q**uotient
MIC	**MIC**rophone (sensor)
EGG	**E**lectro**g**lotto**g**raphic (sensor)
FLO	**Flo**w (sensor)
PRE	**Pre**ssure (sensor)
RMS	**R**oot **M**ean **S**quare
SNF	**S**ingle **N**otch **F**ilter
CPIF	**C**losed **P**hase **I**nverse **F**iltering
RCPIF	**R**egularized **C**losed **P**hase **I**nverse **F**iltering
CPIF	**C**losed **P**hase **I**nverse **F**iltering
DYPSA	**D**ynamic **P**rogramming **P**rojected **P**hase-**S**lope **A**lgorithm
YAGA	**Y**et **A**nother **G**CI **A**lgorithm
SEDREAMS	**S**peech **E**vent **D**etection using the **R**esidual **E**xcitation **A**nd a **M**eanbased **S**ignal
LPC	**L**inear **P**rediction **C**ode
NWAE	**N**ormalized **W**eighted **A**bsolute **E**rror
MTD	**M**uscle **T**ension **D**ysphonia
PGO	**P**article **S**warm **O**ptimization
KDE	**K**ernel **D**ensity **E**stimator
RP	**R**ainbow **P**assage
AM	**A**erodynamic **M**easure
CP	**C**losed **P**hase
SD	**S**pectral **D**istance
FFT	**F**ast **F**ourier **T**ransform
DFT	**D**iscrete **F**ourier **T**ransform

Symbols

f_0	fundamental frequency	Hz
T_0	fundamental period	s
u_g	glottal volume velocity (time domain)	cm^3s^{-1}
$\tilde{u}_g$	u_g estimation derived from IF oral volume velocity	cm^3s^{-1}
$\hat{u}_g$	u_g estimation derived from IF neck skin acceleration	cm^3s^{-1}
$\delta(n)$	unitary Kronecker impulse response	
$\hat{P}_x^{-1}(\cdot)$	empirical percentile of vector signal x	
x_{IF}	Inverse filtered vector signal x	
ω	angular frequency	rad^{-1}
α	significance level	dimensionless
$\underline{\alpha}$	Gauss window shape parameter	dimensionless
Δ	discrete-forward derivative operator	
$\delta(n)$	unitary impulse	
a_k	k-th coefficient of an AR model	dimensionless
b_k	k-th coefficient of an MA model	dimensionless
e_n	error at discrete-time n	
θ	Normalized frequency	rad/samples
θ_c	Normalized central frequency of SNF filter	rad/samples
$\lvert\cdot\rvert$	Absolute value operator	
z	complex variable used in z-transform	
z_p	pole associated to θ_c in z-domain	
$H(z)$	Transfer function in z-domain	
X	Spectrum	
$\mathbf{x}$	sample speech vector	

t_1	opening time of glottal pulse	s
t_2	closing time of glottal pulse	s
$\propto$	proportional to	
f_s	sampling frequency	Hz
a	coefficients in an AR model	
c	constrained coefficients in an AR model	
$\mathbf{a}$	vector of a_k coefficients in an AR model	
$\mathbf{c}$	vector of constrained c_k coefficients in an AR model	
$E(\cdot)$	square error of the argument	
Φ	sample covariance matrix	
p	order of AR model, or p-value used in hypothesis test	
l_0	gain value at DC ($\omega = 0$)	
l_ω	gain value at $\omega = \pi$	
$\mathbf{\Gamma}$	constrait matrix for CPIF and RCPIF methods	
$\mathbf{g}$	Lagrange multiplier vector	
$*$	discrete convolution operator	
n	discrete time	s/samples
$\ln(\cdot)$	natural logarithm operator	
$\mathcal{Z}\{\cdot\}$	$z-$transform operator	
$\mathcal{Z}^{-1}\{\cdot\}$	inverse $z-$transform operator	
$\mathcal{F}^{-1}\{\cdot\}$	inverse Fourier-transform operator	
T_{skin}	transfer function of the neck skin	dimensionless
IT_{skin}	inverted transfer function of neck skin properties	dimensionless
$\dot{U}_{skin}(\omega)$	frequency domain neck skin acceleration signal	$\mathrm{cm} \cdot \mathrm{s}^{-2}$
A_{acc}	neck skin accelerometer effective area	cm^2
M_{acc}	mechanical mass	gr
R_m	mechanical resistance per mass unit	$(\mathrm{g} \cdot \mathrm{s}^{-1} \cdot \mathrm{cm}^{-2})$
M_m	mechanical mass per mass unit	$(\mathrm{g} \cdot \mathrm{cm}^{-2})$
K_m	mechanical stiffness per mass unit	$(\mathrm{dyn} \cdot \mathrm{cm}^{-3})$
$L_{trachea}$	trachea length	cm
L_{sub1}	neck skin accelerometer position	cm
$Z_{skin_{acc}}$	neck skin acoustical impedance	$(\mathrm{g} \cdot \mathrm{s}^{-1} \cdot \mathrm{cm}^{-4})$

Z_{rad}	neck skin radiation impedance per unit area	$(\mathrm{g} \cdot \mathrm{s}^{-1} \cdot \mathrm{cm}^{-2})$
U_{sub1}	volume velocity of sub1 system for IBIF model	m^3/s
U_{sub}	volume velocity of sub system for IBIF model	m^3/s
H_{sub1}	transfer function at sub1 system for IBIF model	dimensionless
Z_{sub2}	acoustical impedance of sub2 system for IBIF model	$(\mathrm{g} \cdot \mathrm{s}^{-1} \cdot \mathrm{cm}^{-4})$
Q	scale factor for IBIF model	
$\mathbf{Q}$	Q parameter vector	
$g_r(\cdot)$	synthesized glottal-pulse function	
F_1	true first formant	Hz
$\hat{F}_1$	estimated first formant	Hz
F_L	lower first formant	Hz
F_H	higher first formant	Hz
w_k	Kaiser smooth window	
w_i	i-th weighted coefficient	
d_0	delay amount	samples
r	Pearson correlation coefficient	dimensionless
d	Cohen's effect size value	
D	Mahalanobis distance	
z_i	i-th z-score value	
$\mathcal{L}(\cdot)$	likelihood function of the argument	

Dedicated to Jocelyn, Tomás and Fernanda.
∞

Chapter 1

Aims and Motivation

1.1 Motivation

Voice disorders are a health problem of growing interest in our society. In the United States, nearly 6.6 % of workers are afflicted with voice problems [1] and in Chile is the one of the two most prevalent occupational health disorders each year, mainly affecting women [2]. Vocal fold (VF) pathologies require either surgical procedures or frequent voice therapy, with results largely depends on the ability to identify and modify specific vocal behaviors. Many of the most common vocal pathologies are usually preceded by an alteration known as *vocal hyperfunction* (VH) [3, 4]. VH is associated with many chronic and recurring conditions that are likely to result from inappropriate patterns of vocal behavior. Hyperfunctional subjects may incorrectly compensate for a deficit in voice production (e.g., reduced loudness due to incomplete glottal closure [5, 6], with excessive muscle tension, subglottal pressure, or other mechanisms causing suboptimal operation of the vocal folds [7]. This type of operation contributes to a "vicious cycle" leading, for instance, to the formation of benign vocal fold trauma, such as nodules and polyps [8]. Such lesions often prevent complete closure of the glottis, producing a breathy voice, vocal fatigue [9], and an increased phonation threshold pressure [10], thus aggravating the initial onset condition of VH. Clinical assessment of vocal function is crucial to evaluate the presence of vocal hyperfunction before and after a vocal intervention (e.g., voice therapy, surgery). However, behaviorally-based pathologies are difficult to assess with conventional perceptual methods. Thus, there is a growing interest in the application of objective measures of vocal function [4, 7], for studying vocal hyperfunction. Objective assessment can provide tools to support subjective

clinical impressions and patient self-report, and to improve the assessment of the effectiveness of surgery and vocal therapy [11]. The most consistent objective evidence concerning vocal hyperfunction showing differences between normal and hyperfunctional behavior reported up-to-date had been obtained through aerodynamic measures of the vocal function [12, 13]. The aerodynamic assessment of vocal function has been suggested to be capable of providing insights regarding the detection and nature of vocal hyperfunction [3, 14, 15]. These measures were obtained during sustained vowels and provided sufficient information to differentiate hyperfunctional from normal voice production when contrasted with a normative set using a z-score assessment [3]. However, most findings related to vocal hyperfunction have never been fully validated for homogeneous groups and matched controls that are large enough for formal statistical testing. The aerodynamic assessment uses indirect measures of vocal function that are obtained from a circumferentially vented (CV) mask. The aerodynamic signal from this mask needs to be inverse filtered to eliminate vocal tract resonances to obtain an estimate of the glottal volume velocity (GVV). Multiple inverse filtering methods have been proposed, including filter banks [16], parametric [17, 18], nonlinear [19], multi-channel [20], adaptive [21], iterative [22], and cepstrum-based [23] methods. However, these inverse filtering approaches are based on experimental data and theoretical models that yield a good fit for normal male voices.

These methods have not been successful in more challenging conditions [24, 25], such as female and pathological voices. Female voices have short pitch periods and less harmonic content in the bandwidth of interest, thus making the estimation and removal of the vocal tract effects difficult. Pathological voices do not follow standard physiological and aerodynamic patterns, thus violating the typical assumptions of parametric and other methods.

Under clinical scenarios, a single antiformant filter of conjugated-poles [18, 26] remains as the most comprehensively tested inverse filtering method for pathological and female voices, though the approach has not been validated and its accuracy remains unassessed. In addition, the dynamic nature of speech (e.g., coarticulation, vocal fold configuration, among other factors) severely alters the glottal aerodynamics, thus creating a challenging inverse filtering scenario. Extending the aerodynamic assessment from sustained vowels to continuous speech could provide further insights into the actual relevance of specific parameters, and new approaches and research questions to explore for the aerodynamic assessment of vocal function. Inverse filtering of running speech

sounds continues to be a topic of research and the accuracy of inverse filtering methods has not been tested in this scenario. A different approach referred to as "subglottal inverse filtering", consists of removing the subglottal and neck skin resonances from recordings made on a neck surface acceleration sensor [27, 28]. Subglottal and neck skin resonances are not expected to vary in time as much as those from the supraglottal tract, thus making this method also attractive for the ambulatory assessment and continuous-speech analysis [28]. However, subglottal models for inverse filtering are more complex to compute, need a subject-specific calibration, only provide unsteady (AC) aerodynamic measures, and remain untested in continuous-speech scenarios. Estimating the uncertainties of this method allows for quantifying the impact on the estimated aerodynamic measures, extending it to more sophisticated data analysis scenarios (e.g., time-series).

In addition, the assessment of inverse filtering (IF) methods is mainly based on the minimization of mean square error (MSE), which has well-known problems handling outliers and when model assumptions do not hold (e.g., the normality of residuals). Incorporating automatic IF assessment with cost functions based on trained-user criteria (e.g., reduction of the formant ripple in the closed phase), could improve processing, accuracy, and robustness of IF methods for challenging cases. Some of these conditions include short glottal closed phase, high pitch and pathological voices, and overfit, and are all key to voice assessment for female voices. The assessment of IF accuracy can be accomplished using synthesized waveforms and self-sustained models of the voice production. These bio-inspired models can provide theoretically-based true glottal airflow waveforms to evaluate both proposed IF methods and objective measures to qualify IF methods as well. Also, recent development in self-sustained models of vocal production allows for modeling vocal hyperfunction [5, 6], which could yields insights to enhance IF methods for challenging cases like female/pathological voices. Automatic IF methods have become a necessity in running speech scenarios and when comprehensive databases are analyzed (e.g., machine learning or big-data). Establishing a framework that allows for articulating IF methods and aerodynamic measure estimates in a reliably in reasonably an accurate fashion, and with a reduced computational cost could improve overall assessment of vocal function. Incorporating large data sets can allow enhanced statistical power, novel findings, and pursuit of new directions in the research field. In a relevant prior study (Hillman et al, 1989) [3], several aerodynamic measures were used to differentiate normal from hyperfunctional behavior using univariate hypothesis

tests, simple pair-wise correlations, and z-score analysis. With additional data, other statistical methods become attractive and can take into account the relationship between groups using all features in an n-dimensional space, increasing the discriminatory power of the aerodynamic measures.

Ambulatory monitoring devices have been developed to capture and store voice recordings from a light-weight accelerometer attached to the neck skin, capturing and storing several days of sound pressure level and fundamental frequency data (along with related dose/time measure) in an in-field continuous-speech scenario [29–31]. These devices were recently extended into a voice health monitor (VHM) to record the raw accelerometer signal and extract additional measures of interest [12] including aerodynamic measures [32] using subglottal inverse filtering methods [28], which is an improvement for the voice health evaluation outside of the clinic. In order to validate aerodynamic estimates from the VHM device with this approach, baseline data obtained from the CV mask could be compared with that of the VHM in controlled static (sustained vowels) and dynamic (rainbow passage) scenarios. The latter is relevant to advance toward the task of precisely pinpoint the time and lapse of hyperfunctional behaviors in an ambulatory assessment and objectively assess patient compliance with therapy goals.

1.2 Aims

The general aim of this thesis is to improve the aerodynamic assessment of vocal function in stationary and dynamic conditions, differentiating normal from pathological voices using robust inverse filtering methods and statistical tools.

1. **Specific Aim 1 (SA1)**: To develop robust and automatic methods for supra and subglottal inverse filtering for both sustained vowels and running speech with a focus on female pathological voices.
2. **Specific Aim 2 (SA2)**: To evaluate the proposed supraglottal inverse filtering methods for sustained vowels in clinical experiments that can assess the discriminatory power and clinical relevance of the proposed aerodynamic measures.
3. **Specific Aim 3 (SA3)**: To evaluate the proposed subglottal inverse filtering methods for sustained vowels and running speech in clinical experiments that can provide a first approximation of the clinical relevance.

1.3 Hypotheses

H1: It is possible to enhance the supra and subglottal inverse filtering methods used in voice assessment to allow for consistent estimations of glottal airflow under challenging conditions by limiting the bandwidth of the signals.

H2: It is possible to estimate the uncertainties of glottal airflow from the neck skin acceleration signal by exploring the signal behavior in running speech. The analysis of the signal uncertainty is expected to improve the calibration of the subglottal IBIF scheme.

H3: The aerodynamic assessment allows for discriminating patients with vocal hyperfunction from their matched controls with significant statistical differences.

1.4 Overview of the proposed methods

The estimation of glottal airflow using supra or subglottal methods is challenging topic, especially when considering female voices, pathological cases, and running speech. Female voices have short pitch periods and less harmonic content in the bandwidth of interest (0 to 5kHz), which makes it difficult to estimate the vocal tract resonances [33, 34]. Most inverse filtering studies have avoided pathological voices, since the results are hard to validate and very little is known about their aerodynamic behavior. Likewise, inverse filtering studies have primarily focused on studying sustained vowels under the assumption of stationary conditions and not running speech. Oral methods have to handle rapid time-varying tract configurations, which produce problems in the stationary IF schemes. In addition, most methods require some manual adjustment or expert assessment, which becomes prohibitive when processing larger databases. Several inverse filtering methods for the supra and subglottal systems were developed and studied in this thesis. Known techniques were revisited for their effectiveness in clinical experiments performed with sustained vowels [18], as well as their potential applicability in clinical scenarios [17, 23]. Simple Notch Filter (SNF), Linear Prediction, and Homomorphic Signal Processing approaches were investigated in detail and adapted to the context of oral airflow inverse filtering of pathological female voices. Experiments with parametric synthesized oral waveforms and physically-inspired numerical models were used as a baseline to assess the proposed methods. Deviation from baseline is reported using new error metrics that are proposed to appraise the subjective quality of the IF methods, and to implement the scheme in an automatic framework that can handle large databases. Inspired by prior studies in the field of hyperfunctional voice assessment [3, 18, 25, 26], we performed experiments in a large group of normal and pathological female voices to investigate the proposed methods for the aerodynamic assessment of vocal function. To prevent the influence of outliers or unreliable data, statistical analysis was performed using a robust z-score classifications scheme and hypotheses tests using multivariate methods.

Validation of glottal airflow estimates from a neck-surface accelerometer in continuous speech presents a series of challenges since the reference signal, namely the inverse filtered oral airflow signal, has a number of problems due to its time varying nature. In contrast, subglottal inverse filtering is subject to less temporal variations since subglottal resonances are essentially time-invariant [27, 28, 35], thus providing a greater potential to perform better in a running

speech task. Subglottal inverse filtering was explored in the context of the impedance based inverse filtering (IBIF) scheme [27, 28, 35] and cepstral methods [23, 36–39]. The main challenge for the former was to translate the scheme from a sustained-vowel scenario to a running speech one. In this regard, the estimation of IBIF model parameters was re-designed using multiple tokens and running speech. To account for the variance in the IBIF model parameter estimates, a statistical model was built where uncertainties were investigated along with aerodynamic measures derived from the neck skin acceleration signal.

1.5 Contributions

The contributions of this work are the follows.

Validation of early methods of clinical voice assessment incorporating enhanced signal processing tools and an automatic framework to improve the objective assessment of clinical voice evaluation.

A reference dataset of aerodynamic measures was estimated from both a group of normal and hyperfunctional voices, all for female subjects. These data could be used as a reference for future research in voice assessment, e.g., subject-specific numerical models of voice production towards clinical applications.

Enhanced methods for inverse filtering oral airflow and neck surface acceleration were explored in challenging inverse filtering conditions, such as pathological female voices in running speech. In this regard, three IF methods were studied and enhanced to improve the aerodynamic measure estimation in static and dynamic scenarios. These IF methods were, a single notch filter (SNF) with a proposed set of cost functions, referred as SNF+Metrics, a regularized closed phase inverse filtering (RCPIF), and a non-parametric subglottal IF method. To asses these IF methods, a specific framework was developed. The design of new cost functions to assess IF methods were extensively tested, evaluated, and validated to improve the aerodynamic measures estimation. All the proposed inverse filtering methods were assessed with synthesized and self-sustained numerical models of voice production, which follows the normal/hyperfunctional behavior including high pitch voices, simulated jitter and shimmer, fundamental frequency and first formant interaction, fluid-structure interaction, supra and sub glottal coupling, and posterior opening gap. Contributions in this direction are expected to have a significant impact in clinical voice assessment, where these conditions are ubiquitous. For the objective voice evaluation in clinical scenarios a combination of SPL-normalized and multivariate statistical methods were evaluated and validated to discriminate between subjects with normal and hyperfunctional voices. From the derived aerodynamic measures several statistical tests were performed including univariate and multivariate hypotheses tests, which allow for improvement of the discrimination power between normal hyperfunctional voices in a series of settings including comfortable/loud loudness conditions. Salient aerodynamic measures were identified, which yields to a strong evidence that aerodynamic measures of vocal function that can

discriminate normal subjects from pathological will be an important contribution to the research field.

In addition, and for first time, neck skin acceleration (ACC)-based aerodynamic measures aligned with oral volume velocity (OVV)-based tokens were studied and explored. The results show preliminary evidence that ACC-based aerodynamic measures have the potential to discriminate normal from hyperfunctional behaviour in the same context that the OVV-based measures. Under this initial ACC-derived results, a framework based on statistical principles was proposed to explore the Q parameters estimation in a frame-based approach for continuous-speech. Relevant analysis for the uncertainties of resulting inverse filtering glottal waveforms and the impact on the aerodynamic measures were estimated. These finding, are the first evidence that uncertainties of subglottal IF methods, and their impact on the aerodynamic measures, could be improved with further research.

1.5.1 Publications

Below are the journal and conference papers published during this thesis work.

Journal Papers

- **V. M. Espinoza**, M. Zañartu, D. D. Mehta, J. H. Van Stand, R. E. Hillman, *Glottal Aerodynamic Measures in Women With Phonotraumatic and Nonphonotraumatic Vocal Hyperfunction*, Journal of Speech, Language, and Hearing Research (JSLHR). August, 2017. 60(8), 2159-2169.
- D. D. Mehta, J. H. Van Stan, M. Zañartu, M. Ghassemi, J. V. Guttag, **V. M. Espinoza**, J. P. Cortés, H. A. Cheyne, and R. E. Hillman, *Using ambulatory voice monitoring to investigate common voice disorders: Research update*, Frontiers in Bioengineering and Biotechnology. Section: Bioinformatics and Computational Biology, 3:155, 2015.

Journal Papers in review

- J. P. Cortés, **V. M. Espinoza**, M. Ghassemi, D. D. Mehta, J. H. Van Stan, Robert E. Hillman, John V. Guttag, M. Zañartu. *Assessment of Phonotraumatic Vocal Hyperfunction Using Ambulatory Glottal Airflow Measures*, submitted for publication.
- M. Díaz-Cádiz, S. D. Peterson, G. E. Galindo, **V. M. Espinoza**, and M. Zañartu, *Estimating Vocal Fold Contact Pressure from Raw Laryngeal High-Speed Videoendoscopy*, submitted for publication.

Conference Papers

- J. P. Cortés, **V. M. Espinoza** , M. Ghassemi , D. D. Mehta, J. H. Van Stan, Robert E. Hillman, John V. Guttag, Matias Zañartu. Using aerodynamic features and their uncertainty for the ambulatory assessment of phonotraumatic vocal hyperfunction, IEEE Biomedical and Health Informatics (BHI'18), Las Vegas, NV, USA, March 4-7, 2018.
- **V. M. Espinoza**, M. Zañartu. *Evaluación Clínica de la Voz por medio de Mediciones Aerodinámicas y Acústicas*, Congreso Internacional de Acústica y Audio Profesional, INGEACUS 2017, 22-25 Noviembre, Valdivia, Chile. 2017.
- **V. M. Espinoza**, D. D. Mehta, J. H. Van Stan, and R. E. Hillman, M. Zañartu. *Uncertainty of glottal airflow estimation during continuous speech using impedance-based inverse filtering of the neck-surface acceleration signal*, in 173rd Meeting of the Acoustical Society of America, Boston, USA. Abstract in The Journal of the Acoustical Society of America 141(5):3579-3579. May 2017. DOI: 10.1121/1.4987622.
- **V. M. Espinoza**, M. Zañartu. *Estudio Dinámico de Parámetros de Filtrado Inverso para el Seguimiento Ambulatorio de la Función Vocal*, IX Congreso Iberoamericano de Acústica - FIA 2014, 1-3 de Diciembre de 2014, Valdivia, Chile. 2014.
- M. Zañartu, **V. M. Espinoza**, D. D. Mehta, J. H. Van Stan, H. A. Cheyne II, M. Ghassemi, J. V. Guttag, and R. E. Hillman. *Toward an Objective Aerodynamic Assessment of Vocal Hyperfunction using a Voice Health Monitor*. 8th International Workshop on Models and Analysis of Vocal

Emissions for Biomedical Applications, MAVEBA 2013, December 16 - 18 2013, Firenze, Italy.

- **V. M. Espinoza**, M. Zañartu, J. H. Van Stan, D. D. Mehta, R. E. Hillman. *Differentiating between females with vocal hyperfunction and matched-controls using inverse filtered aerodynamic measures*, in 10th International Conference on Voice Physiology and Biomechanics, 2016, pp. 201202.

- 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*, in 10th International Conference on Voice Physiology and Biomechanics, 2016, pp. 9596.

- M. Díaz, G. G. Galindo, **V. M. Espinoza**, and M. Zañartu. *Vocal fold contact pressure estimation over laryngeal high speed videoendoscopy based on a vocal edge tracking method using Kalman Filter and Hertzian Impact model*, in 10th International Conference on Voice Physiology and Biomechanics, 2016, pp. 105106.

- **V. M. Espinoza**, M. Zañartu. *Estudio Dinámico de Parámetros de Filtrado Inverso para el Seguimiento Ambulatorio de la Función Vocal*, IX Congreso Iberoamericano de Acústica - FIA 2014, 1-3 de December de 2014, Valdivia, Chile.

- **V. M. Espinoza**, M. Zañartu, D. D. Mehta, ; J. H. Van Stan, H. A. Cheyne II, M. Ghassemi, J. V. Guttag, and R. E. Hillman, *Toward an Objective Aerodynamic Assessment of Vocal Hyperfunction using a Voice Health Monitor: Preliminary Results*, Buenos Aires Voice Metting 2013, STIC-AmSud Meeting on Physics-based Voice Produccition Modeling & SAV Meeting of Interdisciplinary dialog in Human Voice November 19 20, 2013, Buenos Aires, Argentina.

Chapter 2

Background

In this chapter the essential background and theoretical framework for voice assessment are introduced. Two types of vocal hyperfunction are described, namely Phonotraumatic (PVH) and Nonphonotraumatic (NPVH), along with the analysis of state of the art and current challenges in voice assessment. Objectives methods to evaluate the vocal function for clinical voice assessment are presented along with common aerodynamic measures derived from a series of oral airlfow and vibro-acoustics sensors. Current inverse filtering methods are discussed in the context of voice assessment and the associated challenges in that context.

2.1 Voice Assessment

2.1.1 Theoretical Framework

Vocal hyperfunction (VH) refers to chronic conditions of abuse and/or misuse of the vocal mechanism due to excessive and/or imbalanced muscular forces [3] and is associated with the common types of voice disorders. Hillman et al. (1989) proposed that two manifestations of VH reflect different underlying pathophysiological mechanisms that were originally referred to as adducted VH and non-adducted VH, and more recently relabeled as phonotraumatic VH (PVH) and non-phonotraumatic VH (NPVH) [40]. PVH is associated with the formation of benign vocal fold lesions due to chronic tissue trauma (e.g., vocal fold nodules). NPVH is associated with chronic dysphonia and vocal fatigue in the absence of vocal fold tissue trauma, and is also often referred to as primary

muscle tension dysphonia (MTD) [41]. In the view of Hillman et al. (1989)[3], both PVH and NPVH involve increased tension and stiffness of the vocal folds due to heightened and/or imbalanced (uncoordinated) levels of laryngeal muscle activity with an associated increase in aerodynamic forces required to produce and sustain phonation. The two conditions are hypothesized to differ from normal vocal function in terms of the impact of VH on vocal fold adduction and abduction. In PVH, adductory forces appear to predominate to maintain tight approximation of the vocal folds and, in combination with increased aerodynamic forces, create sharp vocal fold collision that leads to tissue trauma. In NPVH, an apparent imbalance between adductory and abductory forces hampers the approximation of the vocal folds, thus reducing the potential for phonotrauma even though heightened aerodynamic forces are present. The original Hillman et al. (1989) paper publication that proposed the two types of VH presented results from an initial study involving a small numbers of heterogeneous patients and non-matched controls that provided preliminary support for the proposed pathophysiological mechanism [3]. Further support for proposed pathophysiology of PVH was provided in three subsequent studies in subjects with vocal fold lesions [3, 42–44]. These earlier investigations employed a combination of non-invasive acoustic and aerodynamic measures that were designed to provide insight into underlying pathophysiological mechanisms, along with possible voice therapy and treatment. It was believed that such measures would have the potential to improve the clinical management of these disorders by providing quantitative metrics for basing judgments about the type and severity of VH for making treatment decisions. Several measures were extracted from recordings of the acoustic signal, intra-oral air pressure, and high-bandwidth oral airflow (pneumotachograph mask) during the repeated production of /pae/ syllables. Primary measures included acoustic measures of fundamental frequency (f_0) and sound pressure level (SPL), estimates of subglottal air pressure (SGP) from the intra-oral air pressure, and estimates of glottal airflow waveform parameters extracted from the oral airflow volume velocity using inverse filtering. Glottal airflow measures included peak-to-peak amplitude of the unsteady airflow (also denoted as AC flow, ACFL), maximum flow declination rate (MFDR), and open quotient (OQ). In general, patients with PVH displayed abnormally elevated values for SGP, ACFL, and MFDR, which was interpreted to reflect increased potential for trauma to vocal fold tissue that contributed to the chronic presence of vocal fold lesions and associated dysphonia. Patients with PVH also tended to have elevated OQ values that were attributed to the obstruction of glottal closure due to the presence of vocal fold lesions [3, 44]. Patients with NPVH also

displayed abnormally elevated values for SGP and OQ, but without concomitant increases in ACFL and MFDR, which was associated with inefficient phonation and dysphonia, but decreased potential to cause trauma to vocal fold tissue. In addition, aerodynamic measures appeared to be more sensitive to the presence of vocal pathology than acoustic measures, particularly when the aerodynamic measures were normalized with respect to SPL. These preliminary results were supported by simulations of self-sustained models of PVH showed that increasing subglottal air pressure in the presence of incomplete glottal closure results on increases in ACFL, MFDR and vocal fold collision forces when a given SPL is held [5, 6]. This result essentially confirms what was observed in Hillman et al. (1989) from patients with PVH and is interpreted as reflecting the potential role of compensation (secondary or reactive VH) in perpetuating phonotrauma, i.e., the vicious cycle [3]. Even though the early results reflecting different (quantifiable) pathophysiological mechanisms for PVH and NPVH were promising (and partially corroborated for PVH with modeling), these findings have never been completely validated in studies on homogeneous groups of patients with hyperfunctional voice disorders and well-matched controls that are large enough for formal statistical testing.

2.1.2 Current clinical methods to objectively measure of vocal function

A number of objective measures have proven useful in the diagnosis and treatment of voice pathologies associated with hyperfunction. The use of multiple bio-sensors to measure the oral airflow, static pressure, radiated acoustic pressure, electroglottography [45] and Laryngeal High Speed Video (HSV) [46] provide objective tools to support clinical decisions. The aerodynamic measures are obtained with the so-called *Rothenberg mask* [16], a circumferentially vented (CV) pneumotachograph mask that captures oral volume velocity (OVV) within a bandwidth of approximately 0 Hz to 1.2 kHz [3]. The glottal airflow is obtained by inverse filtering techniques [33, 34] that compensate for the influence of the vocal tract resonances on the speech signal in order to obtain an estimate of the airflow in the glottis. The resulting inverse filtered glottal volume velocity (GVV) signal provides relevant DC (steady) and AC (unsteady) signal components, from which objective measures of vocal function such as peak-to-peak amplitude (ACFL), maximum flow declination rate (MFDR), minimum and peak flow, AC/DC flow ratio, open and closed quotient (among others) are obtained (see details in section 2.1.3). These measures are normally

obtained by a sequence of repeated /pae/ sounds for various loudness efforts and have been shown to be the most prominent and consistent measures to identify vocal hyperfunction for adducted and non-adducted voices with a normative set using a z-score discriminant technique based on a multivariate linear regression [3]. Clinical evaluation of voice has been extended in recent years to ambulatory monitoring [12] using devices that capture and store information from a light-weight accelerometer attached to the neck skin over the collarbone. The first ambulatory voice monitor (APM Model 3200, KayPENTAX) was able to operate as a data logger storing a maximum of 14 hours of the fundamental frequency (f_0), and SPL values derived from the skin acceleration level (SAL) [30]. Related dose/time measure [47] are computed with a time resolution of 50 milliseconds when data from APM is downloaded to a computer for analysis using a proprietary software. Additionally, the APM provides biofeedback through a pager vibrator based on thresholds for f_0 and SPL, and has been shown to have the potential to aid changes in vocal behavioral under voice therapy. However, the APM does not store the raw acceleration signal, and additional improvements and analysis are not possible using this device. This technology was subsequently extended to record the raw accelerometer signal with the recently developed of the Voice Health Monitor (VHM), employing the same accelerometer sensor and smartphone technology [12]. Translating the aerodynamic assessment to ambulatory monitoring scenarios has numerous challenges. Using a model-based signal processing to cancel the effects of sub-glottal and skin resonances has shown the potential to obtain an estimate of glottal airflow [27, 28] to derive relevant aerodynamic measures. Among the enhanced capabilities of the VHM are: 1) to record and store raw acceleration signals exceeding a week of duration at a 11.025 Hz sample rate and with 16-bit quantization, with signal processing capabilities to run sophisticated algorithms, e.g., real-time inverse filtering.

2.1.3 Objective Parameters of Vocal Function

In Figure 2.1 a typical setup used in voice assessment is depicted based on (A) multiple sensors, and (B) signals involved on the voice assessment of vocal function. In (A), the sensors shown are 1) oral volume velocity sensor (FLO), 2) neck skin accelerometer (ACC), 3) microphone (MIC), 4) electroglottographic sensor (EGG), and 5) intraoral pressure sensor (PRE). In (B), all signals are digitized, time-aligned, and calibrated in their physical units. From these signals, several clinically relevant aerodynamic measures are typically derived. From top to bottom, the signals shown are: 1) sound power (smoothed r.m.s. follower of the acoustic radiated pressure), 2) smoothed oral airflow (including AC and DC components) and 3) intra-oral air pressure. In Figure 2.2, five tokens are marked with a thicker line-width (in red) corresponding to each of five /pae/ syllables. The arrows indicate the samples where the measures are normally taken in intensity (sound power) and oral airflow (average flow). Peak values (*) in the intra-oral pressure signal are normally detected and interpolated to estimate subglottal pressure (SGP) during phonation (x).

High bandwidth measures are also commonly obtained from estimates of the glottal airflow cycle after inverse filtering the oral airflow signal. Through the glottal airflow signal and its time-derivative estimates, temporal and amplitude related measures are estimated for each available cycle and then averaged to characterize vocal behavior for the given frame. Selected measures are discussed in more detail together with the indications in Figure 2.3 since they will be useful for subsequent developments in this thesis work. Given $\mathbf{x} \triangleq [x_0 \dots x_{n_0-1}]$ where n_0 is the number of samples in one glottal cycle. Selected time related measures can be described as follows:

- **Fundamental frequency** (f_0) is defined as the first harmonic of the periodic airflow signal, which can be formulated as $f_0 = f_s/n_0$ where f_s is the sample rate. In general, f_0 or n_0 are estimated based on the autocorrelation sequence [48], but there are several other alternatives [49].
- **Speed Quotient (SQ)** refers to the proportion (ratio) of the opening and closing time of glottal pulse,

$$SQ = \frac{t_1}{t_2}. \tag{2.1}$$

- **Open Quotient (OQ)** is the estimated proportion of a cycle period ($T_0 = \frac{1}{f_0} = \frac{n_0}{f_s}$) in which vocal folds are in open phase of the cycle,

$$OQ = \frac{t_1 + t_2}{T_0} = (t_1 + t_2) \cdot f_0. \tag{2.2}$$

- **Normalized Amplitude Quotient (NAQ)** is a measure to parametrize the glottal closing phase using two amplitude-domain measurements from glottal cycle, ACFL and MFDR, normalized to one glottal period. NAQ is more robust estimate of the Closing Quotient (CQ) [50],

$$\text{NAQ} = \frac{\text{ACFL}}{\text{MFDR}} \cdot \frac{1}{T_0} = \frac{\text{ACFL}}{\text{MFDR}} \cdot f_0 \ \propto \ \text{CQ} = \frac{t_2}{T_0}. \tag{2.3}$$

In a similar way the common amplitude related measures are:

1. **Peak-to-Peak glottal airflow (ACFL)**: is the peak-to-peak amplitude of the unsteady glottal airflow and it has been associated with the amplitude of vocal fold vibration [3, 42],

$$\text{ACFL} = \max \mathbf{x} - \min \mathbf{x}, \quad \text{(mL/s)}. \tag{2.4}$$

2. **Maximum Flow Declination Rate (MFDR)**: is the magnitude of the negative peak of time-derivative of glottal cycle and it has been associated with vocal fold closing velocity [3, 42].

$$\text{MFDR} = |\min \Delta \mathbf{x}|, \quad \text{(L/s}^2\text{)}. \tag{2.5}$$

3. **First to Second Harmonic ratio (H1H2)**: is the difference (in dB) between the first and second harmonic magnitudes, which provides an approximation of the spectral tilt. This is estimated through the magnitude of the spectrum of $x[n]$ ($X(f)$), over several glottal cycles, using the Discrete Fourier Transform (DFT) [51].

$$\text{H1H2} = 20 \log_{10} \left(\frac{|X(f_0)|}{|X(2 \cdot f_0)|} \right) \tag{2.6}$$

4. **Cepstral Peak Prominence (CPP)** is defined as the distance between the first-peak prominence over the noise floor of the cepstral sequence, which is calculated using the inverse Fourier Transform of the log function of the

magnitude of the spectrum $X(f)$. CPP is typically understood as a estimate of breathiness and it has been associated with incomplete glottal closure [39].

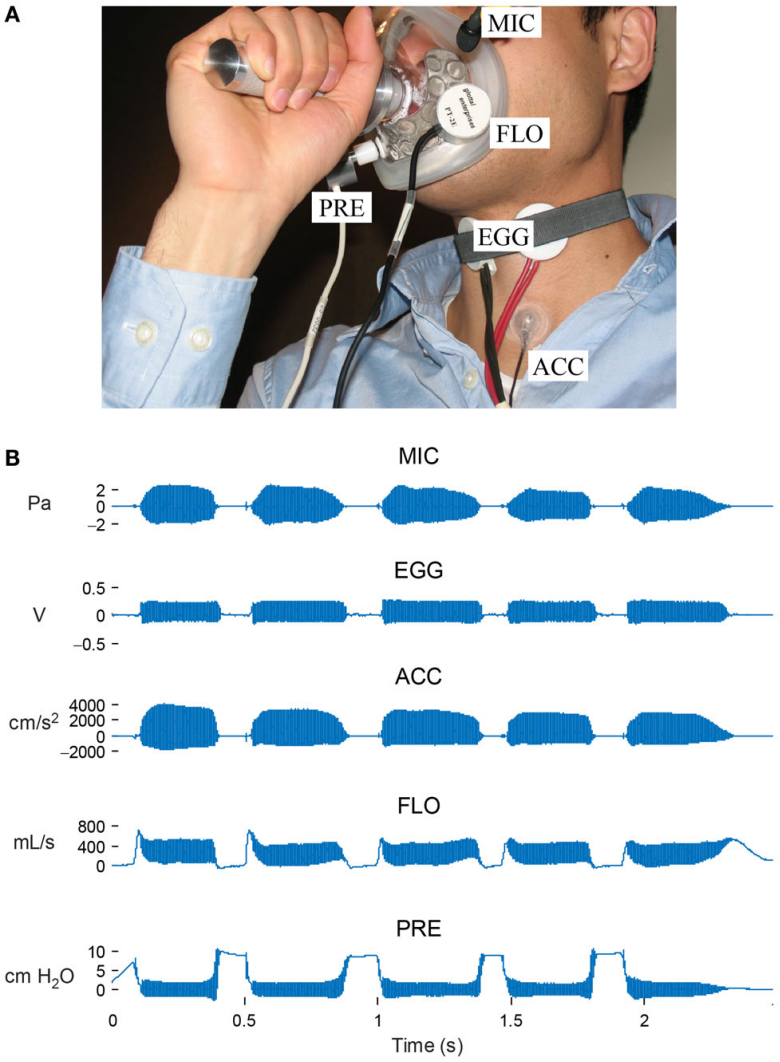


FIGURE 2.1: All voice sensors, with their calibrated signals. A) Typical biosensors and IDs setup for objective voice assessment: MIC = microphone (Pa), EGG = electroglottographie sensor (Volts), ACC = neck skin accelerometer (cm/s^2), FLO = Oral airflow sensor (mL/s), and PRE = Intra Oral Pressure sensor (cm H_2O). Image from [40].

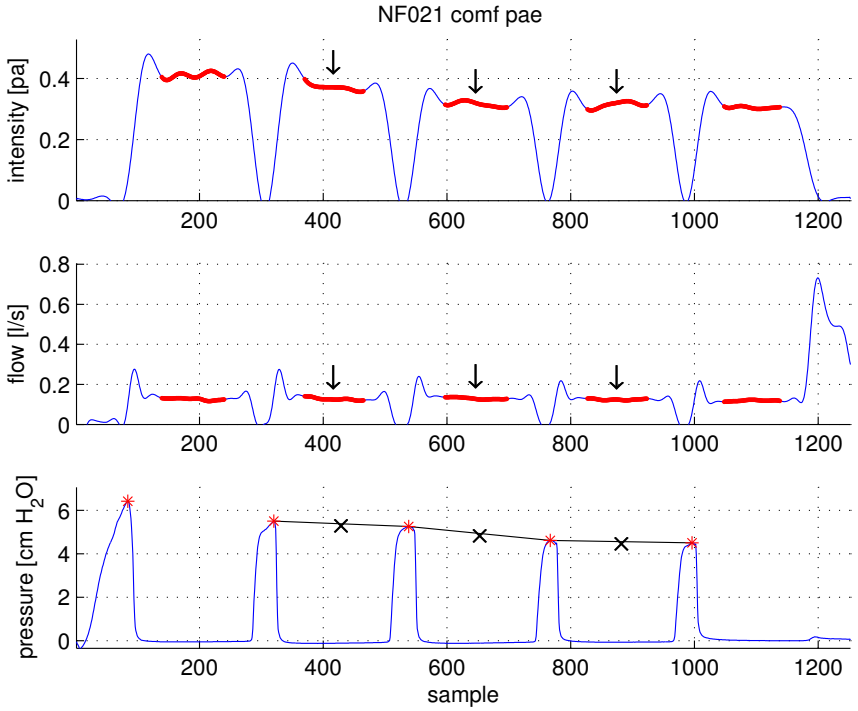


FIGURE 2.2: Low Bandwidth measures. *Top*: Sound intensity (smoothed r.m.s. follower of the radiated acoustic pressure). *Center*: Smoothed DC-component of oral airflow. Arrows indicate mid sample point wherein measures will be extracted). *Bottom*: Intra-oral air pressure for five /pae/ syllables, showing peaks values (*), interpolation lines and mid values (x). Red bold lines are showing selected data (50 % of token middle portion).

2.1.4 ACC derived aerodynamic measures

Vocal measures derived from the acceleration signal are possible to estimate through inverse filtering neck skin and subglottal system properties [28]. Concretely, ACFL, MFDR, OQ, SQ, CQ, NAQ, H1H2, f_0, and CPP can be estimated using same definitions already given in this section. However, aerodynamic measures estimates based on the DC component of the CV mask are not possible to derive from the ACC signal as no DC component is present. Nevertheless, studies are showed that subglottal pressure could be approximated indirectly by using the RMS value of ACC signal [52, 53].

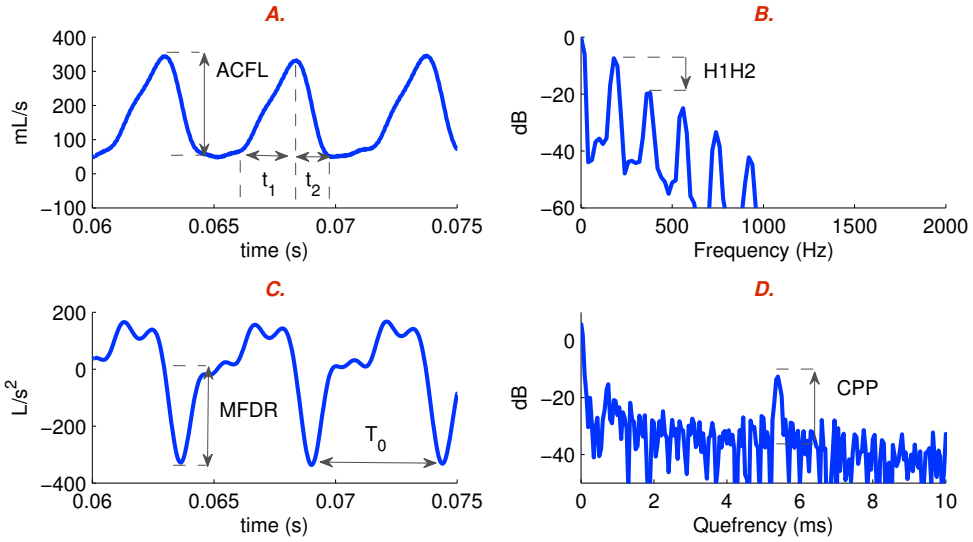


FIGURE 2.3: High bandwidth measures derived from an estimated glottal airflow waveform. (A) Estimated glottal airflow waveform indicating peak-to-peak (ACFL) value, opening time t_1 and closing time t_2 (used to calculate OQ and SQ quotients). (B) Power spectra of the estimated glottal waveform showing H1H2 relative distance between fundamental frequency (1st harmonic) and 2nd harmonic in dB scale. (C) Maximum flow declination rate (MFDR) from the time-derivative estimated glottal airflow as the negative maximum on the waveform. Glottal cycle interval T_0 between two consecutive peaks of the time-derivative of the estimated glottal. (D) The logarithm of the absolute value of the (one-side) real cepstrum indicating first-rhamonic peak and CPP.

2.1.5 Current challenges in voice assessment

One limitation with the current objective approach for voice assessment is related to the validation and accuracy of inverse filtering techniques for pathological voices (most of them have only been tested for normal voices). Only sustained vowels /a/ or /ae/ are used for clinical evaluation due to the difficulty of obtaining IF over other vocal gestures in both males and females voices. Long-term voice evaluations require IF methods to be extended to a continuous speech scenario. Some initial evaluation efforts have been performed for inverse filtering in running speech [23], although not in the context of a clinical evaluation of vocal function under normal and pathological conditions in male and female voices.

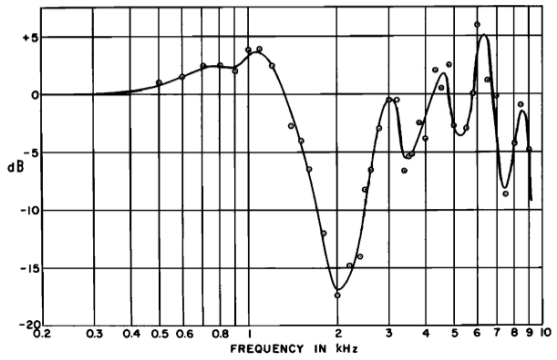


FIGURE 2.4: log(·) magnitude frequency response of Rothenberg mask [16], showing a near flat frequency response below 1 kHz, a strong zero near 2 kHz and fluctuations in higher frequencies

2.2 Inverse filtering of female voices

Most IF methods use the radiated acoustic pressure of the voice (i.e., microphones) [33]. Microphones have a vast number of applications in communication sciences, but are not able to operate down to 0 Hz, and are very sensitive to the acoustic environment and proximity effects [54]. Given that, is it necessary to preserve the DC and AC components of the signal with a high signal-to-noise ratio, which is performed using the so-called *Rothenberg* mask [16] [55]. Even though the mask can operate in DC, it also has some drawbacks. The shape of the mask contributes to a spectral distortion in the recorded signal [16], which is partially reduced by a filter near 1.1 kHz [18]. As shown in Figure 2.4, a deeper zero is present at 2 kHz, as well other fluctuations at higher frequencies [16] that are difficult to compensate.

In voice assessment, inverse filtering refers to the procedure of estimating the aerodynamic flow in the glottis, i.e., the glottal airflow. See an example in Figure 2.5. Voice production begins at the subglottal system level in the lungs, which drives a quasi-continuous pressure to the vocal folds. These pressure, produce self-sustained oscillations that result in acoustic waves that propagate through the vocal tract. The supra-glottal tract provides the time-variating articulation to produce speech [56]. Voice production involves a time-varying complex non linear interaction between aerodynamic flow from the lungs, sub-glottal and supraglottal resonances, and tissue dynamics of the vocal folds [57] [58]. The knowledge of these interactions has multiple applications in speech sciences, speech coding, speaker recognition, speech enhancement, speech synthesis [59] and clinical voice assessment [3].

