

Faculdade de Engenharia da Universidade do Porto



Accurate glottal source estimation and modelling

Bruno Miguel Silva Santos

MASTER'S THESIS

MESTRADO INTEGRADO EM BIOENGENHARIA - ENGENHARIA BIOMÉDICA

Supervisor: Prof. Dr. Aníbal João de Sousa Ferreira, FEUP

Co-Supervisor: Prof. Dr. Jorge Eduardo de Freitas Spratley, FMUP

June 2020

Accurate glottal source estimation and modelling

Bruno Miguel Silva Santos

**MESTRADO INTEGRADO EM BIOENGENHARIA - ENGENHARIA
BIOMÉDICA**

June 2020

Resumo

A fala é uma ferramenta vital utilizada pela maioria das pessoas, tanto na sua vida social como profissional. O facto de o nosso quotidiano depender tanto desta habilidade, leva-nos a encontrar soluções que possam substituir ou auxiliar quando somos impedidos de a utilizar. A fala sussurrada é também uma forma conveniente de comunicação. Permite manter uma conversa restrita a um público próximo ou evitar perturbar ambientes silenciosos. No entanto, estar restringido a esta forma de comunicar pode ser limitador. Portanto, soluções foram criadas para evitar esta situação. Na maioria das vezes, estas passam por métodos invasivos ou desconfortáveis para o utilizador e, além disso, é de se notar a artificialidade do som da voz sintetizada. No passado, isto seria suficiente, mas com os atuais desenvolvimentos na área de processamento de fala e a expectativa do usuário exigem melhores abordagens. O trabalho que foi realizado durante esta dissertação relaciona-se com uma melhor estimação e modelação da fonte glótica com o objetivo de aprimorar as aplicações e superar limitações atuais. Além disso, este trabalho teve como objetivo a caracterização espectral da magnitude e da estrutura de fase do sinal da fonte glótica, o que pode melhorar a inteligibilidade dos sinais sintéticos de fala e a preservar características idiossincráticas do orador. O sinal tratado foi até então considerado inacessível, devido à sua localização limitada. Para o efeito, foram captados dois sinais alinhados temporalmente: um que corresponde ao sinal natural da fonte glótica e outro que corresponde ao sinal natural da fala. Por ser um procedimento inovador, invasivo e complexo, que exige alto nível de especialização, contou com a ajuda de uma equipa multidisciplinar no campo de medicina e de engenharia. Os dados reais recolhidos foram analisados utilizando técnicas de processamento de fala, a fim de recuperar novos conhecimentos sobre a fonte glótica. Além disso, um modelo empírico da fonte glótica foi obtido de acordo com o orador, baseado nos sinais reais da fonte glótica. Os filtros do trato vocal foram caracterizados e estimados com sucesso, preservando as características idiossincráticas dos oradores.

Abstract

Speech is a vital tool often used by most people on their social and professional life. The fact that our daily life depends so much on this ability urges us to find solutions that can replace or assist when prevented from using it. Whispered speech is also a convenient form of communication, since it allows keeping a conversation restricted to a nearby audience and prevents from disturbing silent environments. However, being restrained to this way of conveying information can be limiting. Therefore, solutions have been created to avoid this situation. Most of the time, these are either invasive or unfavorable for the user, creating artificial and synthesized sounding voices. In the past these would be sufficient, but the developments in the field of speech processing and also the user expectations require better approaches. The work carried out during this dissertation focused on finding a more accurate estimation and modelling of the signal generated by the vocal folds, the glottal source. This aims to improve current applications and overcome their limitations. Furthermore, the goal was to characterize the spectral magnitude and phase structure of the glottal source signal, improving the intelligibility of the synthetic signals of speech and preserving the idiosyncratic features of the speaker. The handled signals were thought to be unattainable, due to their location and their limited access. The goal was to capture two time-aligned signals: one correspondent to the natural glottal source signal and another to the natural speech signal. Being this an innovative, invasive and complex procedure that requires a high level of expertise, it counted with the help of a multidisciplinary team in the fields of medicine and engineering. The real data collected was analysed using speech processing techniques in order to retrieve new knowledge about the glottal source. Furthermore, an empirical model of the glottal source was obtained according to speaker based on the real glottal source signals. Vocal tract filters were successfully characterized and estimated, while preserving the idiosyncratic features of the speakers.

Acknowledgments

Gostaria de agradecer de forma breve a todas as pessoas com quem me cruzei nos últimos cinco anos.

Aceitei o desafio de trabalhar neste tópico que me era alheio e, no entanto, nunca aprendi tanto sobre um assunto tão fascinante como este. Estou eternamente grato ao Prof. Doutor Aníbal Ferreira pelo voto de confiança e pela oportunidade única, a qual eu nunca esquecerei. Obrigado pelos seus ensinamentos e pela paciência.

Gostaria de agradecer a colaboração do Prof. Doutor Jorge Spratley e do Doutor Laurentino Mendes que permitiram que a concretização deste trabalho fosse possível. Obrigado pelo profissionalismo e pela boa disposição.

Não poderia deixar de fazer um grande agradecimento a toda a equipa do Projeto "DyNaVoiceR" que mostrou ser a melhor companhia que eu poderia pedir. Obrigado ao João pelas conversas mais aleatórias, ao Marco pelas explicações sempre rebuscadas e à Clara pela motivação. Obrigado a todos pelos inputs e conversas que contribuíram para o desenvolvimento deste trabalho. Queria agradecer a importante ajuda do Prof. Doutor Luis de Jesus na recolha do sinal EGG.

Obrigado aos voluntários que participaram no procedimento e permitiram a realização deste trabalho, mesmo aqueles que entretanto deixaram de ser meus amigos depois desta traição. Um obrigado por darem a vossa voz ao manifesto.

É irónico o facto de ter trabalhado com fala durante os últimos seis meses e mesmo assim não me conseguir expressar da melhor maneira. A quem me acompanhou nos primeiros dois anos, um enorme obrigado por todas as memórias que criamos e por tudo o que me ensinaram. A bila há de ter sempre saudades dos seus m8s e nós dos momentos que passamos na bila. Obrigado pseudo-Raso por partilhares dos meus gostos peculiares e por me adotares nessa tua família disfuncional de MedVet.

Um obrigado à Prof. Doutora Berta Gonçalves pelas palavras incentivadoras e pelo trabalho exemplar que faz pelo curso.

Um obrigado à melhor madrinha (mentira ambos sabemos que arranjava melhor se não fosse o Jantar do Kilt... Está dito, assim, sem pão sem nada). Obrigado por estares sempre presente e por seres a pessoa mais prestável, bem disposta e desenrascada.

Obrigado à minha família de Erasmus, com quem eu me arrependo profundamente de não ter passado mais tempo. Hoje era capaz de trocar todas as viagens que fiz só para poder jantar uma última vez em casa da "mãe".

Obrigado aos Solinca de Matosinhos pela incondicional disponibilidade de treinos de levantamento do copo e sessões para dar à língua, apesar de nem sempre ter comparecido (já se começa a notar). Obrigado por alinharem nos planos mais aleatórios, mesmo sabendo que vai ter sempre massa com atum ao barulho. Obrigado pelo tempo bem passado (assim como aquela posta em Salto). No seguimento, aproveito para agradecer aos melhores exemplos que poderíamos ter tido. Um grande obrigado ao Professor Jorge e à Professora Susana por serem a combinação de exigência e boa disposição (excepto quando as aulas eram sobre calhaus).

Obrigado à minha segunda família, que me adotou quase sem perceber bem como. Agora só tenho a certeza de uma coisa, nem eu, nem eles nos livramos tão cedo uns dos outros. É caso para dizer que foi realmente bom.

Não posso deixar de agradecer a quem esteve sempre presente independentemente da distância ou período. Um obrigado do fundo do coração às minhas velhas e eternas amigas loucas por gatos.

Finalmente, queria deixar um palavra de agradecimento a quem me diz mais neste mundo. Obrigado à minha família por me aturarem tal como sou. Quero dedicar esta tese à minha mãe pelo apoio incondicional, pela paciência e pelo carinho. Obrigado por me teres incutido os princípios que me tornaram na pessoa que sou hoje. Ao meu pai pelos ensinamentos que me deu e por aqueles que ainda estão por vir. Ao meu avô pelos sermões intermináveis em que vira o disco e toca o mesmo. À minha avó materna pelo arroz de tomate e à minha avó paterna pela sopa de legumes. Aos meus tios e primos. Talvez ao meu irmão e ficamos por aqui senão já começa a ficar ridículo. Obrigado.

Este trabalho foi apoiado com uma bolsa de investigação no âmbito do Projeto PTDC/EMD-EMD/29308/2017 - POCI-01-0145-FEDER-029308 - financiado pelo Fundo Europeu de Desenvolvimento Regional (FEDER), através do COMPETE2020 – Programa Operacional Competitividade e Internacionalização (POCI) e com o apoio financeiro da FCT/MCTES através de fundos nacionais (PIDDAC).

Bruno Santos

*“ When life gives you lemonade, make lemons.
Life will be all like, “Whaaaat?” ”*

Phil Dunphy, *Phil’s-osophy*

Contents

Agradecimientos	v
List of Figures	xi
List of Tables	xv
List of Abbreviations	xvii
1 Introduction	1
1.1 Context	1
1.2 Motivation	2
1.3 Goals	2
1.4 Contributions	2
1.5 Document structure	3
2 Background	5
2.1 Speech apparatus	5
2.2 Glottal assessment and visualization	9
2.2.1 Electroglottography	9
2.2.2 Videostroboscopy	9
2.2.3 High quality video examination	11
2.3 Source-Filter theory	12
2.4 Summary	14
3 Literature review	15
3.1 Glottal source estimation	15
3.1.1 Inverse filtering	15
3.1.2 Mixed-phase decomposition	17
3.2 Glottal source models	18
3.2.1 Liljencrants-Fant model	18
3.2.2 Rosenberg model	19
3.2.3 Other models	19
3.3 Glottal source parameterization	20
3.3.1 Time domain features	20
3.3.2 Frequency domain features	21
3.4 Summary	22
4 Data Acquisition and Dataset Characterization	23
4.1 Signal acquisition	23
4.2 Dataset description	25
4.3 Preliminary Analysis	26
4.4 Perceptual tests	28
4.5 Summary	29

5	Glottal Source Characterization	31
5.1	Parametric Spectral Analysis	31
5.1.1	Spectral magnitude analysis	32
5.1.2	Spectral phase analysis	37
5.2	Statistical analysis	42
5.3	Empirical model of the glottal source	45
5.4	Summary	46
6	Vocal Tract Characterization	47
6.1	Estimation of the vocal tract filter	47
6.1.1	Deconvolution approach	47
6.1.2	Adaptive filtering approach	48
6.1.3	Holistic filter design approach	49
6.2	Perceptual tests	51
6.3	Summary	54
7	Conclusions and Future Work	55
	References	57
A	Data Acquisition and Dataset Characterization	65
B	Glottal Source Characterization	79
C	Statistical Analysis	111
D	Vocal Tract Characterization	131

List of Figures

2.1	Illustration of the phonatory system (sagittal cut).	6
2.2	Illustration of the phonatory system (transversal cut).	6
2.3	Illustration of the vocal folds and glottis.	7
2.4	Illustration of the different configurations of the vocal folds that result in different phonation forms: a) normal voice; b) whispered voice; c) voicelessness. The numerated structures correspond to: 1. glottis; 2. arytenoid cartilages; 3. vocal folds; and 4. epiglottis.	7
2.5	Glottal pulse and its relation with the glottal cycle.	8
2.6	EKG signal waveform of a healthy adult male subject.	9
2.7	Fundamental principle of videostroboscopy.	10
2.8	Fundamental process of videokymography, where is depicted the line selected for frame by frame analysis (top) and output image (bottom).	11
2.9	Sequence of high-speed videoendoscopy frames.	12
2.10	Descriptive scheme representing speech production process by the light of the source-filter model and the different contributions based on the phonatory system to the development of the speech signal.	13
3.1	Waveshape of the glottal source derivative (left) and the glottal source (right) according to the LF model.	18
3.2	Waveshape of the glottal source according to the Rosenberg model.	19
3.3	Illustrative example of the waveshape of the glottal source (top) and the glottal source derivative (bottom), where are depicted the different phases and events of the glottal cycle.	21
4.1	Scheme of the positioning for the internal microphone (blue) with an approximated distance to the vocal folds and the external microphone (orange) with an approximated distance to the mouth.	24
4.2	Illustration of the acoustic triangle for the oral tonic vowels /a/, /i/ and /u/ of the standard EP.	25
4.3	Example of the time-aligned stereo recording of speaker 6, containing three repetitions of the three sustained vowels (/a/, /i/ and /u/). The upper figure corresponds to the signal recorded near the vocal folds and the lower figure corresponds to the signal recorded outside the mouth.	26
4.4	Shape of the acoustic pressure signals, the upper signal recorded near the vocal folds and the bottom signal recorded close to the mouth of speaker 3 for the sustained vowel /i/.	27
4.5	Shape of the acoustic pressure signals, the upper signal recorded near the vocal folds showing clear influence of two distinct signals and the bottom signal recorded close to the mouth of speaker 5 (male) for the sustained vowel /i/.	27
4.6	Results obtained for the perceptual tests describing the percent correct identification of the signals given. The lighter bars correspond to the first part and the darker to the second part of the perceptual tests.	28