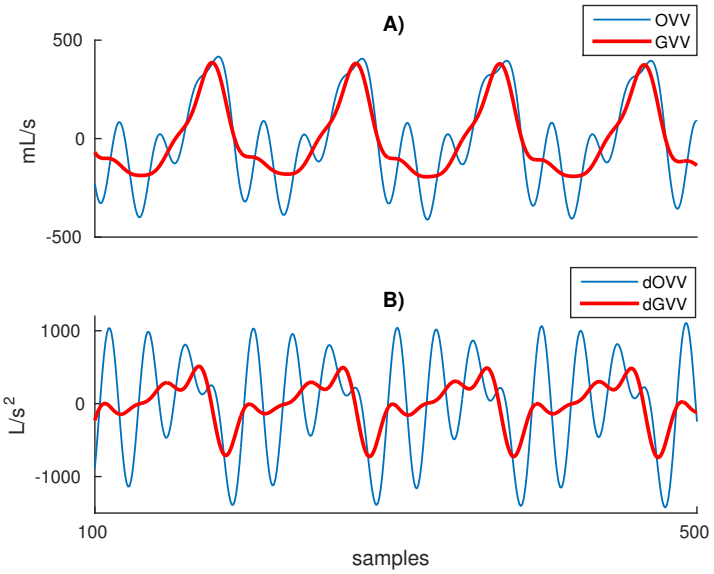


FIGURE 2.5: Example of an inverse filtering application. A) Oral airflow before IF (thin blue), and glottal airflow estimate after IF (solid red). B) Time-derivatives of A). Note the degree of oscillation in the closed phase (usually called formant ripple) for OVV and time-derivative OVV (dOVV) signals are effectively reduced by the IF process.

Research in inverse filtering techniques is extensive. Proposed methods have include analog [60] and digital filters [61], cepstrum-based [23, 38], closed-phase covariance linear prediction [17, 62], physical modeling [21], fuzzy logic techniques [63], and Bayesian estimation [64] to estimate the glottal airflow signal and/or its time-derivative. A detailed review of these techniques can be found in [33, 34, 65].

One common system model for representing the effects of vocal tract on the oral airflow and radiated acoustic pressure uses a series of concatenated tubes to describe the pharynx and nasal cavities [56, 66, 67]. This system model is equivalent to an auto-regressive (AR) model [68, 69], under the assumption of a closed boundary at the glottis, i.e., with no coupling with the subglottal tract, and no contribution of the glottal source [17]. However, the vocal folds are closed (or partially closed [6, 70]) only for a portion of the time cycle as shown in Figure 2.6. Some of the more recent IF methods [17, 23] need exact times instant closures for a reliable estimation of AR model. Although the EGG signal has been widely used to get the exact time of glottal closure, in presence of reduced vocal fold contact (e.g., breathy voices), the estimation of instant closure could be affected [71]. Algorithms such as DYPSA [72] and YAGA [73] make use of the same oral airflow or microphone signals to achieve glottal closure detection and have shown good performance in speech databases [74]. The YAGA algorithm

was used to detect the closed phase for oral airflow measures [55], although they have not yet been used in clinical voice evaluation. For detecting the closure of the vocal folds, a simultaneous electroglottographic (EGG) signal [45] is usually recorded with the oral flow, although other approaches are common too [74]. After identification of the closed phase, covariance analysis could be used to calculate a parametric AR model of the vocal tract only in the closed portion of the cycle [75]. The Covariance method of LPC has the benefit over the more common autocorrelation method [75], in that it provides a more physiological-correct representation of the vocal tract. This method is discussed in detail in section 2.2.1 since it serves as one of the key references for the thesis work.

Additionally, voices with high f_0 present difficulties with the inverse filtering process. The closed phase in these cases is very short in time, and the usual analysis is less robust [17]. In clinical voice evaluation, women have a higher f_0 (on average) than men, are more affected by voice pathologies [76], and it is still a challenging task to achieve a glottal flow estimation in these conditions.

Furthermore, the model assumption of the closed-open tube for the vocal tract model is not very realistic. Video laryngoscopy (VL) and HSV observation of the vocal folds have shown that closure of the vocal folds (the closed side of the tube) is incomplete due to a posterior gap [6, 77], coupling sub and supra-glottal tracts adding a more mixed aerodynamic interaction between them [78].

Another issue in the estimation of glottal flow is the inability of knowing the *true* glottal airflow. Under this scenario, numerical voice modeling [57] has been used to provide a *theoretical true* glottal airflow for benchmarking IF methods [17, 79–81] using sustained vowels. Most of the experiments for testing IF methods have used sustained vowels, and little attention has been paid to the estimation of glottal airflow in running speech. The causal/anticausal glottal source decomposition from Drugmann et al. [23] must be *integrated* in the time domain to obtain the glottal flow waveform and it is susceptible to windowing effects [23], spectral leakage [51] and the presence of zeros inside and outside the vicinity of the unit circle [82].

When detection of the time of closure is not available or not reliable, a common practice is to adjust the inverse filter anti-resonances using subjective criteria to reduce waveform ripple (due to formants) and produce a near flat amplitude in the closed phase [18] [35]. This method has been used in various research papers for objective clinical evaluation of vocal hyperfunction [25, 26, 31], and in spite

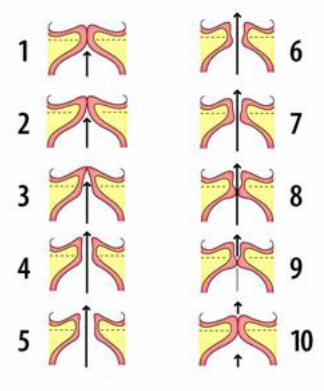


FIGURE 2.6: Vocal phase fold oscillation. (1) Static flow from lungs pushes the vocal folds upwards in (2). The open phase starts in (3) and finishes in (8), and the closed phase starts in (8) and finishes in (10), where the cycle starts again [76]

of its simplicity remains as the most accepted method to deal with pathological voices. However, this method is based on the subjective user experience, and is not suitable for analyzing larger voice databases [83], since no formal mathematical expression has been proposed to accomplish this task. On the other hand, with a sub-glottal inverse filtering scheme, is possible to obtain glottal airflow waveforms from a neck surface accelerometer placed above the clavicle [27, 28]. This approach is known as impedance-based inverse filtering (IBIF) and it is a mechano-acoustic impedance representation from a physiological-based transmission line model with a lumped mechanical model of the neck skin [84]. One of the advantages of this approach is that sub-glottal resonances and mechanical properties of the skin are considered nearly time-invariant, and the neck surface accelerometer is more robust to environmental noise, and the approach does not need to account for supra-glottal coupling. However, the IBIF model needs subject-specific parameters to obtain the subglottal system and mechanical properties of the neck skin [28], which are derived from inverse filtering glottal airflow from the CV mask using a sustained vowel /a/. Despite the initial results of IBIF, validation in different vocal gestures, e.g., continuous speech and pathological voices, has not been accomplished.

2.2.1 Inverse Filtering Methods

In this section, mathematical details of selected inverse filtering methods are summarized since they will serve as a basis for further developments in this thesis.

2.2.1.1 Single Notch Filter

The Single Notch Filter (SNF) [18] is the inverted version of a resonant conjugate-pair of poles following the assumption that each vocal tract resonance can be modeled as a second order system with a transfer function given by

$$\begin{aligned} H(z) &= \frac{gz^2}{(z-z_p)(z-z_p^*)} \\ &= \frac{gz^2}{z^2-(z_p+z_p^*)z+1} \\ &= \frac{g}{1-(z_p+z_p^*)z^{-1}+z^{-2}} \\ &= \frac{g}{1-(2\beta\cos\theta_c)z^{-1}+z^{-2}} \end{aligned} \tag{2.7}$$

where $\theta_c = 2\pi\frac{f_c}{f_s}$ (the formant frequency), $0 < f_c < \frac{fs}{2}$, with the pole in the z-plane at $z_p = \beta e^{j\theta_c}$, with $\beta = e^{-\pi B/f_s}$, where B is the bandwidth in Hz measured at -3 dB. Putting coefficients names in standard form, $H(z)$ yields,

$$H(z) = \frac{b_0}{a_0 + a_1 z^{-1} + a_2 z^{-2}} \tag{2.8}$$

where $a_0 = 1$, $a_1 = -2\beta\cos\theta_c$, and $a_2 = 1$. To obtain b_0, the DC gain is constrained to $H(z_{\theta=0}) = 1$ and $\angle H(z_{\theta=0}) = 0$ [18, 76]. Evaluating at $\theta = 0$ (DC) $\rightarrow z_{\theta=0} = e^{j0} = 1$ in (2.8) yields

$$H(z_{\theta=0}) = \frac{b_0}{a_0 + a_1 + a_2} \tag{2.9}$$

To satisfy the DC constraint, b_0 must be equal to $\sum_{i=0}^{2} a_i$, then $b_0 = 2(1-\beta\cos\theta_c)$.

Since the VT is a stable system, $H(z)$ is well defined and the SNF filter is simply the inverted version of equation (2.7), i.e.,

$$H_{SNF}(z) = \frac{1-(2\beta\cos\theta_c)z^{-1}+z^{-2}}{2(1-\beta\cos\theta_c)} \quad . \tag{2.10}$$

SNF take its name since is used under limited bandwidth conditions, where in which only one formant is considered.

2.2.1.2 Closed-Phase Inverse Filtering

The Closed-Phase Inverse Filtering (CPIF) [17] is also assumes that the VT is an all-pole model and estimates it using a covariance method of linear prediction. The approach has a DC constraint and forces a minimum-phase response by mirroring any conjugated roots that are located outside the unit circle. The method has two important features suitable for glottal airflow estimation: 1) The gain at DC is set to 1, which is very important to keep a meaningful steady aerodynamic component (in magnitude and phase), and 2) the inverse filter is always minimum-phase, which prevents the occurrence of abrupt changes in the closing point [17]. Thus, the AR-based filter is derived as follows. An error is calculated between the signal and the predicted one such as,

$$e_n = x_n + \sum_{k=1}^{p} a_k x_{n-k} = \sum_{k=0}^{p} a_k x_{n-k} = \mathbf{a}^T \mathbf{x}_n \quad , \tag{2.11}$$

where $\mathbf{a} = [a_0, \ldots, a_p]^T$ are the AR model coefficients, with $a_0 = 1$, and signal speech vector $\mathbf{x}_n = [x_n, \ldots, x_{n-p}]^T$.

Using same notation that in Alku et al. [17], then if $0 < n \leq N-1$, the prediction energy of the error $E(\mathbf{a})$ can be calculated as,

$$E(\mathbf{a}) = \sum_{n=0}^{N-1} e_n^2 = \sum_{n=0}^{N-1} \mathbf{a}^T \mathbf{x}_n \mathbf{x}_n^T \mathbf{a} = \mathbf{a}^T \left(\sum_{n=0}^{N-1} \mathbf{x}_n \mathbf{x}_n^T \right) \mathbf{a} = \mathbf{a}^T \mathbf{\Phi} \mathbf{a} \quad , \tag{2.12}$$

where the matrix $\mathbf{\Phi} \in \mathbb{R}^{(p+1)\times(p+1)}$ is the sample covariance matrix.

Incorporating the constraints, we get a new set of constrained coefficients $\mathbf{c} = [c_0, c_1, \ldots, c_p]^T$. Then, for the DC gain ($\omega = 0$) constraint, we use the following statement,

$$C(z) = \sum_{k=0}^{p} c_k z^{-k} \Rightarrow C(e^{j0}) = C(1) = \sum_{k=0}^{p} c_k = l_0 \quad , \tag{2.13}$$

and for $\omega = \pi$

$$C(z) = \sum_{k=0}^{p} c_k z^{-k} \Rightarrow C(e^{j\pi}) = \sum_{k=0}^{p} c_k \cdot (-1)^k = l_\pi \quad . \tag{2.14}$$

The above constraint can be summarized in the following matrices

$$\mathbf{\Gamma} = \begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & 1 & \dots & 1 \\ 1 & -1 & \dots & \pm 1 \end{bmatrix}^T \quad ; \quad \mathbf{b} = \begin{bmatrix} 1 & l_0 & l_\pi \end{bmatrix}^T \Longrightarrow \mathbf{\Gamma}^T\mathbf{c} = \mathbf{b} \quad , \tag{2.15}$$

Then, the optimization problem is rewritten as:

$$\text{minimize} \quad \mathbf{c}^T\mathbf{\Phi}\mathbf{c} \tag{2.16}$$

$$\text{subject to} \quad \mathbf{\Gamma}^T\mathbf{c} - \mathbf{b} = 0 \quad , \tag{2.17}$$

which is a quadratic programming problem with equality constraints and can be solved using the Lagrange multiplier method [85]. Thus, the constrained function is

$$\mathbf{F}(\mathbf{c}, \mathbf{g}) = \mathbf{c}^T\mathbf{\Phi}\mathbf{c} - 2\mathbf{g}^T\left(\mathbf{\Gamma}^T\mathbf{c} - \mathbf{b}\right) \quad , \tag{2.18}$$

where $\mathbf{g} = [g_1 \ g_2 \ g_3]^T > 0$ is the Lagrange multiplier vector.

Taking partial derivative [86] respect to $\mathbf{c}$ and $\mathbf{g}$ and equals to zero, yields

$$\frac{\partial \mathbf{F}}{\partial \mathbf{c}} = \left(\mathbf{\Phi} + \mathbf{\Phi}^T\right)\mathbf{c} - 2\mathbf{\Gamma}\mathbf{g} = 0 \tag{2.19}$$

$$\frac{\partial \mathbf{F}}{\partial \mathbf{g}} = 2\left(\mathbf{\Gamma}^T\mathbf{c} - \mathbf{b}\right) = 0 \quad , \tag{2.20}$$

but $\mathbf{\Phi} = \mathbf{\Phi}^T$ [75], then

$$\mathbf{\Phi}\mathbf{c} - \mathbf{\Gamma}\mathbf{g} = 0 \Rightarrow \mathbf{c} = \mathbf{\Phi}^{-1}\mathbf{\Gamma}\mathbf{g} \tag{2.21}$$

$$\mathbf{\Gamma}^T\mathbf{c} - \mathbf{b} = 0 \Rightarrow \mathbf{g} = \left(\mathbf{\Gamma}^T\mathbf{\Phi}\mathbf{\Gamma}\right)^{-1}\mathbf{b} \quad , \tag{2.22}$$

and using above equations let $\mathbf{c}$ vector of coefficients

$$\mathbf{c} = \mathbf{\Phi}^{-1}\mathbf{\Gamma}\left(\mathbf{\Gamma}^T\mathbf{\Phi}^{-1}\mathbf{\Gamma}\right)^{-1}\mathbf{b} \quad . \tag{2.23}$$

Figure 2.7 shows an example of CPIF method. Note that the DC component of glottal airflow is not altered by the CPIF method, and closed phase ripples are negligible. This method has shown good results even in the presence of incomplete glottal closure [79] and good fitting with simulations using voice numerical models [17, 79].

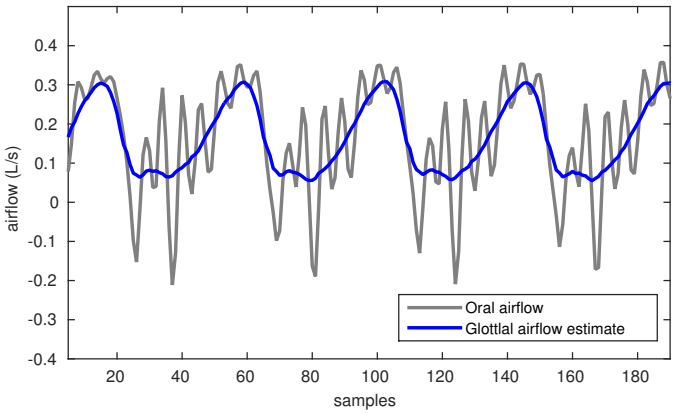


FIGURE 2.7: Close phase inverse filtering example. In gray, oral airflow signal from the CV mask. In blue, glottal airflow estimate after CPIF method [17].

2.2.1.3 Non Parametric Inverse Filtering: Cepstrum

Cepstrum is a non-parametric homomorphic technique used for removing of vocal tract resonances, through a blind deconvolution process that separates the source signal from the impulse response of the vocal tract [75]. As multiple approaches to calculating the cepstrum exists [37, 39, 87–89], proofs of essential cepstrum properties are provided. Cepstrum can be defined as the inverse z-transform of the logarithm of the z-transform of a signal [88], among other definitions [59]. The approach is based on the principle of superposition in a class of nonlinear systems [36], and it allows for separating each system or signal response from (2.27) only knowing the mixed sequence $x[n]$ [88]. Thus, consider an oral airflow signal $x[n]$ with its z-transform $X(z)$, where magnitude and phase have been separated and shown in (2.24)

$$X(z) = |X(z)|e^{j\angle H(z)}. \tag{2.24}$$

Now, applying $\ln(\cdot)$ to (2.24), yields

$$\begin{aligned} \ln\left(X(z)\right) &= \ln\left(|X(z)|\right) + \ln\left(e^{j\angle X(z)}\right) \\ &= \ln\left(|X(z)|\right) + j\angle X(z) \quad , \end{aligned} \tag{2.25}$$

Taking the inverse z-transform of (2.25) (represented by $\mathcal{Z}^{-1}\{\cdot\}$ operator), yields the cepstrum of $x[n]$, therefore

$$\mathcal{Z}^{-1}\left\{\ln\left(X(z)\right)\right\} = \mathcal{Z}^{-1}\left\{\ln\left(|X(z)|\right)\right\} + \mathcal{Z}^{-1}\left\{j\angle X(z)\right\} \quad . \tag{2.26}$$

The inverse z-transform of the real part of (2.26) is called *real cepstrum* of sequence $x[n]$, and a *complex cepstrum* is obtained when the imaginary part is added. Note that imaginary part must consider the unwrapped phase version of $\angle X(z)$ [88].

To get insights into the cepstrum for inverse filtering speech signals consider the following model example through linear convolution,

$$x[n] = s[n] * v[n] * g[n] \quad \overset{z}{\leftrightarrow} \quad X(z) = S(z) \cdot V(z) \cdot G(z) \quad , \tag{2.27}$$

where

$$s[n] = \sum_{k=0}^{N-1} \delta[n - kT] \quad \overset{z}{\leftrightarrow} \quad S(z) = \sum_{k=0}^{N-1} z^{-kT} = \frac{1 - z^{-NT}}{1 - z^{-T}} \quad , \tag{2.28}$$

is an impulse train which with period of T samples, with $0 < |z| < 1$, and

$$v[n] \quad \overset{z}{\leftrightarrow} \quad V(z) = \frac{1}{\prod_{k=1}^{p}(1 - a_k z^{-1})} \quad , \tag{2.29}$$

is an all-pole transfer function related to vocal tract with a_k real-valued coefficients, and

$$g[n] \quad \overset{z}{\leftrightarrow} \quad G(z) = \prod_{k=1}^{r}(-b_k^{-1}) \prod_{k=1}^{r}(1 - b_k z) \quad , \tag{2.30}$$

with b_k is an all-zero transfer function related to glottal pulse, and $\overset{z}{\leftrightarrow}$ denote z-transform operation.

Based on the above signals and their frequency transform, a $\ln(\cdot)$ is applied to the right side of (2.27). Thus, we get

$$\ln\left(X(z)\right) = \ln\left(S(z)\right) + \ln\left(V(z)\right) + \ln\left(G(z)\right) \quad , \tag{2.31}$$

and calculating the inverse z-transform $\rightarrow \mathcal{Z}^{-1}\{\cdot\}$ of (2.31) we obtain

$$\mathcal{Z}^{-1}\{\ln\left(X(z)\right)\} = \mathcal{Z}^{-1}\{\ln\left(S(z)\right)\} + \mathcal{Z}^{-1}\{\ln\left(V(z)\right)\} + \mathcal{Z}^{-1}\{\ln\left(G(z)\right)\} \quad , \tag{2.32}$$

The cepstrum of $x[n]$, i.e., $\hat{x}[n]$, is the summation of cepstrum of $s[n]$, $v[n]$ and $g[n]$, and can be rewritten as,

$$\hat{x}[n] = \hat{s}[n] + \hat{v}[n] + \hat{g}[n] \tag{2.33}$$

For those terms under $\ln(\cdot)$ function, it is possible to replace $\ln(\cdot)$ by its Taylor series in (2.34).

$$\ln(1-B) = -\sum_{n=1}^{\infty} \frac{B^n}{n} \ , \ |B| < 1 \quad , \tag{2.34}$$

where $B \in \mathbb{C}$ and $n \in \mathbb{Z}^+$. Now, taking the $\ln(\cdot)$ for (2.28)

$$\begin{aligned} \ln\left(\frac{1-z^{-NT}}{1-z^{-T}}\right) &= \ln\left(1-z^{-NT}\right) - \ln\left(1-z^{-T}\right) \\ &= -\sum_{k=1}^{\infty} \frac{(z^{-NT})^k}{k} - \left(-\sum_{k=1}^{\infty} \frac{(z^{-T})^k}{k}\right) \quad , \end{aligned} \tag{2.35}$$

using (2.35) and taking its z-transform, the cepstral sequence of impulse train is calculated [90], such that

$$\mathcal{Z}^{-1}\left\{-\sum_{k=1}^{\infty} \frac{(z^{-NT})^k}{k} - \left(-\sum_{k=1}^{\infty} \frac{(z^{-T})^k}{k}\right)\right\} = \left(-\sum_{k=1}^{\infty} \frac{\delta[n-kNT]}{k}\right) + \left(\sum_{k=1}^{\infty} \frac{\delta[n-kT]}{k}\right). \tag{2.36}$$

In the same way, for (2.29) we get:

$$\begin{aligned} \sum_{k=1}^{q} \ln\left(1-a_k z^{-1}\right) &= \sum_{k=1}^{q} \left(-\sum_{n=1}^{\infty} \frac{\left(a_k z^{-1}\right)^n}{n}\right) \quad , \quad |a_k z^{-1}| < 1 \\ &= -\sum_{k=1}^{q}\sum_{n=1}^{\infty} \frac{a_k^n}{n} z^{-n} \\ &\overset{(a)}{=} -\sum_{n=1}^{\infty}\sum_{k=1}^{q} \frac{a_k^n}{n} z^{-n} \\ &\overset{(b)}{=} \sum_{n=1}^{\infty} \left(-\sum_{k=1}^{q} \frac{a_k^n}{n}\right) z^{-n} \quad , \end{aligned}$$

where in (a) summation were swapped, and in (b) one summation term was grouped. Then, using inverse z-transform, yields

$$\begin{aligned} \mathcal{Z}^{-1}\left(\sum_{n=1}^{\infty} \left(-\sum_{k=1}^{q} \frac{a_k^n}{n}\right) z^{-n}\right) &= \begin{cases} 0 & , \ n \leq 0 \\ -\sum_{k=1}^{q} \frac{a_k^n}{n} & , \ n > 0 \end{cases} \\ &= u[n-1] \cdot \left(-\sum_{k=1}^{q} \frac{a_k^n}{n}\right) \quad , \end{aligned} \tag{2.37}$$

and for (2.30)

$$\begin{aligned}\sum_{k=1}^{r}\ln\left(1-b_k z\right) &= \sum_{k=1}^{r}\left(-\sum_{n=1}^{\infty}\frac{\left(b_k z\right)^n}{n}\right) \quad , \quad |b_k z|<1\\ &= -\sum_{k=1}^{r}\sum_{n=1}^{\infty}\frac{b_k^n}{n}z^n\\ &= -\sum_{n=1}^{\infty}\sum_{k=1}^{r}\frac{b_k^n}{n}z^n\\ &= \sum_{n=\infty}^{-1}\left(-\sum_{k=1}^{r}\frac{b_k^{-n}}{-n}\right)z^{-n}\\ &= u[-n-1]\cdot\left(\sum_{k=1}^{r}\frac{b_k^{-n}}{n}\right) \quad ,\end{aligned} \tag{2.38}$$

Adding up (2.36), (2.37) and (2.30) results in

$$c[n] = \underbrace{\left(-\sum_{k=1}^{\infty}\frac{\delta[n-kNT]}{k}\right)}_{c_1[n]} + \underbrace{\left(\sum_{k=1}^{\infty}\frac{\delta[n-kT]}{k}\right)}_{c_2[n]}\ldots \tag{2.39}$$

$$+\underbrace{u[n-1]\cdot\left(\sum_{k=1}^{q}\frac{a_k^n}{n}\right)}_{c_3[n]} + \underbrace{u[-n-1]\cdot\left(\sum_{k=1}^{r}\frac{b_k^{-n}}{n}\right)}_{c_4[n]} \quad . \tag{2.40}$$

After analyzing these results, we can observe that: 1) sequences $c_1[n]$, $c_2[n]$ and $c_3[n]$ are causal and $c_4[n]$ is not causal in the *quefrency* (a pseudo-time) domain, 2) all sequences are bounded in amplitude by a function of the form β^u/m, 3) the $c[n]$ is an infinite sequence, and 4) the $c[n]$ sequence in $n=(0,T)$ is only involved by sequence $c_3[n]$ with a_k coefficients and for $n=[-\infty,0)$ is only involved by sequence $c_4[n]$ with b_k coefficients, and both cases are separable from the periodic sequences $c_1[n]$ and $c_2[n]$ by means of *liftering* (filtering in *quefrency* domain). An example with conjugate pair $a_{1,2}=0.7\pm j0.5$ and periodic sequence with $T=23$, $N=35$ and truncate at $L=148$ is shown in Figure 2.8, for visual analysis of the mentioned observations.

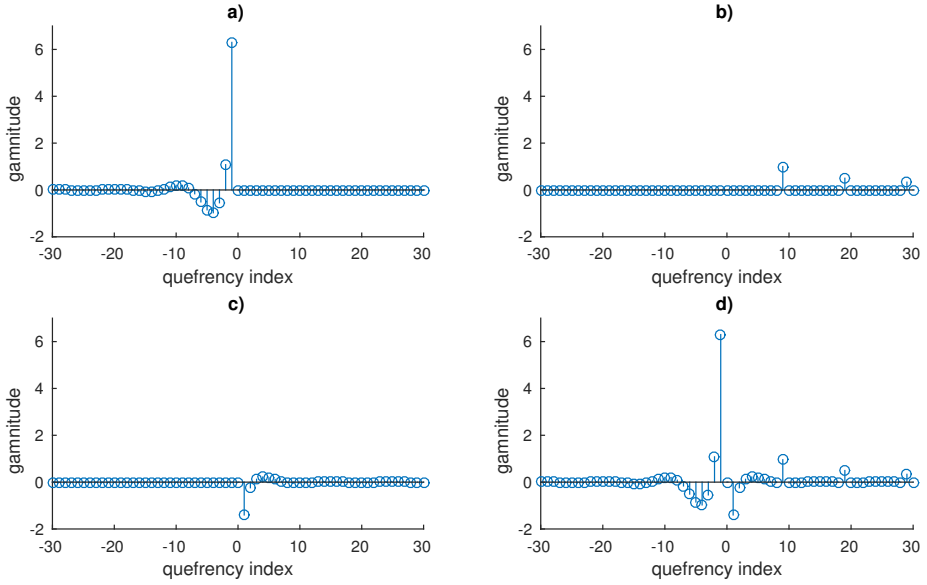


FIGURE 2.8: Cepstrum sequence for: a) Glottal Pulse $c_4[n]$, b) Truncated pulse train $c_2[n]$, c) Vocal tract transfer function $c_3[n]$, d) Sum of $c_2[n]$, $c_3[n]$ and $c_4[n]$

2.2.1.4 Impedance Based Inverse Filtering

The Impedance Based Inverse Filtering (IBIF) is a subglottal inverse filtering method to estimate glottal airflow through an accelerometer attached to the neck surface [28]. The method uses a sub-glottal model based on mechano-acoustic transmission line principles, including neck skin properties. Subject-specific parameters need to be estimated to adjust the model, which is performed by contrasting the IBIF estimates with glottal airflow recordings of a pneumatograph mask [16] flow sensor and acceleration signal filtered with IBIF. This model-based subglottal scheme is illustrated in Figure 2.9, where the electrical equivalent circuit shows the interconnection between the *sub1* and *sub2* systems (the tracts above and below of the location accelerometer) and the impedance of the skin Z_{skin}, which also includes the impedance load of the accelerometer Z_{rad}. The glottal airflow estimate from the accelerometer signal [12] is calculated by equation (2.41),

$$\hat{u}_g(t) = \mathcal{F}^{-1}\left(\frac{-\dot{U}_{skin}(\omega)}{j\omega} \cdot A_{acc} \cdot \frac{1}{T_{skin}(\omega)}\right), \tag{2.41}$$

with

$$\frac{1}{T_{skin}(\omega)} = \frac{Z_{sub2}(\omega) + Z_{skin_{acc}}(\omega)}{H_{sub1}(\omega) \cdot Z_{sub2}(\omega)}, \tag{2.42}$$

$$Z_{skin_{acc}}(\omega) = \frac{R_m + j\left(\omega M_m - \dfrac{K_m}{\omega}\right) + Z_{rad}(\omega)}{A_{acc}}, \tag{2.43}$$

$$Z_{rad}(\omega) = \frac{j\omega \cdot M_{acc}}{A_{acc}}, \tag{2.44}$$

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*, 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. The terms Z_{sub2} and H_{sub1} are calculated by an anatomically based, acoustic model of the sub-glottal system [91], and a transmission-line for the trachea segments above and below of the accelerometer location [28]. The subject-specific parameters of IBIF are scale factors of a mechanical impedance model of neck skin surface, length of the trachea, and accelerometer location from glottis. The parameters are represented by an electrical equivalent circuit in a set $\mathbf{Q} = \{Q_i\}_{i=1,\ldots,5}$ for mechanical resistance R_m, mechanical mass M_m, mechanical stiffness K_m (as shown in equation (2.43)), and lengths $L_{trachea}$ and L_{sub1} as we see in the schematic in Figure 2.9. The magnitude terms in equations (2.43) are the default values for each parameter [84], which are scaled for normalized Q factors as, $R_m = 2320 \cdot Q_1$ in $(\text{g} \cdot \text{s}^{-1} \cdot \text{cm}^{-2})$, $M_m = 2.4 \cdot Q_2$ in $(\text{g} \cdot \text{cm}^{-2})$, $K_m = 491000 \cdot Q_3$ in $(\text{dyn} \cdot \text{cm}^{-3})$, and for $Z_{sub2}(\omega)$ in equation (2.42), $L_{trachea} = 10 \cdot Q_4$, and $L_{sub1} = 5 \cdot Q_5$ are in (cm). Note that default model parameters are obtained for $\mathbf{Q} = [1, 1, 1, 1, 1]$ [28].

In all previous efforts related to IBIF [27, 28], it was required to use recordings of oral airflow from a *Rothenberg* (or CV) mask [16] for a sustained vowel /a/. This signal is used for calibrating the subject-specific model, i.e., find the optimal values of Q parameters of neck skin model that "best" match the CV mask glottal estimate. According to [28], for a given glottal flow $\tilde{u}_g(nT)$ estimation from the *Rothenberg* mask, and an accelerometer inverse filtered signal $\hat{u}_g(nT)$, Q parameters are estimated by minimizing a mean square error (MMSE) function of several aerodynamic measures including the constrains $\mathbf{D} = \{D_i\}_{i=1,\ldots,5}$,

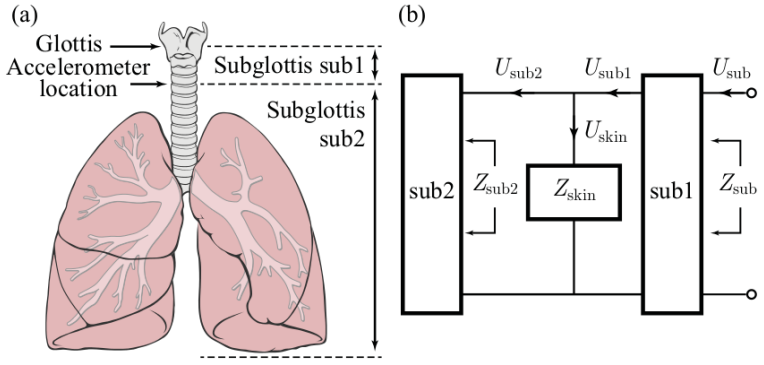


FIGURE 2.9: Representation of the subglottal system (a) Accelerometer position and *sub1* and *sub2* system parts. (b) A mechano-acoustic analogy of the sub-glottal system including load impedance from skin. Image is from Zañartu et al. (2013) [28].

corresponding to each parameter of set $\mathbf{Q}$ as shown in:

$$D_i = \begin{cases} [0.1, 10] & i = 1, 2, 3 \\ [0.8, 1.2] & i = 4, 5, 6 \end{cases} . \tag{2.45}$$

Then, the constrained optimization problem to solve is,

$$\mathbf{Q}^* = \arg\min_{\mathbf{Q}} E(\mathbf{Q}), \text{ subject to } \mathbf{Q} \in \mathbf{D} \quad , \tag{2.46}$$

In order to solve this optimization problem, a Particle Swarm Optimization (PSO) algorithm [92] has been proposed for a point estimate of Q parameters. PSO is an evolutionary computational method based on the social behavior of individuals [92]. PSO has the advantage of not requiring any special assumptions in the sense of classic convex optimization problems, and it is very efficient in solving large scale linear and non-linear constrained problems. However, PSO does not guarantee an optimal solution will ever be found. IBIF performance has been shown with sustained vowels, at constant pitch and comfortable loudness, which is a static test scenario from normal voices [12] [79] [28]. An example of IBIF method is shown in Figure 2.10.

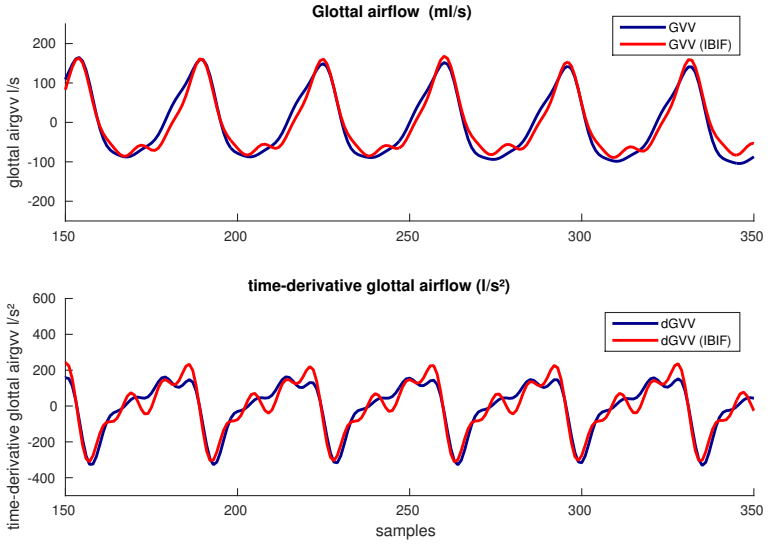


FIGURE 2.10: IBIF example using sustained vowel /a/. In blue, estimates from inverse filtered oral airflow signal, and in red from inverse filtered accelerometer signal. Top, glottal airflow. Bottom, time derivative of glottal airflow.

2.2.2 Assessing Glottal Inverse Filtering

Two approaches to assess inverse filtering process based in objective judgment that have been used in previous studies [83, 93, 94] are reviewed in this section. The first one is based on the group-delay (a frequency domain approach) calculation of the glottal airflow estimation [83], and the second one is based on phase-plane plots of the glottal airflow estimation and its time-derivative [93, 94]. Both methods share similar criteria, which is to minimize the formant ripple in the closed phase portion of the resulting inverse filtered waveform.

Flat Group-delay: To check the suppression of the resonant effects of vocal tract, the group delay ($g.d.$) of the resulting inverse filtered glottal pulse is expected to be constant [83], i.e.,

$$g.d. = -\frac{d\phi}{d\omega} = constant \quad , \tag{2.47}$$

where ϕ is the phase of the frequency response, and ω the frequency variable. In such condition of a flat group-delay, glottal pulse can be characterized with only zeros outside of the unit circle in the z-plane. This is equivalent to a maximum-phase glottal source, the assumption of the inverse filtering methods such as homomorphic signal processing [88] and zero z-transform approach [82],

which are based on the separation of min-max phase components of the unfiltered signal. Figure 2.11 shown the changes in group delay spectra and its corresponding glottal flow estimation, as discusssed in [83].

FIGURE 2.11: Inverse filtering assessment analysing group delay of glottal airflow estimate. Left panel shows glottal pulses after IF. Right panels shown group delay for each glottal pulse on the left. Note the more flat the group delay, more well defined the glottal pulses are. Image is from Alku et. al. (2005) [83]

Phase-plane plots: A phase-plane plot is a graphical representation of some classes of differential equations. In the inverse filtering literature, it is used to assess the glottal airflow estimates [93, 94]. This technique is based on the premise that the vocal tract can be modeled as a cascade of second order resonators [76], for which, the glottal pulse can be viewed as a second-order harmonic equation such that

$$\frac{d^2x}{dt^2} + x = 0 \ . \tag{2.48}$$

Using the substitutions $x = x_1$ and $\frac{dx}{dt} = x_2$, equation (2.48) can be written as a system of first-order differential equations,

$$\frac{dx_1}{dt} = x_2 \quad ; \quad \frac{dx_2}{dt} = -x_1 \tag{2.49}$$

which establishes a relationship between x_1 and x_2 (x against dx) that could be plotted in a phase-plane graph. Glottal waveforms without formant ripple should be cyclic in the phase-plane, i.e., simply closed loops (not necessarily circular). Vocal tract resonances introduce different dynamics, showing sub-cycles (or self-intersect) that turn inside the closed loop of the glottal cycle. Figure 2.12 shows an example of phase-plane applied to inverse filtering.

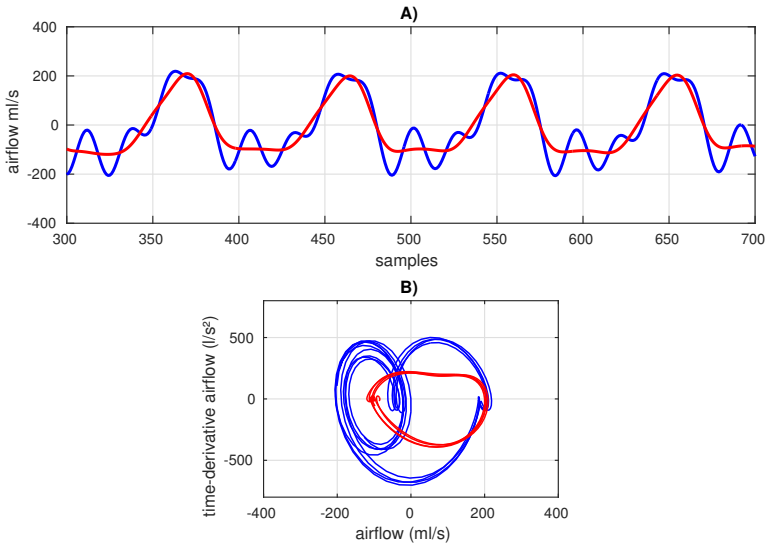


FIGURE 2.12: Example of inverse filtering assessment using the phase-plane plot. A) Airflow waveforms: oral (blue) and glottal (red). B) Phase-plane plots for the waveforms in A). In B), Blue-lines have several sub-cycles (containing resonance information about the vocal tract, i.e., the formant ripples). Instead, the red-line which have only glottal information.

Chapter 3

Enhanced methods for supraglottal and subglottal inverse filtering

This chapter describes enhanced methods to perform inverse filtering in the context of clinical voice assessment related to specific aim 1 of this thesis. The oral airflow and neck skin acceleration signals are considered for these purposes. The results show similar trends as those reported for sustained vowels in past experiments and outperform other methods under challenging conditions when a formant frequency is closer to fundamental frequency (e.g., vowel /i/). The results for aerodynamic measures show less than 10 % of relative error for vowels with one formant below 1kHz (/a/, /e/ , /i/). Exhaustive testing of aerodynamic measures are reported and their sensitivity is discussed in detail.

3.1 Single Notch Filter Inverse Filtering

An approach to inverse filtering (IF) oral airflow is presented in this section. The method consists of minimizing a cost function that control the parameters of a Single Notch Filter (SNF) that yields an estimate of glottal airflow from the oral airflow. Several cost functions (hereafter referred to as *Metrics*) are evaluated and compared with baseline data based on synthesized oral airflow simulations. The proposed metrics are inspired in common subjective criteria to assess IF, which have been reported in the context of clinical voice analysis [3, 25, 44, 95]. This

approach aims to provide an automatic and reliable IF scheme for large databases and running speech scenarios.

3.1.1 SNF+Metrics approach to inverse filtering oral airflow

In the context of clinical voice assessment, when detection of instant closure or closed phase is not available, or not reliable due to pathological conditions, inverse filtering with a single notch filter (see section 2.2.1.1 to further details), has been the preferred choice to reduce waveform ripple (due to formants) to produce a near flat amplitude in the closed phase [18] [35]. This method has been used in several studies in objective clinical evaluation of vocal hyperfunction and abnormal conditions [25] [31] [26], and remains as the most accepted method to IF pathological voices. Regardless of its simplicity and effectiveness, the approach relies on a subjective assessment that is time consuming and not suitable for analyzing large voice databases [83]. In particular, the most common goal of IF is to reduce formant ripple or obtain a flat closed phase, but no formal mathematical framework has been developed for such task.

In this section, a series of metrics are proposed based on heuristic criteria. Two criteria are defined to construct the set of metrics: 1) The criteria of *Formant Ripple*, which refers to oscillations in the closed phase, and *Flat closed phase* which refers to a straight line in the waveform in closed phase. Although these two criteria appear to be similar, they are not necessarily the same. For example, a resulting glottal estimation can be absent of ripple (due to formant), but the closed phase maybe not be a constant straight line (e.g., could have a slope). Table 3.1 shows six metrics based upon these criteria and their mathematical formulation. The signals involved on metrics calculations are defined as $x_{IF}(n)$ for the inverse filtered signal candidate, and $x_{OVV}(n)$ for the unfiltered OVV signal. Both signals must be absent of the DC component of the CV mask and low-pass filtered @1.1 kHz (see section 4.1 for details) to proper calculation of the metrics.

For the first and second metrics (items 1 and 2 in Table 3.1), N is the number of samples, and the $\Delta^{1,2}$ is the first(second)-order time-derivative operator [51] defined as,

$$\Delta x\,(n) \stackrel{\triangle}{=} (h * x)\,(n) \tag{3.1}$$

$$\Delta^2 x\,(n) \stackrel{\triangle}{=} \Delta\left\{\Delta x\,(n)\right\} \quad , \tag{3.2}$$

TABLE 3.1: Proposed Metrics to improve IF process using SNF. $x_{IF}(n)$ is the inverse filtered signal, $x_{OVV}(n)$ the unfiltered OVV signal and f_c is the central frequency of antiformant (SNF).

Item	Criterion	Mathematical formulation
1	Formant Ripple	$\sum_{n=0}^{N-1}(\lvert\Delta^2 x_{IF}\rvert)$
2	Formant Ripple	$\sum_{n=0}^{N-1}(\lvert\Delta x_{IF}\rvert)$
3	Flat closed phase	$(\min(x_{OVV}) - \min(x_{IF}))$
4	Flat closed phase	$-\lvert\hat{P}^{-1}_{x_{IF}}(0.99)/\hat{P}^{-1}_{x_{IF}}(0.01)\rvert$
$5^{(a)}$	Flat closed phase	$-\max\hat{p}(x_{IF})$
6	Formant Ripple	phase-plane plot

(a) Subject to $-\max\hat{p}(x_{IF}) < 0$

with $h(n) = \{\delta(n) - \delta(n-1)\} \cdot f_s$, and $*$ indicating the convolution operator.

The $\Delta^{1,2}$ operator emphasizes ($\approx 6(12)$ dB/8^{ve}) the effects of first formant (F_1) of $x_{IF}(n)$, and is hypothesized to reach a minimum when the central frequency of SNF (f_c) is approximately the true first formant F_1. These two metrics are exemplified in 3.1 and 3.2. For the third metric (item 3 in Table 3.1 and depicted in Figure 3.3), the difference between the minimum values of the filtered and unfiltered signals reach a maximum when the central frequency f_c is approximately F_1. For the fourth metric (item 4 in Table 3.1 and depicted in Figure 3.5) the relation between maximum negative value (using the 1% percentile, $\hat{P}^{-1}_{x_{IF}}(0.01)$ [96]) and maximum positive deviation (using the 99% percentile, $\hat{P}^{-1}_{x_{IF}}(0.99)$ [96]), captures the asymmetry of these peaks values from the glottal waveform. When format ripple is present (i.e., the unfiltered OVV signal) the negative peak amplitude is higher than the negative peak when is filtered (i.e.,the glottal airflow estimate). Therefore, this metric would be lower than -1 in most of the cases. This *asymmetry ratio* is believed to reach a minimum when f_c is closets to F_1. The fifth metric (item 5 in Table 3.1 and depicted in Figure 3.4), is based on an approximate probability density function (pdf) $\hat{p}(x_{IF})$ of the amplitudes of x_{IF}. When f_c approaches F_1, a *peaky* distribution centered around the amplitude values of the closed phase interval emerges [94]. Using a Kernel Density Estimation (see details in section C.2.1) an example is pictured in Figure 3.5 as the $\max\hat{p}(x_{IF})$. The sixth metric, the phase plane plot, was already explained in section 2.2.2, picking out the lower f_c, in absence of resonant sub-cycles.

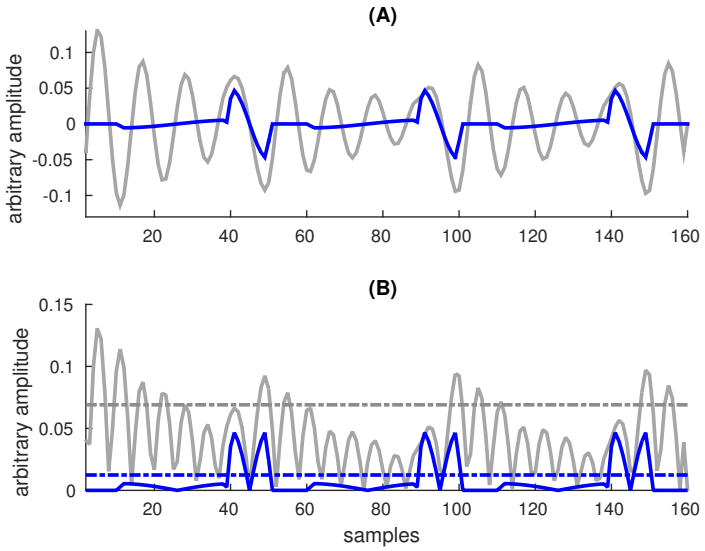


FIGURE 3.1: First metric explanation (according to Table 3.1). In (A), blue-bold, the *best* filtered $\Delta^2 x_{IF}$ (when $f_c \approx F_1$), and in gray, the $\Delta^2 x_{OVV}$ signal. In (B) the absolute value of signals in (A) with the addition of horizontal lines indicating the value when equation $\sum_{n=0}^{N-1}(|\Delta^2 x_{IF}|)$ is evaluated. Note the reduction of formant ripple.

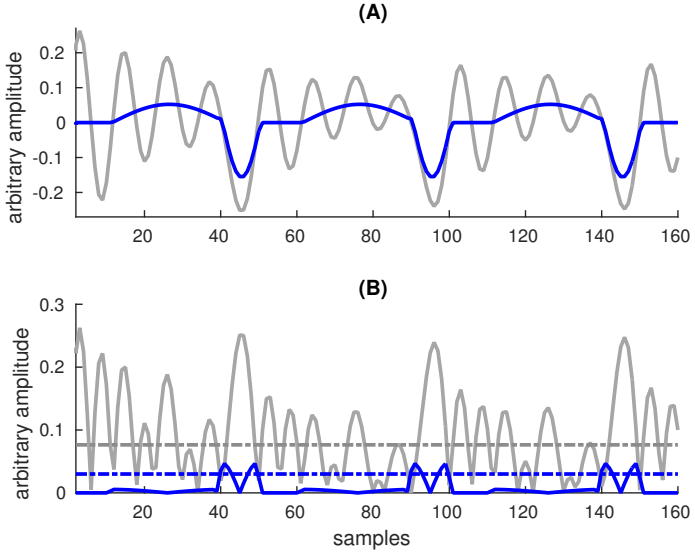


FIGURE 3.2: Second metric explanation (according to Table 3.1). In (A), blue-bold, the *best* filtered Δx_{IF} (when $f_c \approx F_1$), and in gray, the Δx_{OVV} signal. In (B) the absolute value of signals in (A) with the addition of horizontal lines indicating the value when equation $\sum_{n=0}^{N-1}(|\Delta x_{IF}|)$ is evaluated. Note the reduction of formant ripple.