5.1	Block diagram with the main steps of the analysis and parametric modelling performed by the DyNaVoiceR framework, where the LPC represents the spectral magnitude and the NRD represents the spectral phase of a periodic signal. . . .	32
5.2	Typical waveform of the internal acoustic pressure signal uttered by Speaker 3 for vowel /i/ (top), respective magnitude power spectrum (mid) and the respective harmonic structure alongside with the estimated peaks of the harmonic (red triangles) and wrapped NRD representation (bottom).	32
5.3	Magnitude power spectrum and the respective harmonic structure alongside with the estimated peaks of the harmonic (red triangles) for vowel /i/ from speaker 3.	33
5.4	Individual frames spectral magnitudes of the internal (blue) and external (orange) signals, and their respective 95% CIs, obtained from repetition 3 of speaker 3 uttering the sustained vowel /a/.	34
5.5	Mean spectral magnitude slope (brown) and individual frames spectral magnitudes (blue) and their respective 95% CIs, obtained from repetition 3 of speaker 3 uttering the sustained vowel /i/.	35
5.6	Linear regression model that represents the mean spectral magnitude found for all repetitions collected from different speakers uttering the sustained vowel /i/.	36
5.7	Linear regression model that represents the mean spectral magnitude found for all repetitions collected from different speakers uttering the sustained vowel /u/.	36
5.8	Spectral magnitude mean and the representation of the magnitude power spectra for all frames and its corresponding mean for repetition 3 from speaker 3 uttering the sustained vowel /i/.	37
5.9	Spectral NRD slope mean, the unwrapped NRD values (blue), and the corresponding mean and the 95% CIs of the internal signal for repetition 3 from speaker 3 uttering the sustained vowel /i/.	38
5.10	Spectral NRD slope mean, the unwrapped NRD values (orange), its corresponding mean and the 95% CIs of the external signal for repetition 3 from speaker 3 uttering the sustained vowel /i/.	39
5.11	Spectral NRD slope mean found for the internal signal (blue) and the external signal (orange). The unwrapped NRD for the internal signal (blue) and the external signal (orange) are depicted, its corresponding mean and the 95% CIs for repetition 3 from speaker 3 uttering the sustained vowel /i/.	41
5.12	Difference between the external NRD mean (orange) and the internal NRD mean (blue) for repetition 3 of speaker 3 uttering the sustained vowel /i/. The equation that corresponds to the linear regression model of the NRD difference is represented in brown.	41
5.13	Differences between the external and internal NRD means for all signals. Also depicted the equations that correspond to the linear regression model of the NRD difference for each vowel.	42
5.14	Boxplots of all the spectral magnitude values for the first 20 harmonics of the internal signal for repetition 3 from speaker 3 uttering vowel /i/.	43
5.15	Boxplots of all the spectral magnitude values for the first 20 harmonics of the internal signal for repetition 3 from speaker 3 uttering vowel /u/.	43
5.16	Boxplots of all the unwrapped NRD values for the first 20 harmonics of the internal signal for repetition 3 from speaker 3 uttering vowel /a/.	44
5.17	Boxplots of all the unwrapped NRD values for the first 20 harmonics of the external signal for repetition 3 from speaker 3 uttering vowel /a/.	44
5.18	Boxplots of all the unwrapped NRD values for the first 20 harmonics of the internal signal for repetition 3 from speaker 3 uttering vowel /i/.	44
5.19	Boxplots of all the unwrapped NRD values for the first 20 harmonics of the external signal for repetition 3 from speaker 3 uttering vowel /i/.	44
5.20	Boxplots of all the unwrapped NRD values for the first 20 harmonics of the internal signal for repetition 3 from speaker 3 uttering vowel /u/.	44

5.21	Boxplots of all the unwrapped NRD values for the first 20 harmonics of the external signal for repetition 3 from speaker 3 uttering vowel /u/.	44
5.22	Waveform of the empirical model of the glottal source for vowel /i/ obtained for Speaker 2 (orange), for Speaker 3 (yellow) and for Speaker 4 (purple), alongside with the waveform of theoretical model LF (blue) and their derivatives.	46
6.1	Block diagram representing the structure of a general adaptive filtering processing for system identification.	48
6.2	Spectral envelopes of the internal (blue) and external (orange) harmonic structure and frequency response of the prototype filter (red).	49
6.3	Frequency response of the prototype filter and the corresponding frequency response of the IIR and FIR filters computed for the repetition 3 of the vowel /i/ from speaker 3.	50
6.4	Frequency response of the prototype filters and the corresponding frequency response of the IIR and FIR filters obtained for all three repetitions of vowel /i/ from speaker 2.	50
6.5	Frequency response of the prototype filters and the corresponding frequency response of the IIR and FIR filters obtained for all the repetitions of vowel /i/ from all speakers.	51
6.6	Results obtained with the perceptual tests regarding the accuracy in identifying the correct speaker. The blue bars correspond to Speaker 2, the green bars correspond to Speaker 3 and the red bars correspond to Speaker 4.	52
6.7	Results obtained with the perceptual tests regarding the degree of similarity that the participants gave to the chosen sample when compared to the reference sample. The blue bars correspond to Speaker 2, the green bars correspond to Speaker 3 and the red bars correspond to Speaker 4.	52

List of Tables

3.1	Main theoretical glottal source models with a brief description, the number of parameters required and their improvements.	20
4.1	Characterization of the 6 volunteer speakers regarding gender and age.	25
4.2	Results for both parts of the perceptual tests where the values shown correspond to the success rate in identifying the vowel recorded. The p-values obtained regarding the statistical difference between results for Part I and Part II are also depicted.	29
5.1	Spectral magnitude slope values obtained for different repetitions of each sustained vowel (/i/ and /u/) according to the speaker. The values below correspond to the mean value ($\bar{x}$) and to the standard deviation (σ). The bottom values correspond to the mean value of all spectral magnitude slope values.	35
5.2	Spectral NRD slope values for different repetitions of each sustained vowel (/a/, /i/ and /u/) according to the speaker. The mean spectral NRD slope values ($\bar{x}$) and the standard deviation (σ) are also depicted according to vowel and location, alongside with the difference values between the internal and the external NRD means according to vowel.	40
6.1	p-values obtained for the statistical analysis between the degree of similarity for different vowels for the same speaker.	53
6.2	p-values obtained for the statistical analysis between the degree of similarity for different speakers for the same vowels.	53
6.3	Results of the perceptual tests where the values shown correspond to the average success rate in identifying the reference and average similarity to the reference according to speaker. The last column corresponds to the mean value ($\bar{x}$).	53

List of Abbreviations

ARX	AutoRegressive eXogenous
CALM	Causal-Anticausal Linear Model
CC	Complex Cepstrum
CES	Comissão de Ética para a Saúde
CI	Confidence Interval
CHUSJ	Centro Hospitalar e Universitário e Hospitalar de São João
CPIF	Closed Phased Inverse Filtering
DAP	Discrete All Poles
EGG	Electroglottography
FIR	Finite Impulse Response
FNPL	Fiber Naso-Pharyngo-Laryngoscopy
GCI	Glottal Closure Instant
HNR	Harmonic to Noise Ratio
HRF	Harmonic Richness Factor
HSV	High-Speed Videoendoscopy
IAIF	Iterative Adaptive Inverse Filtering
IDFT	Inverse of the Discrete Fourier Transform
IIR	Infinite Impulse Response
LF	Liljencrants-Fant
LMS	Least Mean Squares
LPC	Linear Predictive Coefficient
NAQ	Normalized Amplitude Quotient
NRD	Normalized Relative Delay
ODFT	Odd-Frequency Discrete Fourier Transform
OQ	Open Quotient
ORL	Otorhinolaryngology
PSD	Power Spectral Density
PSP	Parabolic Spectrum Parameter
QOQ	Quasi-Open Quotient
SQ	Speed Quotient
VKG	Videokymography
VTF	Vocal Tract Filter
ZZT	Zeros of the Z-Transform

Chapter 1

Introduction

1.1 Context

VOICE has been a tool of utmost importance for human development. The ability to produce sounds allowed for the evolution of what were once mere noises into a high complexity speech. Furthermore, it represents one of the most efficient methods for exchanging information between speakers, such as ideas or emotions. Whispering is another useful form of communicating, whether when conveying private information or when avoiding disturbing quiet environments. However, some people are restricted to this way of communication, which is characterized by its reduced perceptibility and lack of idiosyncrasy.

Millions of people suffer from voice-related problems, either due to efforts related to their careers, such as teachers and singers [1], or for natural reasons. In fact, in the United States of America, it is estimated that at least 1 out of 50 people has or will have a voice-related disorder during their lifetime [2, 3]. In recent years, research has been conducted with the purpose of converting whispered speech to voiced speech [4, 5, 6]. However, current solutions are far from ideal, since these are often invasive or sound synthetic and artificial. These applications depend highly on trustworthy models that represent the glottal excitation and on the support of acoustic features that describe the vocal tract response. Nevertheless, due to the difficulty in accessing the real source signal, the glottal excitation is usually estimated from the available speech signal, which requires complicated processing. Consequently, these models fall short in describing accurately the real glottal source.