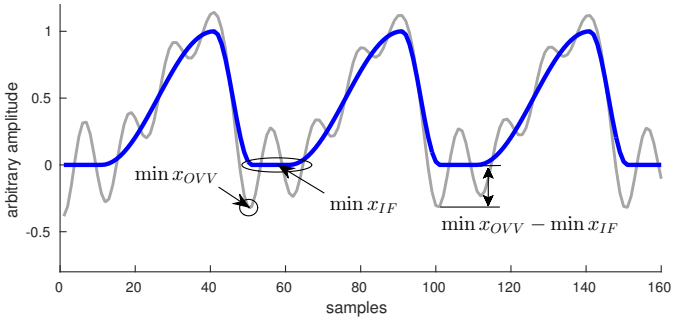


FIGURE 3.3: Third metric explanation (according to Table 3.1). In blue (bold), the *best* filtered x_{IF} (when $f_c \approx F_1$), and the indication of its $\min x_{IF}$. In gray, the x_{OVV} and its minimum value $\min x_{OVV}$. The difference of these minima is pictured as well.

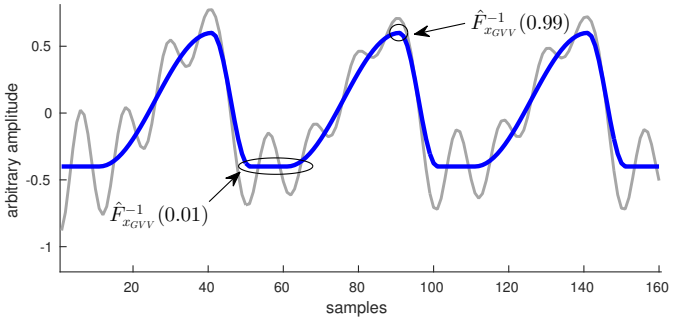


FIGURE 3.4: Fourth metric explanation (according to Table 3.1). In blue (bold), the filtered x_{IF}, and the indication of its $\hat{P}^{-1}_{x_{IF}}(0.99)$(positive peak) and $\hat{P}^{-1}_{x_{IF}}(0.01)$(negative peak) values. In gray, the x_{OVV} is pictured as a reference.

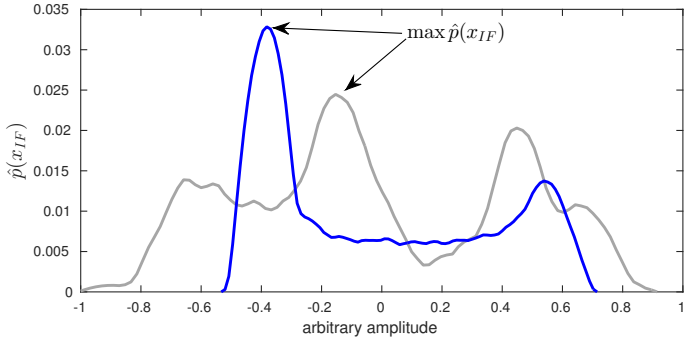


FIGURE 3.5: Fifth metric explanation (according to Table 3.1). In gray, the x_{OVV} has a multimodal p.d.f. (with two prominent univariate ones), compared with the x_{IF} (in blue) showing a more peaky pdf corresponding to the amplitude values centered in the closed phase.

3.1.2 Evaluating the SNF+Metrics approach

Before validating the proposed metrics, examples are presented as an illustration of the potential of the SNF+Metrics approach. A female subject with healthy voice was asked to sustain two vowels, /a/ and /i/, using comfortable loudness condition. It is well known that performing IF for the vowel /i/ is more challenging than for the vowel /a/. For the recorded signals, a stable (in pitch and amplitude) voice segment was selected and all metrics were calculated from the OVV signal. For the vowel /a/, the results are pictured in Figures 3.8, 3.9, and 3.7. In Figure 3.8 the normalized metrics are plotted versus the central frequency of the SNF with a fixed bandwidth of 70 Hz. The minimum values are close of the expected frequency range of typical first formant (700-800 Hz). A manual adjustment of the SNF finds that 710 Hz was the best user choice in the preceding interval. In Figures 3.9 and 3.7 all the proposed SNF+Metrics reduce the ripple in the closed phase with slight differences, in both glottal airflow estimate (Figure 3.8) and its time-derivative (Figure 3.9). For the vowel /i/, normalized metrics are pictured in Figure 3.10. Minimum values are lower and more dispersed than for the vowel /a/, and centered in the range of 350 to 550 Hz. A manual adjustment of SNF finds that 390 Hz was the best user choice. In Figures 3.11 and 3.7 each SNF+Metrics approach was applied, wherein glottal pulses related to metrics 1, 2, 4, and 6 seem to be a good approximation. These proof-of-concept examples show that the SNF+Metrics approach has the potential to perform automatic inverse filtering, even in challenging cases like vowel /i/. In the next section, formal validation of SNF+Metrics is described using synthesized glottal waveforms and numerical models of voice production.

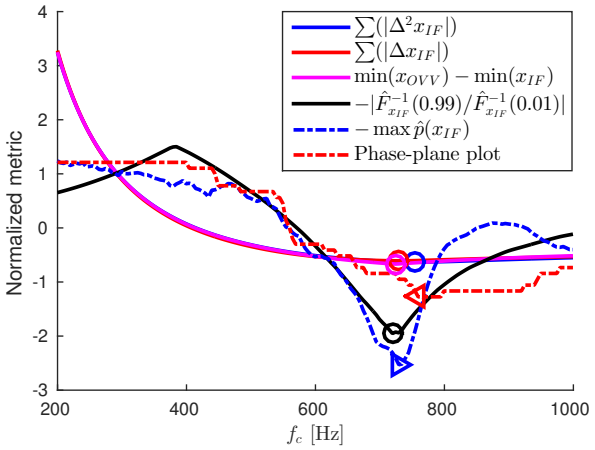


FIGURE 3.6: Normalized metrics curves showing minimum values (colored marker). For this example, the vowel /a/ from a female voice was analyzed. The minimum values of the notch frequency f_c are between 719 to 759 Hz, which is in the expected range for the first formant for vowel /a/. Circles and triangles are indicating the minimum value for a given metric.

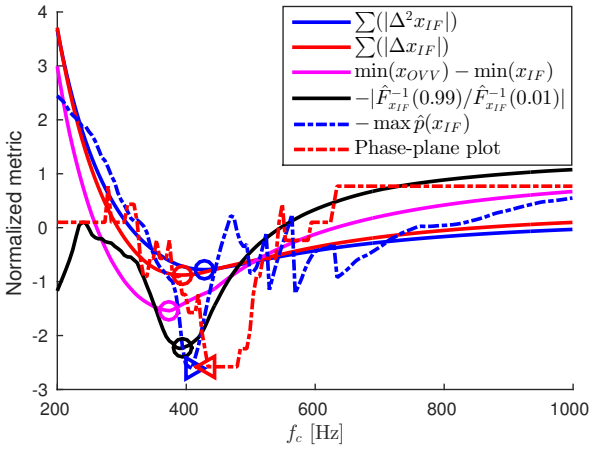


FIGURE 3.7: Normalized metrics curves showing minimum values (colored marker). For this example, the vowel /i/ from a female voice was analyzed. The minimum values are between 374 to 564 Hz, a wider range than expected (vowel /i/ first format is usually in the range of 250 to 350 Hz [76]). Circles and triangles are indicating the minimum value for a given metric.

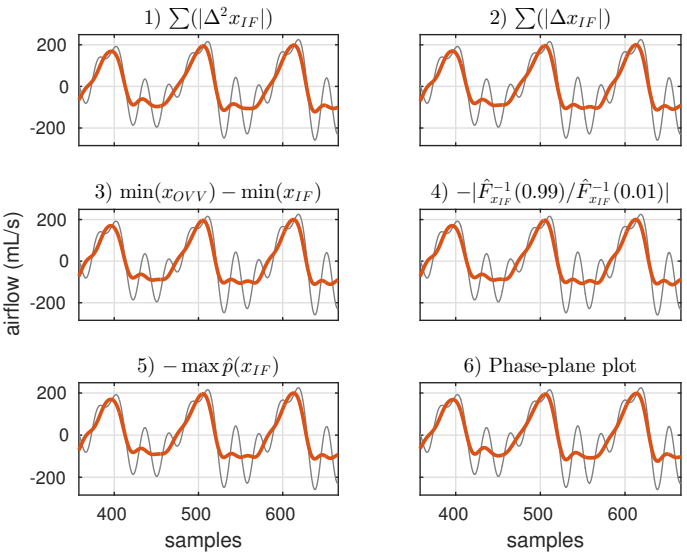


FIGURE 3.8: Example for the estimated glottal airflow using SNF+Metrics approach (solid orange) for a vowel /a/ (female voice). For this example, the formant residuals are small, and the *expected* glottal shapes are similar for all cases. In gray, oral airflow is drawn as a reference.

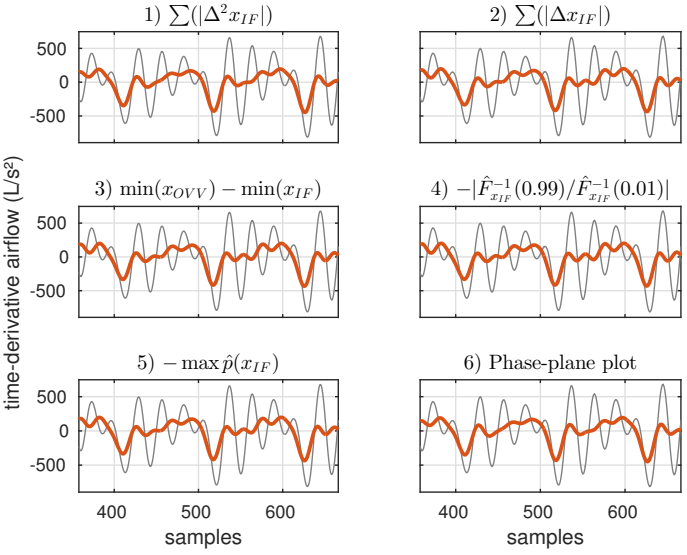


FIGURE 3.9: The time-derivative glottal airflow of Figure 3.8.

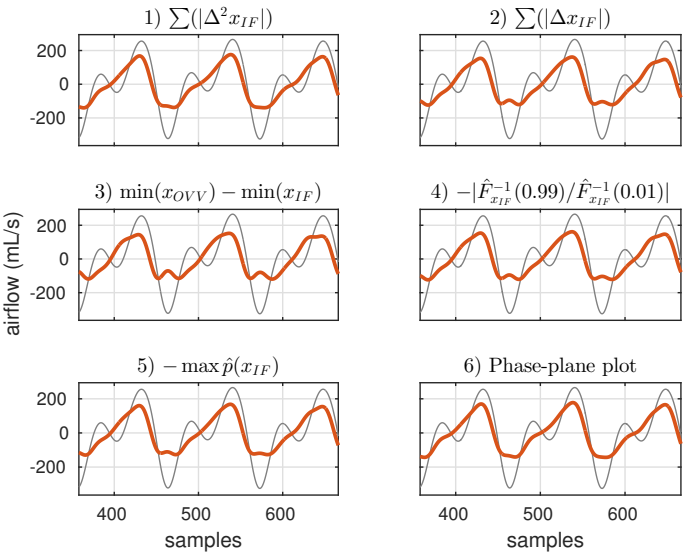


FIGURE 3.10: Example for the estimated glottal airflow using SNF+Metrics approach (solid orange) for a vowel /i/ (female voice). For this example, the formant residuals are reduced and the *expected* glottal pulses with the flattest closed phase are achieved with metrics 1 and 6. In gray, oral airflow is drawn as a reference.

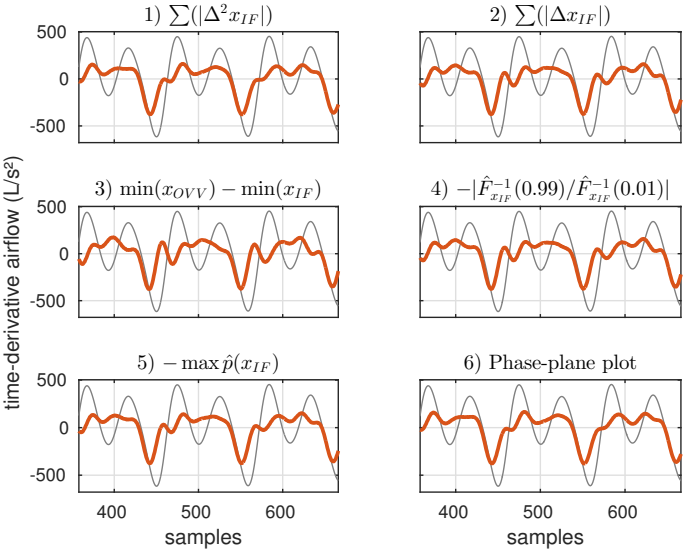


FIGURE 3.11: The time-derivative glottal airflow of Figure 3.10.

3.1.3 Validating SNF+Metrics with Synthesized Glottal Waveforms

Synthesized glottal waveforms with a single formant are designed to incorporate different glottal shapes and voice frequencies to validate the SNF+Metrics approach under various conditions. The Rosenberg model [97] is used to parameterize glottal pulses in the opening, closing, and closed phases. In the

Rosenberg 1971 study [97], multiple glottal pulse shapes are reported but a particular one is selected (equation (3.3)) for not having discontinuities (more prone to our limited bandwidth measures of 1 kHz, approximately), and is presented in equation (3.3),

$$g_R(t) = \begin{cases} \dfrac{a}{2}\left(1 - \cos\left(\pi \dfrac{t}{T_P}\right)\right) & ; \quad 0 \leq t \leq T_P \\ \\ \dfrac{a}{2}\left(1 + \cos\left(\pi \dfrac{t - T_P}{T_N}\right)\right) & ; \quad T_P \leq t \leq T_P + T_N \quad , \end{cases} \tag{3.3}$$

wherein T_P, T_N, a and t are the opening time, closing time, peak-to-peak amplitude and time, respectively. An example is shown in Figure 3.12 for multiple opening and closing times. However, it is required to remark that this type of pulse shape, may not reflect pathological speech.

Several glottal pulses are generated with the Rosenberg model, varying the opening time (T_P), closing time (T_N), and peak-to-peak amplitude (a). Random perturbation ($\pm 1\%$) are incorporated in the time-amplitude parameters to simulate quasi-periodicity and mimick signal to noise ratio found in real measures of normal speech. Formant filter at a given central frequency F_1 (baseline) is applied to all synthesized glottal pulses. The DC component was removed and glottal pulses were low-pass filtered @1.1 kHz (see details in section 4.1.3). The automated SNF+Metrics approach was applied and all formant candidates ($\hat{F}_1$) are stored to further analysis. Each synthesized glottal pulse was characterized with a combination of lower (160 Hz) and higher (320 Hz) pitch, and lower (350 Hz) and higher (650 Hz) formant. For each vowel a total of N=329 synthesized glottal waveforms were simulated.

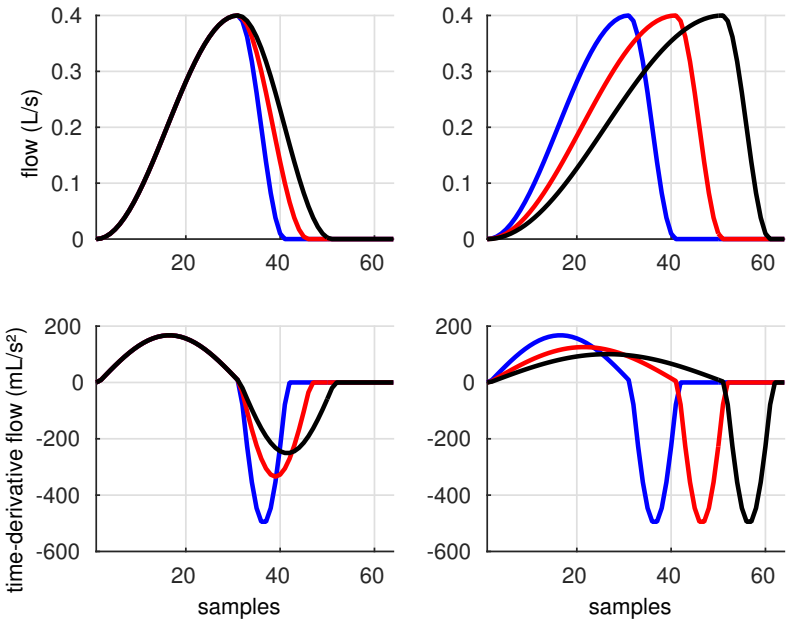


FIGURE 3.12: Rosenberg glottal pulse examples. Top-Left: glottal pulses for several closing times. Bottom-Left: time-derivative of glottal pulses from Top-Left. Top-Right: glottal pulses for several opening times. Bottom-Right: time-derivative of glottal pulses from Top-Right.

For the $\hat{F}_1$ formant candidate, a formant deviation from baseline, $(F_1 - \hat{F}_1)$, is calculated. For each metric, median values and percentile 25% and 75% of this deviation are computed, and are presented in Figures 3.13 and 3.14, for the Lower Formant (F_L) and Higher Formant (F_H), respectively. Both results, F_L and F_H, show small deviation (< 15 Hz) from median values (filled circles) in all cases except for F_H, in which the phase-plane metric has a greater deviation (> 50 Hz). The dispersion (percentile 25% and 75%) looks asymmetric with a small range ($<$ 50 Hz) except for F_H, in which the phase-plane metric have a greater range (> 100 Hz). In this analysis, the phase-plane metric seems to be the weakest feature of the group. Metrics (4) $\min(x_{OVV}) - \min(x_{IF})$, (5) $-|\hat{P}^{-1}_{x_{IF}}(0.99)/\hat{P}^{-1}_{x_{IF}}(0.01)|$, and (6) $-\max \hat{p}(x_{IF})$ exhibit the best performance in this experiment.

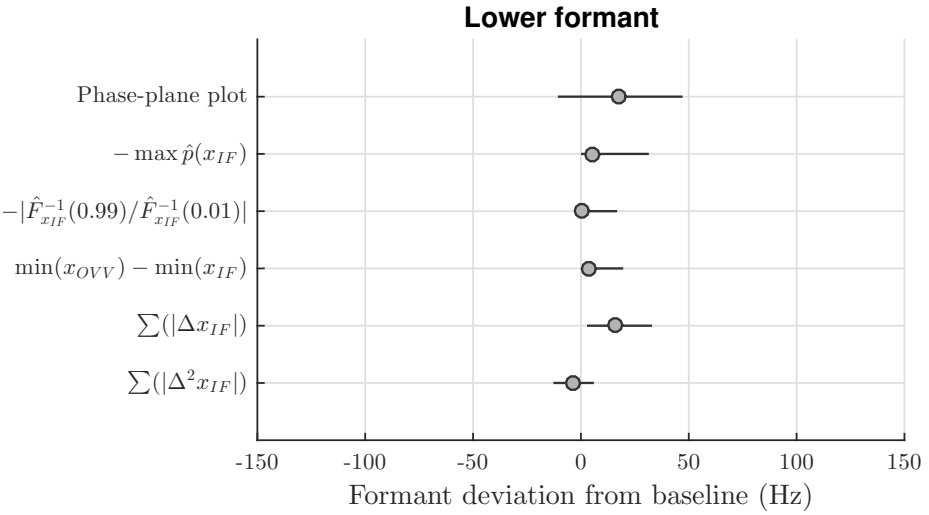


FIGURE 3.13: Metrics performance of the lower formant for the SNF+Metrics approach. Filled circles represent the median values, and horizontal lines the range for percentiles 25% and 75%)

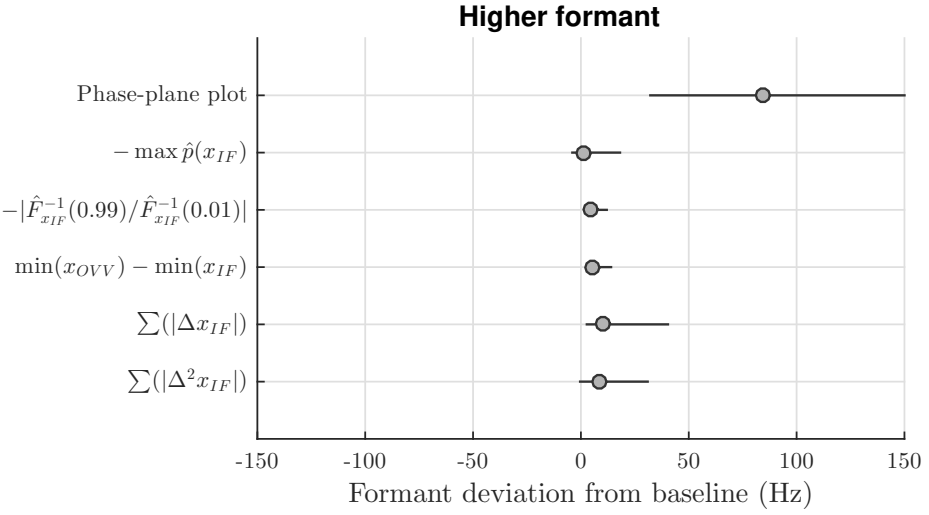


FIGURE 3.14: Metric performance of the higher formant for the SNF+Metrics approach. Filled circles represent for the median values, and horizontal lines the range for percentile 25% and 75%)

3.1.4 Validating SNF+Metrics with self-sustained numerical models

Research in self-sustained numerical models of voice production is extensive [6, 57, 98–100]. They are based on a range of approaches from low-dimensional models [98] to finite element methods [100]. In this thesis, the choice of modeling approach was based on 1) simplicity (i.e., low-dimensional model), 2) flexibility (i.e., a desired set of model parameters representing the physiology of the vocal folds), and 3) the ability to account for source-filter interactions [101, 102], with the aim to obtain more realistics simulations, (e.g., vocal hyperfunction), inspired by the work of Story in [17]. The voice production model presented in Zañartu et al. [6], including posterior glottal opening (PGO), was used for our simulations with the aim to validate the proposed inverse filtering methods. The details of such model above can be found in [6], wherein the key parameters to control posturing are given for the cricothyroid activation (a_{CT}), thyroarytenoid activation (a_{TA}), lateral cricoarytenoid activation (a_{LC}) [103], all values related to muscle evolved in vocal fold phonation. Using the muscle activation plots in [103] allows for selecting adequate parameter values for our simulations. A baseline signal (glottal airflow) was generated with the numerical model of the voice production to test the proposed inverse filtering schemes. Several vowels (/a/, /e/, /i/, /o/ and /u/) and vocal fold configurations were selected, including variations of posterior glottal opening (PGO $\in [0.043, 0.050, 0.058]$), and muscle cricoteroid activation ($a_{CT} \in [0.34, 0.40, 0.46]$), to obtain nine glottal configurations with high fundamental frequencies (in the range of a female voice [76]) for each vowel. The subglottal pressure was fixed for all simulations at 800 Pa and the lateral cricoarytenoid activity (a_{LC}) and thyroarytenoid activity (a_{TA}) to 0.5 and 0.33, respectively. For each vowel and vocal fold simulation, a stable segment was selected to perform inverse filtering analysis. Each segment was resampled to 8 kHz, selecting 512 time samples for analysis. Both signals, the baseline glottal waveform from numerical models and the estimated using the SNF-Metrics approach were low pass filtered (cutoff @1.1 kHz) and the DC component was removed, following same methods that in section 4.1.3. Before performing inverse filtering, a window (Tukey with $\alpha = 0.5$) was applied to avoid abrupt changes at the start and end of the voice frame. The frequency range of simulations was from 155 Hz to 235 Hz approximately, with a mean($\pm$std) of 198($\pm$22.4) Hz.

A normalized error was calculated for the glottal airflow and its time-derivative as,

$$\text{GVV error } \equiv 100\% \cdot \frac{\sum_{n=0}^{N-1} |x_0(n) - x_{IF}(n)|}{\sum_{n=0}^{N-1} |x_0(n)|} \tag{3.4}$$

$$\text{dGVV error } \equiv 100\% \cdot \frac{\sum_{n=0}^{N-1} |\Delta x_0(n) - \Delta x_{IF}(n)|}{\sum_{n=0}^{N-1} |\Delta x_0(n)|} \tag{3.5}$$

where x_0 was the baseline from synthesized glottal waveforms, and x_{IF} the inverse filtered signal.

The estimated glottal flow waveform was shifted in time to compensate for group-delay effects after IF processing to obtain a synchronized GVV and dGVV error estimation.

To evaluate similarities of the estimated glottal airflow and the ground truth reference, the correlation coefficient (equations (3.6) and (3.7)) are calculated for both signals, glottal airflow, and its time-derivative, as follows:

$$r_{GVV} = \frac{\sum_{n=0}^{N-1} \left(x_{IF}(n) - \overline{\mathbf{x_{IF}}}\right)\left(x_0(n) - \overline{\mathbf{x_0}}\right)}{\sqrt{\sum_{n=0}^{N-1} \left(x_{IF}(n) - \overline{\mathbf{x_{IF}}}\right)^2} \sqrt{\sum_{n=0}^{N-1} \left(x_0(n) - \overline{\mathbf{x_0}}\right)^2}} \tag{3.6}$$

$$r_{dGVV} = \frac{\sum_{n=0}^{N-1} \left(\Delta x_{IF}(n) - \overline{\mathbf{\Delta x_{IF}}}\right)\left(\Delta x_0(n) - \overline{\mathbf{\Delta x_0}}\right)}{\sqrt{\sum_{n=0}^{N-1} \left(\Delta x_{IF}(n) - \overline{\mathbf{\Delta x_{IF}}}\right)^2} \sqrt{\sum_{n=0}^{N-1} \left(\Delta x_0(n) - \overline{\mathbf{\Delta x_0}}\right)^2}} \tag{3.7}$$

where,

$$\overline{\mathbf{x_{IF}}} = \frac{1}{N} \sum_{n=0}^{N-1} x_{IF}(n) \tag{3.8}$$

$$\overline{\mathbf{x_0}} = \frac{1}{N} \sum_{n=0}^{N-1} x_0(n) \tag{3.9}$$

$$\overline{\mathbf{\Delta x_{IF}}} = \frac{1}{N} \sum_{n=0}^{N-1} \Delta x_{IF}(n) \tag{3.10}$$

$$\overline{\mathbf{\Delta x_0}} = \frac{1}{N} \sum_{n=0}^{N-1} \Delta x_0(n) \tag{3.11}$$

TABLE 3.2: Mean (std) values of the SNF+Metrics inverse filtering approach using numerical models. Fundamental frequency (f_0) and formant frequency (F_1) estimates, GVV and dGVV normalized absolute error (in units of % to evaluate waveforms deviations), and correlation coefficient r_{GVV} / r_{dGVV} (to evaluate waveform similarities) are reported. In bold, r_{GVV} and r_{dGVV} values that are greater than 0.9 are highlighted as a good indicator of waveform similarity.

Metric	vowel	f_0 (Hz)	F_1 (Hz)	GVV error	dGVV error	r_{GVV} / r_{dGVV}
$\sum_{n=0}^{N-1}(\lvert\Delta^2 x_{IF}\rvert)$	/a/	208.7 (18.6)	819.2 (48.4)	16.6 (2.1)	44.3 (4.1)	**0.97(0.00)** / 0.86(0.02)
	/e/	210.8 (19.1)	711.2 (74.4)	22.3 (2.1)	45.7 (3.6)	**0.96(0.01)** / 0.86(0.01)
	/i/	169.6 (10.7)	361.0 (23.4)	57.1 (3.6)	76.9 (3.0)	0.63(0.12) / 0.48(0.07)
	/o/	209.2 (17.4)	848.4 (41.2)	38.3 (5.4)	65.0 (4.2)	**0.92(0.02**) / 0.69(0.04)
	/u/	189.6 (13.0)	494.5 (34.4)	56.7 (4.6)	76.1 (3.9)	0.79(0.04) / 0.56(0.03)
$\sum_{n=0}^{N-1}(\lvert\Delta x_{IF}\rvert)$	/a/	208.7 (18.6)	739.2 (22.0)	12.4 (0.8)	31.7 (3.8)	**0.98(0.00) / 0.91(0.01)**
	/e/	210.8 (19.1)	471.2 (10.1)	8.3 (1.3)	18.3 (3.7)	**0.99(0.00) / 0.94(0.01)**
	/i/	169.6 (10.7)	203.2 (3.7)	14.5 (3.3)	19.3 (2.6)	**0.96(0.01) / 0.91(0.01)**
	/o/	209.2 (17.4)	714.0 (35.9)	34.6 (6.0)	58.3 (5.8)	**0.93(0.02)** / 0.72(0.07)
	/u/	189.6 (13.0)	279.0 (14.0)	27.9 (3.5)	36.2 (1.9)	**0.94(0.01)** / 0.85(0.01)
$-\lvert\min(x_{OVV}) - \min(x_{IF})\rvert$	/a/	208.7 (18.6)	732.5 (23.7)	12.0 (0.9)	30.7 (4.6)	**0.98(0.00) / 0.91(0.01)**
	/e/	210.8 (19.1)	384.5 (48.3)	31.2 (17.7)	49.1 (19.9)	**0.94(0.05)** / 0.80(0.11)
	/i/	169.6 (10.7)	192.1 (3.0)	18.6 (4.3)	23.3 (4.4)	**0.96(0.01) / 0.91(0.01)**
	/o/	209.2 (17.4)	675.1 (43.5)	33.9 (5.3)	58.8 (6.8)	**0.93(0.02)** / 0.69(0.10)
	/u/	189.6 (13.0)	230.1 (4.4)	31.5 (8.0)	47.1 (5.7)	**0.93(0.03)** / 0.82(0.03)
$-\lvert\hat{P}^{-1}_{x_{IF}}(\alpha)/\hat{P}^{-1}_{x_{IF}}(1-\alpha)\rvert$	/a/	208.7 (18.6)	209.2 (18.9)	99.0 (0.2)	97.6 (0.5)	0.09(0.01) / 0.25(0.01)
	/e/	210.8 (19.1)	302.3 (88.9)	60.1 (30.2)	72.0 (19.7)	0.65(0.40) / 0.58(0.23)
	/i/	169.6 (10.7)	254.3 (48.8)	37.7 (21.2)	49.3 (26.6)	0.77(0.20) / 0.71(0.20)
	/o/	209.2 (17.4)	221.8 (38.2)	98.0 (1.7)	97.9 (0.6)	0.13(0.09) / 0.13(0.13)
	/u/	189.6 (13.0)	240.1 (17.0)	27.6 (5.4)	43.4 (4.8)	**0.95(0.02) / 0.84(0.02)**
$-\max\hat{p}(x_{IF})$	/a/	208.7 (18.6)	728.1 (121.3)	14.4 (6.4)	36.1 (17.6)	**0.97(0.01)** / 0.87(0.08)
	/e/	210.8 (19.1)	540.0 (247.4)	29.9 (29.5)	43.7 (26.5)	**0.90(0.18)** / 0.81(0.18)
	/i/	169.6 (10.7)	178.8 (5.3)	37.0 (6.9)	38.2 (5.0)	**0.91(0.03)** / 0.87(0.02)
	/o/	209.2 (17.4)	696.2 (94.3)	33.8 (8.1)	57.6 (7.5)	**0.93(0.03) / 0.72(0.06)**
	/u/	189.6 (13.0)	245.6 (37.8)	42.3 (25.1)	46.9 (17.3)	0.85(0.15) / 0.80(0.08)
phase-plane	/a/	208.7 (18.6)	731.4 (81.4)	12.4 (3.4)	32.5 (6.8)	**0.98(0.01) / 0.90(0.02)**
	/e/	210.8 (19.1)	563.4 (74.9)	15.1 (8.6)	27.3 (14.1)	**0.98(0.02) / 0.92(0.04)**
	/i/	169.6 (10.7)	225.4 (9.0)	30.0 (3.7)	35.8 (5.0)	0.89(0.03) / 0.85(0.02)
	/o/	209.2 (17.4)	704.0 (71.4)	34.1 (7.0)	57.7 (7.0)	**0.93(0.02)** / 0.72(0.07)
	/u/	189.6 (13.0)	296.7 (9.0)	34.4 (3.1)	40.8 (1.5)	**0.92(0.01)** /0.83(0.01)

In Table 3.2, the results for the proposed inverse filtering method based on the SNF+Metrics approach are presented. The results with numerical models highlighted the differences between the metrics to attempt inverse filtering. However, The metric with best performance using numerical models was $\sum_{n=0}^{N-1}(|\Delta x_{IF}|)$, with its detailed results provided in Table 3.3. First formant estimate ($\hat{F}_1$) was consistent (i.e., small variability) across the multiple voice simulations. Waveform similarities (valued with r_{GVV} and r_{dGVV}) are greater for vowel /a/, /e/ and /i/, and lower for vowel /o/ and /u/. A similar trend is shown in normalized errors (GVV error and dGVV error). In all cases, the time-derivative signal has the higher dGVV errors (and lower correlation coefficients). A remarkable observation is the performance in vowel /i/. With a reasonable accuracy, the SNF+Metric approach is capable of detecting a very low (and very near f_0) formant frequency, a scenario in which other inverse filtering methods have not shown good results [17, 104].

The uncertainties of the SNF+Metrics approach on the aerodynamic measures by using $\sum_{n=0}^{N-1}(|\Delta x_{IF}|)$ are reported in Table 3.4. A normalized error is computed as,

$$\text{AM } \% = 100\% \cdot \frac{|\text{AM}_{baseline} - \text{AM}_{estimated}|}{\text{AM}_{baseline}}, \tag{3.12}$$

where AM stands for **A**erodynamic **M**easure, e.g., for ACFL,

$$\text{ACFL } \% = 100\% \cdot \frac{|\text{ACFL}_{baseline} - \text{ACFL}_{estimated}|}{\text{ACFL}_{baseline}}, \tag{3.13}$$

and same calculations are performed with the other aerodynamic measures. The best performance is for OQ and ACFL in vowels /a/, /e/, and /i/, followed for CPP, MFDR and NAQ in the same vowels. A reduced performance is observed for H1H2. For vowels /o/ and /u/, NAQ, CPP and OQ are within 10% to 20% of error. For all others instances the error was higher than 20%.

TABLE 3.3: Detailed results of SNF+Metric using $\sum_{n=0}^{N-1}(|\Delta x_{IF}|)$. Parameters a_{CT} and PGO of the numerical models, fundamental frequency (f_0) and formant frequency (F_1) estimates, GVV and dGVV normalized absolute error, and correlation coefficient r_{GVV} / r_{dGVV} are reported. In bold, r_{GVV} and r_{dGVV} values that are greater than 0.9 are highlighted as a good indicator of waveform similarity.

vowel	a_{CT}	pgo	f_0	$\hat{F}_1$	GVV error	dGVV error	r_{GVV} / r_{dGVV}
/a/	0.34	0.043	189.1	709.1	12.6	35.0	**0.98** / 0.89
		0.050	187.7	727.7	12.7	36.5	**0.98** / 0.89
		0.058	186.3	746.3	12.7	37.8	**0.98** / 0.89
	0.40	0.043	210.2	710.2	11.3	27.7	**0.98 / 0.92**
		0.050	208.5	738.5	12.1	31.8	**0.98 / 0.91**
		0.058	207.1	737.1	11.3	29.6	**0.98 / 0.92**
	0.46	0.043	233.1	743.1	13.1	28.5	**0.98 / 0.91**
		0.050	231.2	771.2	13.8	30.4	**0.98 / 0.91**
		0.058	229.3	769.3	12.2	28.1	**0.98 / 0.92**
/e/	0.34	0.043	190.5	490.5	6.8	13.8	**0.99 / 0.95**
		0.050	189.1	479.1	6.8	14.0	**0.99 / 0.95**
		0.058	187.5	477.5	6.6	14.1	**0.99 / 0.95**
	0.40	0.043	212.9	472.9	8.7	18.2	**0.99 / 0.94**
		0.050	210.8	470.8	8.1	17.5	**0.99 / 0.94**
		0.058	209.1	459.1	8.5	19.4	**0.99 / 0.94**
	0.46	0.043	235.5	465.5	9.7	22.1	**0.98 / 0.93**
		0.050	233.5	463.5	9.6	22.5	**0.98 / 0.93**
		0.058	231.4	461.4	10.0	22.8	**0.99 / 0.93**
/i/	0.34	0.043	159.5	209.5	15.2	19.5	**0.96 / 0.92**
		0.050	157.6	207.6	12.8	17.4	**0.97 / 0.92**
		0.058	155.5	205.5	10.6	15.9	**0.97 / 0.93**
	0.40	0.043	171.5	201.5	12.5	17.7	**0.96 / 0.91**
		0.050	170.2	200.2	11.9	17.3	**0.97 / 0.91**
		0.058	168.2	198.2	12.4	18.4	**0.97 / 0.92**
	0.46	0.043	183.6	203.6	20.4	22.7	**0.94** / 0.88
		0.050	182.2	202.2	18.0	22.2	**0.95** / 0.89
		0.058	180.8	200.8	17.1	22.7	**0.95** / 0.89
/o/	0.34	0.043	192.0	672.0	29.6	54.0	**0.94** / 0.76
		0.050	189.8	659.8	27.1	52.5	**0.95** / 0.78
		0.058	187.4	697.4	26.4	54.4	**0.95** / 0.78
	0.40	0.043	211.8	741.8	38.7	62.2	**0.91** / 0.69
		0.050	209.6	749.6	36.0	56.5	**0.93** / 0.74
		0.058	207.3	747.3	32.4	50.8	**0.94** / 0.78
	0.46	0.043	231.8	681.8	42.5	65.3	**0.90** / 0.60
		0.050	229.3	739.3	40.3	65.2	**0.91** / 0.65
		0.058	226.9	736.9	38.8	63.5	**0.92** / 0.68
/u/	0.34	0.043	176.7	296.7	29.9	38.3	**0.93** / 0.84
		0.050	174.8	284.8	25.4	34.9	**0.95** / 0.86
		0.058	172.7	282.7	23.3	32.2	**0.96** / 0.87
	0.40	0.043	192.2	282.2	30.3	37.0	**0.93** / 0.84
		0.050	190.3	290.3	32.2	38.3	**0.93** / 0.84
		0.058	188.0	288.0	30.0	37.2	**0.94** / 0.85
	0.46	0.043	207.1	267.1	30.1	36.5	**0.93** / 0.85
		0.050	205.4	265.4	28.3	35.1	**0.94** / 0.86
		0.058	203.3	253.3	22.1	36.3	**0.96** / 0.87

TABLE 3.4: Results of aerodynamic measures from the proposed SNF+Metrics inverse filtering using metric 2 ($\sum_{n=0}^{N-1}(|\Delta x_{IF}|)$). Normalized absolute error of ACFL, MFDR, OQ, H1H2, CPP and NAQ are reported (in percentage from 0 to 100 %). Mean (and standard deviation) for each aerodynamic measure and vowel are calculated as well.

vowel	ACFL (%)	MFDR (%)	OQ (%)	H1H2 (%)	CPP (%)	NAQ (%)
/a/	4.2(3.1)	13.5(3.9)	2.2(1.7)	19.9(5.3)	12.7(7.8)	15.4(4.6)
/e/	2.3(0.7)	15.8(4.2)	3.3(2.3)	16.4(8.4)	12.2(11.1)	15.6(2.8)
/i/	11.7(6.8)	10.1(6.5)	10.1(8.3)	33.5(10.2)	10.3(9.9)	3.1(1.9)
/o/	21.2(4.6)	38.5(21.4)	13.3(8.1)	46.8(12.4)	17.6(10.6)	13.0(8.7)
/u/	26.1(6.9)	16.0(12.8)	20.5(4.1)	79.4(28.5)	12.8(10.1)	12.2(6.6)

3.2 Regularized Closed Inverse Filtering

In close phase analysis [17, 88], the covariance matrix $\mathbf{\Phi}$ is sensitive to the position and length of frame analysis, yielding unjustified inverse filter poles estimations at low frequencies [17], and considerable error in the estimated parameters. In addition, in voices with a higher pitch, the closed phase of the glottal cycle is too short, and CP analysis could erroneously fit poles near the fundamental frequency, as the opening phase raise within the frame. It is proposed a way to control the sensitivity of $\mathbf{\Phi}$ by imposing a weighting on the coefficients, with a regularized factor λ to estimate a more accurate solution [105].

A Regularized Closed Inverse Filtering (RCPIF) method is proposed and was derived from a *Maximum a Posteriori* (MAP) approach as follows. The MAP is obtained by solving equation (3.14)

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} p(\mathbf{a}|\mathbf{x_n}) \quad , \tag{3.14}$$

where $p(\mathbf{a}|\mathbf{x_n})$ is the probability function of $\mathbf{a}$ given $\mathbf{x_n}$, $\mathbf{a} = [a_0, a_1, a_2, \ldots, a_p]^T$ is the parameter vector, and $\mathbf{x}_n = [x_{n-0}, x_{n-1}, \ldots, x_{n-p}]$ the vector of voice data. Using Bayes theorem, we get

$$p(\mathbf{a}|\mathbf{x_n}) = \frac{p(\mathbf{x_n}|\mathbf{a})p(\mathbf{a})}{p(\mathbf{x_n})} \quad ,$$

then, equation (3.14) can be rewritten as,

$$\hat{\mathbf{a}} = \arg\max_{\mathbf{a}} p(\mathbf{x_n}|\mathbf{a})p(\mathbf{a}),$$

where $p(\mathbf{x_n}|\mathbf{a})$ is known as the *likelihood*, and $p(\mathbf{a})$ as the *a priori* distribution of $\mathbf{a}$. The *evidence* $p(\mathbf{x_n})$ is dropped because it does not depend on $\mathbf{a}$. Both probability functions, $p(\mathbf{x_n}|\mathbf{a})$ and $p(\mathbf{a})$, are derived as follow. The joint density [69, 106] of $(x_0, x_1, \ldots, x_{N-1})$ is

$$p(\mathbf{x_n}|\mathbf{a}) \equiv p(x_0, x_1, \ldots, x_{N-1}) = p(x_0) \prod_{i=1}^{N-1} p(x_i|\mathbf{X_{i-1}}) \quad , \tag{3.15}$$

where $\mathbf{X_{i-1}}$ denotes the observations $(x_0, \ldots, x_{i-1})$. Instead of using (3.15), and under the assumption that $x_{-1} = 0$ (i.e., the initial condition), an equivalent *likelihood* function [69, 106] can be stated using

$$p(e_0, e_1, \ldots, e_{N-1}) = \prod_{i=0}^{N-1} p(e_i) \quad , \tag{3.16}$$

where e_i is assumed to be independent and identically distributed. Then, defining the error e_n as,

$$e_n = x_n - \hat{x}_n \quad , \tag{3.17}$$

equation (3.16) becomes

$$\prod_{i=0}^{N-1} p(e_i) = \prod_{i=0}^{N-1} p(x_i - \hat{x}_i) \quad , \tag{3.18}$$

where $\hat{x}_n$ is the *forward* prediction of x_n given by the linear combination of past samples as

$$\hat{x}_n = -\sum_{k=1}^{p} a_k \cdot x_{n-k} \quad , \tag{3.19}$$

which is replaced in (3.17),

$$\begin{aligned} e_n &= x_n + \sum_{k=1}^{p} a_k \cdot x_{n-k} \\ &= \sum_{k=0}^{p} a_k \cdot x_{n-k} \\ &= \mathbf{a}^T \mathbf{x}_n \end{aligned} \tag{3.20}$$

where $a_{k=0} = 1$. Thus, assuming $e_n \sim \mathcal{N}(0, \sigma_e^2)$, $p(e_n)$ is given by,

$$\begin{aligned} p(e_n) &= \prod_{i=0}^{N-1} \frac{1}{(2\pi\sigma_e^2)^{1/2}} \exp\left(-\frac{1}{2\sigma_e^2}(x_i - \hat{x}_i)^2\right) \\ &= \frac{1}{(2\pi\sigma_e^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma_e^2}\sum_{n=0}^{N-1}(x_i - \hat{x}_i)^2\right) \\ &= \frac{1}{(2\pi\sigma_e^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma_e^2}\sum_{n=0}^{N-1}\mathbf{a}^T\mathbf{x}_n\mathbf{x}_n^T\mathbf{a}\right) \\ &= \frac{1}{(2\pi\sigma_e^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma_e^2}\mathbf{a}^T\left\{\sum_{n=0}^{N-1}\left(\mathbf{x}_n\mathbf{x}_n^T\right)\right\}\mathbf{a}\right) \\ &= \frac{1}{(2\pi\sigma_e^2)^{\frac{N}{2}}} \exp\left(-\frac{1}{2\sigma_e^2}\mathbf{a}^T\mathbf{\Phi}\mathbf{a}\right) \quad , \end{aligned} \tag{3.21}$$

where $\mathbf{\Phi} \in \mathbb{R}^{N\times N}$ is the sample covariance matrix.

For the *a priori* distribution we get $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{C_a})$, with $\mathbf{C_a} = \sigma_a^2 \cdot \mathbf{I}$. Then,

$$p(\mathbf{a}) = \frac{1}{(2\pi\sigma_a^2)^{\frac{p}{2}}} \exp\left(-\frac{\mathbf{a}^T\mathbf{a}}{2\sigma_a^2}\right) \quad . \tag{3.22}$$

Applying $\ln(\cdot)$ function to (3.21) and (3.22),

$$\begin{aligned} \ln p(\mathbf{x}|\mathbf{a}) &= -\frac{1}{2\sigma_e^2}\mathbf{a}^T\mathbf{\Phi}\mathbf{a} - (2\pi\sigma_e^2)^{\frac{N}{2}} \quad , \\ \ln p(\mathbf{a}) &= -\frac{1}{2\sigma_a^2}\mathbf{a}^T\mathbf{a} - (2\pi\sigma_a^2)^{\frac{p}{2}} \quad , \end{aligned}$$

where the constant terms that do not depend on parameter vector $\mathbf{a}$ are no needed in the further sections. Then, the MAP [107] can be written as,

$$\begin{aligned} \hat{\mathbf{a}} &= -\arg\max_{\mathbf{a}} \left[\ln p(\mathbf{x}|\mathbf{a}) + \ln p(\mathbf{a})\right] \\ &= \arg\min_{\mathbf{a}} \left[\frac{1}{2\sigma_e^2}\mathbf{a}^T\mathbf{\Phi}\mathbf{a} + \frac{1}{2\sigma_a^2}\mathbf{a}^T\mathbf{a}\right] \\ &= \arg\min_{\mathbf{a}} \left[\mathbf{a}^T\mathbf{\Phi}\mathbf{a} + \frac{\sigma_e^2}{\sigma_a^2}\mathbf{a}^T\mathbf{a}\right] \\ &= \arg\min_{\mathbf{a}} \left[\mathbf{a}^T\mathbf{\Phi}\mathbf{a} + \lambda\mathbf{a}^T\mathbf{a}\right] \quad , \end{aligned} \tag{3.23}$$

where we switched $-\arg\max_{\mathbf{a}}(\cdot)$ to $\arg\min_{\mathbf{a}}(\cdot)$, and $\lambda = \frac{\sigma_e^2}{\sigma_a^2} > 0$, as the regularization factor.

By incorporating gain constraints at $\omega = 0$ (DC) and $\omega = \pi$ (Nyquist frequency) [17], the following equation statement is used for the new vector of constrained coefficients $\mathbf{c} = [c_0, c_1, \ldots, c_p]^T$,

$$C(z) = \sum_{k=0}^{p} c_k z^{-k} \Rightarrow C(e^{j0}) = C(1) = \sum_{k=0}^{p} c_k = l_0 \quad , \tag{3.24}$$

and

$$C(z) = \sum_{k=0}^{p} c_k z^{-k} \Rightarrow C(e^{j\pi}) = \sum_{k=0}^{p} c_k \cdot (-1)^k = l_\pi \quad , \tag{3.25}$$

which can be summarized in matrix notation as

$$\mathbf{\Gamma} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 1 & \cdots & 1 \\ (-1)^0 & (-1)^1 & \cdots & (-1)^p \end{bmatrix}^T \quad ; \quad \mathbf{b} = \begin{bmatrix} 1 & l_0 & l_\pi \end{bmatrix}^T \quad , \tag{3.26}$$

$$\mathbf{\Gamma}^T \mathbf{c} = \mathbf{b} \quad , \tag{3.27}$$

with l_0 and l_π set to unitary gain, and $\mathbf{\Gamma} \in \mathbb{R}^{3 \times p+1}$. Writing as a formal optimization problem, we get

$$\text{minimize} \quad \mathbf{c}^T \mathbf{\Phi} \mathbf{c} + \lambda \mathbf{c}^T \mathbf{c} \tag{3.28}$$

$$\text{subject to} \quad \mathbf{\Gamma}^T \mathbf{c} - \mathbf{b} = 0 \quad , \tag{3.29}$$

which is a quadratic programming problem with equality constraints. The covariance matrix $\mathbf{\Phi}$ and $\lambda \mathbf{I}$ in $\lambda \mathbf{c}^T \mathbf{c} \rightarrow \mathbf{c}^T (\lambda \mathbf{I}) \mathbf{c}$, are both positive definite. Therefore, the quadratic programming problem is convex and can be solved using the Lagrange multiplier method [85]. Thus, the *Lagrangian* [105] is

$$\mathcal{L}(\mathbf{c}, \mathbf{g}) = \mathbf{c}^T \mathbf{\Phi} \mathbf{c} + \lambda \mathbf{c}^T \mathbf{c} - 2\mathbf{g}^T \left(\mathbf{\Gamma}^T \mathbf{c} - \mathbf{b} \right) \quad , \tag{3.30}$$

where $\mathbf{g} = [g_1 \ g_2 \ g_3]^T > 0$ is the Lagrange multiplier vector.

Taking partial derivative [86] respect to $\mathbf{c}$ and $\mathbf{g}$ and equating to zero, yields

$$\frac{\partial \mathcal{L}}{\partial \mathbf{c}} = \left(\mathbf{\Phi} + \mathbf{\Phi}^T \right) \mathbf{c} + 2\lambda \mathbf{c} - 2\mathbf{\Gamma} \mathbf{g} = 0 \tag{3.31}$$

$$\frac{\partial \mathcal{L}}{\partial \mathbf{g}} = 2 \left(\mathbf{\Gamma}^T \mathbf{c} - \mathbf{b} \right) = 0 \quad , \tag{3.32}$$

but $\mathbf{\Phi} = \mathbf{\Phi}^T$ [75, 86], then

$$(\mathbf{\Phi} + \lambda\mathbf{I})\,\mathbf{c} - \mathbf{\Gamma}\mathbf{g} = 0 \Rightarrow \mathbf{c} = (\mathbf{\Phi} + \lambda\mathbf{I})^{-1}\,\mathbf{\Gamma}\mathbf{g} \tag{3.33}$$

$$\mathbf{\Gamma}^T\mathbf{c} - \mathbf{b} = 0 \Rightarrow \mathbf{g} = \left(\mathbf{\Gamma}^T\,(\mathbf{\Phi} + \lambda\mathbf{I})\,\mathbf{\Gamma}\right)^{-1}\mathbf{b} \quad , \tag{3.34}$$

and using the above equations for the $\mathbf{c}$ vector of coefficients $\mathbf{c}$, we get

$$\mathbf{c}_\lambda = (\mathbf{\Phi} + \lambda\mathbf{I})^{-1}\,\mathbf{\Gamma}\left(\mathbf{\Gamma}^T\,(\mathbf{\Phi} + \lambda\mathbf{I})^{-1}\,\mathbf{\Gamma}\right)^{-1}\mathbf{b} \quad , \tag{3.35}$$

which has the same form of equation (2.23), with the added regularization matrix ($\lambda\mathbf{I}$) to the sample covariance matrix $\mathbf{\Phi}$. Having this added term λ reduce the chance to erroneously fit poles near the fundamental frequency.

3.2.1 Validating RCPIF with self-sustained numerical models

The speech material to validate RCPIF consists of the same simulations obtained with the self-sustained model of voice production described in section 3.1.4. However, only the DC component of the airflow waveform simulations was removed to allows RCPIF to fit higher formants. To produce comparable results between RCPIF and the SNF+Metrics approach, a 1.1 kHz low-pass filter is applied after inverse filtering with RCPIF. To emphasize the resonances of the vocal tract, a pre-emphasis filter [89] is used. The glottal closure instant (GCI) was estimated with the SEDREAMS algorithm [74] included on the COVAREP package [108]. For the analysis, the normalized errors (GVV and dGVV error equations (3.4) and (3.5)) and waveform similarities ratios (r_{GVV} and r_{dGVV} equations (3.6) and (3.7)) described in section 3.1.4 are calculated for RCPIF and shown in Table 3.5. Several values for the regularization factor (λ) are reported for the set of simulations.

The results indicate that the first formant estimation ($\hat{F}_1$) is in the range of values reported in the literature [76]. However, when λ value is increased, the regularization factor shown an overall improvement in normalized GVV-dGVV errors and waveform similarities by using RCPIF instead CPIF (or RCPIF with $\lambda = 0$). The best results are obtained for vowel /a/ and /e/ with $\lambda = 0.010$, which both have higher formants. In other words, inverse filtering for lower formants frequencies is still challenging for these methods [17, 104]. The results of the aerodynamic measures are presented in Table 3.6. There is an overall improvement in the aerodynamic measures using regularization, especially when

$\lambda = 0.010$ in vowel /a/. Across all λ values, the best performance is for OQ in vowels /a/ and /e/, followed by ACFL in the same vowels. This is an indication of the robustness of these aerodynamic measures to errors on inverse filtering. The worst performance is for H1H2 in almost all cases which is a direct consequence of the poor performance of these methods to cancel out lower formants near f_0 (e.g. vowel /i/).

TABLE 3.5: Mean (std) values of the RCPIF inverse filtering approach using self-sustained numerical models. Fundamental frequency (f_0) and formant frequency (F_1) estimates, GVV and dGVV normalized absolute error (to evaluate waveforms deviations), and correlation coefficient r_{GVV} / r_{dGVV} (to evaluate waveform similarities) are reported. The RCPIF order was $p = 12$, same as that in [17].

λ	vowel	f_0 (Hz)	$\hat{F}_1$ (Hz)	GVV error (%)	dGVV error (%)	r_{GVV} / r_{dGVV}
0.000	/a/	208.7(18.6)	506.7(168.7)	34.0(25.3)	51.0(20.6)	0.75(0.06) / 0.54(0.11)
	/e/	210.8(19.1)	401.5(139.3)	34.6(30.5)	51.9(21.1)	0.77(0.09) / 0.58(0.14)
	/i/	169.6(10.7)	269.7(63.6)	40.8(15.7)	65.9(10.1)	0.48(0.16) / 0.27(0.10)
	/o/	209.2(17.4)	423.1(49.7)	16.8(3.2)	35.2(5.6)	0.75(0.04) / 0.53(0.04)
	/u/	189.6(13.0)	307.3(75.1)	45.5(20.5)	63.0(14.0)	0.55(0.13) / 0.26(0.12)
0.001	/a/	208.7(18.6)	558.0(128.7)	16.0(5.7)	36.1(5.4)	0.84(0.07) / 0.66(0.03)
	/e/	210.8(19.1)	426.9(101.7)	31.1(28.9)	49.8(20.6)	0.78(0.13) / 0.62(0.15)
	/i/	169.6(10.7)	299.0(24.0)	50.9(3.3)	72.8(3.9)	0.38(0.15) / 0.23(0.10)
	/o/	209.2(17.4)	452.8(27.8)	18.6(2.8)	39.4(3.9)	0.83(0.06) / 0.59(0.04)
	/u/	189.6(13.0)	349.3(16.6)	44.9(4.4)	68.0(5.8)	0.54(0.07) / 0.33(0.07)
0.003	/a/	208.7(18.6)	660.1(31.8)	12.0(1.1)	32.4(1.9)	0.82(0.04) / 0.68(0.03)
	/e/	210.8(19.1)	454.4(81.5)	22.5(15.5)	43.5(12.2)	0.85(0.08) / 0.68(0.09)
	/i/	169.6(10.7)	313.8(15.5)	54.0(4.1)	71.8(2.2)	0.43(0.18) / 0.25(0.11)
	/o/	209.2(17.4)	440.8(35.9)	22.1(2.7)	44.5(4.5)	0.86(0.05) / 0.59(0.04)
	/u/	189.6(13.0)	344.3(13.5)	44.9(5.1)	65.8(4.5)	0.60(0.06) / 0.36(0.05)
0.010	/a/	208.7(18.6)	674.4(25.2)	11.4(1.1)	32.0(2.8)	0.80(0.04) / 0.67(0.03)
	/e/	210.8(19.1)	482.7(76.9)	18.9(6.2)	41.1(5.1)	0.88(0.05) / 0.71(0.04)
	/i/	169.6(10.7)	331.0(11.1)	58.6(4.4)	71.9(1.5)	0.50(0.16) / 0.29(0.09)
	/o/	209.2(17.4)	438.8(69.2)	29.7(3.6)	54.4(6.1)	0.89(0.03) / 0.57(0.05)
	/u/	189.6(13.0)	344.0(10.1)	48.8(5.4)	66.0(3.2)	0.68(0.05) / 0.39(0.03)
0.030	/a/	208.7(18.6)	685.1(25.4)	12.1(1.3)	33.7(3.2)	0.81(0.04) / 0.68(0.02)
	/e/	210.8(19.1)	486.5(73.5)	21.0(5.7)	45.1(5.2)	0.91(0.05) / 0.73(0.02)
	/i/	169.6(10.7)	342.9(9.0)	62.2(4.5)	73.1(1.4)	0.56(0.14) / 0.31(0.07)
	/o/	209.2(17.4)	446.1(115.5)	35.8(5.9)	61.9(9.5)	0.88(0.02) / 0.54(0.05)
	/u/	189.6(13.0)	350.2(7.7)	55.5(4.9)	69.8(2.7)	0.74(0.04) / 0.42(0.03)

TABLE 3.6: Results of aerodynamic measures from the proposed RCPIF approach with multiples regularization factors. ACFL, MFDR, OQ, H1H2, CPP and NAQ normalized absolute error are reported. Mean (and standard deviation) for each aerodynamic measure and vowel are calculated as well.

λ	vowel	ACFL (%)	MFDR (%)	OQ (%)	H1H2 (%)	CPP (%)	NAQ (%)
	/a/	20.5(11.5)	10.8(3.2)	7.2(7.1)	17.6(13.1)	15.1(12.3)	14.5(12.4)
	/e/	19.2(21.4)	25.4(30.8)	6.6(6.9)	44.1(26.3)	23.2(16.5)	10.9(11.7)
0.0000	/i/	34.9(35.5)	26.2(12.7)	18.6(19.2)	163.5(110.4)	11.3(9.3)	76.8(31.8)
	/o/	8.2(4.7)	12.2(9.5)	8.8(3.5)	26.9(15.2)	13.5(7.2)	12.5(12.4)
	/u/	27.9(11.8)	29.2(11.9)	31.6(15.3)	137.7(98.4)	11.8(9.2)	66.3(44.3)
	/a/	14.3(5.4)	9.4(7.2)	4.7(3.7)	13.2(7.3)	15.8(14.7)	17.3(11.0)
	/e/	12.3(12.0)	10.8(10.5)	7.3(9.1)	42.4(21.2)	18.3(10.3)	8.4(8.8)
0.0010	/i/	53.2(18.9)	26.5(8.1)	19.7(15.8)	280.2(37.6)	12.3(6.7)	109.2(11.4)
	/o/	2.5(2.1)	19.2(18.9)	8.1(4.0)	29.9(18.7)	12.2(8.7)	14.6(13.7)
	/u/	33.6(15.0)	30.8(13.2)	36.3(6.0)	243.5(30.5)	13.7(9.3)	95.9(15.2)
	/a/	10.8(2.3)	6.5(3.5)	3.9(2.5)	12.2(6.2)	11.8(6.7)	11.3(6.5)
	/e/	8.6(7.8)	6.0(6.3)	4.1(2.9)	40.8(14.4)	18.8(9.7)	8.7(7.9)
0.0030	/i/	67.3(25.2)	19.1(8.2)	20.7(16.0)	271.6(28.2)	14.6(8.4)	106.5(11.9)
	/o/	2.5(1.9)	25.8(21.4)	7.9(4.8)	36.2(15.0)	10.8(10.0)	20.0(12.2)
	/u/	36.4(15.8)	25.5(12.7)	37.4(4.9)	207.9(18.6)	13.8(10.0)	84.9(10.5)
	/a/	10.6(2.3)	6.0(2.5)	3.1(1.5)	11.5(4.4)	11.0(4.0)	7.6(7.1)
	/e/	3.8(2.8)	10.0(4.5)	4.8(4.5)	40.0(11.3)	14.3(7.4)	11.8(3.4)
0.0100	/i/	88.5(32.8)	10.3(6.8)	22.6(18.4)	247.8(39.7)	13.2(7.4)	103.7(14.6)
	/o/	6.3(4.0)	43.3(24.1)	11.1(5.3)	46.2(9.7)	11.5(8.8)	24.3(10.5)
	/u/	46.7(18.6)	15.4(11.6)	43.1(4.9)	169.0(31.8)	12.7(13.2)	73.3(4.4)
	/a/	10.4(2.3)	4.9(3.3)	3.2(2.1)	12.7(5.1)	10.4(5.1)	9.9(6.4)
	/e/	3.4(3.2)	17.0(7.8)	5.0(4.8)	43.0(11.9)	21.5(14.2)	12.4(3.2)
0.0300	/i/	107.5(40.3)	9.3(7.1)	22.2(19.1)	230.9(42.0)	13.6(10.1)	102.7(16.4)
	/o/	19.0(10.3)	57.3(34.9)	13.2(8.0)	51.8(7.7)	12.1(10.3)	22.4(10.6)
	/u/	63.8(23.7)	12.1(3.1)	45.1(3.9)	146.4(43.4)	17.3(13.7)	68.9(5.4)

3.3 Enhancing subglottal inverse filtering

3.3.1 Enhanced Subglottal Inverse Filtering using a Non-Parametric Cepstral Approach

As was described in section 2.2.1.4, an *Impedance Based Inverse Filtering* (IBIF) technique [28] is currently used to estimate glottal airflow from skin acceleration recordings. This approach need an initial calibration process, using an inverse filtered oral airflow signal from the Rothenberg mask as a reference. This task is challenging due to several issues with the mask and the inverse filtering procedure.

In this section a method to approximate the skin+subglottal system with a non-parametric approach is reported with the aim of reducing the dependency of oral airflow estimation in the calibration process of Subglottal Impedance Based Inverse Filtering (IBIF) [28].

The non-parametric approach is based on spectrum envelopes calculated via homomorphic signal processing (mathematical details are provided in section 2.2.1.3) and compared with a physical-based model of IBIF and neck-skin acceleration signal recordings.

In theory, homomorphic filtering in cepstral domain allow us to estimate the inverse filtering transfer function without any specific model assumption (i.e., non-parametric), and without any reference signal (i.e., blind). In this work, a non-linear, non-parametric transformation (known as *Cepstrum* [59]) is used to separate the system response from the mixed source-system signal. Proof of concept experiments are performed with synthesized and recorded acceleration signals. Preliminary results show moderate and good agreement in our experiments.

Two experiments are performed to obtain insights about the proposed subglottal cepstrum approach described in section 2.2.1.3. For the first experiment randomly generated synthesized glottal waveforms from Rosenberg model were generated, using the same approach described in section 3.1.3. A set of IBIF parameters $Q_{1,2,3} = [1, 1, 1]$, trachea length 10 cm ($Q_4 = 1$) and accelerometer location at 5 cm ($Q_5 = 1$)were fixed for this experiment. The goal of this experiment is to imitate several glottal airflow shapes observed in continuous speech. The second experiment, consist of recording a neck skin acceleration signal from a female voice while reading the Rainbow Passage. With this second experiment, a similar result is expected compared with the first one, as several

configurations of. Baseline parameters for this case were estimated from a sustained vowel /a/ for the same recording session and participant, and yield IBIF parameters $Q_{1,2,3} = [0.56, 1.1, 4.4]$ with trachea length 10 cm and accelerometer location 4 cm.

A voice activity detection based on the autocorrelation function is used to detect voiced frames and estimate fundamental frequency f_0 [12]. For a given f_0, a quefrency cut off is defined for liftering in cepstral domain [48, 88] From signals of both experiments, frame-based *cepstral* analysis (512 samples, Gauss window with $\underline{\alpha} = 2$, and no overlap) is performed and *liftered* for each frame. Each cepstral analysis frame was transformed to the frequency domain via the inverse cepstrum transform [38, 88]. Mean values and percentile (5% and 95%) are estimated for each bin of the frequency response. A spectral distortion (SD) measure [59] is used to quantify the (dis)similarities between IBIF frecuency response and the non-parametric approach, wherein SD formula is:

$$SD = \frac{1}{N}\sqrt{\frac{1}{2\pi}\int_{-\pi}^{\pi}\left[10\log_{10}\left(\frac{X_{NP}}{X_{IBIF}}\right)\right]^2 d\omega} \tag{3.36}$$

where X_{NP} is the spectrum magnitude derived by cepstral analysis, X_{IBIF} is the spectrum magnitude derived from the baseline IBIF model. In addition, differences in peak frequency peaks (approximated subglottal resonances) are estimated as well.

Results are shown in Figure 3.15 and Table 3.7 for experiment 1, and Figure 3.16 and Table 3.8 for experiment 2. The SD values are only calculated between 40 Hz and 1.5 kHz for both experiments, to match the bandwidth of the accelerometer used in this work. For experiment 1, the shapes show a moderate fit with SD=25. In Figure 3.15 arrows are indicating four resonance peaks detected with the cepstrum approach (blue). In Table 3.7 are showed the frequency differences with the IBIF model (black) with smallest difference for the first and fourth subglottal resonance. For the experiment 2, the shapes shows a closer fit, with a SD=6.22, indicating better similarities with the estimated IBIF model. Only 2 frequency peaks are evidenced with the cepstrum approach, suggesting a SNR reduction beyond 2 kHz.

TABLE 3.7: Detected resonances in the experiment 1: IBIF (ground-truth) and Cepstrum-based, respectively. Units in Hz.

Id Peak	IBIF	Cepstrum	Difference
1	625	641	16
2	1453	1531	78
3	2156	2109	57
4	2844	2828	16

TABLE 3.8: Detected resonances in experiment 2: IBIF and Cepstrum-based. Units in Hz.

Id Peak	IBIF	Cepstrum	Difference
1	625	703	78
2	1484	1562	78

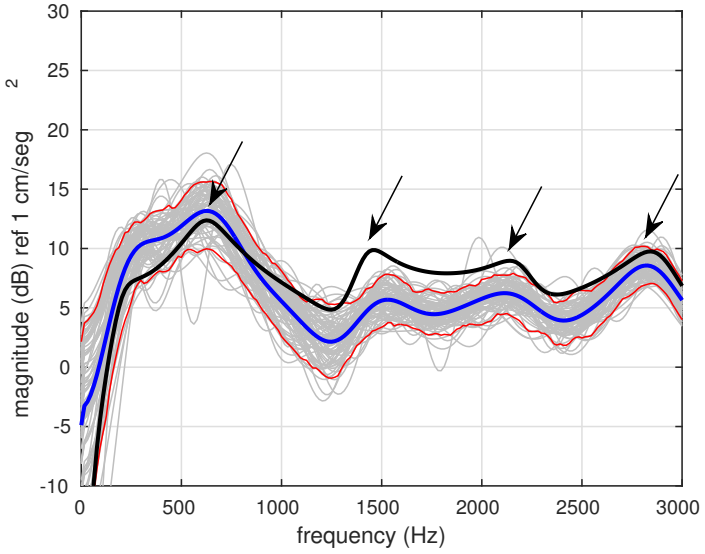


FIGURE 3.15: Simulated Rosenberg glottal pulses filtered with $T_{skin}(\omega_k)$. Black: IBIF model. Blue: Liftered Cepstrum. Red: percentile 5% and 95%. Gray: each liftered frame.

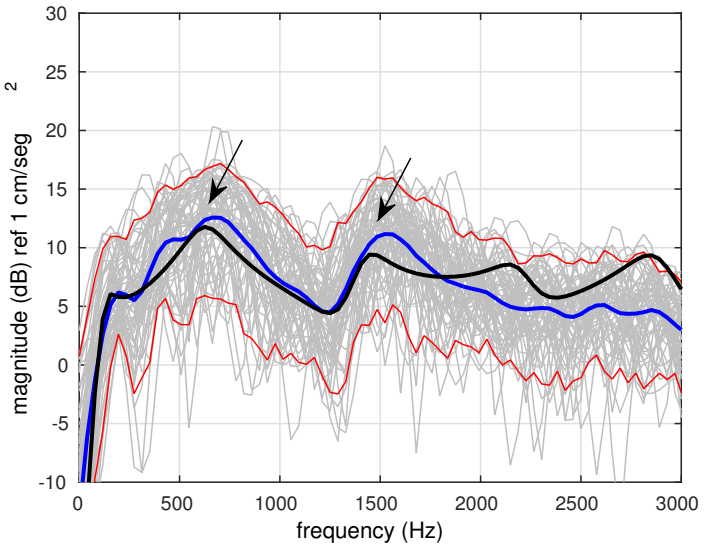


FIGURE 3.16: Inlab recording results for the non-parametric subglottal system. Legend sames as figure 3.15.

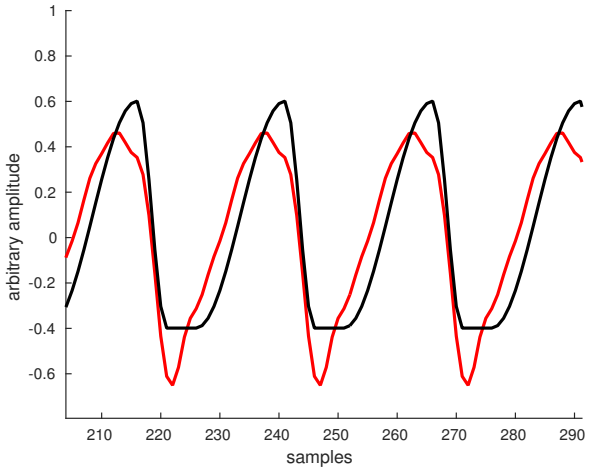


FIGURE 3.17: Comparing baseline glottal synthesized waveform (solid black) and a minimum-phase version of IBIF via cepstrum (solid red). It is clear that the closed phase is distorted by using the minimum-phase version of IBIF.

As is observed in Figure 3.15 and 3.16, the non-parametric approach does not perfectly fit the $T_{skin}(\omega_k)$ response. However, estimated resonances are closer to resonances of the $T_{skin}(\omega_k)$ baseline. In figure 3.17 a simulated experiment using synthesized glottal waveforms and the non-parametric approach is shown. The closed phase of the inverse filtered signal using cepstrum is not straight compared with the original glottal pulse. This results is similar with those in [81] but using the approach described in [23].

It is well known that this first subglottal resonance mainly depends on the trachea length [91, 109]. This allows for rethinking the utility of the non-parametric approach to be used as a trachea length estimator. Thus, a relationship between the trachea length and first sub-glottal resonance was tracked from several $T_{skin}(\omega_k)$ simulations and showed in Figure 3.18 with a suitable model. Following, an estimate of trachea length by the first resonance of subglottal system is possible. With this approach, is no longer needed to include trachea length estimation in the PSO algorithm which results in a substantial reduction of the computational load (and time) when Q parameters are estimated in more larger data sets (e.g. rainbow passage paragraph).

Another advantage of the non-parametric approach was that it is no longer needed the oral airflow estimation to catch the trachea length needed for IBIF. Only the neck skin acceleration signal performing the rainbow passage paragraph is enough with the approximation presented here.

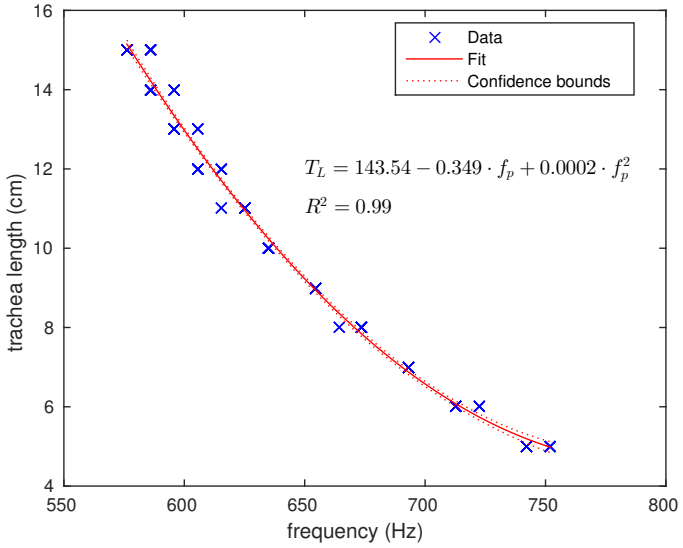


FIGURE 3.18: Simulated relationship between the trachea length and first sub-glottal resonance from several $T_{skin}(\omega_k)$ realizations. The fitted model is pictured and his equation as well.

3.3.2 Smooth inversion of IBIF

The transfer function T_{skin} from the IBIF algorithm is based upon a physiological transmission line that has a strong zero at DC due to the impedance of subglottal system. When T_{skin} is inverted, instabilities occurs at low-frequency. To deal with this problem, a zero-phase, unit gain (or 0 dB in log-scale), ad-hoc modification, hereafter, the zero-phase approach, was proposed in the original IBIF paper [28]. As in any zero-phase methods, edge effects can occurs with this low frequency ad-hoc correction. Applying a properly Tukey window to the ACC-signal reduce this effect in the inverse filtered signal. In addition, the zero-phase approach in [28] could trigger group-delay distortion for the filtered ACC-signal. To investigate these effects, a different approach to deal with inversion of T_{skin} is presented in this section.

A smooth-weighting at low-frequency is applied to $|1/T_{skin}|$ for gradually reduce the slope magnitude towards unitary gain at DC, and the phase response of $1/T_{skin}$ is simply the T_{skin} phase response multiplied by -1.

The smooth-weighting at low-frequency is applied to the log-magnitude inverted $T_{skin}(\omega_k)$ response (i.e., $1/T_{skin}$), $IT_{skin}(\omega_k)$. The $\log_{10}|IT_{skin}(\omega_k)|$ (two-sided spectrum from $\omega \in [0 \dots 2\pi k/N]$ with $k \in [0 \dots N-1]$), is filtered in frequency domain with a symmetric window $w(k)$ for a frequency below $\omega_{k_c} \approx \omega_c$, as shown in equation (3.37),

$$|IT_{skin_{smooth}}(\omega_k)| = 10^{(\log_{10}|IT_{skin}(\omega_k)|)\cdot w(k)}, \tag{3.37}$$

then $|IT_{skin_{smooth}}(\omega_k)|$ and $\angle IT_{skin}(\omega_k)$ are used to reconstruct the frequency response as follow,

$$IT_{skin_{stable}}(\omega_k) = |IT_{skin_{smooth}}(\omega_k)|e^{j\angle IT_{skin}(\omega_k)}. \tag{3.38}$$

For the smooth-weighted a Kaiser window [110] is applied in frequency domain because its ability of parametrize the curve slope by a single parameter by the following equations:

$$w_K(k) = \begin{cases} w_{K0}(k) & , \quad 0 \dots (k_c - 1) \\ 1 & , \quad k_c \dots (N-1-k_c) \\ w_{K0}(N-1-k) & , \quad (N-k_c) \dots (N-1) \end{cases} \tag{3.39}$$

with

$$w_{K0}(k) = \begin{cases} \frac{J_0\left(\beta_S\sqrt{1-\left(\frac{2(k-k_c+1)-1}{k_c-1}\right)^2}\right)}{J_0(\beta)}, & 0 \leq k \leq k_c - 1 \\ 0 & \text{otherwise} \end{cases}, \tag{3.40}$$

where β_S is the smooth parameter, $J_0(\cdot)$ the Bessel function of order zero, k the frequency bin, k_c the corresponding frequency bin for ω_{k_c}, and N the total number of frequency bins.

Under the constraint that any given frequency and amplitude of the inverted signal must be unaltered by this smoothing process, then we should consider that $\omega_c \ll 2\pi f_0$. For female voices the average lower pitch $f_0 \approx 250$ Hz (see Table 4.8), then a $f_c = 80$ Hz seems to be a reasonable first choice.

An interpretation of the zero-phase approach is pictured in Figures 3.19 and 3.20, for a Q set of $Q = [1, 1, 1, 1, 1]$. In the zero-phase approach, any log-magnitude greater than 0 dB is set to 0 dB (Figure 3.19), and the phase value is set to 0 degree (Figure 3.20). In the proposed smooth-weighted approach, the transition of the log-magnitude is smoothly altered until is set to 0 dB at DC. Here, the phase is not altered, as pictured in Figure 3.20. Note that both proposals change the IT_{skin} frequency response below 90 Hz. Results are identical for both approaches for a synthesized waveform of 150 Hz, as pictured in Figure 3.21, since the phase differences are below the frequency content of the signal. Changing the Q set to alter this condition, e.g., $Q = [2, 2, 2, 1, 1]$, results in differences between methods since we now overlap the frequency content of the signal within the range of the phase changes. Note that for this example, both approaches change the $IT_{skin}(\omega_k)$ frequency response below 160 Hz (see Figures 3.22 and 3.23), and the synthesized waveform of 150 Hz is expected to change with both methods. Note that the difference of approximately 0.3 dB for the magnitude and greater than 70 degrees for the phase at 150 Hz is observed. However, the zero-phase approach produces a more evident phase distortion in the low-frequency component of the glottal pulse, as is pictured in Figure 3.24. Then, preserving the phase response of an inverted $T_{skin}(\omega_k)$, yields minimum phase distortion and a more robust glottal waveform integrity.

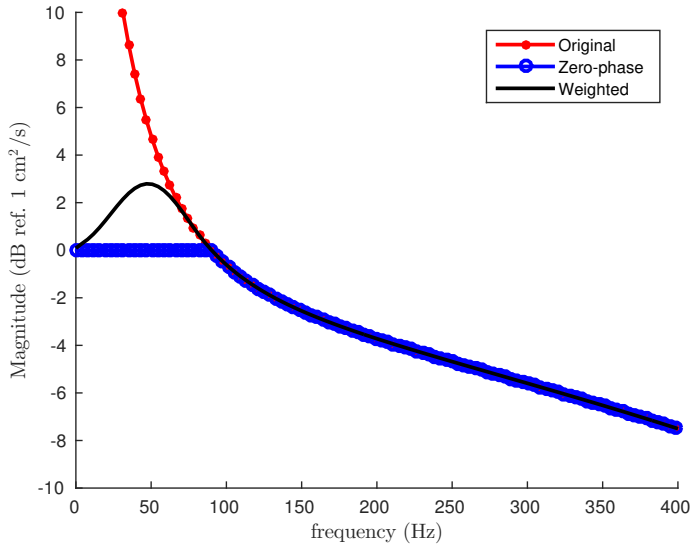


FIGURE 3.19: Superimposed magnitude response for the original, zero-phase, and smooth-weighted approach of $IT_{skin}(\omega_k)$ for a Q set of $Q = (1, 1, 1, 1, 1)$. Note that below $f_c < 90$ Hz the magnitude is progressively weighted to zero dB at DC.

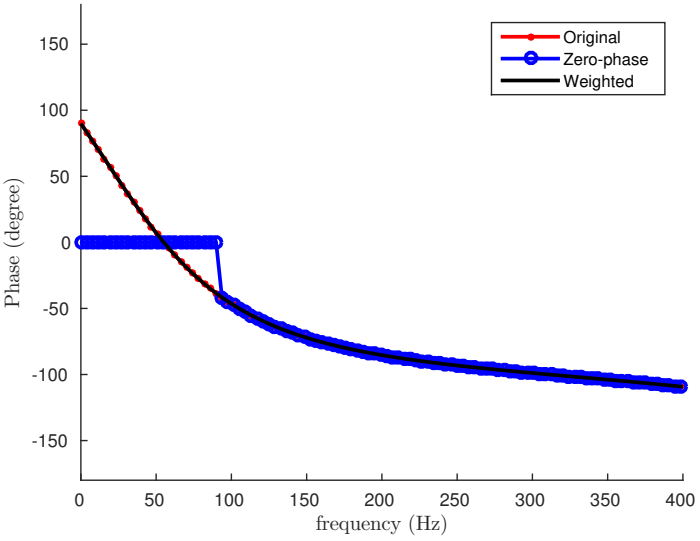


FIGURE 3.20: Superimposed phase response for the original, zero-phase, and smooth-weighted approach of $IT_{skin}(\omega_k)$ for a Q set of $Q = (1, 1, 1, 1, 1)$. Note the phase differences of both approaches below $f_c < 90$.

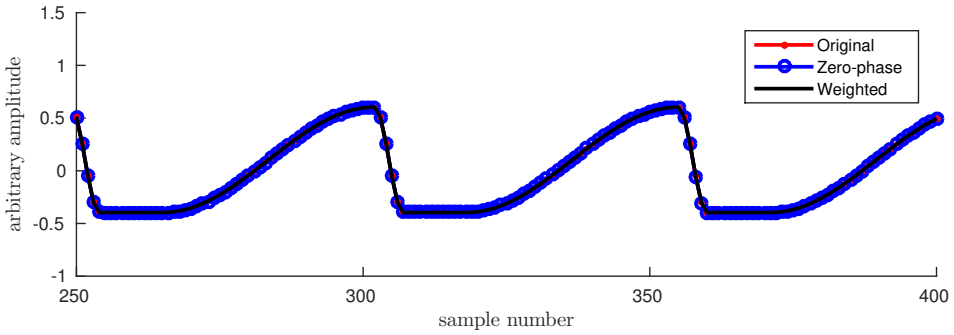


FIGURE 3.21: Glottal airflow estimation after the IBIF smoothed-weighted in low frequencies for a Q set of $Q = (1, 1, 1, 1, 1)$. Note how is preserved the shape of glottal pulses using both approaches of $T_{skin}(\omega_k)$.

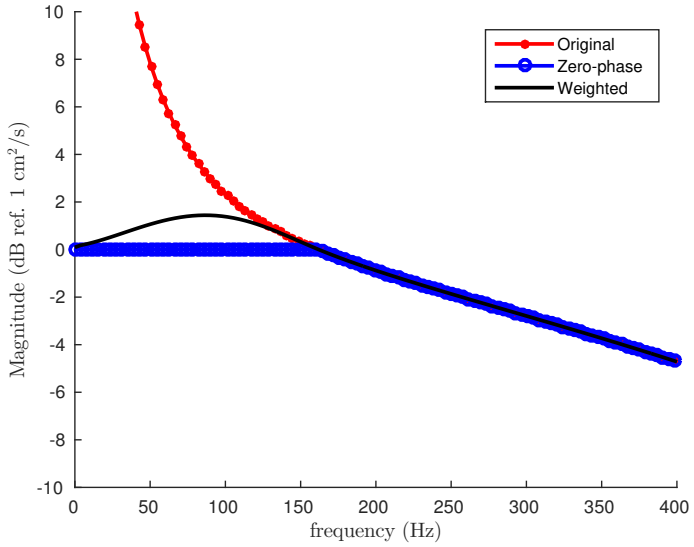


FIGURE 3.22: Superimposed magnitude response for the original, zero-phase, and smooth-weighted approach of $IT_{skin}(\omega_k)$ for a Q set of $Q = (2, 2, 2, 1, 1)$. Note that below $f_c < 160$ Hz the magnitude is progressively weighted to zero at DC.

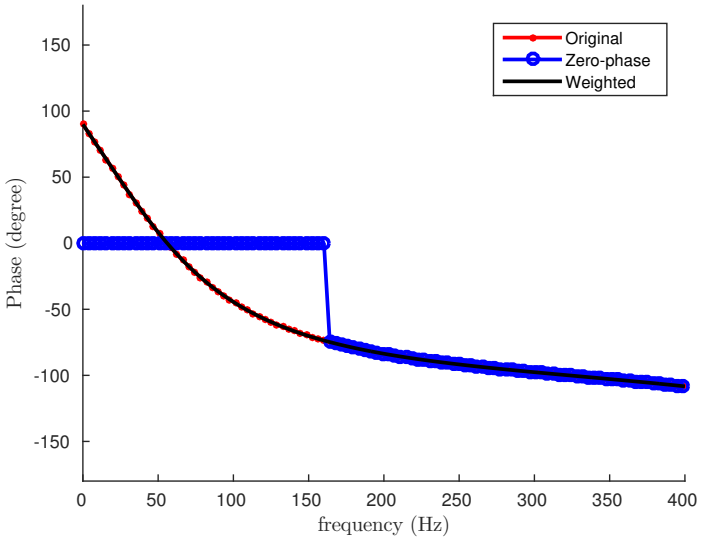


FIGURE 3.23: Superimposed phase response for the original, zero-phase, and smooth-weighted approach of $IT_{skin}(\omega_k)$. Note the phases differences of both approaches below $f_c < 160$.

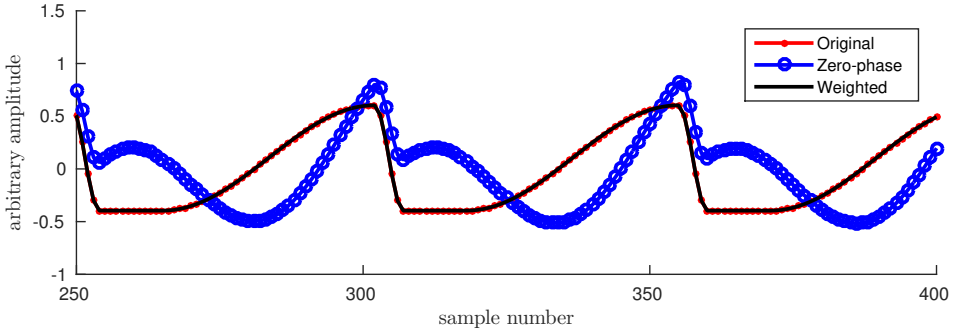


FIGURE 3.24: Glottal airflow estimation after the IBIF smoothed-weighted in low frequencies for a Q set of $Q = (2, 2, 2, 1, 1)$. Note how is preserved the shape of glottal pulses using the smooth-weighted approach of $T_{skin}(\omega_k)$.

3.3.3 IBIF weighted-error and computational cost reduction strategy

In this section, we present a variation of the cost function for the parameters obtained for the IBIF algorithm. The goal of this is to reduce the computational cost in the optimization process, improve the robustness of the estimation and normalize the error fit for further analysis.

Using the estimated glottal airflow from the accelerometer signal $\tilde{u}_g$, the IBIF model is calibrated to obtain subject-specific parameters by minimizing the following normalized weighted absolute error (NWAE) function,

$$\text{NWAE}(\mathbf{Q}) = \sum_{i=1}^{3} w_i \cdot e_i(\mathbf{Q}), \tag{3.41}$$

with

$$\sum_{i=1}^{3} w_i = 1, \quad 0 < w_i < 1, \tag{3.42}$$

where each w_i was set to $0.\bar{3}$, and

$$e_i(\mathbf{Q}) = \frac{\sum_{n=0}^{N-1} \left|\Delta^{(i-1)}\tilde{u}_g - \Delta^{(i-1)}\hat{u}_g\right|}{\sum_{n=0}^{N-1} |\Delta^{(i-1)}\tilde{u}_g|}, \tag{3.43}$$

where $\tilde{u}_g$ is the glottal airflow signal, $\hat{u}_g$ the synchronized accelerometer inverse filtered signal (see subsection 3.3.4 for further details), e_i the error value between signals, w_i the weighted coefficient, and $\Delta^{(i-1)}$ the time-derivative operator of order $(i-1)$. The goal of use different time-derivative signal versions is for balancing the energy of higher harmonics (adding a 6 dB/8^{ve} emphasis for each order) in NWAE to avoid over-fitting in the low frequency range. Therefore, the optimization problem is stated as:

$$\hat{\mathbf{Q}} = \arg\min_{\mathbf{Q}} \text{NWAE}(\mathbf{Q}), \text{ subject to } \mathbf{Q} \in \mathbf{D} \quad , \tag{3.44}$$

where $\mathbf{D} = \{D_i\}_{i=1,\ldots,5}$, is a set of constrains corresponding to each parameter of $\mathbf{Q}$ set. To solve this optimization problem, a Particle Swarm Optimization (PSO) algorithm [92] is used for point estimate of Q parameters. To reduced the computational load of this task two procedures are proposed. The first one was already described in section 3.3.1 by using a model to approximate the trachea length from the subglottal first resonance with a non-parametric approach applied to the neck skin acceleration signal. Acceleration location will be

approximated to half of this trachea length. Such approach is suitable for a high load with the rainbow passage paragraph or larger databases. The second approach, suitable to work in short voice segments (e.g., sustained vowels), consist of several pre-calculated (i.e., before PSO algorithm started) configurations of sub-glottal systems for a set of equally spaced values of trachea length and accelerometer location. Each pre-calculated (Z_{sub} and H_{sub1}) transfer functions was indexed and retrieved inside the PSO algorithm. As a consequence, both approaches substantially reduce the computational time of the optimization process.

3.3.4 Synchronized shifted signals

To minimize NWAE, oral airflow and acceleration signal must be time aligned. A rough approximation is to align them using the sample cross-correlation function [51] and find the maximum peak shifted in the neighborhood of mid lag position [75]. To improve this initial approximation, a delay parameter d is added in PSO algorithm by shifting the indexes of signals vectors (oral airflow and neck acceleration). As 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 (3.43) by

$$\hat{u}_{gt}(nT) \quad ; \quad n \in [d_0, N-1-d_0], \text{ and} \tag{3.45}$$

$$\tilde{u}_{gtd}(nT) \quad ; \quad n \in [d_0+d, N-1-d_0+d]. \tag{3.46}$$

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. A initial value for d_0 value, could be 0.5 times of the average glottal cycle time in the frame, as other based on the group-delay differences between both inverse filtering response, i.e., oral airflow and neck skin acceleration IF methods.

3.3.5 Discussion and Conclusions

In this chapter, were proposed and tested several improvements over selected methods of inverse filtering. A new inverse filtering method for supraglottal resonances approach based on several metrics was evaluated using a Single Notch Filter of band-passed synthesized oral waveforms. Deviation from baseline of several *Metrics* shows good results over multiple simulations of glottal pulses, fundamental frequencies (with higher and lower pitch) and simulated formants (higher and lower resonances). Some Metrics perform better than others but general performance similar. The same approach was tested using simulated waveforms based on self sustained models of voice production. The results show differences between metrics and one of them outperforms across simulations. This novel approach open a new way to look at the inverse filtering problem where predominant methods based on autoregressive models perform poorly. Further research is needed to accomplish the situation when two or more formants are present in the signal to be inverse filtered, which is not the case for oral airflow due to limited bandwidth of the signals.

For the RCPIF approach, both waveform similarities and normalized error for the aerodynamic measures were improved when $\lambda \in (0.003, 0.01)$, mainly for vowels /a/ and /e/. The results for vowel /i/ were the weakest in the group. and he SNF+Metrics approach shows better results in almost all cases compared with the RCPIF method.

The potential application of a blind and non-parametric homomorphic estimation of $T_{skin}(\omega_k)$ directly from the neck skin acceleration signal was explored with moderate results, but subglottal resonances were easily traceable and covariated with the trachea length. A model with high accuracy was derived in that regard to reduce computational cost of the optimization algorithm which one IBIF is evolved.

A weighed error function is proposed for the PSO algorithm, and delay adjustment between signals are incorporated to improve the calibration fit of the IBIF model.

Chapter 4

Glottal aerodynamic measures in adult females with phonotraumatic and non-phonotraumatic vocal hyperfunction

This chapter addresses the core of the specific aim 2 of the thesis. The goal was validate preliminary studies [3] that suggested that glottal aerodynamic measures could differentiate pathophysiological phonatory mechanisms for phonotraumatic and nonphonotraumatic vocal hyperfunction. Using time and frequency domain signal processing, several aerodynamic measures from inverse filtered approximations of the glottal airflow were estimated noninvasively using a pneumotachograph mask. Recordings from 78 women with normal voices, 16 women with organic vocal fold lesions, 16 women with muscle tension dysphonia, and 2 associated matched control groups with normal voices were used to differentiate vocal hyperfunction. Using multivariate statistical methods and regression models of a group of SPL-Normalized measures showed statistically significant differences between a group of normal voices compared with pathological group with vocal hyperfunction. The results from this chapter further confirm previous hypotheses and preliminary results indicating that SPL-normalized estimates of glottal aerodynamic measures can be used to describe the different pathophysiological phonatory mechanisms associated with phonotraumatic and nonphonotraumatic vocal hyperfunction. These results also

shed light into the relevance of aerodynamic features in the context of ambulatory voice monitoring.

4.1 Methods

4.1.1 Participants

Two groups of adult female participants with voice disorders were analyzed: 16 patients with PVH (vocal fold nodules or polyps) and 16 patients with NPVH (primary MTD). Diagnoses were based on a complete team evaluation by laryngologists and speech-language pathologists at the Massachusetts General Hospital (MGH) Voice Center. Each participant was evaluated with 1) a complete case history, 2) endoscopic imaging of the larynx, 3) aerodynamic and acoustic assessment of vocal function, 4) a patient-reported Voice-Related Quality of Life (V-RQOL) questionnaire [111], and 5) clinician-administered Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) assessment [112]. All patients were enrolled before the administration of any voice treatment. Table 4.1 shows group-based averages for the V-RQOL and CAPE-V assessments. Table 4.2 summarizes demographics of the patients.

TABLE 4.1: Mean (standard deviation) of Voice-Related Quality of Life (V-RQOL) and Consensus Auditory-Perceptual Evaluation of Voice (CAPE-V) ratings for the patient groups with phonotraumatic (PVH) and non-phonotraumatic (NPVH) vocal hyperfunction.

	PVH	**NPVH**
V-RQOL		
Social-Emotional	76.3 (21.7)	70.9 (29.1)
Physical Functioning	61.4 (22.0)	65.4 (21.9)
Total Score	67.5 (19.5)	67.8 (23.2)
CAPE-V		
Overall Severity	34.3 (13.2)	25.4 (21.2)
Roughness	19.8 (13.1)	12.8 (10.5)
Breathiness	17.1 (14.1)	5.6 (8.6)
Strain	23.7 (14.4)	16.3 (20.9)
Pitch	9.8 (13.3)	6.8 (9.8)
Loudness	6.3 (11.0)	5.6 (9.0)

TABLE 4.2: Demographics of the two patient groups: phonotraumatic (nodules and polyps) and non-phonotraumatic (muscle tension dysphonia, MTD) vocal hyperfunction.

Occupation	No. Subject Pairs	Patient Diagnoses
Actor	2	MTD
Administrator	3	MTD (2), Nodules (1)
Admin. Assistant	1	MTD
College student	2	MTD
Consultant	2	Nodules (1), Polyp (1)
Event planner	1	Polyp
Fitness instructor	2	MTD (1), Nodules (1)
Full-time mother	3	MTD
Marketing	1	Nodules
Media relations	1	Nodules
Music teacher	1	Nodules
Psychologist	1	Nodules
Registered nurse	3	MTD (1), Nodules (1), Polyp (1)
Sales	1	Polyp
Social worker	1	MTD
Systems analyst	1	MTD
Talent recruiter	2	Nodules
Teacher	4	MTD (2), Nodules (2)

Each patient aided in identifying her control subject with normal vocal status who was matched for sex, occupation, and approximate age (5 years). The normal vocal status of all 32 control subjects was confirmed by a licensed speech-language pathologist specializing in voice disorders via interview (subjects reported no difficulties with their voices in daily life), laryngeal videostroboscopic examination, and CAPE-V assessment. The ages (mean standard deviation) of the PVH and matched-control groups were 32.3 12.8 years and 32.9 12.9 years, respectively, thus not statistically different. Similarly, the ages of the NPVH and matched-control groups were not statistically different at 42.1 14.2 years and 40.7 13.5 years, respectively. Note that the patient groups were only matched in age and occupation with their respective control groups; the study was not designed to compare PVH and NPVH patient groups. Informed consent was obtained from all the subjects participating in this study, and experimental protocols were approved by the institutional review board of Partners HealthCare System at Massachusetts General Hospital. Subjects were enrolled in a larger study on smartphone-based ambulatory voice monitoring [40]. For this study, only data from adult females were used because of the higher incidence of

female patients with VH in the study sample, which reflects the incidence in the population [113] and the desire to control for sex-specific voice characteristics.