The primary source of excitation, the glottal source, carries interesting information that can improve the naturalness of these speech synthesis applications. In the short-term, this dissertation is focused on the realization of *DyNaVoiceR* project's sub-task for accurate glottal source estimation and modelling. Alongside with better understanding of the human phonation process, it can add valuable information that can be applied in multiple areas such as speech synthesis, expressivity analysis, pathology detection and speaker recognition. These findings will translate in excellent quality and performance once integrated into the diverse voice technology applications. The expected long-term outcome of this project is to restore natural voiced speech from whispered or dysphonic speech based on speech processing techniques.

1.2 Motivation

In everyday life, different situations require the use of different modes of speech. However, in order to be heard, the speech mechanism is able to compensate when the surrounding conditions demand so. Being restricted from using its full potential can be socially unfavourable and limiting for patients. Even more so if their voice is a valuable asset to their professional life, jeopardising their aptness to fulfil their role. Having this in mind, the long-term goal of this project is to assist people who suffer from voice disorders or that have undergone medical procedures that have restrained them to whispered voicing as a way of communicating. The desired solution is a non-invasive and real-time application that preserves idiosyncrasies and converts whispered speech into voiced speech. Thereby, fitting this existing gap between the need for a more natural and effective way of communicating for voice disabled people. One of the critical aspects of this approach is the faithful characterization of the real glottal source in its spectral magnitude and phase structure, which will be the aim of this dissertation.

1.3 Goals

The aim of this dissertation is the estimation and modelling of the glottal source, a sub-task of the project *DyNaVoiceR*. The project is focused on converting whispered speech into voiced speech for a real-time application. For this reason, the approach to be considered has to be feasible in a limited period of time, i.e., it should be fast enough in order to avoid the "lip sync" problem.

With this work, it is expected that an accurate characterization of the glottal source is achieved and, consequently, the reconstruction of the vocal tract filter that is needed for easily converting the glottal source signal into a speech signal. For this, special recordings of acoustic data near the vocal folds will be carried out. Insight and models extracted from these recordings are expected to improve the understanding of the real glottal excitation.

1.4 Contributions

The work has major relevance in the field of speech processing and enhances the scientific knowledge about the involvement of the vocal folds in the phonation process. Furthermore, a dataset of 6 speakers (3 males and 3 females) of two time-aligned signals of 3 sustained vowels (\a\, \i\ and \u\) was developed, one from the glottal excitation source within the phonatory system (as close as possible to the vocal folds) and another from the speech signal outside (close to the mouth). Finally, the characterization of the magnitude and phase glottal pulse will be the highlight of this work and, subsequently, the estimation of a vocal tract filter that can transform the glottal source signal into a natural speech signal. The application of these models to the project's developed framework will contribute for improving the whispered speech to voiced speech conversion system.

1.5 Document structure

Throughout this dissertation, fundamental concepts and technical terms are defined. Moreover, basic knowledge is provided in order to promote a coherent and rational thought process. The structure of this document is the following:

- Chapter 1: Introductory chapter, where it is explained the need and the pertinence of having a more accurate glottal source model based on empirical data and its applications;
- Chapter 2: Background information, giving fundamental concepts on the physiological and anatomical aspects of the speech apparatus and the current medical procedures for assessing the phonatory system, as well as the Source-Filter model used as a fundamental basis for speech processing;
- Chapter 3: Covers the most recently developed approaches for analysing and estimating the glottal source, the existing glottal pulse models and their parameterization;
- Chapter 4: Describes the signal acquisition procedure and characterizes of the dataset obtained for further study of the recorded signals. Also, a preliminary study of the linguistic content of the glottal source signals is performed;
- Chapter 5: In this Chapter, the spectral analysis performed to the signals recorded is explained and the characterization of the glottal source is described. Furthermore, a statistical analysis of the recorded signals is made and the empirical models obtained from the real glottal source signals are described;
- Chapter 6: This chapter addresses the approaches used for the vocal tract estimation and the evaluation of the synthetic signals generated;
- Chapter 7: Conclusions are presented on what has been accomplished in this dissertation and future steps.

Chapter 2

Background

SPEECH is defined as the faculty or power of speaking through oral communication, the ability of expressing thoughts or emotions using speech sounds. Speech production is a complex process that requires the use of the vocal apparatus and results in the utterance of intelligible speech.

This chapter addresses some fundamental concepts about the anatomy and physiology of the speech apparatus and the current methods for assessing and visualizing the organ responsible for the glottal excitation, the vocal folds. Finally, a light is shone on the source-filter theory and the way the glottal excitation interacts with the vocal tract.

2.1 Speech apparatus

The voice organ is an intricate structure that allows the production of speech, depicted in Figure 2.1. Its study will be relevant for further understanding the function of each component and how it contributes for the occurrence of this phenomenon. The phonetic system can be functionally divided into three major parts: the breathing apparatus (subglottal system), the larynx (more precisely, the vocal folds) and the vocal tract (supraglottal system) [7]. The larynx, which corresponds to the boundary of these two systems, can be clearly visualized in Figure 2.2.

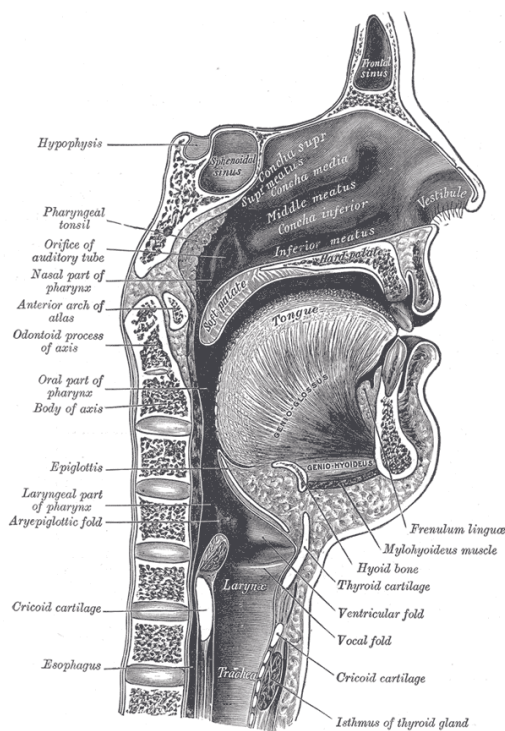


Figure 2.1: Illustration of the phonatory system (sagittal cut) [8].

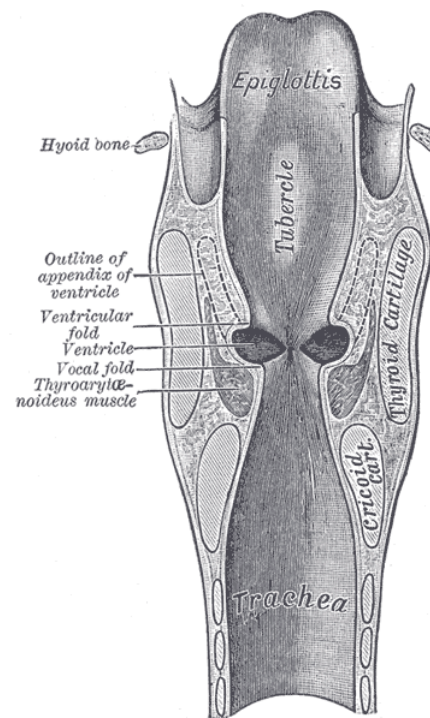


Figure 2.2: Illustration of the phonatory system (transversal cut) [8].