4.1.2 Data Acquisition Protocol

The data acquisition protocol was based on methods used in previous studies [3] [25] [95], enabled the non-invasive estimation of glottal airflow (from the oral airflow), subglottal air pressure (from intra-oral air pressure), and acoustic measures (from radiated pressure) of vocal function. Subjects were asked to produce three sets of five consecutive /pae/ syllables in two different loudness conditions (comfortable and loud). Subjects were free to choose levels that were most natural for them without any prescribed levels of absolute pitch and loudness (however, subjects were instructed to maintain a constant pitch and loudness within each syllable string). A posteriori analysis showed that the SPL of the loud condition was approximately 6 dB higher (on average) than that of the comfortable loudness condition. During the syllable production, simultaneous recordings were obtained of the 1) oral airflow volume velocity (OVV) using a circumferentially vented high-bandwidth pneumotachograph mask (Glottal Enterprises, Syracuse, NY) with an effective bandwidth of approximately 0 Hz to 1.2 kHz, 2) intra-oral pressure (IOP) using a catheter passed between the lips and connected to a low-bandwidth pressure sensor with an effective bandwidth of approximately 0 Hz to 80 Hz, and 3) the acoustic signal using a condenser microphone (MIC; MKE104, Sennheiser, Electronic GmbH, Wennebostel, Germany) placed 10 cm from the lips and having a bandwidth greater than 10 kHz. Acoustic and aerodynamic signals were low-pass filtered with an 8 kHz cutoff frequency (CyberAmp Model 380, Axon Instruments, Inc.) and synchronously sampled at a rate of 20 kHz and 16-bit quantization (Digidata 1440A, Axon Instruments, Inc., Union City, CA). OVV, IOP, and MIC signals were calibrated to physical units. The OVV signal was calibrated to units of mL/s using reference airflow levels (MCU-4 Pneumotach Calibration Unit, Glottal Enterprises). The IOP transducer was calibrated using a closed syringe system that provided reference levels of 0, 5, 10, 15, and 20 cm H2O. The MIC signal was calibrated using a Cooper-Rand electrolarynx sound source that generated multiple reference tones at increasing intensity levels measured by a Class 2 sound level meter (NL-20, RION, Tokyo, Japan) to map the uncalibrated voltage signal to units of Pascal and dB SPL at 10 cm.

4.1.3 Data Analysis

The OVV signal was low-pass filtered at 1100 Hz with a 10th-order Chebyshev Type II filter and then decimated to 8192 Hz to simplify the inverse filtering procedure (focusing only on the first formant) and to avoid the antiresonance in the frequency response of the pneumotachograph mask [16]. The IOP signal was low-pass filtered at 80 Hz with a 5th-order Butterworth filter and then decimated to 256 Hz. The MIC signal was low-pass filtered at 80 Hz with a 4th-order Butterworth filter and then decimated at 256 Hz to yield a root-mean-square (RMS) envelope [18]. All filtering processes were applied to the signals in both forward and reverse directions to yield zero-phase distortion and thus maintain time-alignment with the other physiological signals. Airflow and acoustic measures were computed from the middle three syllables in each string to avoid voice initiation and termination effects [95], yielding a total of nine sets of measurements per loudness condition [3] [25, 95]. To avoid onset and offset effects for the vowels, 25% from the beginning and end of each vowel sample was discarded to yield a stable mid-vowel segment (see black arrows in Figure 4.1). These pre-processing steps allowed vocal function measures that could be compared to previous studies of subjects with normal voices and VH [3, 25, 42].

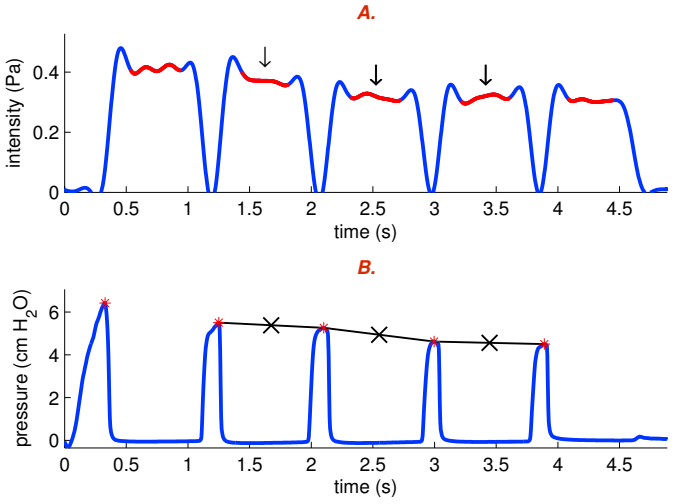


FIGURE 4.1: Definition of low-bandwidth glottal airflow waveform measures. (A) Sound intensity (smoothed root-mean-square of the radiated acoustic pressure). Black arrows (↓) indicate mid-vowel segments (red line) during which glottal aerodynamic measures were computed. (B) Intra-oral pressure for five /pae/ syllables showing peak values as red asterisks and interpolation lines indicating estimated subglottal pressure halfway between peaks (black Xs).

The inverse filtering (IF) technique described in chapter 3, section 3.1, was

utilized to cancel out the effects of the first formant for the OVV signal and estimate the glottal airflow from which measures were extracted to characterize the glottal volume velocity voicing source [25, 26]. Most of the many IF algorithms that decompose voice source and vocal tract filter components rely on estimating the vocal tract transfer function during the closed phase of vocal fold vibration [34, 114]. These techniques commonly suffer in the context of high fundamental frequencies (limiting the number of samples within the closed phase of each glottal cycle), nonlinear source-filter interaction, and pathological phonation [34, 114–117]. The present study could be impacted by such issues because all subjects were females (generally having higher fundamental frequencies than that of male speakers), and half of them had voice disorders (pathological phonation with reduced closed phase and increased non-linear source-filter interaction), making the detection of glottal closure instants and the application of closed-phase IF methods challenging. To counter these challenges, the single notch filter (SNF) IF technique described in chapter 3 was used to reduce waveform ripple and produce nearly flat amplitude in the closed phase [18, 35]. Similar single formant methods have applied successfully in previous IF studies of both normal and pathological voice production [3, 25, 26].

However, such approaches have required that the user makes interactive expert judgments by visually assessing the IF waveform and spectrum, with the goal of minimizing formant ripple [83]. However, individual user interaction is time-consuming and not suitable for analyzing large numbers of voice samples [83]. Thus, we developed a simple approach to determine an initial first formant (F1) candidate in the OVV signal based on minimizing the formant ripple using the following error criterion:

$$\sum_{n=0}^{N-1} \left| \Delta^2 x(n) \right| \tag{4.1}$$

where $x(n)$ is the inverse filtered OVV signal at sample, Δ^2 is the second-order time-derivative operator and N is the number of samples. The Δ^2 operator emphasizes (12 dB/octave) the high-frequency ripple related to the first formant whose energy decreases as the center frequency of the SNF approaches F1. The SNF is applied by sweeping the center frequency from 200 Hz to 1000 Hz in 1 Hz steps. In this initial step, the filter bandwidth was fixed to 70 Hz to follow past procedures [3, 18, 25] and also due to the challenges of bandwidth estimation. This decision was supported by recent simulation experiments that showed less than 20 Hz variability in using an LP approach to estimate the bandwidth of F1

for the /ae/ vowel [118], which has little influence on the IF waveform. The initial glottal airflow estimate was achieved when equation (4.1) reached a minimum value, which is visually confirmed as follows. For each loudness condition, we selected one vocalic token to inverse filter for which SPL was closest to the mean SPL across the nine vocalic segments per subject. A custom MATLAB graphical user interface (GUI), shown in Figure 4.2, provided the ability to visually confirm and fine-tune the SNF parameters if this were judged necessary after the automatic process described above. Several simulations of this automated process were performed and reported in chapter 3. Slider controls in the GUI allowed the user to adjust the SNF center frequency (F1) and bandwidth (BW1) to minimize any evidence of residual formant activity (e.g., ripple) based on the visual examination of multiple displays, including 1) direct comparisons between the original OVV and resulting IF (glottal airflow) waveforms and first derivatives to check for the amount of reduction in format ripple during the closed phase, 2) spectral displays of the linear predictionbased estimates of F1 for the OVV signal and resulting notch filter using the auto-correlation method [75] to check for evidence of residual formant energy, and 3) power density spectrum of the estimated glottal airflow signal using a Hann window of 512 samples to ensure there was a decrease in the spectral tilt.

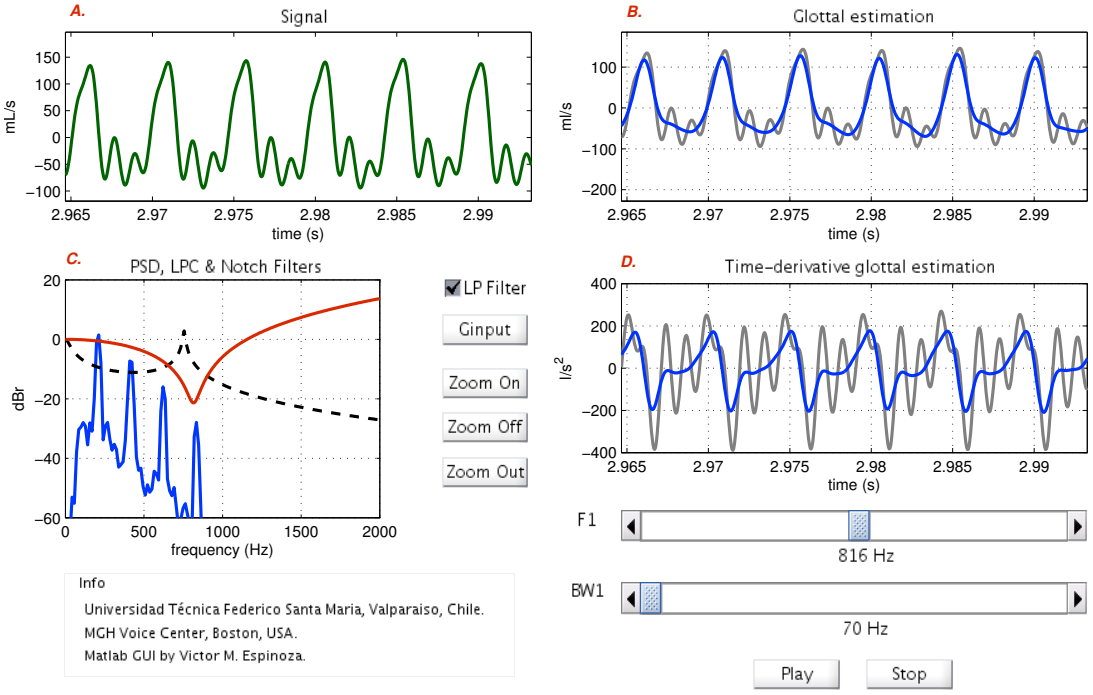


FIGURE 4.2: Graphical user interface to aid in verifying the automatic inverse filtering algorithm. (A) Original oral airflow waveform. (B) Estimated glottal airflow waveform after inverse filtering superimposed on oral airflow. (C) Power spectral density of the estimated glottal airflow waveform (solid blue), linear prediction spectrum (dashed black), and single notch filter frequency response (solid red). (D) Time-derivative of the estimated glottal airflow waveform superimposed on time-derivative of oral airflow waveform. Slider controls dynamically change the center frequency (F1) and bandwidth (BW1) of the inverse filter. Sound player buttons (play and stop controls) provide audio feedback to user.

The initial SNF parameter candidates (using the automatic approach) performed sufficiently well in approximately 70% of the cases. For the remaining data, SNF center frequencies and bandwidths were minimally adjusted, with bandwidth variation having a small impact on the measures of interest. The linear prediction-based resonance was less efficient as a visual reference for voices with higher pitch and higher spectral tilt. Regardless, the combination of automatic and interactive approaches provided a reasonable and efficient system that also reduced the degree of uncertainty associated with the IF process, which is of particular concern when dealing with high-pitched and pathological female voices.

4.1.4 Measures

Low-bandwidth and high-bandwidth measures were extracted from the processed data. As shown in Figure 4.1, the low-bandwidth measures were taken at mid-vowel and included estimates of 1) average dB SPL from the root-mean-square (RMS) envelope of the acoustic signal and 2) average SGP from the average of the peak intra-oral air pressures during lip closure for the /p/ sounds before and after each vowel. Figure 4.3 shows the high-bandwidth measures taken from the IF estimates of glottal airflow and included 1) ACFL, defined as the peak-to-peak amplitude of the waveform; 2) MFDR, defined as the absolute negative peak of the first derivative of the waveform; and 3) OQ, defined as the ratio of the open phase to the total cycle duration, wherein the open and closure time points were obtained at 5% amplitude between minimum and peak flow to minimize the effect of closed-phase ripples.

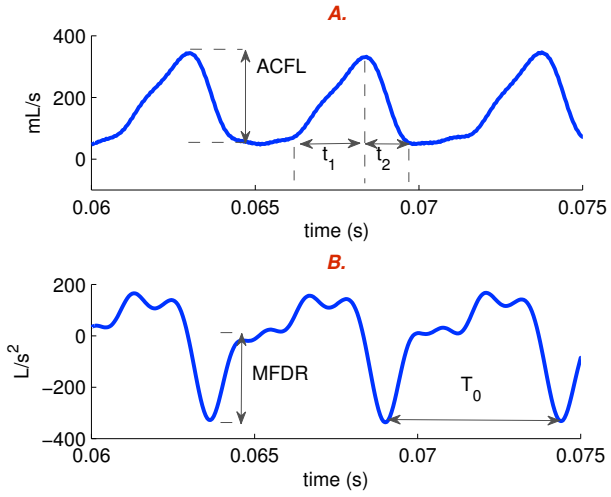


FIGURE 4.3: Definition of high-bandwidth glottal airflow waveform measures. (A) Estimated glottal airflow waveform, where ACFL is defined as the peak-to-peak waveform amplitude and $OQ = (t_1 + t_2)/T_0$, where t_1 is the opening phase duration, t_2 is the closing phase duration, and T_0 is the time interval between two consecutive peaks of the (B) time-derivative of the estimated glottal airflow waveform. MFDR is defined as the maximum negative peak in the derivative waveform.

4.2 Statistical Analysis

Descriptive statistics (means and standard deviations) were computed for all the measures of vocal function for each of the two loudness conditions (comfortable and loud) within each of the four subject groups of 16 participants. Previous work has shown that all of the aerodynamic measures in the present study are highly correlated with SPL [14, 25] and that adjusting/normalizing for this relationship improves the sensitivity of the measures for differentiating normal and pathological vocal function [3]. Thus, each measure was also normalized with SPL (dividing SPL by a given measure) to facilitate comparisons between groups (i.e., to increase the sensitivity of the measures by controlling for the impact of variations in SPL). For SGP, ACFL, and MFDR, the normalization process entailed first converting the measured values into logarithm scales (20 times the common logarithm of each measure) before computing the ratios. Log-scaling linearized the relationship between these glottal aerodynamic measures and SPL. This step was not performed for OQ because OQ is essentially a percentage representing a time-based relationship. Note that the normalization process produced ratios that may be interpreted as larger values reflecting more efficient voice production, i.e., higher SPL relative to a given aerodynamic measure. In the following, note normalized parameters use prime notation, e.g., MFDR'. Before applying SPL normalization, the strength of these relationships was confirmed by employing the same methods adopted in this study to analyze data from a larger group of 78 adult female subjects with normal voices that included the matched-control subjects in the current study. The strength of relationships between each measure and SPL was evaluated by using pair-wise linear regression to yield the coefficient of determination (R^2) and slope of the regression line. The results, shown in Table 4.3, confirmed that SPL is highly correlated with almost all of the aerodynamic measures being employed in this study and thus justify the use of SPL-based normalization. OQ is the exception, but we include it because better results were found using normalization. It is also reassuring that the changes in SPL that were observed for a doubling in ACFL, MFDR, and SGP were in very close agreement with results reported in the literature [25, 83, 119, 120]. In particular, Björklund and Sundberg (2015) found for adult females an 11.1 dB increase in SPL per doubling in SGP, which compares well with the 11.0 dB increase in SPL observed in the present study (see Table 4.3). Pair-wise linear regression was also used to assess the relationships between f_0 and each measure in the larger group of

normal females but did not yield substantial linear correlations ($R^2 < 0.49$) and thus were not included in the normalization process. Statistical testing was performed on the SPL-normalized data. Since the two groups with voice disorders were each carefully matched to separate groups of normal subjects, comparisons were only carried out between each patient group (PVH or NPVH) and its respective control group. Group-based comparisons were first evaluated with a multivariate paired-sample Hotelling's T^2 (a brief description is in appendix C) [121, 122] using all features in a four-dimensional space (ACFL', MFDR', SGP', and OQ'). If statistical significance was found using the multivariate t-test or if associated effect size magnitudes were large (Cohens $|d| > 0.6$) [123], follow-up comparisons were performed using one-tailed paired t-tests (all hypotheses predicted larger parameter values for the disordered groups, $p < 0.05$) to determine the individual contribution of each feature.

TABLE 4.3: Coefficient of determination (R^2), slope, and change in SPL per measure doubling based on pairwise linear regressions between each aerodynamic measure and SPL for a group of 78 adult females with normal voices. Doubling is defined as +6 dB for ACFL', MFDR' and SGP'. pp = percentage points.

Measure	**R^2**	**Slope**	**Change in SPL per doubling of measure**
ACFL'	0.61	1.46 dB/dB	8.7 dB
MFDR'	0.68	1.17 dB/dB	7.0 dB
SGP'	0.73	1.85 dB/dB	11.0 dB
OQ'	0.42	-0.40 dB/pp	20 dB

4.3 Matched control results

Table 4.4 reports the descriptive statistics for the original (un-normalized) measures within each subject group and loudness condition. In general, it appears that the PVH group displayed higher values across all measures than those in its control group and that the measures for the NPVH cohort tended to be approximately equivalent or slightly lower than those obtained for its control group. Table 4.5 displays the descriptive statistics for the SPL-normalized measures within each subject group and loudness condition. Table 4.6 summarizes results from statistical tests of the SPL-normalized data in Table 4.5. Overall, the PVH group displayed statistically lower SPL-normalized values than those in its control group: comfortable condition $F(4, 11) = 6.45$ ($p = 0.006$), loud condition $F(4, 11) = 6.69$ ($p = 0.006$), with large associated effect size magnitudes. Follow-up paired t-tests demonstrated statistically significant

differences for all of the aerodynamic measures in both loudness conditions, with large effect size magnitudes for all comparisons except for MFDR' in the comfortable voice condition (moderate effect size magnitude of 0.53). Overall, the NPVH group displayed statistically lower SPL-normalized values than those in its control group: comfortable condition $F(4, 11) = 3.19$ ($p = 0.057$), loud condition $F(4, 11) = 4.91$ ($p = 0.008$), with large associated effect size magnitudes. Follow-up paired t-tests demonstrated statistically significant differences and large effect size magnitudes for OQ' in the comfortable and loud conditions and for SGP' in the comfortable loudness condition.

TABLE 4.4: Group mean (standard deviation) for aerodynamic and SPL measures from the /pae/ syllable productions in comfortable (upper values for each measure) and loud (lower values for each measure) voice for the PVH and NPVH patient groups and associated matched control groups with normal voices.

Measure	**PVH controls**	**PVH group**	**NPVH controls**	**NPVH group**
ACFL (mL/s)	205 (63)	296 (102)	271 (94)	220 (77)
	264 (90)	400 (141)	340 (123)	302 (112)
MFDR (L/s^2)	306 (131)	415 (177)	386 (204)	269 (128)
	418 (189)	648 (309)	573 (314)	491 (248)
SGP (cm H_2O)	8.2 (1.6)	12.7 (4.5)	8.6 (2.7)	8.8 (1.6)
	11.5 (1.8)	17.6 (5.2)	13.2 (3.8)	13.4 (3.4)
OQ (%)	67.9 (10.7)	87.0 (8.3)	70.3 (8.6)	78.1 (10.3)
	65.8 (12.8)	81.1 (10.1)	58.7 (8.6)	63.0 (7.3)
SPL (dB SPL)	83.0 (5.0)	84.4 (4.6)	84.2 (5.4)	81.8 (5.9)
	89.2 (4.9)	91.3 (4.6)	92.4 (4.1)	90.1 (5.3)

TABLE 4.5: Group mean (standard deviation) for SPL-normalized/log-scaled aerodynamic measures from the /pae/ syllable productions in comfortable (upper values for each measure) and loud (lower values for each measure) voice for the PVH and NPVH patient groups and associated matched control groups with normal voices. pp = percentage points.

Measure	**PVH controls**	**PVH group**	**NPVH controls**	**NPVH group**
ACFL' (dB/dB)	1.81 (0.10)	1.73 (0.10)	1.75 (0.09)	1.77 (0.08)
	1.86 (0.10)	1.78 (0.11)	1.85 (0.10)	1.84 (0.10)
MFDR' (dB/dB)	1.70 (0.09)	1.65 (0.1)	1.67 (0.10)	1.72 (0.08)
	1.73 (0.09)	1.66 (0.11)	1.72 (0.13)	1.72 (0.11)
SGP' (dB/dB)	4.62 (0.41)	3.96 (0.45)	4.67 (0.60)	4.38 (0.33)
	4.24 (0.28)	3.75 (0.38)	4.24 (0.47)	4.05 (0.25)
OQ' (dB/pp)	1.26 (0.26)	0.98 (0.13)	1.22 (0.20)	1.07 (0.21)
	1.41 (0.32)	1.13 (0.13)	1.61 (0.28)	1.45 (0.20)

TABLE 4.6: Results of between-group statistical comparisons using Table 4.5 data. Reported are effect sizes for the multivariate, paired-sample Hotelling's T^2 tests and univariate, one-tailed paired t-tests (Cohen's d). Negative values for the univariate effect sizes signify that SPL-normalized measures are smaller in the patient groups than in their respective control groups. $^*p < 0.025$, $^+p < 0.05$, $^a p = 0.056$.

Group comparison	Hotelling's T^2	ACFL'	MFDR'	SGP'	OQ'
PVH vs Controls					
Comfortable	1.48*	-0.80*	-0.53$^+$	-1.53*	-1.36*
Loud	1.51*	-0.76*	-0.70$^+$	-1.47*	-1.11*
NPVH vs Controls					
Comfortable	1.04^a	-	-	-0.60$^+$	-0.73$^+$
Loud	1.29*	-	-	-	-0.66$^+$

TABLE 4.7: Mean SPL-normalized glottal aerodynamic measures from current and previously published investigations studying adult female subjects. For each measure, upper values list the mean values from Table 4.5. Lower values list the mean of previously reported data from, for respective columns, 20 subjects with normal voices (Holmberg et al., 1988 [25]), 10 subjects with PVH (Holmberg et al., 2003 [42]), and 2 subjects with NPVH (Hillman et al., 1989 [3]).

Measure	Study	Comfortable			Loud		
		Normal	PVH	NPVH	Normal	PVH	NPVH
ACFL' (dB/dB)	Mean values from Table 4.5	1.78	1.73	1.77	1.86	1.78	1.84
	Holmberg et al., 1988	1.78			1.85		
	Holmberg et al., 2003		1.67			1.73	
	Hillman et al., 1989			1.74			1.73
MFDR' (dB/dB)	Mean values from Table 4.5	1.69	1.65	1.72	1.73	1.66	1.72
	Holmberg et al., 1988	1.72			1.74		
	Holmberg et al., 2003		1.54			1.55	
	Hillman et al., 1989			1.68			1.7
SGP' (dB/dB)	Mean values from Table 4.5	4.65	3.96	4.38	4.24	3.75	4.05
	Holmberg et al., 1988	5.00			4.56		
	Holmberg et al., 2003		3.99			3.86	
	Hillman et al., 1989			4.33			4.15
OQ' (dB/pp)	Mean values from Table 4.5	1.24	0.98	1.07	1.51	1.13	1.45
	Holmberg et al., 1988	1.01			1.17		
	Holmberg et al., 2003		1.39			1.43	
	Hillman et al., 1989			0.9			1.01

4.4 Univariate group statistics results

In this section group results are presented based on early studies in vocal hyperfunction [3, 14, 25, 42, 44, 95]. A normative set of normal females voices is used as a reference to be statistically contrasted with two groups of women with voice disorders. Two methods are used for comparison, the z-score approach [3] along with an hypothesis test using the Bonferroni method [122].

4.4.1 Normative set from a group of normal female voices

Analysis with a normative set contrasting the pairwise matched control analysis, we explore in this section behaviors against a normative set. Thus, a group of 78 normal female voices was used to obtain additional relationships between measures. This group (hereafter, the normative set) is first compared with a similar study from Perkell et al. (1994) [26], and later analyzed in the context of a regressed z-score framework and hypothesis testing.

Table 4.8 shows results for the normative set of normal (healthy voices) women in comfortable and loud loudness conditions. The results from this study (2017) are compared to those in Perkell et al. [26] (1994), where it is noted tthat values for all measures are greater in both means and standard deviation values, than those reported in Perkell et al.(1994). Possible reasons for this increment can be stated. First, an overall increase in loudness (higher SPL) of about 10 dB is noted in both comfortable and loud loudness conditions. This increment in loudness is correlated with the increase of all vocal measures from approximately 20% to 100%, as is shown in Table 4.8. This means that the 2017 scenario represents a louder normal condition that of Perkell et al. (1994) study. Second, the bandpass filter used in this study have a different setting than Perkell et al. (1994) signal (see section 4.1.2 for further details). This affect vocal measures which depend on the harmonic content of the signal [24], e.g., MFDR has the highest difference from Perkell et al. (1994) study ($\approx$ 95% higher). Third, the number of subjects of the present study is larger than the 1994 study, providing more accurate estimates. Finally, even though the methods were similar in both studies, differences can be noted in the low-pass filtering and in the automated IF approach (see details in section 4.1.3), particulary since in the 1994 study the IF process was manually performed.

TABLE 4.8: Parameter values from a previous (1994) group and a current (2018) group of female speakers, for comf, loud & soft voice conditions. Only SPL, MFDR, ACFL and Pressure are compared. Within each voice condition, first two columns show the 1993 and 2017 means (with standard deviations in parentheses), except for soft voice conditions (Table I in Perkell et al. 1994 is only for comfortable and loud voice conditions). In the next column (diff) are differences between 1994 and 2017 means. For SPL, the differences are in dB; for the remaining parameters the differences are in percentage of the 1994 values.

	Comf Voice			Loud Voice			Soft
	1994 (n=15)	2017 (n=78)	diff	1994 (n=15)	2017 (n=78)	diff	2017 (n=71)
SPL (dB)	74.0 (3.3)	84.3 (5.1)	10.3 dB	80.8 (3.8)	91.8 (4.1)	11.0 dB	75.2 (5.6)
f_0 (Hz)	203.70 (21.60)	249.37 (40.91)	22.4%	216.00 (23.60)	264.28 (41.26)	22.4%	246.01 (45.9)
MFDR (l/s^2)	184.20 (62.50)	358.7 (156.5)	94.7%	373.50 (130.20)	536.64 (240.92)	43.7%	174.47 (92.7)
ACFL (l/s)	160.00 (50.00)	242.7 (43.3)	51.7%	220.00 (60.00)	329.49 (110.31)	49.8%	147.33 (67.1)
SGP (cm H_2O)	5.50 (1.30)	9.03 (2.55)	64.2%	7.60 (1.80)	13.81 (3.65)	81.7%	6.02 (1.7)

TABLE 4.9: Additional vocal measures of the present study (2017) for normal female voices in different loudness conditions with number of subjects in parenthesis. Means (with standard deviations in parentheses) are showed for H1H2, SQ, OQ, CPP and NAQ. .

	Comf (n=78)	Loud (n=78)	Soft (n=71)
H1H2 (dB)	11.2 (3.5)	8.9 (3.2)	16.8 (4.2)
SQ (pp)	237.11 (60.75)	227.28 (67.81)	188.82 (45.7)
OQ (pp)	74.02 (11.21)	65.66 (11.68)	87.01 (7.01)
CPP (dB)	17.25 (2.96)	19.46 (2.70)	12.92 (2.8)
NAQ	0.177 (0.039)	0.171 (0.040)	0.218 (0.039)

4.4.2 Regressed z-score and hypothesis test analysis

In this section two strategies to differentiate between the normal and pathological datasets are presented. These approaches use regressed z-score statistics, and multiple hypothesis tests.

These procedures are based on fitting multiple regression models to the normative set of normal voices. The linear model that best describes the normative set (hereafter, the normative model) was estimated by a robust strategy [105, 124] with the aim of weighting down outliers to minimize the bias of the standard (non-robust) linear model. To detect outliers candidates, we use a Cook distance [125], a Leverage criterion, and three-sigma edit rule. This allowed for excluding from the regression analysis the outlier beyond a given threshold [124]. The sill of these criteria were conservative and did not exclude more than the 5% of data. The regressed z-scores approach [3] is calculated as follows:

$$z_i = \frac{y_i - \bar{y}}{\sigma_m} \quad , \tag{4.2}$$

where y_i is the vocal parameter to be tested, $\bar{y}$ the predicted bivariate value for a given SPL and f_0 of y_i, and σ_m the standard error of the model. When an absolute value of z-score is greater than two times the standard error σ_m, the value of x_i was labeled as a pathological [3]. For the multiple hypothesis tests an independent sample Welch t-test [96, 126] was calculated. In addition, the Bonferroni method [122] was used as a first attempt to carry out a multivariate statistical analysis (see appendix C.1.1 for a brief explanation).

Normative model: A set of Normative models were determined to perform both z-score analysis and hypothesis tests. These models resemble early approaches to differentiate vocal hyperfunction [3, 25, 95]. The vocal parameters (or features) used for each normative model were: Subglottal Pressure (SGP), maximum flow declination rate (MFDR), Peak to Peak airflow (ACFL), Open Quotient (OQ), Speed Quotient (SQ), First to Second harmonic ratio (H1H2), Cepstral Peak Prominence (CPP), and Normalized Amplitude Quotient (NAQ)[3, 25, 50]. Each feature was covariate with SPL and fundamental frequency (f_0) with a robust regression model [124] across all loudness conditions, considering interaction and second order predictors as well. For ACFL, MFDR and SGP a log transformation was performed under the evidence of heteroscedasticity [127] and to simplify the expected normative model because

SPL is already in log scale. The log-scaled features were relabeled as ACFL_L, MFDR_L and SGP_L, respectively. All possible models were tested following a *Best Subset* approach [105]. At the same time diagnostics calculations were executed to finding outliers candidates. The "best" normative model ($\mathcal{M}$) was selected by finding the maximum Adjusted-R^2 which $|t_{\beta_n}| > 1,96$ and $p_n < 0.05$ $\forall \beta_n$, to avoid overfitting and collinearity [127].

TABLE 4.10: Linear regression coefficients fitted from 78 normal voices based on /pae/ gestures across **all loudness** condition. Scaled features of SPL and f_0 were used to analyze the contribution to each coefficient.

Parameter	intercept	SPL	f_0	$SPL \cdot f_0$	SPL^2	f_0^2	Adjusted-R^2
SGP_L (dB)	19	3.3	0.41	0	0	0	0.74
MFDR_L (dB)	50	4.8	0	0	-0.92	0	0.71
ACFL_L (dB)	-13	3.7	-1.1	0	-0.67	0	0.68
SQ (pp)	220	27	-28	-17	0	0	0.31
OQ (pp)	75	-10	4.3	2.2	-1.4	0	0.54
H1H2 (dB)	12	-3.6	1.6	0	0	0	0.61
CPP (dB)	17	3.1	-1	0	0	0	0.59
NAQ	-15	-1.3	1.4	0.43	0	0	0.6

The regression models in Table 4.10 had good performance related to adjusted-R^2, excepting SQ which does not qualify as a good response because adjusted-$R^2 = 0.31 < 0.49$, i.e., less than 49% of the variance is explained by the model [127]. The most relevant predictor is SPL (which further support SPL-normalized features in section 4.2), and f_0. Higher order predictors (square and interaction) were less relevant, excepting OQ which shows a mixed relationship with SPL and f_0 . SGP_L, MFDR_L and ACFL_L show the strongest dependency with SPL in these groups of parameters. H1H2, CPP, and NAQ show a mixed dependency with SPL and f_0.

Z-score for phonotraumatic and nonphonotraumatic voices: Using the normative models, z-score statistics are calculated for 16 subjects with PVH and 16 with NPVH voices, the very same data that was used for the SPL-normalized multivariate analisys in section 4.3. Table 4.11 shows the results of the z-score analysis. The results indicate that OQ and SGP_L (with a maximum of 50% discrimination) are the measures with greater discriminatory power in the case of PVH. In addition, $ACFL_L$ and $MFDR_L$ presented the weakest results for both types of hyperfunction. These findings are different to those in Hillman et al. (1989) wherein glottal waveform parameters of ACFL and MFDR were salient, and time-based vocal measures (e.g., OQ) were weak. In Hillman et al. (1989) the number of subjects was only a few cases (mostly men), and even women were only 2 subjects, so definitive evidence was not possible to achieve.

However, the lack of raw data from the previously reported research was improved here and results from the z-score analysis appears to be aligned with the SPL-normalized analysis as well, although with less statistical power.

TABLE 4.11: Z-score analysis to detect vocal hyperfunction. For each column, the number of subjects whose voices are greater than two standard deviation (|z-score| > 2) and its z-score efficiency (in parentheses) for 2 groups (of 16 subjects each) with phonotraumatic and non-phonotraumatic hyperfunction.

	Phonotraumatic		Non-Phonotraumatic	
	Comf	Loud	Comf	Loud
$ACFL_L$ (dB)	1 (6.25 %)	0 (0 %)	0 (0%)	1 (6.25%)
$MFDR_L$ (dB)	0 (0 %)	2 (12.5 %)	0 (0%)	2 (12.5%)
H1H2 (dB)	3 (18.75 %)	3 (18.75 %)	3 (18.75%)	0 (0%)
SGP_L (dB)	8 (50 %)	6 (37.5 %)	2 (12.5%)	3 (18.75%)
OQ (pp)	4 (25 %)	8 (50 %)	2 (12.5%)	1 (6.25%)
CPP (dB)	4 (25 %)	4 (25 %)	3 (18.75%)	3 (18.75%)
NAQ	2 (12.5 %)	5 (31.25 %)	5 (31.35%)	2 (12.5%)

Bonferroni-based hypotheses test for phonotraumatic and nonphonotraumatic voices: Both sets (control and pathological) were *regressed* using each normative model (see section 4.4.2 for details), i.e., the predicted response of the model is subtracted from both datasets (normal and pathological). In Figure 4.4 a graphical explanation is depicted. In the presence of outliers, we perform a conservative cleanup criteria of 3 standard deviation, using trimmed means (20 %) and MADN [128]. An independent sample t-test (Welch t-test [96, 126]) for means is performed between normal and pathological groups, in both *regressed* and *non-regressed* (the later, without SPL and f_0 covariates) versions. A two-tailed test based on the Bonferroni method was performed [122] (see section C.1.1 for details). Effect-sizes were calculated using d_{Cohen} [129] for each hypothesis test. Results for this analysis are presented in Table 4.12. Using the regressed data, the results we observe statistically significant differences between the normative and phonotraumatic group of female voices in 3 of 8 parameters. However, only 1 of 8 parameters was significantly different between the normative and non-phonotraumatic group. All differences represented medium and large effect sizes using Cohen's approach [130]. The non-regressed data yielded less strong findings when comparing the normative group to the phonotraumatic (2 of 8 were significantly different) and non-phonotraumatic groups (1 of 8 were significantly different). The most salient features were SGP_L and OQ. In terms of effect sizes magnitude OQ and SGP_L presents the best results for the phonotraumatic group. SGP_L, and H1H2 presents the best results for the non-phonotraumatic group.

TABLE 4.12: Effects sizes d_{Cohen} for significant mean differences (p_{value} was Bonferroni corrected) for regressed and non-regressed data versions. Sample size of the normative set was n_1=215, and for the pathological set was n_2=45.

Aerodynamic measure	Phonotraumatic	Non-Phonotraumatic	d_{Cohen}	p_{value}
SGP_L (dB)	Regressed	- -	-1.08	<0.0001
SGP_L (dB)	- -	Regressed	-0.50	0.0022
SGP_L (dB)	Non-Regressed	- -	-0.65	<0.0001
OQ (pp)	Regressed	- -	-1.28	<0.0001
OQ (pp)	Non-Regressed	- -	-0.78	<0.0001
H1H2 (dB)	- -	Non-Regressed	0.36	0.0062
H1H2 (dB)	Regressed	- -	-0.48	0.0012

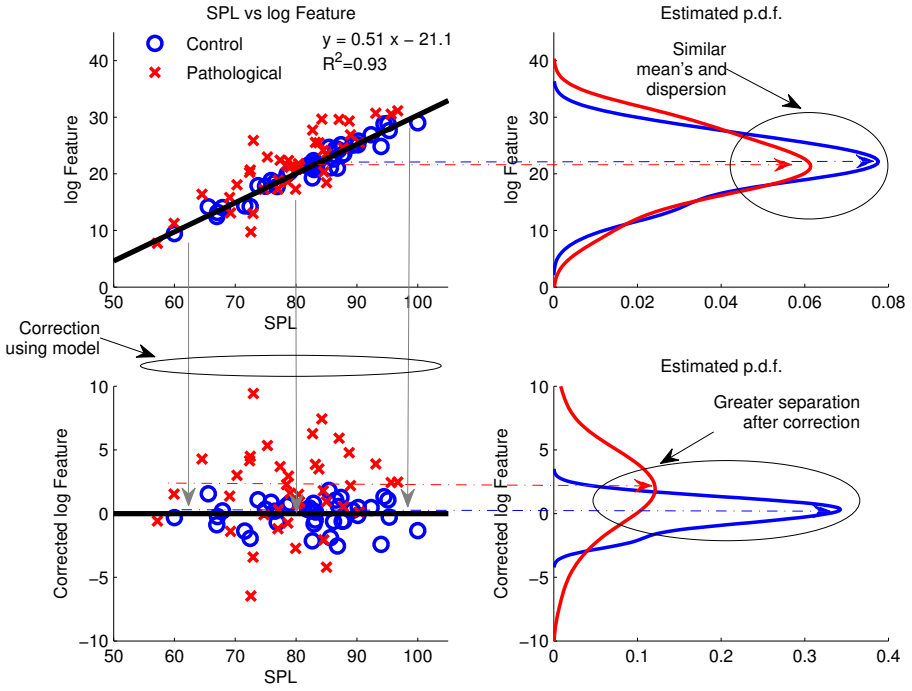


FIGURE 4.4: A conceptual example correcting vocal parameters as a function of SPL. Top-Left: SPL vs logged feature, with a fitted line for controls, i.e., the model. Top-Right: Estimated pdf for the log feature without any correction (the *non-regressed* version of datasets), shows no differentiation in central tendency and dispersion. Bottom-Left: Both sets (control and pathological) are *regressed* using the model, i.e., the predicted response of the model is subtracted to both datasets. Bottom-Rigth: Estimated pdf shows greater separation after correction (the *regressed* version of datasets). Vertical gray lines indicate the transition of non-regressed to regressed data. Red and Blue dash-dot line, point out pdf peak. The pdf was estimated using a Gaussian kernel density estimator [131].

4.5 Discussion

This study sought to confirm preliminary evidence that glottal aerodynamic measures could identify pathophysiological mechanisms for PVH and NPVH (Hillman et al., 1989) that are each distinctly different from normal vocal function. Statistically significant results from data collected in carefully selected cohorts of patients with PVH and NPVH and well-matched control groups provide support for the observations and hypotheses that were based on this earlier work. Namely, PVH patients displayed significantly reduced SPL-normalized values for the aerodynamic measures of SGP, ACFL, MFDR, and OQ. This means that higher than normal levels of these parameters are needed to attain a given SPL (reduced vocal efficiency), thus reflecting increased potential for trauma to vocal fold tissue that would contribute to the chronic presence of vocal fold lesions and associated dysphonia in this group. These findings, including the relatively larger effect size magnitudes for SGP_L and $ACFL_L$ than for $MFDR_L$, are in agreement with the results of modeling VH reported in [6] that demonstrated that compensatory mechanisms could account for the increases in aerodynamic measures. Specifically, increasing SGP to maintain a given SPL when there is reduced glottal closure (e.g., obstruction of glottal closure by vocal fold pathology) results in an elevation of ACFL and MFDR, with a concomitant increase in vocal fold collision forces. The combined results from the present study and modeling work [6] reflect the vicious cycle that is associated with PVH in which a compensatory increase in vocal effort could also cause additional vocal fold trauma [3]. In contrast, NPVH patients displayed abnormally lower SPL-normalized values only for SGP and OQ. This means that while higher than normal levels of these two parameters are needed to attain a given SPL (reduced vocal efficiency), the lack of a concomitant increase in ACFL and MFDR reflects decreased potential to cause trauma to vocal fold tissue.

The data for this study came entirely from adult female subjects because of the higher incidence of VH in women [113] and the desire to maintain homogeneity of the groups. However, most of the experimental (e.g., excised larynges, magnetic resonance imaging, and mechanical models) and theoretical (e.g., numerical models) studies of voice production have been based on data from male subjects. Inverse filtering the first formant was therefore more challenging than that reported in much of the literature because the higher fundamental frequency and spectral tilt associated with female voices increased the

uncertainty of the estimation process. For example, as fundamental frequency increases, the number of harmonics is limited by the bandwidth of our measures (1 kHz). Therefore, a common female fundamental frequency of 300 Hz or higher will only provide two harmonics (600 and 900 Hz), which will have a direct influence on parameters that include high-frequency information (e.g., MFDR). Furthermore, recent results from numerical vocal fold models show that the closed phase is not completely flat [6]; thus, even in the presence of incomplete glottal closure, the ripple-minimization approach [132] may not be the best physiological criteria for inverse filtering. For signals with higher spectral tilt, we observed small differences in the resulting waveforms when varying the center frequency and bandwidth of the SNF used to cancel out the first formant. The waveform for these cases follow a sinusoidal pattern which may influence the accuracy of parameters like OQ and MFDR.

As a historical reference, Table 4.7 compares the SPL-normalized aerodynamic measures computed in the present study and those derived (i.e., estimated using their SPL and log-transformed parameters) from published data [3, 25, 43]. The general observation here is that the ratios in the current study are in similar numerical ranges to those derived from previously published data. Also, the estimated SPL-normalized aerodynamic measures from past studies decrease for both types of VH compared with measures in subjects with normal voices, which is in agreement with the current results.

The positive results of the present study support the potential for glottal aerodynamic measures to objectively quantify pathophysiological mechanisms for PVH and NPVH that are each distinctly different from normal vocal function. Treatment-related investigations should also be undertaken to further assess the value of the measures, as clinically meaningful measures would ideally migrate toward normal values after successful therapeutic intervention. For example, better insights into etiological mechanisms (based on empirical evidence that is currently lacking) might be attained by using these measures to assess vocal function in PVH patients before and after surgical removal of lesions and subsequent to voice therapy. Based on prevailing clinical assumptions [133], it is possible that post-surgical measures would continue to reflect the persistence of VH (potential for recurrence of vocal fold trauma) and that the measures would only show significant migration toward normal following vocal retraining (voice therapy). Such expectations are somewhat tempered by previous evidence that the within-subject variability of some glottal aerodynamic measures (e.g., for repeated pre-treatment baseline measures) may limit their sensitivity to

treatment effects [42]. However, it is hoped that the signal analysis framework used in the current study will reduce the portion of this reported variability that may have been related to the use of older methodologies. The recent development of subglottal impedance-based inverse filtering [28, 32, 134] offers the capability to extract estimates of glottal airflow waveform parameters from a neck-surface acceleration signal, thus providing the opportunity to unobtrusively obtain these promising measures as individuals go about their normal daily activities [40]. Such capabilities could provide a much more accurate assessment of an individuals typical function (e.g., the prevalence and severity of VH during a typical day) and potentially supply physiologically-based biofeedback targets to aid in reducing VH [32]. The descriptive statistics for the group of 78 normal voices (see Tables 4.8 and 4.9), have an approximated 10 dB difference in SPL increased values for almost all vocal measures in Table 4.8, compared with Perkell et al. (1994). This increment was fundamental to propose SPL-normalized features to compensate the aerodynamic measures for differences in loudness. For the group results in section 4.4, the normative model appears to improve the discrimination capabilities of aerodynamic measures for both z-score analysis and hypothesis test. This suggests that a model-normalized feature (instead of SPL-normalized) could improve the overall power discrimination. Higher order predictors (in this case the interaction between predictors and squared ones) appears to not improve the overall model response. However, covariation with SPL and f_0 seems to improve it. As SPL and f_0 were the predictors for a given aerodynamic measures model (see Table 4.10), no relationship between aerodynamic measures was analyzed for the regressed z-score approach, however, a exploratory analysis was performed based on a Fourier-series model (reported in appendix B) which evidence common factors and boundary values between ACFL and MFDR. The regressed z-score poorly resemble the positive results in early studies of vocal hyperfunction. The power discrimination is weak. However, OQ and $\mathrm{SGP_L}$ prevails as salient features. The hypothesis test using the Bonferroni method performs similarly to z-score. Again, only OQ and $\mathrm{SGP_L}$ looks as salient features. Remarkably, effect sizes are in similar tendency and range from those estimated by SPL-Normalized comparison from section 4.3, but with a lesser magnitude.

4.6 Conclusion

The results of this study confirm previous hypotheses and preliminary results indicating that SPL-normalized estimates of glottal aerodynamic measures (SGP, ACFL, MFDR, and OQ) can be used to identify pathophysiological phonatory mechanisms associated with two primary manifestations of VH that are each distinctly different from normal vocal function. PVH is associated with abnormally lower values for all of the SPL-normalized glottal aerodynamic parameters, reflecting lower vocal efficiency and increased potential for trauma to vocal fold tissue. NPVH exhibits abnormally lower SPL-normalized values for SGP and OQ, but without concomitant decreases in SPL-normalized ACFL and MFDR, reflecting inefficient phonation and decreased potential for trauma to vocal fold tissue. The descriptive statistics of normal voices follows the pattern described in similar studies in the past and can be used as a reference data in the field. Regressed z-score analysis poorly follows previous studies in the field. Performing hypothesis test based on the Bonferroni method had a better performance than z-score analysis but was weaker than the SPL-normalized approach.

Chapter 5

Glottal airflow estimation through neck skin acceleration signal for the assessment of vocal hyperfunction

This chapter aims to extent the results shown in Chapter 4 , whereas now an Impedance-based Inverse Filtering (IBIF) algorithm is used to estimate the glottal (airflow) volume velocity (GVV) from a neck-surface acceleration (ACC) signal, including the proposed methods in chapter 3. For the experiments described in chapter 4, a synchronized neck skin acceleration signal was recorded along with oral airflow. The ACC-based aerodynamic measures were estimated by a subject-specific set of Q parameters from sustained vowels /a/ and /i/ in comfortable and loud loudness condition according to the methods presented in [28]. For the analysis, specific voice segments (tokens) were selected in female patients with phonotraumatic and nonphonotraumatic hyperfunction with their matched normal controls. Following the methods from chapter 4, SPL-normalized measures [135] are reported and statistically analyzed. In addition, a sensitivity analysis of Q parameters was carried through to investigate the parametric variations of the IBIF model. In this regard, a probabilistic method to estimate Q parameter in continuous speech is proposed. Employing the neck-surface acceleration signal along with the methods presented in chapter 3, the proposed probabilistic model is used to trace uncertainties in the estimated aerodynamic measures during continuous speech using the neck skin acceleration signal.