The main energy source is provided by the subglottal system, due to the eviction of airflow towards the trachea by the lungs. The lungs are the major organ involved in the crucial gas exchange process, allowing for blood to obtain oxygen and release carbon dioxide. Supported by the diaphragm and aided by the intercostal muscles, the thorax volume decreases and creates an ascendant airflow, that increases the subglottal pressure [9]. Once this airflow reaches the larynx it is modulated by the vocal folds (or cords), which can be seen in Figure 2.3. The space between the vocal folds is defined as glottis and corresponds to the boundary between subglottal and supraglottal systems [10].

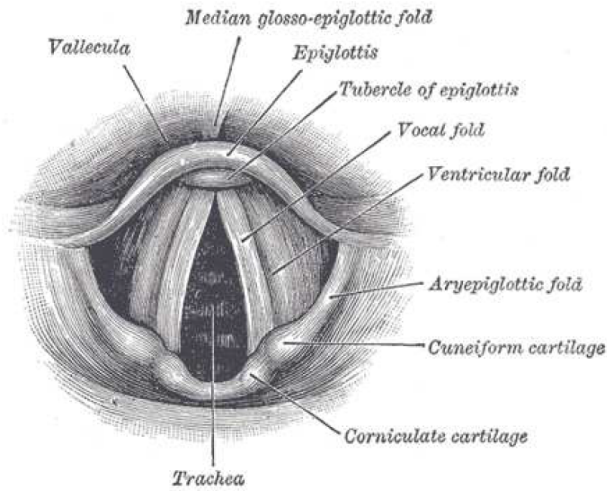


Figure 2.3: Illustration of the vocal folds and glottis [8].

The modulation at the glottal level previously mentioned can be either periodic or noisy, depending on the positioning of the arytenoid cartilages (2) that, subsequently, control the configuration of the vocal folds (3), as shown in Figure 2.4.

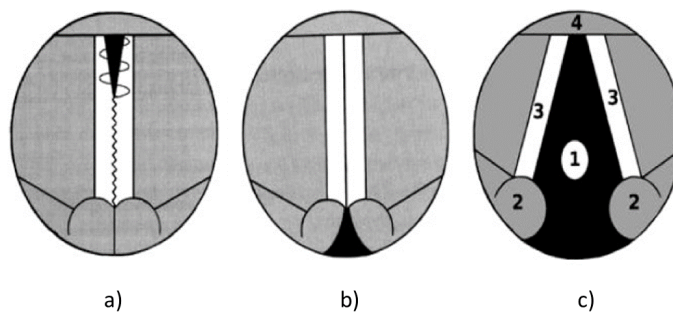


Figure 2.4: Illustration of the different configurations of the vocal folds that result in different phonation forms: a) normal voice; b) whispered voice; c) voicelessness. The numerated structures correspond to: 1. glottis; 2. arytenoid cartilages; 3. vocal folds; and 4. epiglottis. Adapted from [11].

The arytenoid cartilages have different possible dispositions: in adduction, where the vocal folds constrict the air passage; or in abduction, where the vocal folds are wide opened (which is the case when breathing) [12, 13]. The combination of these different positions allows the production of different phonation modes. When in adduction, the rising subglottal pressure forces the air to open and cause a quasi-periodic vibration of the vocal folds. Consequently, the opening and closing at a certain frequency generates a harmonic structured sound wave. In

case of abduction, the vocal folds open and the airflow passes without constrictions and, subsequently, exhibiting a continuous airstream without periodic glottal excitation (noisy behaviour) and resulting in an unvoiced sound [12, 13].

The glottal cycle can be divided into four different moments: the opening phase, the closing phase, the return phase and the closed phase [14]. Each phase can be associated to a certain moment of the glottal pulse, as shown in Figure 2.5.

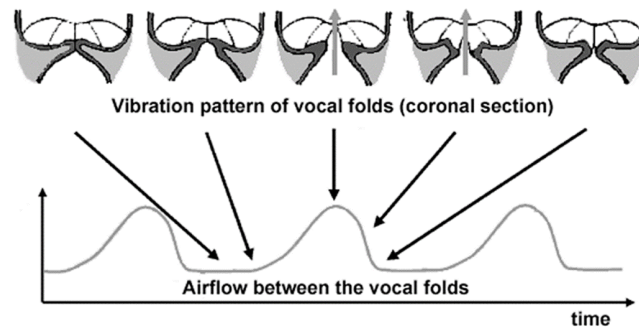


Figure 2.5: Glottal pulse and its relation with the glottal cycle [14].

The speech attributes, such as speech volume, stress pattern and speech duration, can be controlled at the glottal level by adjusting the airstream volume. The glottal behaviour also influences the voice perceived pitch. This property of speech is associated to the fundamental frequency (F_0) of the speech sound, which in turn is conditioned by the length of the speaker's vocal folds. Generally, the length of the vocal folds can be approximately, according to the gender, 9 to 13 mm for healthy females and 15 to 20 mm for healthy males [10].

The oral, nasal and pharyngeal resonant cavities constitute the vocal tract. These are responsible for further modulating the glottal source according to the influence of the supra-laryngeal portions and its articulators. The articulators can be divided into: active, which correspond to the velum (soft palate), jaw, tongue and the lower lip; and passive, such as the upper lip, the teeth, the alveolar ridge and the hard palate. By assuming different configurations, the articulators create distinct levels of constriction that act on different sites and modify the produced sound [15]. This confers the sound its timbre and the intended linguistic information by shaping the airflow spectral properties [16, 17].

The measurements of the vocal tract vary according to each individual. Commonly, the length of a healthy adult male ranges between 17 to 20 cm and its diameter corresponds to 3 cm [10].

Lastly, the labial radiation contributes with the final spectral modifications. The combination of all these elements determines the sound wave that is perceived by the listener as speech.

2.2 Glottal assessment and visualization

The glottal source or voice source corresponds to the acoustic pressure signal generated by the vocal folds at the glottis that serves as excitation in speech. However, the vocal folds are difficult to analyze due to their fast vibration and the difficult access to the signal (inside the larynx) [18, 19].

Having this in mind, multiple devices and procedures were developed in order to overcome these limitations and allow the observation of glottal behaviour. Such approaches rely on different functional methods, such as electrical [20], electromagnetic [21] and visual [22]. The latter is broadly accepted for voice production research and employed by voice clinicians to examine vocal folds with possible disorders. However, in order to obtain visual information the procedures are usually very invasive [23].

2.2.1 Electroglottography

The electroglottography (EGG) is a non-invasive exam that measures the impedance between the vocal folds and does so by transmitting a high-frequency current through two electrodes placed on opposite sides of the throat (on the larynx level). The different positions of the vocal folds are translated in fluctuations of the electrical impedance of the applied current. This is possible due to the fact that soft tissues are good electrical conductors when in comparison with the air that is present within the larynx lumen. Whenever the vocal folds move, the glottis takes a different conformation and changes the volume of the air column, altering the impedance accordingly [24].

The EGG shows a signal throughout time, as can be seen in Figure 2.6, measuring the vocal folds relative contact area. The waveshape of the signal is flat when the vocal folds are held apart due to the lack of impedance variation. Results are susceptible to the effects of skin moistness and involuntary movements of the larynx. Some noise can be also introduced by the inherent distortion levels of the device. Nonetheless, this solution allows the examination of the glottal behaviour. The interpretation of the EGG signal and its correlation with the phonation process is still a subject of great interest in current studies [25]. These measurements may also be used as a reference for comparing pitch tracking methods or Glottal Closure Instant (GCI) detection methods [11].