5.1 Accelerometer-based aerodynamic measures in subjects with hyperfunctional voices with matched controls

Chapter 4 describes methods to detect vocal hyperfunction based on oral airflow recordings of sustained vowels. As a reminder, the recordings were obtained from a series of five consecutive /pae/ utterances, where neck skin acceleration signal was simultaneously recorded as well. Both signals can be used to obtain ACC-based aerodynamic measures and performing same statistical methods than those presented in chapter 4. In this section, we evaluate if ACC-based aerodynamic measures were similar than those derived from the inverse filtered oral airflow reported in section 4.3 for the same selected tokens.

5.1.1 Methods

In this section, ACC-based aerodynamic measures from selected segments (tokens) of multiple /pae/ gestures are reported. The aim is to validate these ACC-based measures against oral airflow based results from chapter 4. During the oral airflow recordings for those experiments the neck skin acceleration signal was recorded simultaneously. Thus, we perform a synchronous analysis of the aerodynamic ACC-based measures.

To obtain an inverse filtered approximation of glottal airflow from the ACC signal, a subject-specific set of Q parameters from the IBIF algorithm. Q parameters were not derived and tested for same tokens. Instead, several sustained vocal gestures were recorded separately and used to estimate IBIF parameters. Synchronous recordings of oral airflow and neck skin acceleration signals were obtained when each subject performed a series of sustained vowels gestures (/a/ and /i/) with a constant pitch in comfortable and loud loudness condition. For each gesture, a bandpass filtered (60-1100 Hz) oral airflow vowel segment was used to perform inverse filtering with a single notch filter (SNF) constrained to unitary gain at DC [18]. The main criteria to set SNF parameters was the minimization of formant ripple in the closed phase using the SNF+Metrics approach described in section 3.1. Once a glottal airflow approximation is obtained from the CV mask, Q parameters were estimated using the optimization scheme depicted in section 3.3. A unique Q set was determined and visually confirmed (i.e., when waveforms from both oral airflow

and accelerometer were reasonable similar) with a Matlab GUI designed to assess inverse filtering by a trained user. Details of Matlab GUI options and capabilities are reported in appendix A. Afterwards, each ACC token was inverse filtered to obtain approximations of glottal airflow and aerodynamic measures for further statistical analysis.

5.1.2 Results for /pae/ gestures using the ACC-signal.

Table 5.1, shows ACC-based aerodynamic measures for ACFL, MFDR, and OQ as described in section 2.1.3. The results from Table 5.1 are similar to those in Table 4.4 for the OVV-based measures, but some differences do emerge. Table 5.2, shows the trimmed mean (20%) and MADN (Median Absolute Deviation Normalized) [124] differences between OVV-based and ACC-based aerodynamic measures, normalized to OVV-based measures. As is observed, MFDR has the greatest differences between -2.7 and 21.1 %, followed by OQ (ranging from -14.5 to 0.3 %). ACFL is the least affected aerodynamic measure with a variation between -1.2 to 10.2 %. Performing same statistical analysis as in Chapter 4 (i.e., paired multivariate and univariate test including effects sizes calculations), ACC-based aerodynamic measures, yields the results shown in Table 5.3. The results indicate that ACC-based aerodynamic measures does differentiate between the two populations. The results also indicate that there is a loss of statistical power by using ACC-based measures, wherein only PVH vs Controls in comfortable loudness, $F(3, 13) = 4.62$, $p = 0.021$, was statistically significant. The follow up hypotheses tests recognize that ACFL' and OQ' appears as salient features with larger effect sizes (showed in Table 5.3), and in similar performance than the InLab results presented in chapter 4.

TABLE 5.1: Group mean (standard deviation) for ACC-based aerodynamic measures for the /pae/ syllable productions in comfortable (upper row for each measure) and loud (lower values for each measure) voice for the PVH (n=16) and NPVH (n=14) patient groups and associated matched control groups with normal voices .

Measure	**PVH controls**	**PVH group**	**NPVH controls**	**NPVH group**
ACFL (mL/s)	205 (70.8)	321 (127)	267 (127)	198 (65.9)
	258 (101)	422 (221)	341 (164)	270 (100)
MFDR (L/s^2)	277 (105)	400 (197)	345 (218)	224 (105)
	399 (161)	657 (375)	560 (410)	381 (155)
OQ (%)	70.5 (11.8)	86.3 (9.96)	70 (5.68)	77.5 (14.5)
	69 (12.6)	82.8 (11.9)	66 (10.1)	67.1 (10.9)

TABLE 5.2: Mean (standard error), using trimmed mean 20% (MADN), of the Relative Differences between OVV-based and ACC-based aerodynamic measures from the /pae/ syllable productions in comfortable (upper row for each measure) and loud (lower row for each measure) voice for the PVH and NPVH patient groups and associated matched control groups with normal voices.

Measure	**PVH controls**	**PVH group**	**NPVH controls**	**NPVH group**
ACFL (% error)	-1.5 (15.6)	-1.2 (18.5)	3.3 (15.2)	5.4 (26.7)
	-0.5 (20.6)	1.2 (26.5)	4.0 (23)	10.2 (20.6)
MFDR (% error)	4.1 (25.9)	7.4 (14.7)	12 (14.4)	17.7 (14.9)
	-2.7 (26.9)	7.2 (25.7)	6.8 (21)	21.1 (16.8)
OQ (% error)	-1.7 (5.81)	-0.5 (12.7)	-0.1 (9.54)	0.3 (10.5)
	-5.0 (14.3)	-2.8 (14)	-14.5 (16.5)	-6.0 (20.9)

TABLE 5.3: Results of between-group statistical comparisons SPL-normalized features using the IBIF filtered ACC signal. Effect sizes are reported for the univariate, one-tailed paired t-tests (Cohens d), and the multivariate Hotelling's T^2 test. Negative values for the univariate effect sizes mean that SPL-normalized measures are smaller in the patient groups than in their respective control groups.$^*p < 0.025$.

Group comparison	Hotelling's T^2	ACFL'	OQ'
PVH vs Controls			
Comfortable	1.41*	-0.74*	-1.14*

Two reasons could be hypothesized for Table 5.3 results. First, the absence of SGP could have softened the power of the multivariate analysis, in consequence, a direct comparison with results reported in chapter 4 could be challenging or even misleading. Incorporating an approximation SGP based on the RMS value of ACC signal [52] is a possible avenue to improve the multivariate analysis in this case. Second, Q parameter estimation from sustained vowel segments (the current golden standard from Zañartu et al. [28]) could be considered as an initial approximation, and thus may need to be corroborated with additional measurements (see a proposal in section 5.4). As an initial approximation to this problem, one strategy is to perform a Q parameter sensitivity analysis that could be helpful to obtain insights on the uncertainty of the resulting measures due to Q error estimates.

5.2 Sensitivity of Q parameters

A sensitivity analysis is performed in this section for the IBIF Q parameters. In particular, how the variation of Q parameters change the frequency response of IBIF. The analysis is divided in magnitude and group delay vs. frequency. For the analysis in magnitude, multiple simulations of $10\log_{10} T_{skin}(\omega)/T_{skin_0}(\omega)$ are shown in Figures 5.1, 5.2, and 5.3 (see details of $T_{skin}(\omega)$ in equation (2.42)). Q parameter were set to range from 1 to 10 in steps of 0.33 for all Q1, Q2 and Q3. T_{skin0} was calculated with all Q parameters set to 1. The magnitude of spectrum for the Q1 sensitivity (related to R_m) is pictured in Figure 5.1. For Q1>1 the most notorious variation is near 70 Hz, and the response above 200 Hz is almost flat. For Q1<1 the variations follow a high-pass filter structure in the entire bandwidth of interest ($<$ 1000 Hz), with a zero near to 70 Hz. These variations could mainly affect low-frequency components of the glottal airflow estimates in magnitude and phase. For the magnitude of spectrum for the Q2 sensitivity (related to M_m), several simulations are shown in Figure 5.2. The variation affects the entire bandwidth of interest (from 200 to 1000 Hz) with amplitude changes varying from -10 dB to 6 dB. The behavior of Q2 resembles a gain factor for Q2<1 and a high-pass filter when Q2>1 in the range of interest (i.e., for $f > f_0$). These variations could affect the high-frequency components of the glottal pulse, which are related to the closing period of the glottal pulse. The magnitude of the spectrum for the Q3 sensitivity is shown in Figure 5.3. The more considerable variation occurs below 200 Hz, or similarly, below the fundamental frequency for most female voices. It is highly plausible that Q3 measurements are noisier due to the absence of signal power in the range $f < f_0$ during the calibration process of IBIF. Frequencies around 200 Hz have a 2 dB peak present. After that point, the response is fairly flat, so no effective changes are expected in the resulting glottal pulse from that component. Regarding the group-delay sensitivity of Q1, Figure 5.4 shows that most of the temporal variability is near the DC region and around 70 Hz. These effects are reduced to zero with increasing frequency, especially above 200 Hz with almost no differences regardless the variations of Q1. Figure 5.5 shows the group-delay changes for Q2, where similar variability than Q1 is observed but with a peak at 30 Hz approximately. These temporal variabilities are reduced to zero as frequency increase. Finally, Figure 5.6 presents the Q3 sensitivity for the group-delay where greater variability and more phase distortion is expected around f_0 region (200 Hz). These greater temporal variations due to Q3 in terms of the group-delay provides evidence that Q3 is hardest to estimate and has the potential to influence the

resulting inverse filtering of ACC-based measures. Most of the variations in these parameters occur at low frequencies, below the expected fundamental frequency. However, the Q parameter estimation or calibration process of the IBIF system is not excited in the frequency range of skin resonances, suggesting that the origin of uncertainties in the inverse filtering system could be attributed to this problem.

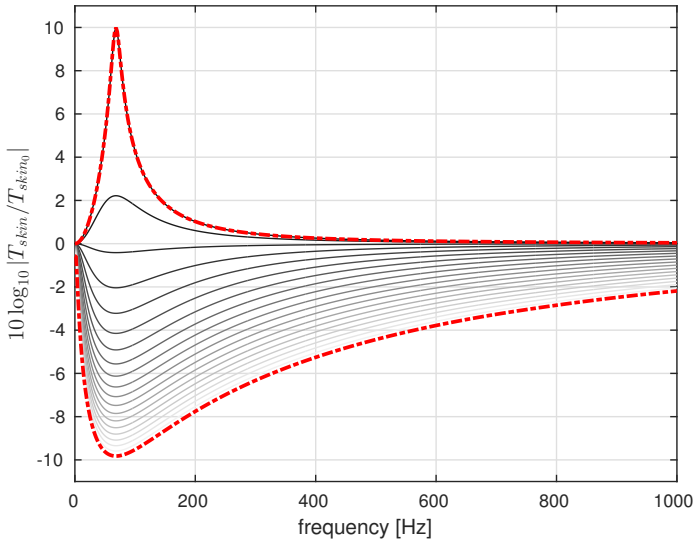


FIGURE 5.1: **Spectral magnitude differences due to Q1 Sensitivity**. Boundaries of (spectral magnitude) T_{skin}/T_{skin0} responses in red dash line. Lines between the boundaries from smaller (lighter) to higher (darkest) values of Q1. All other Q parameters are fixed to 1. The largest variability is ± 10 dB, near to 70 Hz.

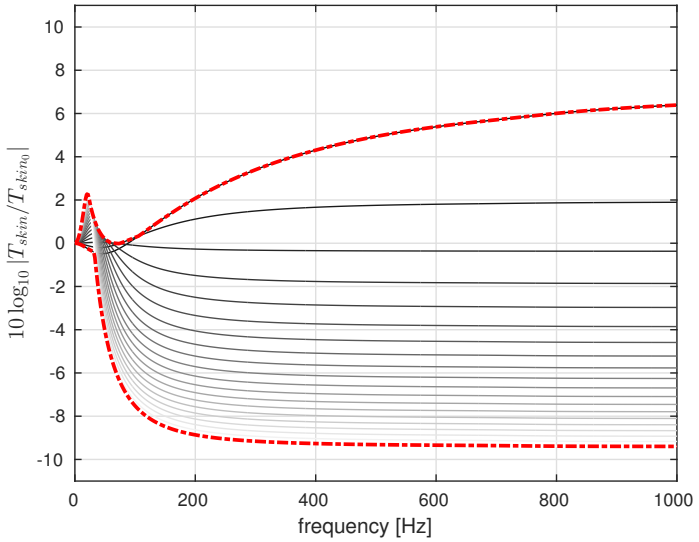


FIGURE 5.2: **Spectral magnitude differences due to Q2 Sensitivity**. Boundaries of magnitude responses in red dash line. Lines between the boundaries from smaller (lighter) to higher (darkest) values of Q2. All other Q parameters are fixed to 1. T_{skin} Variability is present almost in the entire bandwidth, with ranges between [-10,6] dB

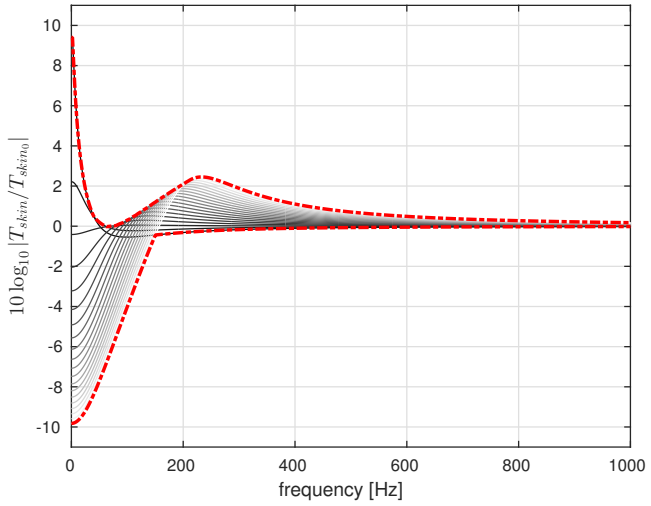


FIGURE 5.3: **Spectral magnitude reponse due to Q3 Sensitivity**. Boundaries of magnitude responses in red dash line. Lines between the boundaries from smaller (lighter) to higher (darkest) values of Q3. All other Q parameters are fixed to 1. T_{skin} Variability is present below 600 Hz with ranges between [-10,2] dB

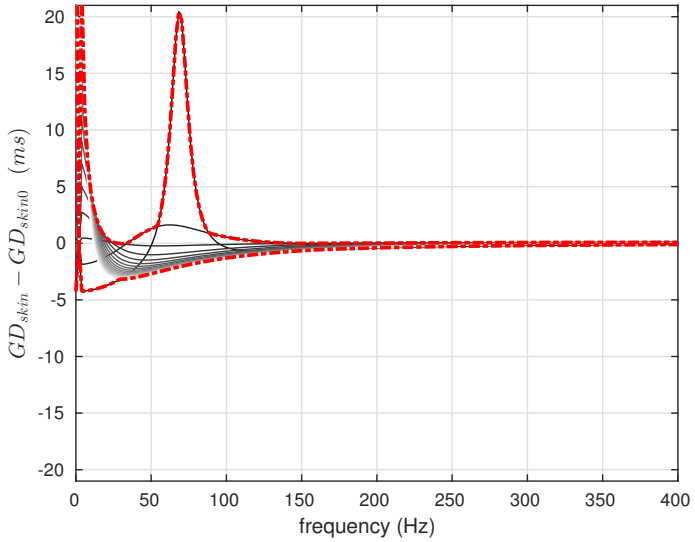


FIGURE 5.4: **Group delay differences due to Q1 Sensitivity**. Boundaries of group-delay differences in red dash line. Lines between the boundaries from smaller (lighter) to higher (darkest) values of Q1. All other Q parameters are were fixed to 1.

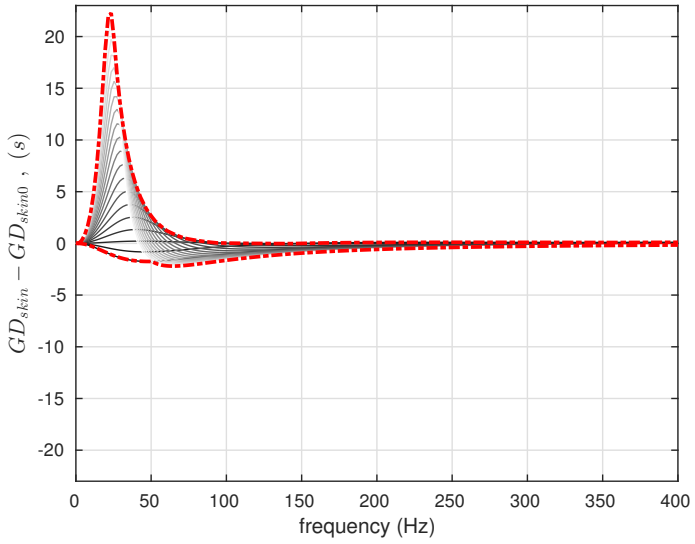


FIGURE 5.5: **Group delay response due to Q2 Sensitivity (related to M_m)**. Boundaries of group-delay differences in red dash line. Lines between the boundaries from smaller (lighter) to higher (darkest) values of Q2. All other Q parameters are were fixed to 1.

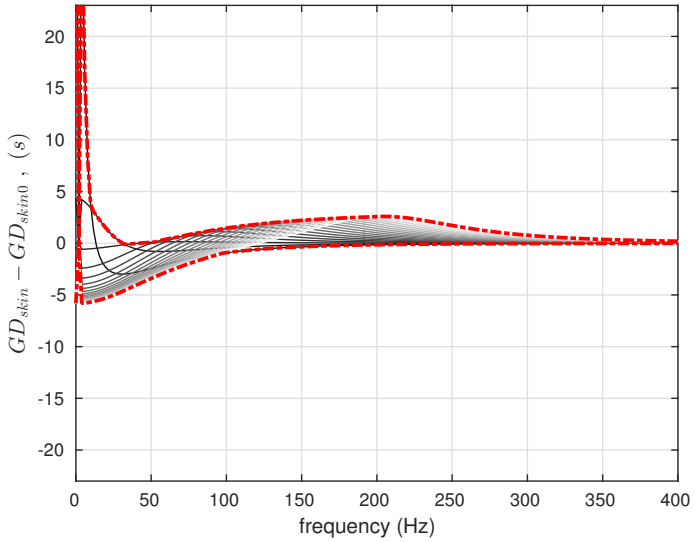


FIGURE 5.6: **Group delay differences due to Q3 Sensitivity (related to K_m).** Boundaries of group-delay differences in red dash line. Lines between the boundaries from smaller (lighter) to higher (darkest) values of Q3. All other Q parameters are were fixed to 1.

5.3 Q parameters estimation from sustained vowels: a case study

The initial method to estimate Q parameters used a sustained vowel /a/ in modal voice during the IBIF calibration process, which yielded reliable results in other vowels and pitch conditions [28, 134]. However, when the reference signal was recorded at a much higher pitch, the harmonic density of the voice is reduced. Such signal cannot excite completely the skin+subglottal system in the frequency range of interest, for which the resulting Q parameter estimation can be biased. Additionally, when subjects have hyperfunctional voices, some of the assumptions for inverse filtering methods (e.g., closed glottis), may not hold. Using several sustained vowels in non-connected speech instead of a single token in the IBIF calibration process can provide insights on how IBIF Q parameters vary with respect to the vocal gestures. Following same procedure described in section 5.1, each vowel segment was used to estimate a Q set in a fully automated way. In Table 5.5, all estimated Q parameters are reported for four subjects: Patients with phonotraumatic and nonphonotraumatic hyperfunction and their respective matched controls. As a first observation, Q sets obtained from different gestures do exhibit differences, as indicated by the standard deviation value. Given these differences, it is desired to analyze how the aerodynamic measures may have deviated from the OVV-based measures (the baseline). To achieve this task, we proceed as follows. From Table 5.5 the Q set with the largest deviation is used to cross-validate the remaining vocal gestures. As an example, consider the Q sets for the subject Normal 01 shown in Table 5.5; the largest deviation occurs for vowel /i/ in comfortable loudness condition. Subsequently, H1H2, ACFL, MFDR, and OQ are computed. From these aerodynamic measures, the maximum deviation is reported in Table 5.4. As observed, there is a substantial deviation when Q set from the worst case scenario within each subject is applied to the remaining gestures. These results suggest that IBIF-Q parameter estimation from a single sustained vowel may have problems in a more general context, e.g., continuous speech scenarios, which is analyzed in the following section.

TABLE 5.4: Maximum deviation values of the relative error for the aerodynamic measures.

Subject	H1H2 %	ACFL %	MFDR%	OQ%
Normal 01	-35.2	30.2	26.5	-30.2
Phonotraumatic 01	-67.2	37.5	57.7	-30.3
Normal 02	-303.2	-27.3	43	-43.5
Non Phonotraumatic 02	-126.4	29.3	8.5	20.7

TABLE 5.5: Q parameters derived from sustained vowels for four subjects: Two normal paired with two hyperfunctional (phonotraumatic and nonphonotraumatic) voices. The mean and standard (std) error is calculated, along with the standard error of the mean based on bootstrap samples (see section C.2.4 for further details).

	vowel	loudness	Q1	Q2	Q3	Q4	Q5
	/a/	comf	1.17	1.18	6.91	1.40	0.80
Normal		loud	1.15	1.22	17.69	1.00	1.00
01	/i/	comf	0.57	0.71	5.81	1.40	1.40
		loud	1.11	1.13	12.08	1.00	1.00
		mean	1.00	1.06	10.62	1.20	1.10
		std	0.29	0.23	5.45	0.23	0.30
		std (mean)	0.12	0.10	2.26	0.10	0.11
	/a/	comf	0.88	2.08	9.68	1.25	1.00
Phonotraumatic		loud	0.93	2.23	14.48	1.45	1.50
01	/i/	comf	0.10	0.58	0.10	0.75	1.1
		loud	0.97	0.85	3.07	1.55	1.3
		mean	0.72	1.43	6.83	1.25	1.30
		std	0.42	0.84	6.48	0.36	0.20
		std (mean)	0.19	0.42	3.19	0.16	1.00
	/a/	comf	1.09	1.59	10.46	1.55	1.00
Normal		loud	1.08	1.86	20.00	0.85	0.40
02	/i/	comf	1.30	0.80	0.10	1.55	0.50
		loud	1.52	0.72	6.19	1.55	1.30
		mean	1.25	1.24	9.19	1.38	0.80
		std	0.21	0.57	8.37	0.35	0.40
		std (mean)	0.09	0.24	3.52	0.15	0.20
	/a/	comf	0.10	0.53	8.32	1.25	1.1
Non-Phonotraumatic		loud	0.30	0.44	11.04	0.70	0.7
02	/i/	comf	0.10	0.56	8.48	1.00	1.00
		loud	0.26	0.10	7.08	1.55	0.8
		mean	0.19	0.41	8.73	11.25	0.9
		std	0.10	0.21	1.66	3.62	0.2
		std (mean)	0.05	0.10	0.70	1.58	0.1

5.4 Frame-based aerodynamic measures and their uncertainties

In order to increase the statistical power and robustness of the calibration method, more Q estimates are needed, and a frame-based estimation of IBIF parameters for continuous speech could be designed. In this regard, the rainbow passage (RP) paragraph, is an appropriate candidate as a vocal gesture for obtaining several samples of Q parameters. In the RP paragraph, the speech co-articulation produces complex dynamic changes that allow for enriching the sample space of estimating Q parameters. However, performing IF from the CV mask in running speech is still a challenging task. Thus, a strategy to minimize the influence of unreliable IF performance in the context of estimating Q parameters dynamically is proposed in this section. Therefore, a method to determine the uncertainty of glottal aerodynamic estimates from a neck skin acceleration signal is introduced. We extend previous analyses of sustained vowels using the neck skin + subglottal scheme from section 5.1.2 toward continuous speech in a frame-based approach. The underlying hypothesis is that when Q parameters are estimated in a frame-based approach, a probabilistic model with a distinct central tendency and dispersion can be determined. Thus, selected voiced frames of both oral-airflow and acceleration signals recorded during the rainbow passage gesture are used to build a probabilistic model of IBIF Q parameters. The model is used to run multiple random realizations of the inverse-filtered neck-surface acceleration signal, with the goal to propagate the uncertainties of the IBIF Q parameters to the aerodynamic measures estimates. The probabilistic model is estimated using data from patients with vocal hyperfunction and normal matched-controls at a comfortable pitch and loudness. Note that we observed variations for loudness and pitch, but in this first approach, only a single condition was explored (comfortable, modal voice). Statistical analysis are reported and compared with those obtained with sustained vowels.

5.4.1 Methods

Estimates of glottal airflow in running speech are explored here in a frame-based approach for both oral airflow and accelerometer signals. Inverse filtering of oral airflow was performed using the SNF+Metrics approach described in section 3.1.1, and aerodynamic measures (ACFL, MFDR, H1H2, and OQ) were estimated for each voiced frame. IBIF parameters Q1, Q2, and Q3, were determined with the PSO algorithm described in section 3.3 for each frame, altogether with aerodynamic measures from ACC filtered signal. However, previous to apply the frame-based approach, and using methods described in section 3.3.1, a unique set of Q4 and Q5, was estimated for each RP paragraph. The later was justified assuming that uncertainties for Q4 and Q5 are challenging to set in the same probabilistic framework that neck skin related Q parameters, and the variability is expected to be minimal. Therefore, Q4 and Q5 were fixed for all voiced frames. Only the voiced frames were considered in this analysis where a voice activity detection algorithm based on the autocorrelation function was used.

A cleaning procedure is performed to eliminate outliers and/or unreliable data, if any of the following criteria is met:

1. Avoid soft loudness (SL) conditions: Frames where the skin acceleration level (SAL) was less than percentile 25% were discarded, i.e., when $SAL < \hat{P}_{25}(SAL)$. Reasons to do this are: 1) Greater loudness tends to highlight the effects of the vocal tract and skin properties in both signals, and 2) SL were excluded in recent studies using /pae/ gesture [135] for being more unreliable than others loudness conditions.

2. Discard frames with extreme values (0.1 and 10) in any Q parameter. A ± 1 order of magnitude margin is considered a physiologically reasonable boundary for Q parameters [28]. These extreme values are already limited (constrained) in the PSO algorithm meaning that they are outside of the physiological range.

3. Avoid high pitch voiced frames, i.e., a higher fundamental frequency (f_0) of 350 Hz is proposed, to get (at least) three harmonics within the 1.1 kHz of bandwidth of the CV mask. Limiting f_0 is hypothesized to reduce bias in the estimation of Q parameters.

4. Avoid extreme values of formant (F_1): only include values between percentile 25% and 95%, where F_1 is a vector containing the first formant of each voiced frame. As mentioned before, inverse filtering lower formants is challenging due to the f_0 and F_1 interaction, even with the improvements presented in section 3.1.

5. Only retain frames with large similarities of the Pearson coefficient, i.e., $r_{GVV} \geq 0.7$ and $r_{dGVV} \geq 0.7$, where the correlation coefficient were calculated from both OVV and ACC inverse filtered signals with equations (3.6)(for r_{GVV}) and (3.7)(for r_{dGVV}).

6. Discard frames with mean square error in the calibration process larger than a given threshold for either glottal airflow and its time-derivative.

5.4.2 Statistical analysis of Q parameters estimates

Several statistical analysis are performed in this section for the frame-based approach. However, only Q1, Q2, and Q3 analyzed are performed, assuming that trachea length (related to Q4) and the accelerometer position (related to Q5) were fixed as was mentioned in section 5.4.1. Descriptive statistics, and a statistical model are analyzed and reported. For this purpose, Q1, Q2, and Q3 parameters are assumed to be uncorrelated. Despite that this assumption simplifies the analysis for each Q parameters more complex analysis can be done as further research. Even thought, it is difficult to test this assumption, in practice, no significant linear correlation between Q parameters was detected when using the Pearson correlation coefficient with a threshold $|r| < 0.7$.

For the probabilistic scheme, histograms, kernel density estimation (KDE) [96], and statistical model are calculated. Under the assumption of unimodality, a Gamma density distribution (hereafter the PDF-based model) is fitted using the Maximum Likelihood method. The fitted probability density estimations are plotted in Figure 5.7, 5.8, 5.9, and 5.10, each of which correspond to subject Normal 01, Phonotraumatic 01, Normal 02, and Non-Phonotraumatic 02, respectively. Here in, only histograms follow the total probability rule, (i.e., $\sum P(x) = 1$), and KDE and the statistical model are scaled arbitrarily for visual comparison.

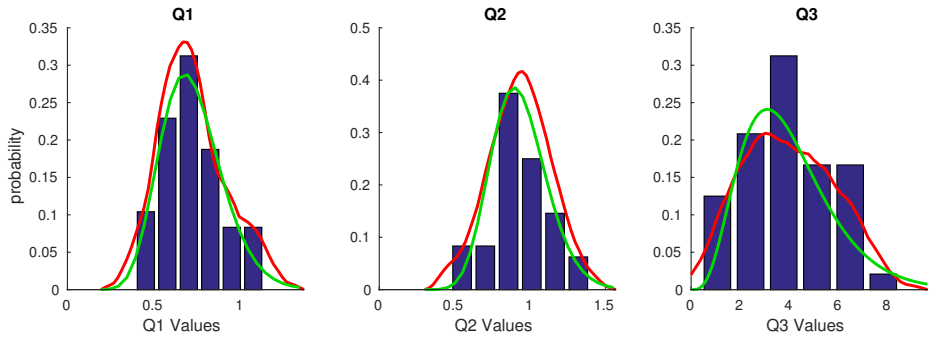


FIGURE 5.7: Histograms (bars), nonparametric p.d.f. estimation (red), and Gamma p.d.f. fit (green) for subject Normal 01.

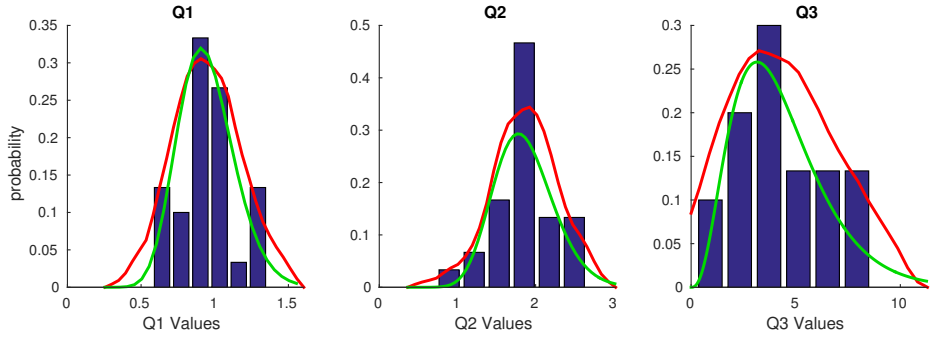


FIGURE 5.8: Histograms (bars), nonparametric p.d.f. estimation (red), and Gamma p.d.f. fit (green) for subject Phonotraumatic 01.

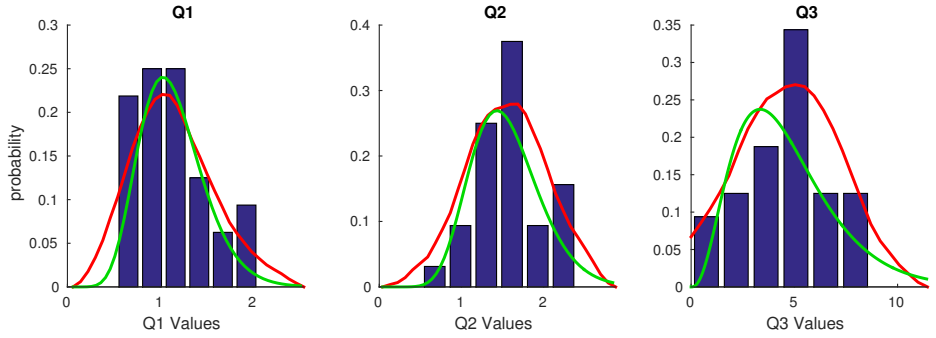


FIGURE 5.9: Histograms (bars), nonparametric p.d.f. estimation (red), and Gamma p.d.f. fit (green) for subject Normal 02.

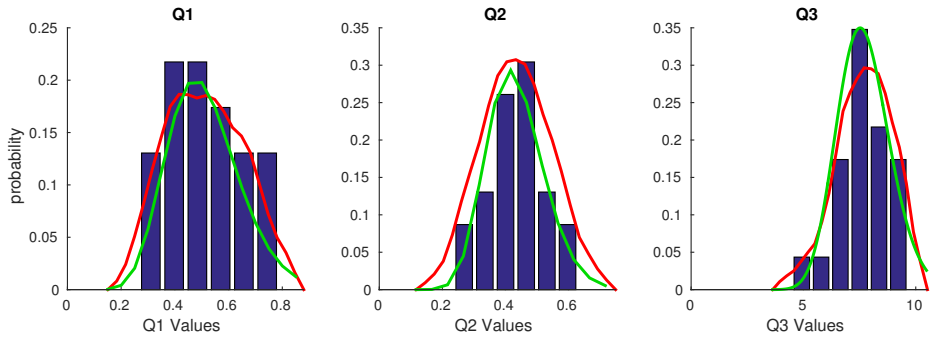


FIGURE 5.10: Histograms (bars), nonparametric p.d.f. estimation (red), and Gamma p.d.f. fit (green) for subject Non-Phonotraumatic 02.

As an initial observation of the Figures 5.7, 5.8, 5.9, and 5.10, the Q parameters histograms and KDE follow approximately an exponential or bell-like distributions, allowing to estimate a properly defined central tendency and dispersion. Descriptive statistics are summarized in Table 5.6 for several models, including normal (Gaussian), Gamma, and bootstrap trimmed mean (20%).

When comparing the mean values reported in Table 5.6 with those in Table 5.5 an absolute difference is observed which has significant bias. The differences between these values are showed in Table 5.7. The same observation is valid between the standard error listed in Table 5.8.

TABLE 5.6: IBIF parameters statistics from multiple vowels utterances. Central tendency and standard error in parenthesis are reported for the four subjects. For the Gamma based model the central tendency values is the mode, for the Gaussian the sample mean, and for the bootstrap the trimmed mean at 20 %.

Parameter	Model	Normal 01	Phonotraumatic 01	Normal 02	Non Phonotraumatic 02
Q1	Gaussian	0.72 (0.18)	0.95 (0.19)	1.1 (0.36)	0.51 (0.13)
	Gamma	0.68 (0.18)	0.91 (0.19)	1 (0.35)	0.48 (0.13)
	Bootstrap	0.71 (0.052)	0.94 (0.063)	1.1 (0.14)	0.51 (0.062)
Q2	Gaussian	0.93 (0.18)	1.9 (0.37)	1.5 (0.41)	0.43 (0.091)
	Gamma	0.9 (0.18)	1.8 (0.38)	1.4 (0.42)	0.42 (0.09)
	Bootstrap	0.94 (0.046)	1.9 (0.12)	1.6 (0.15)	0.43 (0.04)
Q3	Gaussian	4 (1.8)	4.3 (2.1)	4.6 (2)	7.7 (1.1)
	Gamma	3.1 (1.8)	3.2 (2.2)	3.3 (2.4)	7.5 (1.2)
	Bootstrap	3.9 (0.55)	4.2 (0.86)	4.7 (0.86)	7.8 (0.44)

TABLE 5.7: IBIF mean parameters differences between Q sets derived from sustained vowels and a frame-based Gaussian model. The absolute difference of the mean values are reported

Parameter	Model	Normal 01	Phonotraumatic 01	Normal 02	Non Phonotraumatic 02
Q1	Gaussian	0.28	0.23	0.14	0.31
Q2	Gaussian	0.13	0.47	0.13	1.11
Q3	Gaussian	6.62	2.43	5.59	1.13

TABLE 5.8: IBIF standard deviation (std) parameters differences between Q sets derived from sustained vowels and a frame-based Gaussian model. The absolute difference of the std values are reported

Parameter	Model	Normal 01	Phonotraumatic 01	Normal 02	Non Phonotraumatic 02
Q1	Gaussian	0.11	0.23	0.08	0.03
Q2	Gaussian	0.05	0.47	0.16	0.12
Q3	Gaussian	3.65	4.35	6.37	0.53

In summary, the results in this section show a significant bias in both mean and standard error values between Q parameters estimated from sustained vowels and the proposed frame-based approach. As more sample Q set are available from the frame-based approach, a probabilistic model based on the Gamma distribution was possible to obtain. The effect of this analysis on the resulting inverse filtering glottal waveform, now along with confidence intervals derived by multiple realization of the probabilistic model, is explored in the following section.

5.4.3 Uncertainties of glottal waveforms

In order to improve the statistical significance of the parameter estimate for exploring the uncertainties, we perform Monte Carlo (MC) random simulations to obtain multiple (n=200) Q sets based on preceding pdf's described in section 5.4.2. Each Q MC set is used to filter the ACC signal, from which the uncertainties of the estimates is quantified using a re-sampling technique called bootstrap (see section C.2.4 for further details). Then, for each waveform, mean and percentile ($\hat{P}_5$ and $\hat{P}_{95}$, hereafter referred as confidence intervals) values are calculated. The results for a given segment are presented in Figures 5.11, 5.12, 5.13, and 5.14, and correspond to subjects Normal 01, Phonotraumatic 01, Normal 02, and Non Phonotraumatic 02, respectively. For each figure, left panels show the simulations based on the PDF-model, and right panels the bootstrapped trimmed mean (20 %). For simulations with the PDF-model, the waveform uncertainty is greater and the distribution for each sample is not perfectly symmetric (not shown) around the values calculated with Q parameters estimates from the model (solid blue). In addition, during the closing phase of the glottal pulse, the uncertainties are greater than during the opening phase of the glottal pulse (see the negative peak for the time derivative waveform). This is an important finding not reported before since it could have incidence in vocal measures like MFDR, SQ, and CPP which ones depend on the harmonic and spectral balance. However, the mean value (solid blue) looks like a good approximation in the closing phase, indicating that the estimated Q parameters from the PDF-model approach, capture a Q mean set candidate for the subject.

For subject Normal 01 (Figure 5.11), the inverse filtered signal has a normal glottal pulse shape and it is similar to the OVV-based filtered waveform (solid gray). Note that the later mainly occurs between the confidence intervals (CI), which is coherent with the proposed statistical framework, and could be considered one (of many) possible glottal waveform realization. We highlight that for first time is possible to bound the inverse filtered waveform and to establish the most likely glottal waveform using the proposed statistical scheme. For the matched hyperfunctional voice, Phonotraumatic 01 (Figure 5.12), the glottal pulse is also well defined. Interestingly, it is clearly observable that there is a notch during the opening phase, suggesting that source-filter interactions [57, 58, 102, 136] are captured with the IBIF-approach.

For the non-phonotraumatic matched pair, subject Normal 02 (Figure 5.13) exhibits the glottal waveform has a normal shape too, and even more than the

OVV-based filtered waveform (solid gray). This suggests that the ensemble of multiple measures of Q parameters may improve the robustness of glottal estimates based on the ACC signal, even when the estimated glottal airflow from CV mask is less reliable in the frame.

For the subject Non Phonotraumatic 02 (Figure 5.14), some ripple is present in the closed phase, and in the neighbor of the peak amplitude. Its shape has a plateau rather than a sharp form. The residual ripple can be explained as a poor suppression of the first subglottal resonance, which is confirmed with the same pattern in the confidence intervals. As a reminder, trachea length and accelerometer position were estimated using methods early described in section 3.3.1 and fixed before the frame-based analysis presented in this chapter, as the computational cost is too high to perform multiple random realizations. This limitation may have played a role in this case.

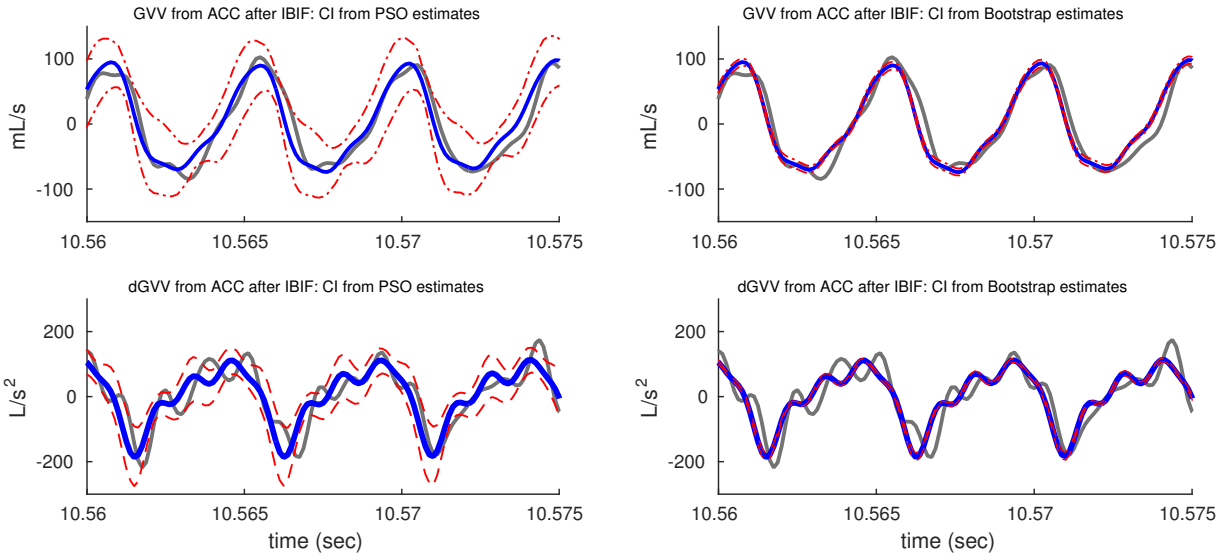


FIGURE 5.11: Uncertainties of glottal waveforms for subject **Normal 01** (red dot-dashed line). Glottal airflow waveforms (solid blue) from ACC filtered signal closely follows the estimated glottal airflow from OVV signal (solid gray). CI: Confidence Interval. PSO: Particle Swarm Optimization.

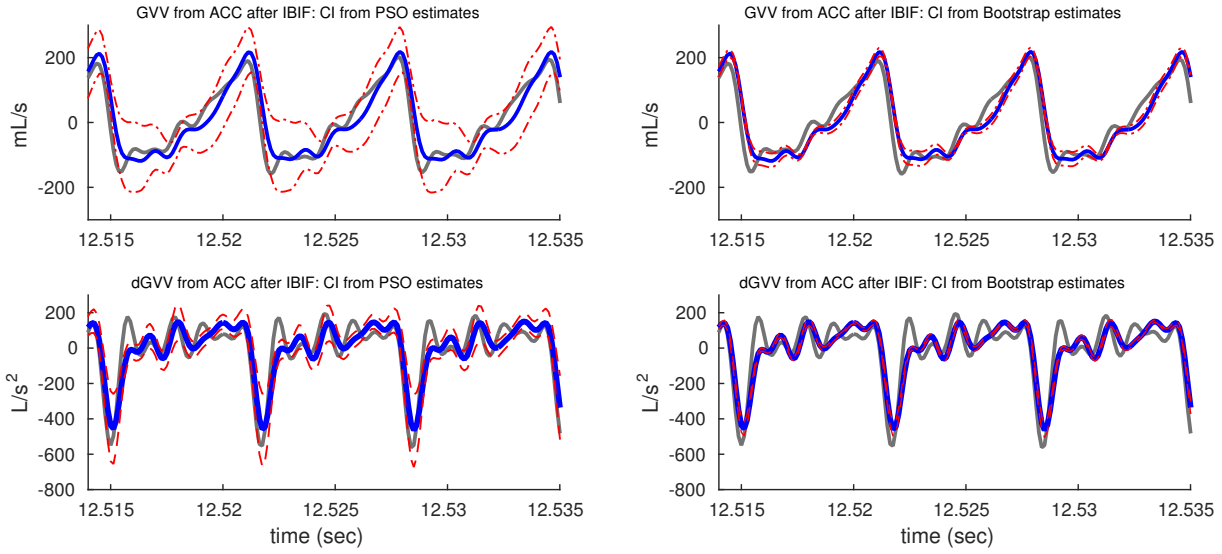


FIGURE 5.12: Uncertainties of glottal waveforms for subject **Phonotraumatic 01** (red dot-dashed line). Glottal airflow waveforms (solid blue) from ACC filtered signal closely follows the estimated glottal airflow from OVV signal (solid gray). CI: Confidence Interval. PSO: Particle Swarm Optimization.

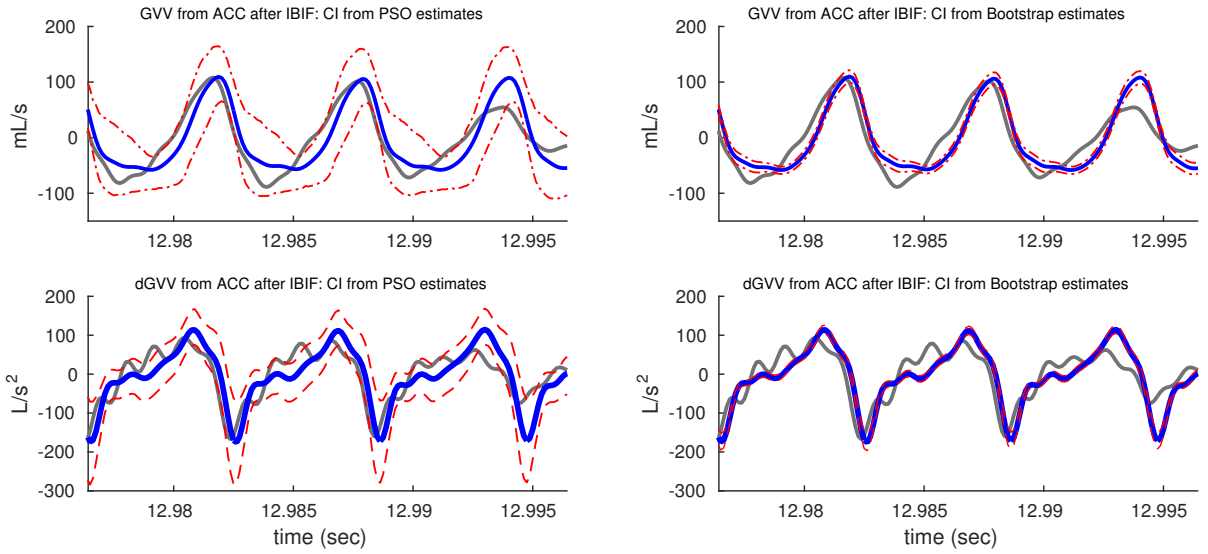


FIGURE 5.13: Uncertainties of glottal waveforms for subject **Normal 02** (red dot-dashed line). Glottal airflow waveforms (solid blue) from ACC filtered signal closely follows the estimated glottal airflow from OVV signal (solid gray). CI: Confidence Interval. PSO: Particle Swarm Optimization.

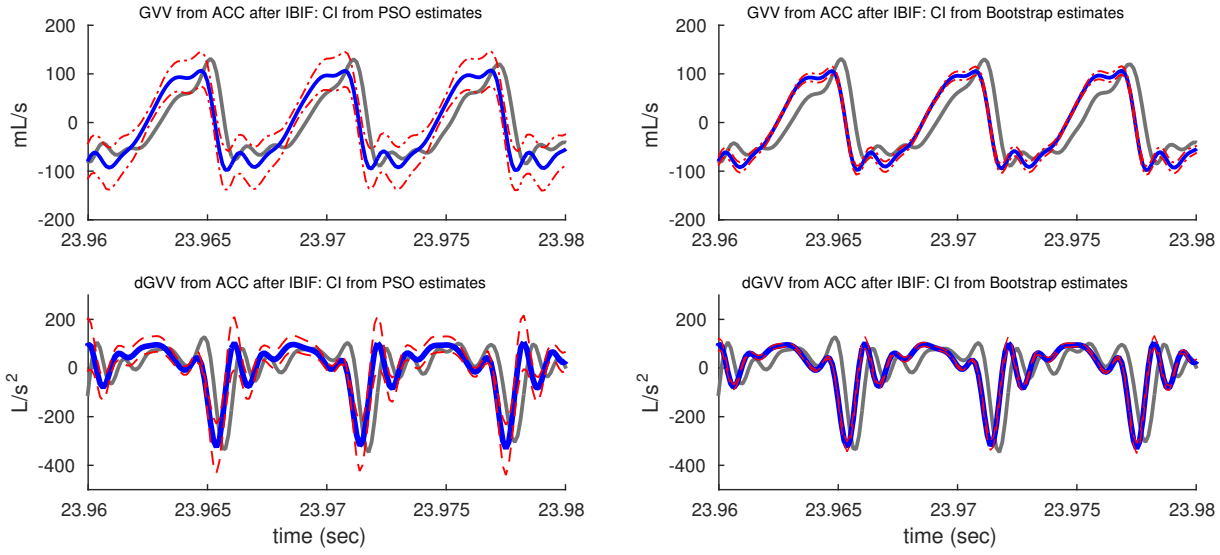


FIGURE 5.14: Uncertainties of glottal waveforms for subject **Non Phonotraumatic 02**. Glottal airflow waveforms (solid blue) from ACC filtered signal closely follows the estimated glottal airflow from OVV signal (solid gray). CI: Confidence Interval. PSO: Particle Swarm Optimization.

5.4.4 Relating uncertainties of Q parameters to aerodynamic measures

For each realization of the inverse filtered ACC signal from section 5.4.3, selected aerodynamic measures (ACFL, MFDR, H1H2 and OQ), were estimated for each voiced frame, thus allowing for relating the Q parameter uncertainties to those measures. The results for the PDF-based model are shown in Figure 5.15, 5.16, 5.17 and 5.18 and are described as follows. At the top of each sub-Figure, the median value (bold circle) and percentile range (5% and 95%, the vertical line behind bold circle) were drawn to elucidate confidence intervals (CI) for each voiced frame. Note that the log-scale in ACFL and MFDR allows for a nearly constant CI across the voiced frames. In a linear scale, these uncertainties appear more amplitude dependent, i.e., lower values have smaller dispersion. Under this observation, a relative error is proposed to evaluate the uncertainty and to cancel out the amplitude dependency, which is computed according to equation (5.1).

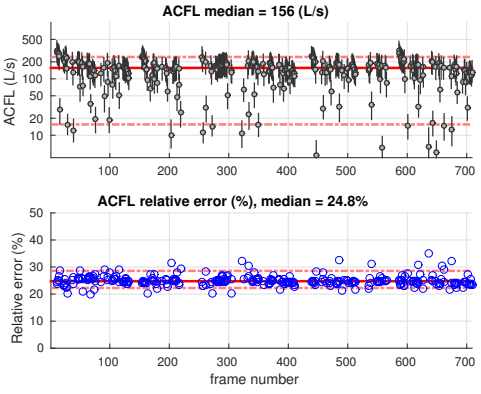
$$\%\text{R.E.}_{AM} = \frac{\text{MADN(AM)}}{\text{median(AM)}} \cdot 100\% \quad , \tag{5.1}$$

where MADN is the Median Absolute Deviation Normalized [124], and AM is the given aerodynamic measure (e.g., ACFL, MFDR, among others). For OQ and H1H2, uncertainty is simply quantified with the MADN value within the frame as a measure of dispersion. The additional horizontal lines in red show median and percentile range (5% and 95%, dash-dot line) for the median values (bold circle), aiming to summarize the voice behavior in the rainbow passage paragraph. Results reported in Figures 5.15 to 5.18, allow for describing the influence of the inverse filtering uncertainties on the aerodynamic measures derived form the ACC signal. The results of the $\%\text{R.E.}_{AM}$ for ACFL and MFDR measures are very similar, with a small number of outliers (beyond the limits of percentiles 5% and 95%). For H1H2 the dispersion (valued with MADN statistics) a ranges from 1.2 to 3.1 dB (on average), with a moderate number of outliers. Open quotient (OQ) is between 3.22% to 8.42 %, with a few outliers as well. Table 5.9 summarize the average $\%\text{R.E.}_{AM}$ for each subject and aerodynamic measure.

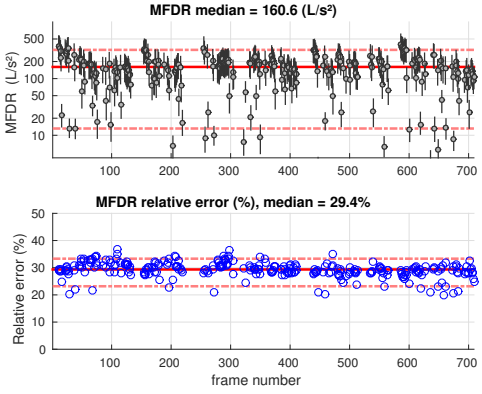
Analyzing the results in Table 5.9, the lower $\%\text{R.E.}_{AM}$ are for the aerodynamic measures ACFL and MFDR for the Phonotraumatic 01 and Non Phonotraumatic 02 cases, with approximately 20 % of variation. OQ have the lower mean relative error, but the larger standard error (std) of the group of measures.

TABLE 5.9: Mean and standard error (std) variability of the traced aerodynamic measures from the $Q_{1,2,3}$ uncertainties. Subscript indicating the given relative error according to equation (5.1). Note that these statistics are derived from the data illustrated in the bottom panels of each subfigure of the Figures 5.15 to 5.18

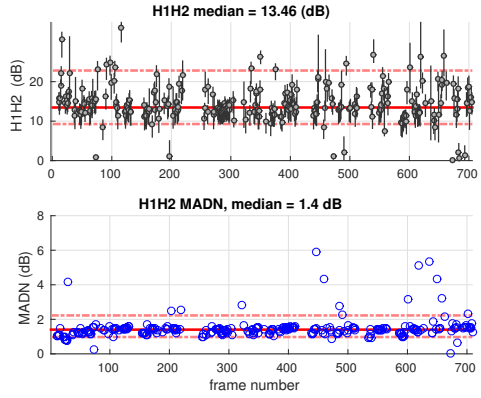
Subject	$ACFL_{\%}$	$MFDR_{\%}$	$H1H2_{dB}$	$OQ_{\%}$
Normal 01	24.8 (1.9)	29.4 (2.8)	1.40 (0.6)	5.63 (3.6)
Phonotraumatic 01	21.1 (1.0)	21.4 (2.4)	1.20 (0.7)	8.42 (7.1)
Normal 02	28.0 (2.9)	31.6 (4.3)	1.45 (0.9)	5.44 (5.6)
Non Phonotraumatic 02	20.3 (2.1)	20.1 (1.7)	3.09 (0.8)	3.22 (3.9)



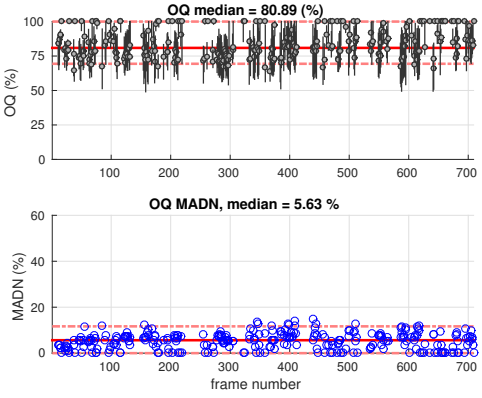
(A) ACFL Analysis



(B) MFDR Analysis

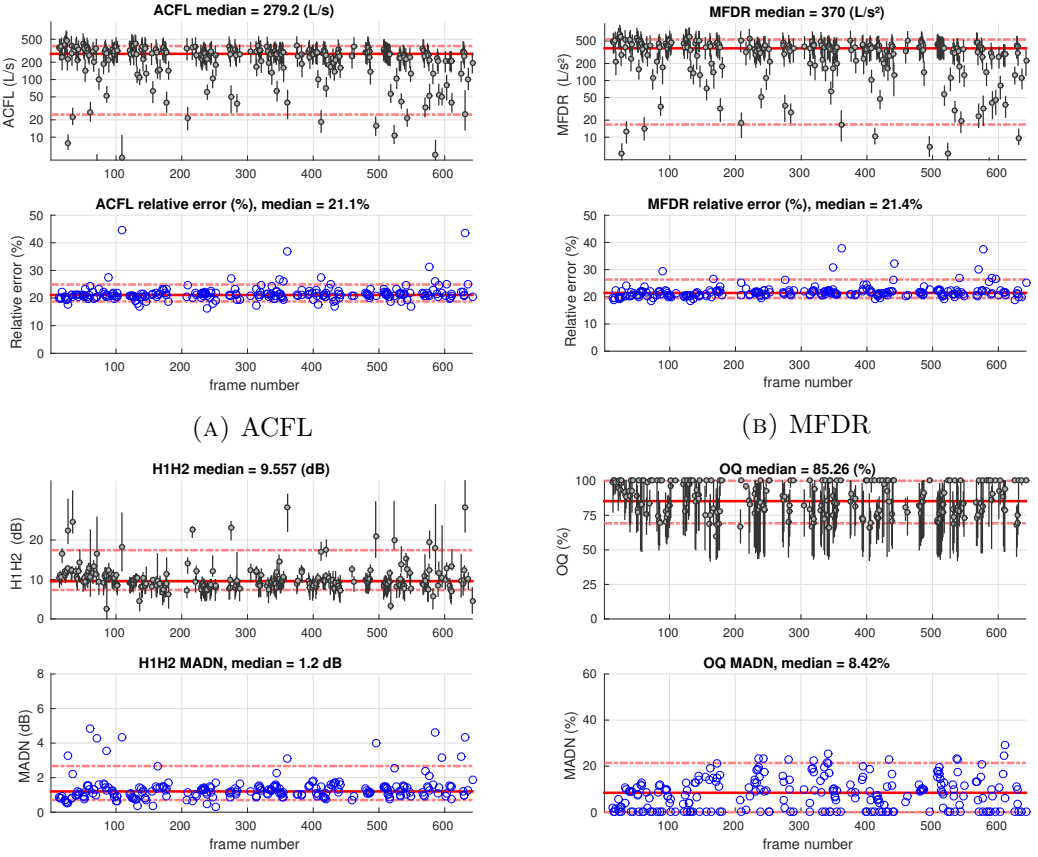


(C) H1H2 Analysis



(D) OQ Analysis

FIGURE 5.15: Mean values and uncertainties of IBIF-based aerodynamic measures of voiced frames of the entire rainbow passage paragraph for subject **Normal 01**. At the top of each figure the given aerodynamic measures (bold circle) and their uncertainties from IBIF measures (line behind bold circle) are shown. At the bottom of each subfigure, the relative error %R.E. for each voiced frame is shown. Horizontal red lines show median value (solid) and percentile 5% and 95% (dot-dashed) of bold circles.



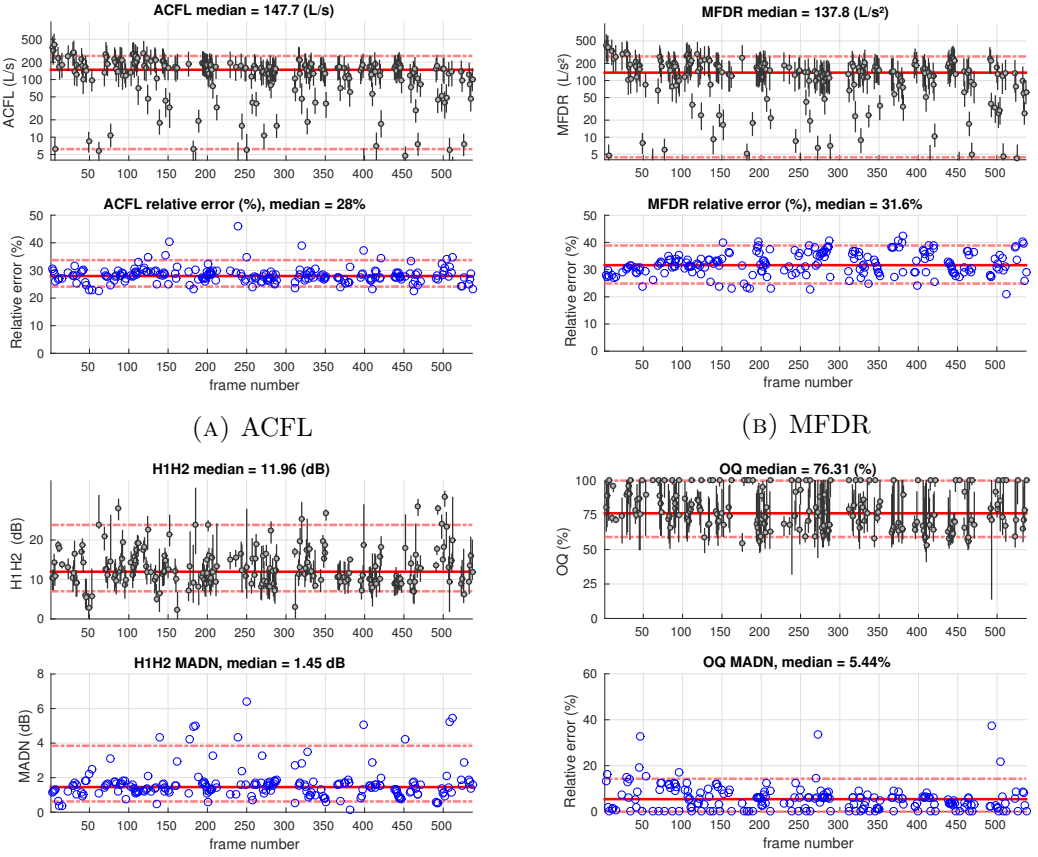
(A) ACFL

(B) MFDR

(C) H1H2

(D) OQ

FIGURE 5.16: Mean values and uncertainties of IBIF-based aerodynamic measures of voiced frames in continuous speech for subject Phonotraumatic 01. At the top of each figure the given aerodynamic measures (bold circle) and their uncertainties from IBIF measures (line behind bold circle). At the bottom of each subfigure the relative error %R.E. for each voiced frame is shown. Horizontal red lines show median value (solid) and percentile 5% and 95% (dot-dashed) of bold circles.



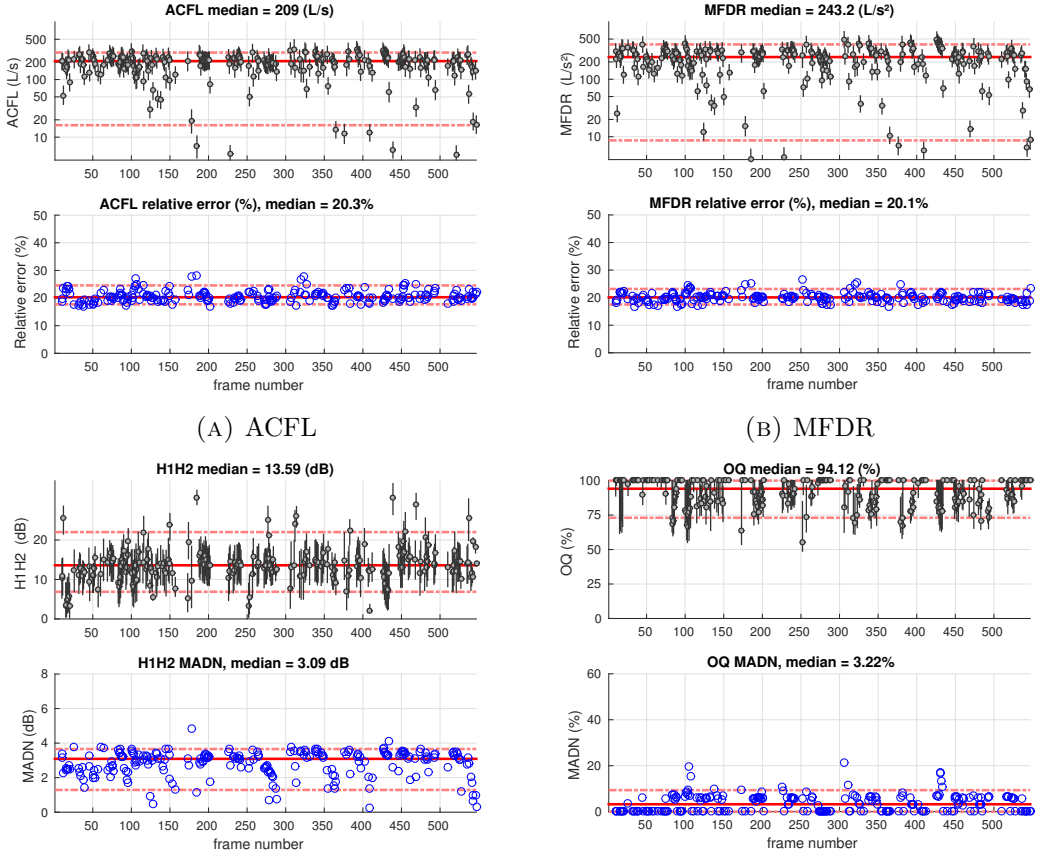
(A) ACFL

(B) MFDR

(C) H1H2

(D) OQ

FIGURE 5.17: Mean values and uncertainties of IBIF-based aerodynamic measures of voiced frames in continuous speech for subject Normal 02. At top of each figure the given aerodynamic measures (bold circle) and their uncertainties from IBIF measures (line behind bold circle) are shown. At the bottom of each subfigure the relative error %R.E. for each voiced frame. Horizontal red lines show median value (solid) and percentile 5% and 95% (dot-dashed) of bold circles.



(A) ACFL

(B) MFDR

(C) H1H2

(D) OQ

FIGURE 5.18: Mean values and uncertainties of IBIF-based aerodynamic measures of voiced frames in continuous speech for subject Non Phonotraumatic 02. At top of each figure the given aerodynamic measures (bold circle) and their uncertainties from IBIF measures (line behind bold circle) are shown. At the bottom of each subfigure the relative error %R.E. for each voiced frame. Horizontal red lines show median value (solid) and percentile 5% and 95% (dot-dashed) of bold circles.

5.4.5 Aerodynamic measures evaluation for the three inverse filtering methods.

To compare the aerodynamic measures estimates from the three methods, namely supraglottal OVV-based (labeled in Figures 5.19 to 5.22 as *from OVV*), subglottal ACC-based using the RP paragraph (labeled *from RP*), and subglottal ACC-based from sustained vowels utterances (labeled *from Vowels*), boxplots with multiple statistics are presented. ACFL, MFDR, H1H2 and OQ are calculated for each voiced frames. The boxplots describe the sample mean (diamond), interquartile 25% & 75% (box), median (horizontal red-line), and quartile 5% & 95% for the outer limits. Outliers are excluded as the occurrence was not relevant for the cases.

The results indicate that median values are consistent between methods, except for MFDR from subjects Normal 01, Phonotraumatic 01, and Normal 02, which were underestimated when Q parameters were derived from sustained vowels. The H1H2 was larger using Q parameters derived for sustained vowels for the subject Non Phonotraumatic 02. The H1H2 parameter shows less dispersion by estimating from RP. This result is encouraging due to the normalized nature of the parameter himself. This suggests that H1H2 could be capable to differentiate normal control subjects from their hyperfunctional matched pair in continuous speech scenarios, without using the SPL-Normalized features presented in chapter 4. ACFL and MFDR show a symmetric dispersion as H1H2 and OQ show tail asymmetry, that could be explained for the occurrence of sporadic events or outliers.

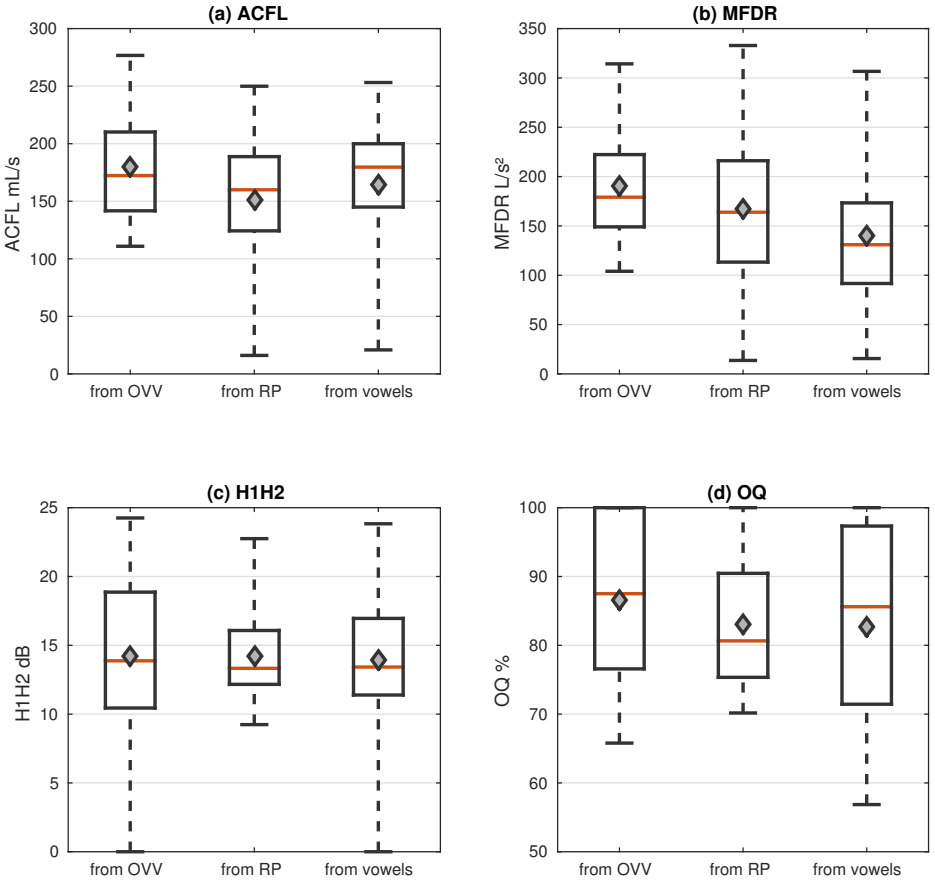


FIGURE 5.19: Boxplots of aerodynamic measures for both IBIF calibration methods: 1) from the rainbow passage paragraph, and 2) from vowels. As a reference aerodynamic measures derived from oral airflow inverse filtering (from OVV) is pictured as well. The data was from subject **Normal 01**. For diamond: sample mean, box: interquartile (25% & 75%), red-line: median, and outer limits: quartile 5% & 95%.

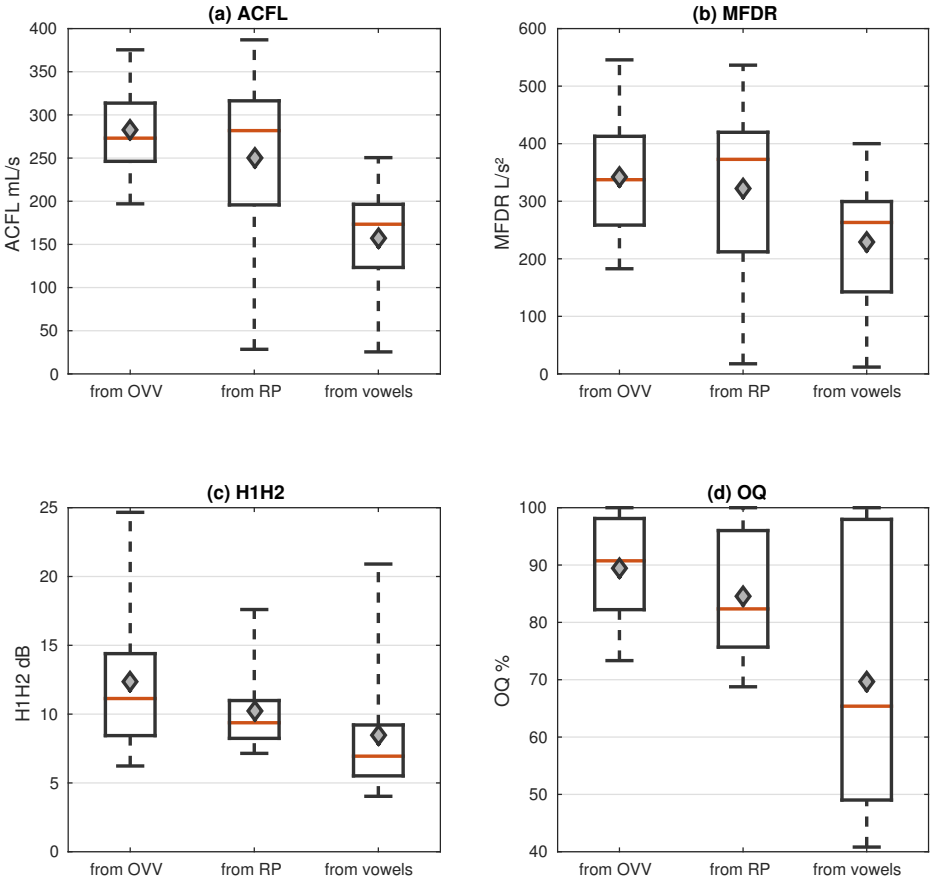


FIGURE 5.20: Boxplots of aerodynamic measures for both IBIF calibration methods: 1) from the rainbow passage paragraph, and 2) from vowel /a/. As a reference aerodynamic measures derived from oral airflow inverse filtering (from OVV) is pictured as well. The data was from subject **Phonotraumatic 01**. Boxplot description same as Figure 5.19.

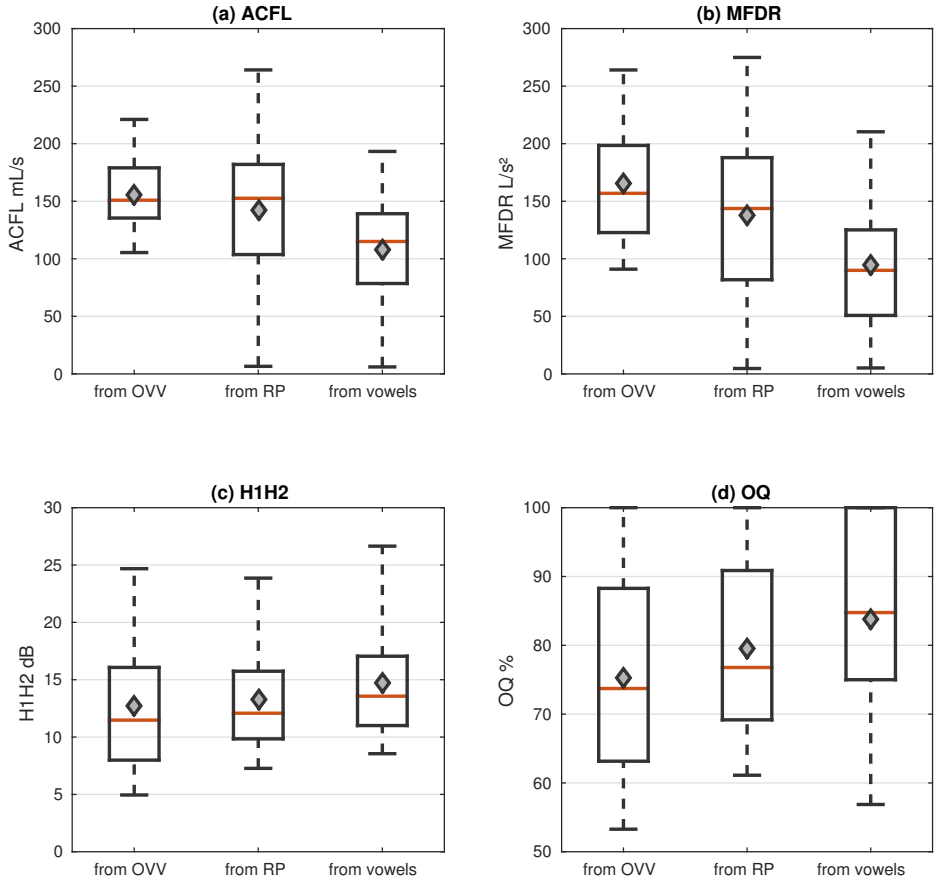


FIGURE 5.21: Boxplots of aerodynamic measures for both IBIF calibration methods: 1) from the rainbow passage paragraph, and 2) from vowel /a/. As a reference aerodynamic measures derived from oral airflow inverse filtering (from OVV) is pictured as well. The data was from subject **Normal 02**. Boxplot description same as Figure 5.19.

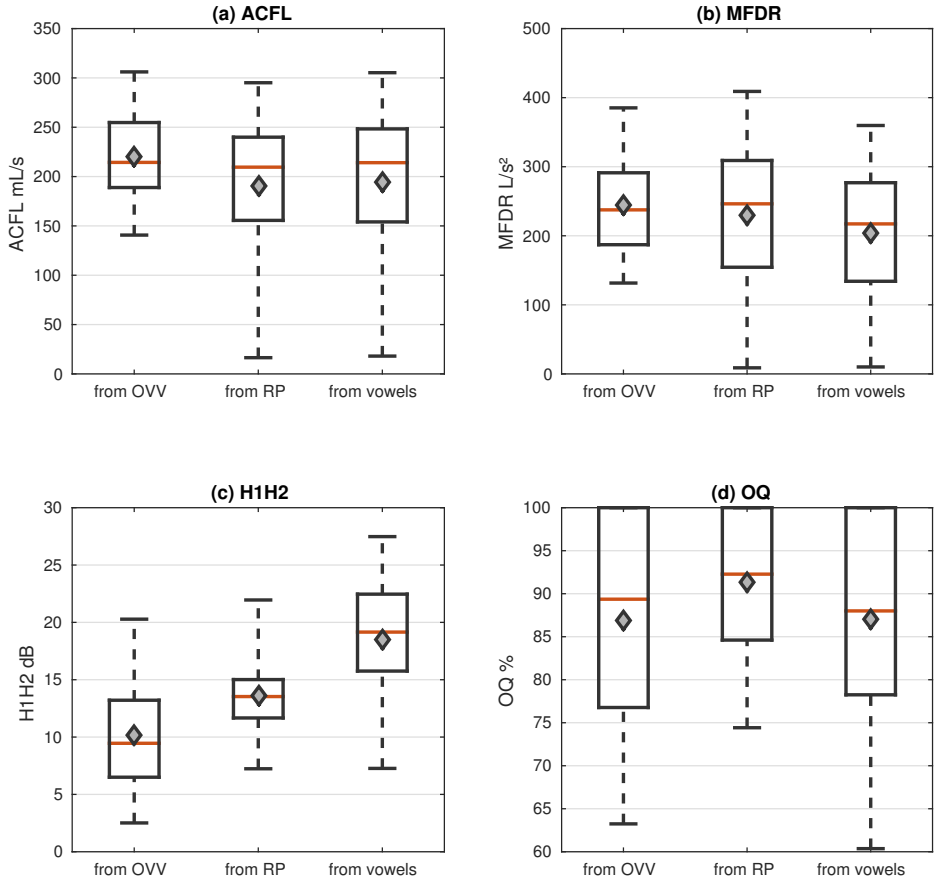


FIGURE 5.22: Boxplots of aerodynamic measures for both IBIF calibration methods: 1) from the rainbow passage paragraph, and 2) from vowel /a/. As a reference aerodynamic measures derived from oral airflow inverse filtering (from OVV) is pictured as well. The data was from subject **Non Phonotraumatic 02**. Boxplot description same as Figure 5.19.

5.4.5.1 Differences between estimated aerodynamic measures

Statistical significance of the resulting aerodynamic measures is tested using several paired Wilcoxon rank sum test [96] for each subject. Wilconox rank sum test was preferred instead of classical t-test, because we assume non-normality. Independent and identically distributed (iid) samples are achieved by pairing frames from OVV and/or ACC-based aerodynamic measures. Results for each possible test are reported in Tables 5.10, 5.11, 5.12, and 5.13. Comparing hypotheses tests between aerodynamic measures derived from Q parameters estimated from the PDF-based model (labeled as *from IBIF-RP* in the aforementioned Tables), the only instances where hypothesis H_1 is not rejected was for the H1H2 parameter in subject Normal 01 (Figure 5.19 panel (c)). The other test combinations always reject at least two test hypotheses, indicating that the resulting aerodynamic measures do not come from identical populations among methods to determine Q parameters, according to Wilconox test approach.

TABLE 5.10: Hypotheses test for subject **Normal 01**. Checklist symbol indicate which dataset is used for the hypotheses test. H_{test} column indicates if test rejects the null hypothesis (1) or not (0).

Aerodynamic measure	from OVV	from IBIF (RP)	from IBIF (vowels)	H_{test}	p_{value}
ACFL	✓		✓	0	0.481
		✓	✓	1	<0.001
	✓	✓		1	<0.001
MFDR	✓		✓	1	<0.001
		✓	✓	1	<0.001
	✓	✓		1	<0.001
H1H2	✓		✓	0	0.938
		✓	✓	0	0.422
	✓	✓		0	0.574
OQ	✓		✓	1	<0.001
		✓	✓	0	0.476
	✓	✓		1	0.003

TABLE 5.11: Hypotheses test for subject **Phonotraumatic 01**, Checklist symbol indicate which dataset is used for the hypotheses test. H_{test} column indicates if test rejects the null hypothesis (1) or not (0).

Aerodynamic measure	from OVV	from IBIF (RP)	from IBIF (vowels)	H_{test}	p_{value}
ACFL	✓		✓	1	<0.001
		✓	✓	1	<0.001
	✓	✓		0	0.205
MFDR	✓		✓	0	0.795
		✓	✓	1	<0.001
	✓	✓		1	<0.001
H1H2	✓		✓	1	<0.001
		✓	✓	1	<0.001
	✓	✓		1	<0.001
OQ	✓		✓	1	<0.001
		✓	✓	1	<0.001
	✓	✓		1	<0.001

TABLE 5.12: Hypotheses test for subject **Normal 02**. Checklist symbol indicate which dataset is used for the hypotheses test. H_{test} column indicates if test rejects the null hypothesis (1) or not (0).

Aerodynamic measure	from OVV	from IBIF (RP)	from IBIF (vowels)	H_{test}	p_{value}
ACFL	✓		✓	1	<0.001
		✓	✓	1	<0.001
	✓	✓		0	0.179
MFDR	✓		✓	1	0.001
		✓	✓	1	<0.001
	✓	✓		1	<0.001
H1H2	✓		✓	0	0.081
		✓	✓	1	0.002
	✓	✓		1	<0.001
OQ	✓		✓	1	0.003
		✓	✓	1	<0.001
	✓	✓		1	<0.001

TABLE 5.13: Hypotheses test for subject **Non Phonotraumatic 02**. Checklist symbol indicate which dataset is used for the hypotheses test. H_{test} column indicates if test rejects the null hypothesis (1) or not (0).

Aerodynamic measure	from OVV	from IBIF (RP)	from IBIF (vowels)	H_{test}	p_{value}
ACFL	✓		✓	1	<0.001
		✓	✓	1	<0.001
	✓	✓		0	0.179
MFDR	✓		✓	1	0.001
		✓	✓	1	<0.001
	✓	✓		1	<0.001
H1H2	✓		✓	0	0.081
		✓	✓	1	<0.001
	✓	✓		1	<0.001
OQ	✓		✓	1	0.01
		✓	✓	1	0.02
	✓	✓		0	0.697

5.5 Discussion and Conclusions

In this chapter, accelerometer-based aerodynamic measures in subjects with hyperfunctional voices with matched-controls are reported and analyzed. Based on the same methods used in chapter 4, ACC-derived aerodynamic measures were computed and statistically tested. The results shows that only ACFL' and OQ' have a larger effect sizes for the PVH group in the loud loudness condition, and yield values similar to those reported from Table 4.6 for the CV mask. However, the nonexistence of the SGP aerodynamic measures from the acceleration signal and the initial approximation of the IBIF-Q parameters could have a repercussion in the multivariate analysis, and a direct comparison between the discrimination power from the CV mask and the ACC in this regard could be misleading.

The uncertainty of the subglottal inverse filtering methods was also evaluated in this chapter. In this regard, a sensitivity analysis for the IBIF models parameters Q1, Q2, and Q3 was performed. The results show that most of the variability in frequency and group-delay response is below 200 Hz, a range where the reference voice signal to calibrate the IBIF model is not present, thus leading to some difficulties in the Q parameters estimation. Under the previous observation, an alternative method to estimate Q parameters was proposed. The method consists of a frame-based Q parameter estimation, a cleaning (data) procedure, a statistical model approximation, and multiple Monte Carlo simulations to trace uncertainties from the Q set to the aerodynamic measures. Using the Single Notch Filter to inverse filter the OVV signal yielded good results in section 3.1. However, signals with two formants (e.g., vowel /o/ and /u/) performed worse and vocal tract effects were only partially filtered. Conducting a "cleaning" data analysis (i.e. excluding unreliable data before the statistical analysis), allows for reducing the influence of inverse filtering inaccuracies, the number of outliers and their influence on the parameter estimation, and hopefully increasing the precision of the Q parameter estimation. The statistical analysis exhibits that Q1 and Q2 are most stable among frames, but Q3 exhibits a significant variance that could trigger phase distortion (group-delay issues), as shown in the sensitivity analysis of Qs in section 5.2. Results show that the probabilistic-model approach follows an either exponential-family distributions with a well-defined central tendency and dispersion. This allows to estimate confidence intervals from both glottal waveforms and aerodynamic measures using random realizations based on Monte Carlo methods. The resulting

uncertainties of the glottal waveform appear to be larger in the closing phase of the glottal pulse. This finding showed a negative impact in measures like MFDR and SQ, which strongly depend on the energy of higher harmonics and the magnitude of time-derivative of the glottal pulse (see a complementary MFDR harmonic analysis in appendix B). ACFL, H1H2, and OQ are robust with this kind of uncertainties. Regardless the variability in the closing phase, the estimated mean glottal waveform resembles the expected glottal shape, when the ensemble of several Q sets are used to obtain a mean set candidate for a given subject. Thus, for the first time, statistically derived bounds of glottal waveforms and aerodynamic measures are reported in the context of a clinical voice assessment. This new information regarding the estimated uncertainties could be used as *a priori* evidence for more sophisticated analysis, e.g. sequential-data approaches such as the Kalman-filter [69, 137, 138]. When comparing the aerodynamic measures derived from the proposed frame-based approach for IBIF-Q estimation (i.e., based on OVV signal vs. IBIF from rainbow passage and vs. IBIF from vowel /a/), statistically significant differences were evidenced. This suggests that a single realization of a sustained vowel, the current approach utilized for the IBIF scheme, can be considered as a biased estimated the skin+subglottal transfer function of IBIF. The ensemble of several frame-based Q parameter estimates using a statistical framework for a variety phonation gestures (with loudness, pitch, and main formant variations) provides a more robust subglottal inverse filtering mechanism. Vocal gestures with smooth and wider changes in pitch, could be a favorable scenario to estimate Q parameters, using the statistical framework introduced here. Glide-vowels (e.g., /a/ to /e/ to /i/, among others) are also good candidates for that purpose.

Chapter 6

Discussion and Conclusions

This thesis work aimed to advance the aerodynamic assessment of vocal function in static and dynamic scenarios to differentiate normal from pathological voices using improved inverse filtering methods and statistical tools. A critical review of the methods including a theoretical framework for vocal hyperfunction, clinical methods to voice assessment, and signal processing tools to estimate glottal airflow and aerodynamic measures, was presented in chapter 2. In chapter 3, enhanced inverse filtering (IF) methods for both oral airflow and neck skin acceleration signals are introduced. The methods are explored to differentiate vocal hyperfunction in patients, and the uncertainties of supra and subglottal IF estimates. Three inverse filtering methods that in several past studies had shown potential to enhance the glottal airflow estimation and their derived aerodynamic measures are further developed. A combination of a Single Notch Filter (SNF) with several metrics (cost functions), namely SNF+Metrics approach, were evaluated to automatically detect an inverse filtering candidate under constraints of ripple minimization and flat closed phase. Minimizing these metrics was feasible to assess glottal airflow estimation in high pitch voiced with voice disorders, an also enabled to automate the inverse filtering process without the supervision of a trained user. This is especially important when a larger database is collected (e.g., ambulatory monitoring), and the time to process the data is a constraint. Simulated waveforms of oral airflow (the baseline data) based on numerical models of voice production were used to evaluate the performance of the SNF+Metrics approach, along with multiple configurations of glottal shape, higher/lower pitch, simulated jitter and shimmer, and higher/lower formants. From the estimated glottal airflow of the proposed inverse filtering methods, aerodynamic measures were also calculated and compared to those

derived from the simulated waveforms (baseline). The SNF+Metrics approach yields very good results for the simulated vowels, in higher/lower pitch and formant configurations, compared with the baseline data. This suggests that using metrics can effectively improve the inverse filtering process for vowels with one formant below 1.1 kHz (limited by the bandwidth of the CV mask). In addition to these results, a challenging case for IF was achieved for high-pitched, lower-formant voices, (e.g., vowel /i/), where most of the inverse filtering methods in the literature perform poorly. Under these results, SNF+Metrics approach was successfully used to inverse filtering signals from the experiments described in chapter 4 to differentiate subjects with vocal hyperfunction from their matched-normal, and also for a frame-based approach analysis in continuous-speech in chapter 5. However, a validation of the SNF+Metrics approach is still pending for more than one-formant below 1.1 kHz (e.g., vowels /u/ and /o/). Along with the SNF+Metrics approach, a linear prediction method referred to Closed-Phase Inverse Filtering (CPIF) [17] was explored. This method is calculated within the closed portion of the glottal waveform, between closure and opening instant, allowing for a more accurate representation of an open-closed tube (all-pole model) [88], even in the presence of incomplete glottal closure [79]. In this regard, an extension of the inverse filtering scheme based on the covariance method from Alku et al. [17] was developed including a regularization term given an *a priori* information of filter coefficients. The goal was to minimize its sensitivity to closure-instant inaccuracies [17], and its overfitting [105]. Using the same simulated waveforms of the SNF+Metrics approach, RCPIF showed an improved estimation when a tuned-regularization term was selected. Aerodynamic measures derived from RCPIF shows reduced error for ACFL, MFDR, and OQ in vowels /a/ and /e/. However, the SNF + Metrics approach overcomes the RCPIF with a reduced error in almost all values, in both waveform similarities and aerodynamic measures. Incorporating the neck skin acceleration signal into the aerodynamic assessment was explored as well. A complementary approach for the subglottal system using a blind (i.e., without an input reference), non-parametric homomorphic estimation of subglottal + neck skin properties was developed. Using several configurations of simulated glottal waveforms and voiced-frames from a continuous-speech paragraph (rainbow passage), a subglottal + neck skin frequency response was estimated. The results showed that the non-parametric approach was capable to capture the subglottal resonances. These resonances are mainly related to the tracheal length [109], and under this assumption it was possible to indirectly determine the trachea length for a given subject. A simple model (see Figure

3.18) between the first subglottal resonance and the trachea length was determined and used in chapter 5, without performing search-optimization algorithms [28], which are time-consuming when larger datasets are used.

The problem of accurately assessing hyperfunctional voice disorders in the clinical setting is still challenging. Current methods for aerodynamic assessment are built on a 30-year old technology. These methods are based on oral airflow recordings of the vocal function using a pneumatograph CV mask, inverse filtered glottal airflow measures and z-score analysis. Prior efforts have not been validated, and a critical review and update is crucial to advance into more robust approaches for aerodynamic assessment. An extensive study of these methods were addressed for phonotraumatic vocal hyperfunction (PVH) and nonphonotraumatic vocal hyperfunction (NPVH) based voice disorders in chapter 4, including an updated normative dataset of selected aerodynamic measures for normal voices, a SPL-Normalized aerodynamic measures analysis, and both multivariate statistical methods and normative analysis to differentiate well-matched control/pathological subjects. The normative set of aerodynamic measures in Table 4.8 is an updated version of several past studies (e.g., Perkell et al. 1994 [26]), including an additional set of aerodynamic measures (H1H2, SQ, OQ, CPP, and NAQ in Table 4.9) to be used as reference data. In general, similar patterns to those in Perkell et al. [26] were found, e.g., larger values of ACFL are observed when loudness conditions increase, and overall changes in mean and standard deviation values were detected. Carefully selected cohorts of patients with PVH and NPVH were used to highlight differences against their matched control. SPL-Normalized measures were better suited to compensate for loudness differences in SGP, ACFL, MFDR, and OQ, and was a key element to determine statistically significant differences for the hypotheses tests. Multivariate Hotelling's T^2 and Bonferroni-corrected hypotheses tests were evaluated to determine their applicability and power discrimination. The results from these test show that Hotelling's T^2 test outperform revisited z-scores based analysis and Bonferroni-corrected univariate test in the proposed framework. Results for the aerodynamic assessment shows significant differences and large effects sizes for a group of patients with PVH where in descending order of discrimination SGP, OQ, ACFL, and MFDR were salients. For NPVH only SGP and OQ were salient compared with their control group. These results confirm preliminary evidence that glottal aerodynamic measures could identify vocal hyperfunction.

In chapter 5, and for first time, accelerometer-based aerodynamic measures in subjects with vocal hyperfunction were reported and analyzed following same methods described in chapter 4 for OVV-based measures. In particular, the analysis of the ACC-signal was accomplished in tokens that were synchronized with the OVV-based analysis from chapter 4. The initial IBIF-Q parameters were achieved according to the recommendations in Zañartu et al. (2013) [28]. The results show that the power discrimination of the ACC-based measures was moderate compared from those derived from OVV-based measures. Nonetheless, significant differences and large effects sizes of ACFL' and OQ' for the PVH group in loud loudness condition was determined. However, a definitive conclusion is still pending, under the absence of a SGP aerodynamic measures from the acceleration signal and the initial estimate of the IBIF-Q parameters. Given these preliminary results, the uncertainty of the subglottal inverse filtering methods was subsequentially evaluated for two pairs of matched subjects including both PVH and NPVH scenarios. A sensitivity analysis and a frame-based estimation of the Q parameters were proposed to obtain firsts insights of the uncertainties for the inverse filtering process of the ACC-signal. The sensitivity analysis showed high variability for $Q_{1,2,3}$ in frequencies below 200 Hz, a range where the reference signal to calibrate the Q parameters is not well-represented, reinforce the initial observation that Q parameters estimation accuracy could be affected. In this regard, performing a Q parameter estimation in a frame-based approach shown a statistically evident central tendency and dispersion, based on the proposed probabilistic model. However, in some cases, the dispersion of Q parameters is larger, particularly for Q3, which suggest possible phase-distortion issues considering the sensitivity analysis described in section 5.2. In addition, the frame-based probabilistic-model allowed to estimate confidence intervals from both glottal waveforms and aerodynamic measures using Monte Carlo simulations. In general, the estimated glottal waveform derived from the proposed method resembles a conventional glottal shape. For the glottal waveforms uncertainties, the larger variability was detected in the closing phase, that should have a direct impact on aerodynamic measures like MFDR and SQ. Nonetheless, parameters like ACFL, H1H2, and OQ appear more immune to this kind of uncertainties for the calculated results. Traced uncertainties of the inverse filtering process on the aerodynamic measures were estimated as well. These results show that ACFL and MFDR appear to be magnitude-dependent, i.e., higher values have more dispersion under Q parameter uncertainties. However, H1H2 does not shows this dependency due to its log-transformed nature, as well as OQ, which is a time-based parameter.

Comparing OVV-based measures, with both ACC-based on a single reference signal calibration (as recommend Zañartu et al. [28]), and the proposed frame-based approach, several pair-wise statistical test were performed. The results shows statistically significant differences in almost all cases indicating that methods do not estimate the same aerodynamic measures. The proposed framework provides a comprehensive method to get uncertainty information of the inverse filtering process compared to the calibration with a single sustained vowel, and their impact over the estimated aerodynamic measures, that could be used as *a priori* information in a more sophisticated analysis, e.g. time-series analysis.

Further research can be conducted on several fronts after this thesis. Aerodynamic assessment can be improved with the incorporation of new recording methods already available, e.g., gradient-temperature based sensors (a.k.a. Microflown) [139], allowing for estimating radiated pressure and volume velocity in the same point space without using the Rothenberg mask. Inverse filtering methods have not been fully investigated under machine learning framework (e.g., deep neural networks [140]), that have yielded remarkable results in other research fields. Look into different ways to estimate subject-specific subglottal + neck skin frequency response, e.g., using an external source instead of the subject voice source could be an alternative to the current methods. Data derived from ambulatory monitoring devices could be analyzed in a time-series approach to obtain insights into the dynamics of the aerodynamic measures which currently lacks of findings in the research field.