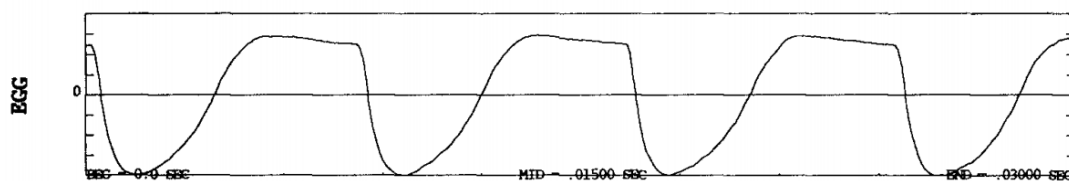


Figure 2.6: EGG signal waveform of a healthy adult male subject [26].

2.2.2 Videostroboscopy

The videostroboscopy is an efficient method for observing and studying the dynamic aspects of the vocal folds in the laryngeal, by using an intermittent light to illuminate the vocal folds, it

is capable of obtaining a series of frames that create a video. Through establishing a flashing frequency lower than the vocal folds vibration frequency, each frame captures the cycle phase shortly after the previously captured.

For an easier understanding, a visualization of the its fundamental principle can be observed in Figure 2.7. This imposes some limitations and requires that the vibration frequency is stationary in order to capture the whole vibration cycle with high accuracy and quality [27, 28, 25].

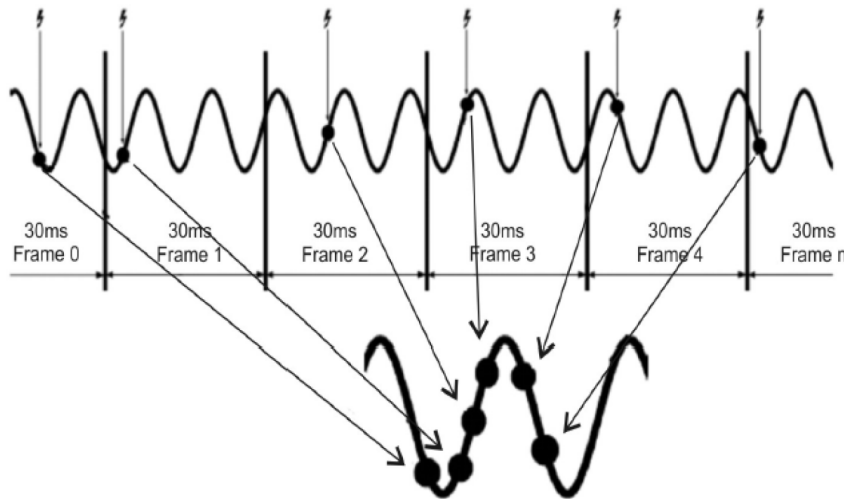


Figure 2.7: Fundamental principle of videostroboscopy [29].

2.2.2.1 Videokymography

Videokymography (VKG) is a digital technique that allows the visualisation of the vocal folds vibration at high-speed. The system utilizes a high-speed camera that captures almost 8 000 images per second. From the standard laryngeal image, the camera selects a horizontal active line, which is transversal to the glottis. The recorded horizontal lines are compiled vertically, forming a top-down image with the evolution of the vocal folds behaviour through time. From the visual analysis of this image, it is possible to extract information related with frequency, amplitude, left-right asymmetries and the phases of the glottal cycle [30]. However, only recently more than qualitative outcomes have been shown from a VKG, such as the segmentation of the vocal folds edges [31].

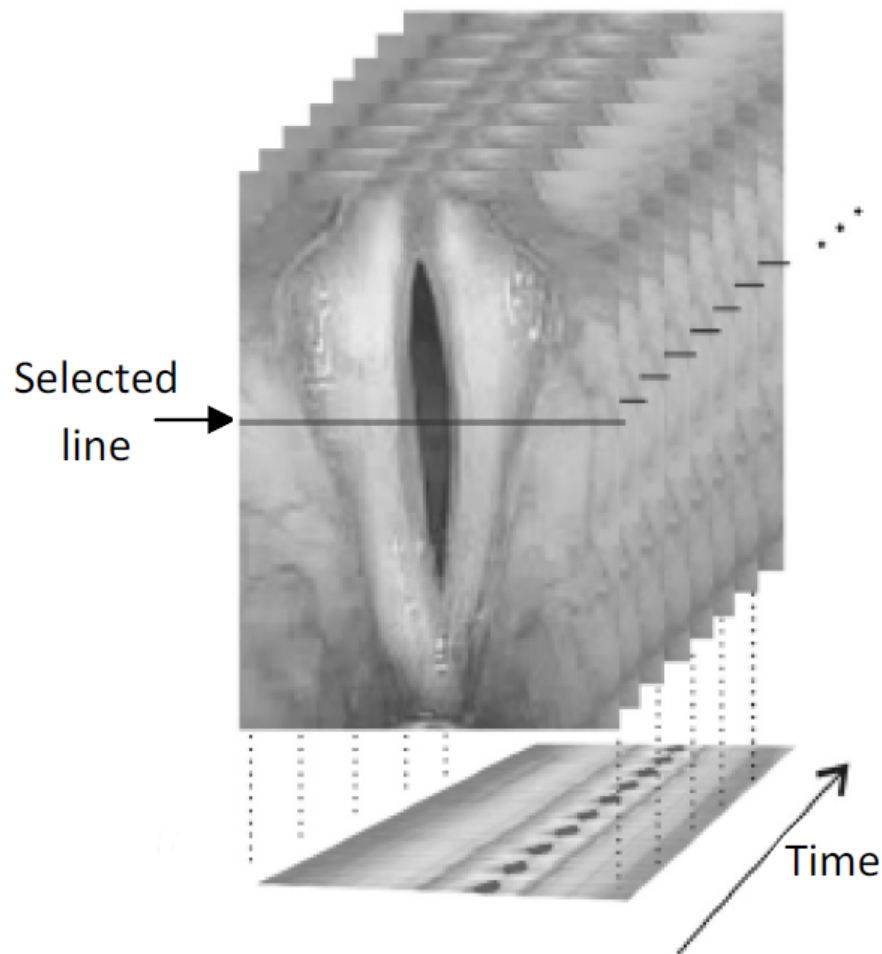


Figure 2.8: Fundamental process of videokymography, where is depicted the line selected for frame by frame analysis (top) and output image (bottom) [29].

2.2.3 High quality video examination

The High-Speed Videoendoscopy (HSV) is considered one the most accurate and precise procedures for visualizing the vibration of the vocal folds due to its high-speed imaging system capable of recording entire high quality images. The endoscope is placed inside the larynx near the vocal folds, where it records from a top-down position, as shown in Figure 2.9. Only recently, the high frame rate procedure was made possible due to the technological progress in the field of signal acquisition.

Nevertheless, there are still improvements to be done in the temporal and spatial resolution. As a consequence, a compromise between image resolution and frame rate needs to be made due to the limited data transfer speed. Another limitation is the light source restriction because of the endoscope work channel small diameter. The typical resolution of a HSV system ranges from 100 up to 300 pixels in each direction and frame rates of 10 000 frames per second. This requires large data storage and, for this reason, the recordings are usually short. A way of circumventing

this issue is by recording gray-scale images, since color images imply heavier files and require more illumination.

This technique is one of the most adequate for studying glottal behaviour related phenomena and for carrying vocal folds dynamics research [25].



Figure 2.9: Sequence of high-speed videoendoscopy frames [32].

A similar technique is the Fiber Naso-Pharyngo-Laryngoscopy (FNPL) which consists in the visualization and acquisition of high quality video of the vocal folds that makes use of a fiber optic system. The major difference to the HSE is the fact that the naso-pharyngo-laryngoscope is inserted through one of the nasal cavities instead of the oral cavity [33].