Appendix A

Matlab GUI to perform Inverse filtering tasks

In this appendix are briefly described the functionality of a custom Matlab GUI to perform inverse filtering tasks. See Figure A.2 for a GUI explanation. At top-left are showing the raw data (blue = OVV signal, gray = ACC signal) which was loaded using the *File-Open* menu. At top-center a *Settings* panel with multiple options for standard low pass filter to be activated and selected. A checkbox to enable IBIF calculation is available as well. At bottom for the panel, a *Get parameter*s button allows obtaining aerodynamic measures for both oral airflow and neck skin acceleration estimates which are listed in the command window of Matlab (not showed). IBIF optimization (with PSO) options are available as well. At top-right, figures for all the waveforms evolved: 1) the oral waveform from the CV mask (solid gray), 2) the inverse filtered CV mask (solid blue), and 3) the inverse filtered ACC signal (solid red). At middle-left, spectral approximations for the inverse filtered signals plus a reference of the SNF. At the right of the spectrum, a phase-plane plot to assess the resulted inverse filtered oral airflow. At bottom of GUI from left to right: 1) Slide controls and popup menus to set up Q parameters, 2) delay and gain compensation of the ACC signal, 3) audio controls, 4) voice selection segments and zoom tools, and 5) slider controls for 3 SNF filters and bandwidth. Path and filename for the loaded Matlab file are visualized in the text boxes. Buttons for automatic inverse filtering are available as well. The work flow to perform inverse filtering is:

1. Load file using *File*→*Open*. The *.mat file have a specific structure that was designed by the MGH Voice Center Research Group.

2. In *Zoom-Ginput* panel you can zoom a segment of the waveforms. All waveforms axes will be zoomed accordingly. Then, push *Ginput* button in the *Zoom-Ginput* panel and pick limits points to select a segment for analysis.

3. As a segment was selected you can perform inverse filtering for the oral airflow signal automatically by push *Auto IF v1.0* button or manually adjust scrollbars accordingly. Use the spectrum and phase-plane plot for visual assessment. You can use more than one filter at the same time. All filters are active, unless you move scroll bars to the right where central frequencies are set to infinity. Bandwidth is only available for F2. F1 and F3 bandwidth are fixed to 70 Hz. Note ACC signal is not showed until you activate IBIF checkbox in panel *Settings*.

4. Using the IF glottal airflow, Q parameters can be estimated. In the *Optimization* panel *PSO IBIF* must be selected. If you want to visualize PSO trend, activate *See PSO trend* check box. Wait until ACC signal is visualized in the waveform axes. You can also manually adjust Q parameter using scroll bars.

5. Pressing *Get Parameters* button, aerodynamic measures are displayed in the command window for both signals.

An example of PSO algorithm trend is presented in Figure A.1. Note that the error fit, rapidly decay to its minimum and slightly improvement is achieved when the number of iteration grows.

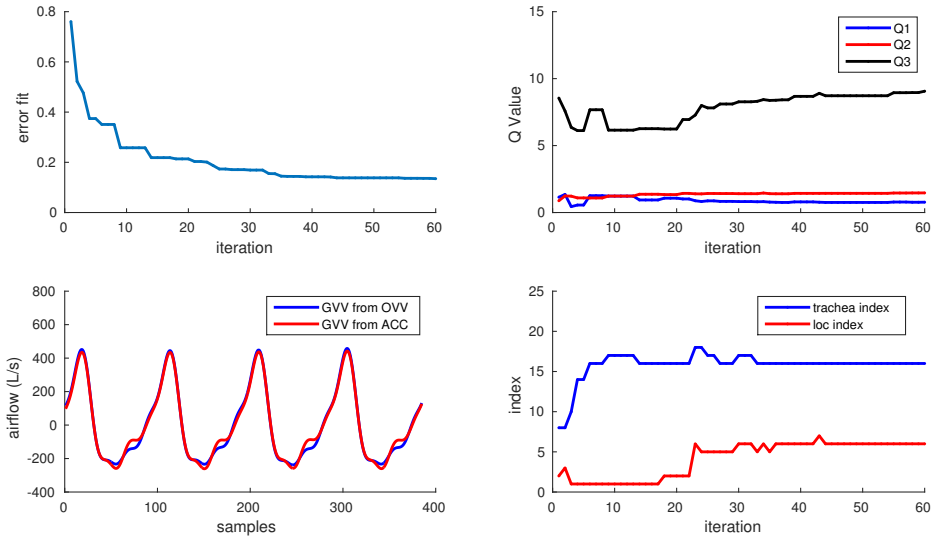


FIGURE A.1: PSO trend example using sustained vowel /a/. Top-Left: Error fit trend based on equation (3.43) along PSO iteration. Top-right: Q1, Q2, and Q3 parameter trend along PSO iteration. Bottom-left: Visual reference of waveforms. Bottom-right: Trachea and accelerometer position index vs iteration.

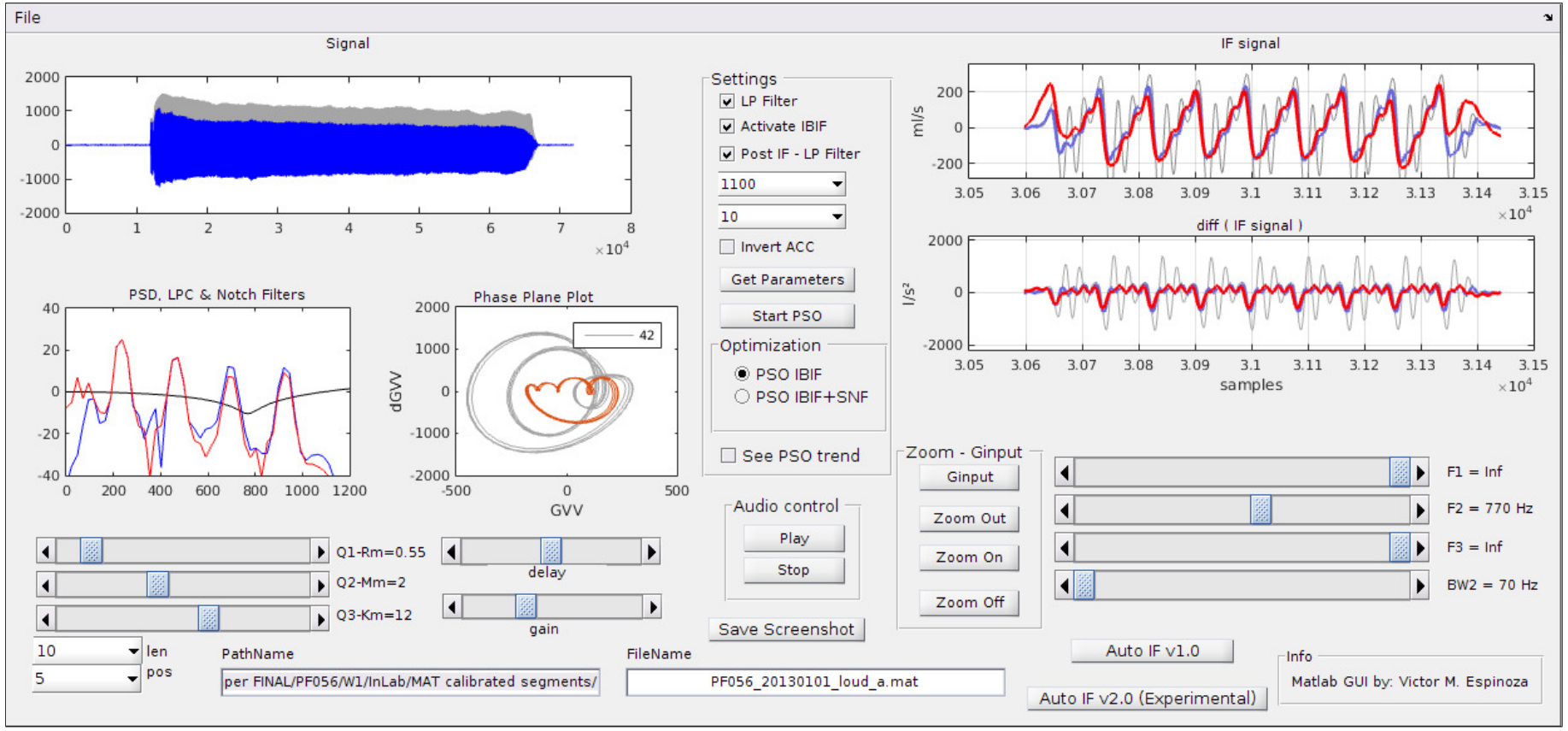


FIGURE A.2: Matlab GUI which perform inverse filtering. Tasks are fully automated but manual control is allowed.

Appendix B

Theoretical analysis of vocal measures

Multiple studies have empirically illustrated the relationship between several vocal parameters (e.g. see Holmberg et al. (1988) [25]). Based on an analytical glottal pulse model we explore theoretical relations between selected measures. The aim is to obtain insights of the relationship between parameter and their physiological and physical interpretation.

Based on an analytical glottal pulse model with an harmonically related sum of sinusoids, we get that the airflow signal $x(t)$ is given by

$$\begin{aligned} x(t) &= \sum_{i=0}^{\infty} a_i \cdot \sin(\omega_i t + \phi_i) \\ &= \sum_{i=1}^{\infty} a_0 \cdot \alpha^{i-1} \cdot \sin(i\omega_0 \cdot t + \phi_i) \end{aligned} \tag{B.1}$$

where $\omega_i = i\omega_0$ (harmonic oscillation), $a_i > 0$, $a_i = a_0 \cdot \alpha^{i-1}$, a_0 the amplitude of the first harmonic, and $0 < \alpha < 1$ as an approximated harmonic decay ratio. We empasize here that the magnitude decay model is approximated to several definitions about the spectral behavior of glottal pulse [56, 76]. In that regard, α can be expressed in dB per octave (dB/8^{ve}) units as

$$20\log_{10}(a_0 \cdot \alpha^{i-1}) = (i-1) \cdot 20\log_{10}\alpha + 20\log_{10} a_0 \tag{B.2}$$

by taking the derivative respect to $(i-1)$-th harmonic we obtain the decay rate in dB/8^{ve} as $20\log_{10}(\alpha)$, e.g., if $\alpha = 0.5$, then $20\log_{10}(0.5) = -6$ dB/8^{ve}. Note

that α fit with the definition of H1H2 as the ratio of the first to second harmonic magnitude as well.

B.1 Maximum Flow Declination Rate (MFDR) analysis

For MFDR, we are interested in the time derivative of (B.1), which is:

$$\begin{aligned}\frac{d}{dt}\{x(t)\} &= \frac{d}{dt}\left\{\sum_{i=1}^{\infty} a_0 \cdot \alpha^{i-1} \cdot \sin(i\omega_0 \cdot t + \phi_i)\right\} \\ &= a_0 \cdot \omega_0 \sum_{i=1}^{\infty} i \cdot \alpha^{i-1} \cdot \cos(i\omega_0 t + \phi_{i-1}) \end{aligned} \tag{B.3}$$

Calculating the superior bound of equation (B.3),

$$\left|\sum_{i=1}^{\infty} i \cdot \alpha^{i-1} \cdot \cos(i\omega_0 t + \phi_{i-1})\right| \overset{(a)}{\leq} \sum_{i=1}^{\infty} \left|\underbrace{i \cdot \alpha^{i-1}}_{\geq 0} \cdot \cos(i\omega_0 t + \phi_{i-1})\right| \tag{B.4}$$

$$\overset{(b)}{\leq} \sum_{i=1}^{\infty} i \cdot \alpha^{i-1} \cdot \underbrace{|\cos(i\omega_0 t + \phi_{i-1})|}_{\leq 1} \overset{(c)}{\leq} \sum_{i=1}^{\infty} i \cdot \alpha^{i-1} \quad , \tag{B.5}$$

wherein in equation (B.4) (a) is for the triangle inequality $|A + B| \leq |A| + |B|$, (b) for $i \cdot \alpha^{i-1} > 0$, and (c) for $|\cos(i\omega_0 t + \phi_{i-1})| \leq 1$. Then, MFDR is (initially) bounded by

$$0 < MFDR \leq a_0 \cdot \omega_0 \sum_{i=1}^{\infty} i \cdot \alpha^{i-1} \tag{B.6}$$

From (B.6), three cases are considerer:

- **Case 1**: If we consider $0 < \alpha < 1$, the summation in (B.6) converge to $\frac{1}{(1-\alpha)^2}$. Thus,

$$MFDR \leq a_0 \cdot \omega_0 \cdot \frac{1}{(1-\alpha)^2} \tag{B.7}$$

  Equation (B.7) indicates that the MFDR superior bound depends on f_0, first harmonic amplitude (a_0), and the decay ratio of harmonics amplitude (α).

- **Case 2**: When considering a finite number of harmonics N (i.e., bandlimited signal) and $0 < \alpha < 1$, the *maximum* value for MFDR is:

$$\begin{aligned} MFDR &\leq a_0 \cdot \omega_0 \sum_{i=1}^{N} i\alpha^{i-1} \\ &\leq a_0 \cdot \omega_0 \cdot \frac{1}{\alpha} \sum_{i=1}^{N} i\alpha^{i} \\ &\leq a_0 \cdot \omega_0 \cdot \frac{1-(N+1)\alpha^N + N\alpha^{N+1}}{(1-\alpha)^2} \end{aligned} \tag{B.8}$$

  Another way to write equation (B.7) is replacing

$$N = I\left(\frac{f_{max}}{f_0}\right) , \tag{B.9}$$

  where $I(\cdot)$ indicate the integer part of the argument. So, equation (B.8) is a function of f_{max} (the upper limit of the pass-band) as well. The later result is relevant as Alku et al. [24] shows empirically how MFDR depends on bandwidth, which is a similar result compared to our theoretical approach.

- **Case 3**: If $\alpha \to 0$ then (B.7) become:

$$MFDR \geq a_0 \cdot \omega_0 \tag{B.10}$$

  which is the *minimum* value for MFDR. Note that equation (B.10) is the MFDR from time-derivative of a sine signal with amplitude a_0 and frequency ω_0.

Now, updating the bounds of MFDR using (B.10) and (B.7), we obtain

$$a_0 \cdot \omega_0 \leq MFDR \leq a_0 \cdot \omega_0 \frac{1}{(1-\alpha)^2} \tag{B.11}$$

If $a_0 \approx ACFL$ (i.e., waveform have a shape that is *more* sinusoidal than glottal), then:

$$MFDR \approx ACFL \cdot \omega_0 \tag{B.12}$$

The equation (B.12) establish a relationship between fundamental frequency, ACFL and MFDR when $\alpha \to 0$. This scenario is relevant because subjects with incomplete glottal closure empirically shows higher spectral tilt, i.e., small α that could be related to a breathy voice quality.

From above analysis, MFDR linearly depends on the amplitude of first harmonic and the fundamental frequency, in both upper and lower limits. The decay ratio of harmonics α limit the upper limit of the MFDR value too. This α value has an approximated relationship with H1H2 as the difference of the first to second harmonic as $\alpha \approx 10^{(\mathrm{H1H2}/20)}$. The bandwidth of the glottal model (i.e., the number of harmonics evolved) has a direct impact on MFDR value, which is in agreement with the empirically derived analysis in [24].

B.2 Peak-to-peak amplitude (ACFL) analysis

Using a glottal pulse model with sum of sinusoids (Fourier series approach), we gets

$$x(t) = \sum_{i=0}^{\infty} a_i \sin(\omega_i t + \phi_i) \tag{B.13}$$

$$= a_0 \sum_{i=1}^{\infty} \alpha^{i-1} \sin(i\omega_0 t + \phi_{i-1}) \tag{B.14}$$

Taking the absolute value and following same steps than in (B.5), we obtain,

$$\left| \sum_{i=1}^{\infty} \alpha^i \sin(i\omega_0 t + \phi_i) \right| \overset{(a)}{\leq} \sum_{i=1}^{\infty} \left| \alpha^{i-1} \sin(i\omega_0 t + \phi_{i-1}) \right| \tag{B.15}$$

$$\overset{(b)}{\leq} \sum_{i=1}^{\infty} \alpha^{i-1} \underbrace{\left|\sin(i\omega_0 t + \phi_{i-1})\right|}_{\leq 1} \overset{(c)}{\leq} \sum_{i=1}^{\infty} \alpha^{i-1} = \sum_{i=0}^{\infty} \alpha^i \quad , \tag{B.16}$$

wherein in equation (B.15) (a) is for the triangle inequality $|A+B| \leq |A| + |B|$, (b) for $\alpha^{i-1} > 0$, and (c) for $|\sin(i\omega_0 t + \phi_{i-1})| \leq 1$. Then,

$$ACFL \leq \sum_{i=0}^{\infty} \alpha^i \tag{B.17}$$

We condsider three cases derived from (B.17):

- **Case 01**: If $0 < \alpha < 1$, the summation in (B.17) converge to $\frac{1}{1-\alpha}$. Replacing on it we gets,

$$ACFL \leq \frac{a_0}{1-\alpha} \tag{B.18}$$

- **Case 02**: If $0 < \alpha < 1$, and the sum is over N harmonics, then

$$ACFL \leq a_0 \frac{1 - \alpha^{N+1}}{1 - \alpha} \tag{B.19}$$

- **Case 03**: if $\alpha \to 0$ then $ACFL \to a_0$.

Then, combining cases, ACFL is bounded by

$$a_0 \leq ACFL \leq a_0 \frac{1}{(1 - \alpha)} \tag{B.20}$$

In summary, Peak-to-peak amplitude (ACFL) linearly depends on the amplitude of the fundamental frequency in both upper and lower limits, and inversely to $(1 - \alpha)$. In the case of limited number of harmonics (or equivalently a limited bandwidth), presented in case 02 of ACFL analysis, upper bound is affected in a similar way that MFDR upper bound described in the previous section.

The two analysis (ACFL and MFDR) show an evident dependency through a_0 value. This suggests a possible way to normalize these bounds and make it level calibration free. Then setting $ACFL_N = \frac{ACFL}{a_0}$, a normalized ACFL bound is,

$$1 < ACFL_N \leq \frac{1}{1 - \alpha} \ , \tag{B.21}$$

and for $MFDR_N = \frac{MFDR}{a_0}$, a normalized MFDR bound yields,

$$\omega_0 \leq MFDR_N \leq \omega_0 \frac{1}{(1 - \alpha)^2} \tag{B.22}$$

Appendix C

Statistical tools

In objective voice assessment, several mathematical and statistical tools are used to validate vocal behavior in both normal and pathological voices [3, 14, 95, 95]. Most of these tools are based on the linear model [107]. A brief review of the commonly used methods is presented in the following [96, 121, 122, 141].

C.1 Hypotheses test

C.1.1 The Bonferroni Method

The Bonferroni method is a multiple hypotheses test that uses simply univariate ones (e.g. Student's t-test, Welch test, among others [126]), instead of multivariate ones (e.g., Hotelling T^2, see next subsection to further details). This method is used as a first attempt to perform a multivariate differentiation between pathological and normal voices in the followings chapters, as in other studies focused in univariate statistical test [3, 25]

Suppose that, multiple simultaneous test C_i confidence statement from a linear combination of $\mathbf{a}_i\mu$ are required. Then

$$\begin{aligned} P[\text{all } C_i \text{ true}] &= 1 - P[\text{at least one } C_i \text{ false}] \\ &\geq 1 - \sum_{i=1}^{m} P(C_i \text{ false}) = 1 - \sum_{i=1}^{m}(1 - P(C_i \text{ true})) \\ &\geq 1 - (\alpha_1 + \ldots + \alpha_m) \\ &\geq 1 - \sum_{i=1}^{m} \alpha_i \end{aligned}$$

where P is the probability of a given event. If $\sum_{i=1}^{m} \alpha_i \leq \alpha_0$ (the overall rate error), and assuming each $\alpha_i = \alpha$, then $\alpha_0 \leq \alpha/m$, which is knowing as the Bonferroni correction [122].

Some highlighted features of the Bonferroni method are 1) do not take into account the correlation structure behind the confidence statement, 2) It is known to be conservative, i.e., is prone to no reject the null hypotheses, and 3) have a good control of Type II error (the overall rate error).

In conclusion, Bonferroni method allows us to perform multiple comparisons using univariate hypotheses test by scaling individuals significance levels α by $1/m$ [122].

C.1.2 Hotelling's T^2

Hotelling's T^2 is the multivariate version of univariate hypothesis test [121, 122]. The advantage to use this multivariate test is allows to perform the hypotheses test in a multidimensional space of parameters, taking to account the linear relationships between them [122]. The mathematical form of the test is described as follows. Assuming data is normally distributed and random variables $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent and identically distributed (iid) with mean vector $\boldsymbol{\mu} = (\mu_1, ..., \mu_p)^T$ and $\mathbf{X}_i \in \mathbb{R}^p$, the general hypothesis test is:

$$H_n : \boldsymbol{\mu} = \boldsymbol{\mu_0} \text{ vs. } H_a : \boldsymbol{\mu} \neq \boldsymbol{\mu_0}. \tag{C.1}$$

where $\boldsymbol{\mu_0} = (\mu_{10}, ..., \mu_{p0})^T$ is a means vector.

For Hotelling's T^2 statistic, this is equivalent to

$$T^2 = n\big(\bar{\mathbf{X}} - \boldsymbol{\mu_0}\big)^T \mathbf{S}^{-1} \big(\bar{\mathbf{X}} - \boldsymbol{\mu_0}\big) \geq \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha), \tag{C.2}$$

with

$$\underset{(p\times p)}{\mathbf{S}} = \frac{1}{n-1}\sum_{j=1}^{n}\left(\mathbf{X}_j - \bar{\mathbf{X}}\right)\left(\mathbf{X}_j - \bar{\mathbf{X}}\right)^T, \tag{C.3}$$

$$\underset{(p\times 1)}{\bar{\mathbf{X}}} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{X}_j \tag{C.4}$$

wherein T^2 is the statistics proposed for Hotelling [142], $\mathbf{S}$ the covariance matrix, $F_{p,n-p}(\alpha)$ denotes a random variable with an F-distribution with p and $n-p$ degree of freedom, at its upper (100α)th percentile.

C.1.3 Effect sizes

Acording to [129], the Cohen's d distance for effect sizes is the mean difference that would obtain if the dependent variable were scaled to have unit variance within the groups of experiments. Mathematically is

$$d = \frac{\mu_e - \mu_c}{\sigma_p} \tag{C.5}$$

where μ_e is the mean true value for the experimental group and μ_e for the control group, and σ_p^2 the common variance, assuming observations are independent and normally distributed within groups of experiments. Approximated values for μ_e, μ_c , and σ_p^2 for the corresponding sample means and the pooled-variance which is given

$$s = \sqrt{\frac{(n_e-1)s_e^2 + (n_c-1)s_c^2}{n_e + n_c - 2}} \tag{C.6}$$

where s_e and s_e are the standard deviation of the experimental and control group based on n_c and n_e sample size, respectively. Values of $|d| > 0.6$ indicate large effect sizes between the groups. For a multivariate version of effect sizes, the Mahalanobis distance D [121, 122] is used and given in the equation

$$D = \sqrt{\left(\bar{\mathbf{X}}_\mathbf{e} - \bar{\mathbf{X}}_\mathbf{c}\right)^T \mathbf{S_p}^{-1}\left(\bar{\mathbf{X}}_\mathbf{e} - \bar{\mathbf{X}}_\mathbf{c}\right)}. \tag{C.7}$$

where $\bar{\mathbf{X}}_{\mathbf{e,c}}$ are the mean vectors under evaluation (experiment and control) and $\mathbf{S}_p$ the sample (pooled) covariance matrix , given by,

$$\mathbf{S_p} = \frac{1}{n_e + n_c}\left\{(n_e-1)\mathbf{S_e} + (n_c-1)\mathbf{S_c}\right\} \tag{C.8}$$

where $\mathbf{S_e}$ and $\mathbf{S_c}$ are the sample covariance of the experimental and control dataset given in equation (C.3).

C.2 Model estimation

Probability distributions is a key mathematical tool to analyze several voice datasets in this work. Parametric and nonparametric estimates are used.

C.2.1 Kernel Density Estimation (KDE)

Non-parametric estimation of a probability density function (pdf) is briefly described in this section. An usual way to obtain approximated pdf's is averaging normalized kernels functions as

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{DE}\left(\frac{x - X_i}{h}\right) \quad , \tag{C.9}$$

where K_{DE} is a given kernel, n the amount of sample data, and h the bandwidth of the kernel [96]. A rule-of-thumb bandwidth estimator h_{opt} for a Gaussian Kernel [131] is defined as follows. Let $X_1, \ldots X_n$ be iid random variables, with density $f(x)$, the optimal bandwidth h_{opt} for KDE using a Gaussian kernel is [131]

$$h_{opt} = \frac{0.9\hat{\sigma}}{n^{1/5}} \tag{C.10}$$

where

$$\hat{\sigma} = \min(s, 1.34 \cdot IQR) \tag{C.11}$$

$$s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{C.12}$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{C.13}$$

and $IQR = Q_3 - Q_1$ is the interquartile area of the data, a measure of statistical dispersion [96].

C.2.2 Maximum Likelihood Estimation

The Maximum Likelihood Estimation (MLE) is the most common method for estimating parameters given a parametric statistical model [96, 107, 141]. Let $X_1, \ldots, X_n$ be iid with pdf $f(x;\theta)$, then the *likelihood function* is defined by

$$\mathcal{L}_n(\theta) = \prod_{i=1}^{n} f(X_i;\theta) \tag{C.14}$$

Maximizing (C.14) with respect to θ the MLE is defined as,

$$\hat{\theta}_n = \arg\max_{\theta} \mathcal{L}_n(\theta) \tag{C.15}$$

Having an enough large number of observations n, the maximum likelihood estimator $\hat{\theta}_n$ has the followings central properties [96, 107].

- The MLE is consistent (asymptotically unbiased): $\hat{\theta}_n \xrightarrow{P} \theta$ denotes the true value of the parameter θ;
- The MLE is equivariant (Invariance property), i.e., if $\hat{\theta}_n$ is the MLE of θ then $g(\hat{\theta}_n)$ is the MLE of $g(\theta_n)$.
- The MLE is asymptotically Normal: $(\hat{\theta}_n - \theta)/\hat{\sigma} \rightarrow \mathcal{N}(0,1)$.

From a practical point of view, an equivalent manner to find the MLE is maximizing the log-likelihood function which is defined as,

$$l_n(\theta) = \log \mathcal{L}_n(\theta) \tag{C.16}$$

This *transformed* function leads to the same outcome as the likelihood, because the log function is always an increasing function.

C.2.3 Robust Estimates

Robust methods try to improve the performance of a given model by minimizing the influence of sporadic or unreliable data point in a dataset [124]. We expect to improve the performance of our models using robust linear regression along with a combination of outliers detection utilizing the Cook distance [125], Leverage criteria and three-sigma edit rule [124] to exclude (or weight) data from the regression analysis.

Classical robust methods are based on the Maximum Likelihood method and a modified loss function criteria [124]. These methods are called M-estimators [124]. For given $\rho-$function (which is the modified loss criterion) equation (C.15) is rewritten as

$$\hat{\theta}_n = \arg\max_{\theta} \sum_{i=1}^{n} \rho(x_i - \theta) \tag{C.17}$$

where

$$\rho = \log f(x;\theta) \tag{C.18}$$

An important family of of these $\rho-$functions are the *Huber functions* [105, 124, 143]

$$\rho_k(x_i,\theta) = \begin{cases} (x_i-\theta)^2 & ; \quad |x_i-\theta| \leq k \\ 2k|x_i-\theta| - k^2 & ; \quad |x_i-\theta| > k \end{cases} \tag{C.19}$$

Note error lost function $\rho_k(x_i,\theta)$ is quadratic and l_1-norm based, so large errors are weighted differently than small errors for the MLE estimate.

C.2.4 Bootstrap

The bootstrap is a method to estimate standard errors and confidence intervals for a given statistic [105, 127]. It belongs to the collection of non-parametric methods in statistics and is based on resampling (with replacement) the original data set of measurements. The procedure is described as follows. Let $\alpha = g(z_1,\ldots,z_n)$ be any function of the data $Z = [z_1,\ldots,z_n]$. From randomly selected n observations from the original data set Z, a bootstrap data set Z^{*1} is produced. From this new data set, a bootstrap estimate for $\hat{\alpha}$, is calculated and called $\hat{\alpha}^{*1}$. This algorithm is repeated B times for each new bootstrap data set $Z^*,\ldots,Z^{*B}$, and B corresponding estimates of $\hat{\alpha}^{*1},\ldots,\hat{\alpha}^{*B}$. Finally, the standard error of these bootstrap estimates are calculated using the formula [96, 105, 127, 144] ,

$$SE_B(\hat{\alpha}) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\alpha}^{*b} - \frac{1}{B}\sum_{r=1}^{B}\hat{\alpha}^{*r}\right)^2}. \tag{C.20}$$

Bibliography

[1] 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.

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

[3] 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. *J Speech Hear. Res.*, 32:373–392, 1989.

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

[5] 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.

[6] M. Zañartu, G. E. 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(6):3262–3271, 2014.

[7] D. D. Mehta and R. E. Hillman. Voice assessment: Updates on perceptual, acoustic, aerodynamic, and endoscopic imaging methods. *Curr. Opin. Otolaryngol. Head Neck Surg.*, 16:211–215, 2008.

[8] J. S. Stemple, Roy N., and B. K. Klaben. *Clinical Voice Pathology: Theory and Management*. Plural Publishing Inc., fifth edition edition, 2014.

[9] M. M. Johns. Update on the etiology, diagnosis, and treatment of vocal fold nodules, polyps, and cysts. *Otolarynol. Head Neck Surg.*, 11:456–461, 2003.

[10] P. Zhuang, A. J. Sprecher, M. R. Hoffman, Y. Zhang, M. Fourakis, J. J. Jiang, and C. S. Wei. Phonation threshold flow measurements in normal and pathological phonation. *Laryngoscope*, 119:811–815, 2009.

[11] R. E. Hillman, W. W. Montgomery, and S. M. Zeitels. Current diagnostics and office practice: Appropriate use of objective measures of vocal function in the multidisciplinary management of voice disorders. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 1997.

[12] 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.

[13] 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.

[14] E. B. Holmberg, R. E. Hillman, J. S. Perkell, P. C. Guiod, and S. L. Goldman. Comparisons among aerodynamic, electroglottographic, and acoustic spectral measures of female voice. *Journal of Speech and Hearing Research*, 38(6):12121223, 1995.

[15] A. I. Gillespie, J. Gartner-Schmidt, E. N. Rubinstein, and K. V. Abbott. Aerodynamic profiles of women with muscle tension dysphonia/aphonia. *Journal of Speech, Language, and Hearing Research*, 56(2):481–488, 4 2013.

[16] 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.

[17] P. Alku, C. Magi, S. Yrttiaho, T. Bäckström, and B. Story. Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. *Journal of the Acoustical Society of America*, 125(5):3289–3305, 2009.

[18] 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.

[19] M. Rothenberg and S. Zahorian. Nonlinear inverse filtering technique for estimating the glottal-area waveform. *The Journal of the Acoustical Society of America*, 61(4):1063–1070, 1977.

[20] D. G. Childers, J. C. Principe, and Y. T. Ting. Adaptive wrls-vff for speech analysis. *IEEE Transactions on Speech and Audio Processing*, 3(3):209–213, May 1995.

[21] Q. Fu and P. Murphy. Robust glottal source estimation based on joint source-filter model optimization. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(2):492–501, March 2006.

[22] Paavo Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11(2–3):109–117, 1992.

[23] T. Drugman, B. Bozkurt, and T. Dutoit. Causal-anticausal decomposition of speech using complex cepstrum for glottal source estimation. *Speech Communication*, 2011.

[24] P. Alku and E. Vilkman. Effects of bandwidth on glottal airflow waveforms estimated by inverse filtering. *The Journal of the Acoustical Society of America*, 98(2):763–767, 1995.

[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. *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] H. A. Cheyne II. *Estimating glottal voicing source characteristics by measuring and modeling the acceleration of the skin on the neck*. PhD thesis, Harvard-MIT Division of Health Sciences and Technology Speech and Hearing Biosciences and Technology Program, 2002.

[28] 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.

[29] Titze I. R. Švec, J. G. and P. S. Popolo. Vocal dosimetry: Theoretical and practical issues. In T. Wittenberg G. Schade, F. Muller and M. Hess, editors, *AQL 2003 Hamburg: Proceeding Papers for the Conference Advances in Quantitative Laryngology, Voice and Speech Research*, 2003.

[30] 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.

[31] R. E. Hillman and D. D. Mehta. Ambulatory monitoring of daily voice use. *SIG 3 Perspectives on Voice and Voice Disorders*, 21(2):56–61, 2011.

[32] A. Llico, M. Zañartu, A. Gonz ález, G. R. Wodicka, D. D. Mehta, J. H. Van Stan, and R. E. Hillman. Real-time estimation of aerodynamic features for ambulatory voice biofeedback. *JASA Express Letters*, 138(1):EL14–EL19, 2015.

[33] 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.

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

[35] H. A. Cheyne II. Estimating glottal voicing source characteristics by measuring and modeling the acceleration of the skin on the neck. In *Medical Devices and Biosensors, 2006. 3rd IEEE/EMBS International Summer School on*, pages 118–121, Sept 2006.

[36] A. V. Oppenheim. *Superposition in a class of nonlinear systems*. PhD thesis, Massachusetts Institute of Technology (MIT), 1965.

[37] A. V. Oppenheim. Speech analysis-synthesis system based on homomorphic filtering. *The Journal of the Acoustical Society of America*, 45(1):309–309, 1969.

[38] A. V. Oppenheim and R. W. Schafer. Homomorphic analysis of speech. *Audio and Electroacoustics, IEEE Transactions on*, 16(2):221–226, Jun 1968.

[39] R. Fraile and J. I. Godino-Llorente. Cepstral peak prominence: A comprehensive analysis. *Biomedical Signal Processing and Control*, 14:42 – 54, 2014.

[40] D. D. Mehta, J. H. Van Stan, M. Zañartu, M. Ghassemi, J. V. Guttag, V. M. Espinoza, J. P. Cortés, H. A. Cheyne, and R. E. Hillman. Using ambulatory voice monitoring to investigate common voice disorders: Research update. *Frontiers in Bioengineering and Biotechnology*, 3(155), 2015.

[41] N. Bhattacharyya. The prevalence of voice problems among adults in the united states. *The Laryngoscope*, 124(10):23592362, 2014.

[42] 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(3):269 – 282, 2003.

[43] C. M. Sapienza and E. T. Stathopoulos. Respiratory and laryngeal measures of children and women with bilateral vocal fold nodules. *Journal of Speech and Hearing Research*, 37(6):12291243, 1994.

[44] R. E. Hillman, E. B. Holmberg, J. S. Perkell, M. Walsh, and C. Vaughan. Phonatory function associated with hyperfunctionally related vocal fold lesions. *Journal of Voice*, 4:52–63, 1990.

[45] D. E. Veeneman and S. L. BeMent. Automatic glottal inverse filtering from speech and electroglottographic signals. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 33(2):369–377, Apr 1985.

[46] J. G. Švec and H. S. Schutte. Videokymography: High-speed line scanning of vocal fold vibration. *Journal of Voice*, 10:201–205, 1996.

[47] 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(4):919–932, 2003.

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

[49] A. Tsanas, M. Za nartu, M. A. Little, C. Fox, L. O. Ramig, and G. D. Clifford. Robust fundamental frequency estimation in sustained vowels: Detailed algorithmic comparisons and information fusion with adaptive kalman filtering. *The Journal of the Acoustical Society of America*, 135(5):2885–2901, 2014.

[50] 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.

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

[52] 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, 2016.

[53] V. S. McKenna, A. F. Llico, D. D. Mehta, J. S. Perkell, and C. E. Stepp. Magnitude of neck-surface vibration as an estimate of subglottal pressure during modulations of vocal effort and intensity in healthy speakers. *Journal of Speech, Language, and Hearing Research*, 60(12):3404–3416, 2017.

[54] L. Beranek and T. Mellow. *Acoustics: Sound Fields and Transducers*. Elsevier (Oxford), 1st edition edition, 2012.

[55] J. Gudnason, D. D. Mehta, and T. F. Quatieri (2014). Closed phase estimation for inverse filtering the oral airflow waveform. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014.

[56] J. L. Flanagan. *Speech Analysis Synthesis and Perception*. Springer-Verlag, 1972.

[57] I. R. Titze. *The Myolelastic Aerodynamic Theory of Phonation*. National Center for Voice and Speech, 2006.

[58] I. R. Titze. Nonlinear source-filter coupling in phonation: Theory. *Journal of the Acoustical Society of America*, 123:2733–2749, 2008.

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

[60] R. L. Miller. Nature of the vocal cord wave. *The Journal of the Acoustical Society of America*, 31(6):667–677, 1959.

[61] M. Nakatsui and J. Suzuki. Method of observation of glottal-source wave using digital inverse filtering in time domain. *The Journal of the Acoustical Society of America*, 47(2B):664–665, 1970.

[62] H. W. Strube. Determination of the instant of glottal closure from the speech wave. *The Journal of the Acoustical Society of America*, 56(5):1625–1629, 1974.

[63] Y. Shapira and I. Gath. A geometrical fuzzy clustering-based solution to glottal wave estimation. *The Journal of the Acoustical Society of America*, 104:3070–3079, 1998.

[64] H. Auvinen, T. Raitio, M. Airaksinen, S. Siltanen, B. H. Story, and P. Alku. Automatic glottal inverse filtering with the markov chain monte carlo method. *Computer Speech & Language*, 28:1139–1155, 2014.

[65] J. Walker and P. Murphy. A review of glottal waveform analysis. In Yannis Stylianou, Marcos Faundez-Zanuy, and Anna Esposito, editors, *Progress in Nonlinear Speech Processing*, volume 4391 of *Lecture Notes in Computer Science*, pages 1–21. Springer Berlin Heidelberg, 2007.

[66] J. Liljencrants. *Speech Synthesis with a Reflection-Type Line Analog*. PhD thesis, Royal Institute of Technology, Stockholm, Sweden, 1985.

[67] I. R. Titze. *Principles of Voice Production*. Prentice Hall, 1994.

[68] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer Texts in Statistics. Springer International Publishing, 2017.

[69] T. Söderström and P. Stoica. *System Identification*. Prentice Hall, New York, 1989.

[70] M. Zañartu, D. D. Mehta, J. C. Ho, R. E. Hillman, and G. R. Wodicka. Observation and analysis of in vivo vocal fold tissue instabilities produced by nonlinear source-filter coupling: A case study. *The Journal of the Acoustical Society of America*, 129:326–339, 2011.

[71] D. G. Childers and C. K. Lee. Vocal quality factors: Analysis, synthesis, and perception. *Journal of the Acoustical Society of America*, 90:23942410, 1991.

[72] P.A Naylor, Anastasis Kounoudes, J. Gudnason, and M. Brookes. Estimation of glottal closure instants in voiced speech using the dypsa algorithm. *Audio, Speech, and Language Processing, IEEE Transactions on*, 15(1):34–43, Jan 2007.

[73] M. R P Thomas, J. Gudnason, and P.A Naylor. Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(1):82–9, Jan 2012.

[74] T. Drugman, M. Thomas, J. Gudnason, P. Naylor, and T. Dutoit. Detection of glottal closure instants from speech signals: A quantitative review. *Audio, Speech, and Language Processing, IEEE Transactions on*, 20(3):994–1006, March 2012.

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

[76] K. N. Stevens. *Acoustic Phonetics*. MIT Press, 2000.

[77] 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. *Journal of the Acoustical Society of America*, 130:3999–4009, 2011.

[78] T. V. Rananthapadmanabha and G. Fant. Calculation of true glottal flow and its components. *STL-QPSR*, 23:1–30, 1982.

[79] 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.

[80] 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.

[81] Y. R. Chien, D. D. Mehta, J. Gunason, M. Za nartu, and T. F. Quatieri. Evaluation of glottal inverse filtering algorithms using a physiologically based articulatory speech synthesizer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(8):1718–1730, Aug 2017.

[82] B. Bozkurt, B. Doval, C. d'Alessandro, and T. Dutoit. Zeros of z-transform representation with application to source-filter separation in speech. *Signal Processing Letters, IEEE*, 12(4):344–347, April 2005.

[83] P. Alku, M. Airas, T. Bäckström, and H. Pulakka. Using group delay function to assess glottal flows estimated by inverse filtering. *Electronics Letters*, 41(9):562–563, April 2005.

[84] 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.

[85] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[86] R. A. Horn and Ch. R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[87] D. G. Childers, D. P. Skinner, and R. C. Kemerait. The cepstrum: A guide to processing. *Proceedings of the IEEE*, 65(10):1428–1443, Oct 1977.

[88] T. F. Quatieri. *Discrete-Time Speech Signal Processing: Principles and Practice*. Pearson Education Inc., 2012.

[89] 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.

[90] R. W. Schafer. *Echo Removal by Discrete Generalized Linear Filtering*. PhD thesis, Massachusetts Institute of Technology, Research Laboratory of Electronics, 1969.

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

[92] 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.

[93] J.A. Edwards and J.A.S. Angus. Using phase-plane plots to assess glottal inverse filtering. *Electronics Letters*, 32(3):192–193, Feb 1996.

[94] T. Bäckström, M. Airas, L. Lehto, and P. Alku. Objective quality measures for glottal inverse filtering of speech pressure signals. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings. (ICASSP '05). IEEE International Conference on*, volume 1, pages 897–900, March 2005.

[95] E. B. Holmberg, R. E. Hillman, J. S. Perkell, and C. Gress. Relationships between intra-speaker variation in aerodynamic measures of voice production and variation in spl across repeated recordings. *Journal of Speech, Language, and Hearing Research*, 37:484–495, 1994.

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

[97] 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.

[98] B. D. Erath, M. Za nartu, K. C. Stewart, M. W. Plesniak, D. E. Sommer, and S. D. Peterson. A review of lumped-element models of voiced speech. *Speech Communication*, 55(5):667 – 690, 2013.

[99] B. H. Story and I. R. Titze. Voice simulation with a body-cover model of the vocal folds. *Journal of the Acoustical Society of America*, 97:1249–1260, 1995.

[100] F. Alipour, D. Berry, and I. R. Titze. A finite-element model of vocal-fold vibration. *Journal of the Acoustical Society of America*, 108:3003–3012, 2000.

[101] I. R. Titze and B. H. Story. Acoustic interactions of the voice source with the lower vocal tract. *Journal of the Acoustical Society of America*, 101:2234–2243, 1997.

[102] I. R. Titze and A. S. Worley. Modeling source-filter interaction in belting and high-pitched operatic male singing. *Journal of the Acoustical Society of America*, 126(3):1530–1540, 2009.

[103] I. R. Titze and B. H. Story. Rules for controlling low-dimensional vocal fold models with muscle activation. *Journal of the Acoustical Society of America*, 112:1064–1076, 2002.

[104] A. El-Jaroudi and J. Makhoul. Discrete all-pole modeling. *IEEE Transactions on Signal Processing*, 39(2):411–423, Feb 1991.

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

[106] J. Durbin and S. J. Koopman. *Time Series Analysis by State Space Methods*. Oxford University Press, 2012.

[107] S. Kay. *Fundamentals of Statistical Signal Processing, Vol. I - Estimation Theory*. Prentice Hall, 1993.

[108] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer. Covarep 2014; a collaborative voice analysis repository for speech technologies. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 960–964, May 2014.

[109] S. M. Lulich and J. R. Morton. Subglottal resonances of adult male and female native speakers of american english. *Journal of the Acoustical Society of America*, 132(4):2592–2602, 2012.

[110] A. V. Oppenheim, R. Schafer, and J. Buck. *Discrete-Time Signal Processing*. 1999.

[111] N. D. Hogikyan and G. Sethuraman. Validation of an instrument to measure voice-related quality of life (v-rqol). *Journal of Voice*, 13(4):557 – 569, 1999.

[112] G. B. Kempster, B. R. Gerratt, K. Verdolini Abbott, 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.

[113] M. Kunduk and A. J. McWhorter. True vocal fold nodules: the role of differential diagnosis. *Current Opinion in Otolaryngology & Head and Neck Surgery*, 17(6):449452, 2009.

[114] T. Koc and T. Ciloglu. Nonlinear interactive source-filter models for speech. *Computer Speech & Language*, 36:365–394, 2016.

[115] P. Jinachitra and J. O. Smith. Joint estimation of glottal source and vocal tract for vocal synthesis using kalman smoothing and EM algorithm. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005.

[116] P. Milenkovic. Glottal inverse filtering by joint estimation of an ar system with a linear input model. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 42(1):2842, 1986.

[117] D. Vincent, O. Rosec, and T. Chonavel. A new method for speech synthesis and transformation based on an arx-lf source-filter decomposition and hnm modeling. In *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP 07*, volume 4, pages 525–528, 2007.

[118] D. D. Mehta and P. J. Wolfe. Statistical properties of linear prediction analysis underlying the challenge of formant bandwidth estimation. *The Journal of the Acoustical Society of America*, 137(2):944–950, 2015.

[119] S. Björklund and J. Sundberg. Relationship between subglottal pressure and sound pressure level in untrained voices. *Journal of Voice*, 2015.

[120] G. Fant. Preliminaries to analysis of the human voice source. *STL-QPSR*, 4:1–27, 1982.

[121] T. W. Anderson. *An Introduction to Multivariate Statistical Analysis.* Wiley & Sons, 2004.

[122] R. A. Johnsons and D. W. Wichern. *Applied Multivariate Statistical Analysis.* Pearson Education Inc., 2013.

[123] J. Cohen. *Statistical power analysis for the behavior science.* Lawrance Eribaum Association., 1988.

[124] R. Maronna, R. Martin, and V. Yohai. *Robust Statistics: Theory and Methods.* John Wiley & Sons, Ltda, 2006.

[125] R. D. Cook. Detection of influential observation in linear regression. *Technometrics*, 19:15–18, 1977.

[126] B. L. Welch. The significance of the difference between two means when the population variances are unequal. *Biometrika*, 29(3/4):350–362, 1938.

[127] G. James, G. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning.* Springer, 2015.

[128] R. R. Wilcox. *Fundamentals of Modern Statistical Methods: Substantially Improving Power and Accuracy.* Springer, 2010.

[129] L. V. Hedges and I. Olkin. *Statistical Methods for Meta-Analysis.* Academic Press, 1985.

[130] J. Cohen. A power primer. *Psychological Bulletin*, 112(1):155–159, 1992.

[131] B. W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, 1986.

[132] J. Lindqvist-Gauffin. Inverse filtering. instrumentation and techniques. *Speech Transmission Laboratory Quarterly Progress and Status Report (STL-QPSR).*, 5(4):1–4, 1964.

[133] R. Leonard. Voice therapy and vocal nodules in adults. *Current Opinion in Otolaryngology & Head and Neck Surgery.*, 17(6):453457, 2009.

[134] 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 vocal hyperfunction 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.

[135] 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.

[136] I. R. Titze, T. Riede, and P. Popolo. Nonlinear source-filter coupling in phonation: Vocal exercises. *Journal of the Acoustical Society of America*, 123(4):1902–1915, 2008.

[137] R. H. Shumway and D. S. Stoffer. *Time Series Analysis and Its Applications: With R Examples*. Springer-Verlag New York, 2011.

[138] J.V. Candy. *Bayesian Signal Processing: Classical, Modern and Particle Filtering Methods*. Wiley-IEEE Press, 2009.

[139] H. E. de Bree. *The Microflown e-book*. Microflown Technologies, 2009.

[140] M. A. Ponti, L. S. F. Ribeiro, T. S. Nazare, T. Bui, and J. Collomosse. Everything you wanted to know about deep learning for computer vision but were afraid to ask. In *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*, pages 17–41, Oct 2017.

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

[142] H. Hotelling. The generalization of student's ratio. *Ann. Math. Statist.*, 2(3):360–378, 08 1931.

[143] P. J. Huber. Robust estimation of a location parameter. *Ann. Math. Statist.*, 35(1):73–101, 03 1964.

[144] B. Efron. Bootstrap methods: Another look at the jackknife. *Ann. Statist.*, 7(1):1–26, 01 1979.