2.3 Source-Filter theory

The source-filter theory attempts to represent the previously described vocal system in a simplified analog model that approximates the complex phonation mechanism by combining different sound sources and acoustic filters. This approach relates the speech articulation with the acoustic signals intrinsic features and is supported by assuming that source and filter are independent [34, 35]. This statement is not entirely correct since interactions between source-tract exist and the glottal flow depends, to a certain extent, on the modifications of the vocal tract impedance. Nevertheless, this theory has been presumed appropriate for the work developed in this dissertation, given that most research studies and technological applications rely on its assumptions [36]. According to this theory, the functional parts of the phonetic system can be represented by the elements of a sound generator: where the lungs represent the power supply, the vocal folds represent an oscillator and the vocal tract represents a resonator [37]. As previously mentioned, the sound sources can be either periodic or non-periodic according to glottal behaviour. When dealing with voiced sounds, the glottal excitation signal is modelled as a periodic impulse train and the unvoiced sounds are modelled as white noise [38].

The supraglottal system is represented by a group of filters, for which the resonances resemble the filter formants and the radiation effects. Therefore, the voice production model is assumed to be the result of a convolution between the excitation source and the set of filters that represent to the vocal tract [35, 39].

The glottal cycle provides a train of air pulses that propagates towards the vocal tract. This train of air pulses is characterized in the Fourier domain by several peaks of energy called the harmonics or partials. The partials are a group of spectrum tones for which the lowest tone corresponds to the fundamental frequency and the remaining are called the overtones. During this process, the shape of the wave is spectrally modified in which several formants are formed, as a result of the influence from the different acoustic resonances of the vocal tract [15]. The

frequency amplitude for each formant is associated with the quality of voice and the type of vowel or consonant of the perceived sound [15].

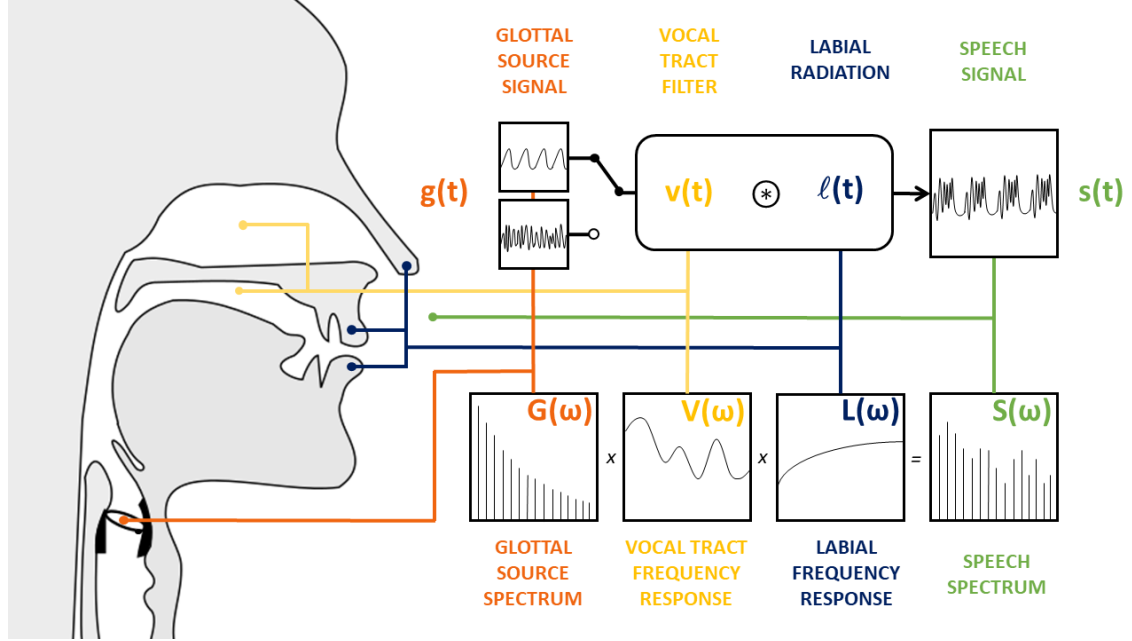


Figure 2.10: Descriptive scheme representing speech production process based on the source-filter model and the different contributions of the phonatory system to the development of the speech signal. Adapted from [23, 40]

This model represented in Figure 2.10 is a simple and convenient approximation that allows the study of voiced and unvoiced speech [41]. Thereby, the source-filter theory can be mathematically represented in the time-domain by:

$$s[n] = g[n] * v[n] * \ell[n] \quad (2.1)$$

where, $s[n]$ corresponds to the speech output signal, $g[n]$ to the glottal source signal, $v[n]$ to the vocal tract response and $\ell[n]$ to the labial radiation. In the Z-domain, the Equation 2.1 can be represented by:

$$S(z) = G(z)V(z)L(z) \quad (2.2)$$

where the Z transform of the glottal source signal corresponds to $G(z)$, the Z transform of the vocal tract response corresponds to $V(z)$, also referred to as Vocal Tract Filter (VTF), and the Z transform of the labial radiation corresponds to $L(z)$.

Accordingly, the solution of the following equation needs to be reached in order to use the glottal inverse filtering:

$$G(z) = \frac{S(z)}{V(z)L(z)}. \quad (2.3)$$

According to Equation 2.3, the impact of the vocal tract must be accounted for in order to determine the waveform of the glottal flow. Specifically when dealing with a voiced speech

signal, the waveform of the glottal flow shows a regular periodic shape. Usually, a p-order all-pole filter is used to model the VTF:

$$V(z) = \frac{1}{1 - \sum_{i=1}^p b_i z^{-i}} \quad (2.4)$$

where the resonances of the VTF correspond to poles and, consequently, to the formant frequencies of the vocal tract. A high pass filter is applied by the labial radiation which is approximated by a first-order time-domain derivative [15], implying that the effective excitation of the vocal tract corresponds to the derivative of glottal flow and thus:

$$L(z) = 1 - \alpha z^{-1}. \quad (2.5)$$

The labial radiation coefficient α assumes a value lower than 1, normally within 0.95 and 0.99 so that the zero lies in the unitary inner circle of the z plane, which can be expressed by the following equation [40]:

$$L(z) \approx \frac{1}{\sum_{k=0}^N \alpha^k \cdot z^{-k}}. \quad (2.6)$$

Since $\alpha < 1$, then N is in practice finite, but in theory it should be infinite. This result implies that a significant number of poles approximates the effect of a zero [42].

Many attempts have been made in order to model in both time and frequency domains the glottal source signal according to these assumptions. However, there are limitations to the linear Source-Filter theory, which falls short to describe the real scenario where source and filter interact.

2.4 Summary

Throughout this chapter, a background is given regarding the anatomy and the physiology of the phonatory system. Additionally, some techniques are presented, in particular the naso-pharyngo-laryngoscopy, for the analysis and observation of the vocal folds.

The chapter ahead describes current techniques used for the estimation of the glottal source and the vocal tract filter. The theoretical glottal source models as well as the time and frequency parameters that characterize them are also mentioned.

Chapter 3

Literature review

HERE some of the most recent developments in the field of glottal source and vocal tract filter estimation are outlined. Moreover, the theoretical glottal pulse models described in the literature are introduced, alongside with the time and frequency parameters used to define them. In this chapter a review is made of the current solutions for glottal source estimation and some of their shortcomings. It is also given a brief summary of the theoretical models, their evolution and improvements throughout time.

3.1 Glottal source estimation

This section describes the main methodologies for estimating the glottal source directly from the speech waveform. It is interesting to note that these methods only requires the speech signal, allowing for a non-invasive approach.

3.1.1 Inverse filtering

These techniques are based on the Source-Filter theory of speech production and on the assumption that a system exists when the transfer function is equal to the inverse transfer function of one filter (or an aggregation of filters) due to the influence of the articulatory components that change speech both in time and frequency [43]. First, a parametric model of the vocal tract is obtained and, secondly, the contribution of the vocal tract is removed.

However, current estimation methods rely on different approaches, either taking into account the glottal cycle particularities or relying on iterative and adaptive processes [44, 45].

3.1.1.1 Closed Phase Inverse Filtering

The Closed Phase Inverse Filtering (CPIF) algorithms that are presented here depend on the information during the closed phase of the cycle. When the glottis is closed, the effects of the subglottal cavities are minimal, allowing a better estimation of the vocal tract transfer function. Accordingly, these CPIF methods estimate a parametric model of the spectral envelope, which is calculated for the duration of the closed phase [46].

Major disadvantage is the imprecision in determining the closed phase period. Some approaches have tried to tackle this problem by extracting information from a EGG signal to identify this period [47].

Another approach attempted to analyze the formant frequency modulation during the glottal cycle and was focused on identifying the transitions between open and closed phases [48].

Improvements to the robustness of CPIF were also described by constraining the direct current gain [49]. A drawback of this approach is the sample rate being insufficient to determine the closed phase in high-pitched voices (with shorter periods) and, therefore, not providing an accurate filter estimation. Having this in mind, an approach was proposed where multiple glottal cycles are examined for a more accurate estimation of the closed glottal phases [50].

Finally, an approach that allows the presence of a non-zero glottal wave over closed glottal phases was also proposed [51].

3.1.1.2 Iterative adaptive inverse filtering

Iterative and adaptive algorithms improved substantially the quality of the glottal source estimation. One approach attempted the integration of the Liljencrants-Fant (LF) within the Autoregressive eXogenous (ARX) model of speech production, where the parameters for both source and filter were collectively estimated and the resultant ARXLF model estimation was optimized by iterative and adaptive processes [52, 53].

Another method consisted in finding, within a speech frame, the best candidate for a glottal waveform estimate through iterative processing without the necessity for detection of Glottal Closure Instants (GCI) [52].

Lastly, one of the most prevailing approaches is the Iterative Adaptive Inverse Filtering (IAIF), which is based on iterative enhancement of the vocal tract and source components [54]. This algorithm is a semi-automatic method that uses speech pressure signal as an input and produces an estimation of the correspondent glottal source signal. The procedure can be divided in analysis, inverse filtering and integration. An iterative process initially estimates the glottal contribution from the speech spectrum. The estimation of the glottal excitation is attained by cancelling the vocal tract effect, using inverse filtering, and the labial radiation, by integration [54].

The previous approach was later improved, showing more accurate results for high-pitched voices by replacing the Linear Predictive Coefficients (LPC) analysis with the Discrete All Poles (DAP) modelling technique [55]. This modification made possible a bias reduction related to the harmonic structure of the speech spectrum in the approximations of the formant frequencies [55].

3.1.1.3 Javkin *et al.* Method

This method relies on the assumption that speech waveforms are a combined result of the phonatory setting and the configuration of the vocal tract. Assuming that it is possible to subtract the effect of the vocal tract from the speech waveform, then the glottal waveform can be examined without requiring invasive procedures. This approach describes a frequency-domain based algorithm, since the spectral analysis conveys more information on the formants produced due to the effects of the vocal tract and the labial radiation [40].

During the open phase of the glottal cycle, formant frequency and bandwidth suffer modifications due to the vocal tract and source interactions. Therefore, the most reliable estimation of the vocal tract parameters should be obtained during the glottal closed phase, which can be obtained from the LPC residual signal. In order to eliminate the vocal tract formants and to cancel the effects of labial radiation, a digital filter with the inverse response was proposed and a model for each formant was developed [40].

3.1.2 Mixed-phase decomposition

A distinct approach for glottal flow estimation methods is based on a speech mixed-phase model [56]. These models assume that speech has a minimum phase (causal) and a maximum phase (anti-causal) constituents. The maximum-phase is related to the glottal open phase, while the minimum-phase is related to the glottal return phase and to the vocal tract impulse response [57]. The idea behind mixed-phase decomposition methods is to isolate the maximum-phase component, which describes the glottal excitation and its contribution, from the speech signal [49].

3.1.2.1 Zeros of the z-transform

The Zeros of the Z-Transform (ZZT) method assumes that speech is a mixed-phase signal. This means that the glottal open phase corresponds to the anti-causal component and both the glottal closed phase and the vocal tract filter correspond to the causal component [58]. The GCI determine the limit between these two glottal cycle phases. On one hand, the vocal tract dominated spectrum presents scarcely the glottal flow component. On the other hand, the glottal dominated spectrum shows ripples of low amplitude due to the vocal tract influence [58]. ZZT is a representation of the z-transform polynomial through its roots and that representation in case of a speech signal is equivalent to the union of the ZZT sets for the pulse, the glottal source and the vocal tract filter [58]. Once the GCI and the roots of the z-transform are obtained, the latter are divided into two sets that correspond to the causal and anti-causal part of the speech signal. By applying the DFT to each set their spectrum is achieved. Therefore, by computing the Inverse of the Discrete Fourier Transform (IDFT) of the anti-causal the estimation of the glottal source is determined [58].

3.1.2.2 Complex cepstrum

The Complex Cepstrum (CC) methodology is based on the same assumptions as the ZZT method. This method is similar to the previously mentioned, although in terms of computation time it is considered to be a faster approach [59]. The decomposition of the speech signal in this case considers that the maximum-phase component and the minimum-phase component relate, respectively, to the glottal open phase and to both the glottal closed phase and the contribution of the vocal tract. The separation of these components and the estimation of the glottal contribution is possible by only considering the negative part of the CC. Later, a new method was proposed that uses the information regarding the identification of GCI. This method takes advantage of the fact that the GCI delimits the glottal closed phase and the glottal open phase and,

therefore, the causal and anti-causal signals. The Anti-causality Dominated Region is demonstrated to approximate accurately the glottal open phase, due to the lack of contribution from the glottal closed phase and the vocal tract filter inside this region [60].

3.2 Glottal source models

The purpose of theoretical models is to describe the glottal source waveform using different analytical parametric expressions that derive from the analysis of physiological measurements [11].

3.2.1 Liljencrants-Fant model

The Liljencrants-Fant (LF) model was suggested as a five parameter model. The LF model can be divided in two parts, the first is related to estimating the glottal flow derivative with information regarding the opening phase and the second describes the closure phase. It complies with the premise that the integration of the function must be null for the complete period [61].

$$g'_{LF}(t) = \begin{cases} E_0 e^{\alpha t} \sin(w_g t) & , 0 \leq t \leq t_e \\ -\frac{E_e}{\beta t_a} (e^{-\beta(t-t_e)} - e^{\beta(t_c-t_e)}) & , t_e < t \leq t_c = T_0 \end{cases} \quad (3.1)$$

for which T_0 is the fundamental period, t_c is the time of the time-derivative minimum, t_e is the period interval, E_0 corresponds to the scale factor required for ensuring area balance, E_e to the amplitude of the negative peak, α corresponds to $C\pi$ (for which C relates to the sinusoid), the frequency of the sinusoid is related to $w_g = 2\pi F_g$ where $F_g = \frac{1}{2t_p}$ and, finally, β is a decrease constant for the exponential recovery period [42].

The LF model has shown its efficiency and is hitherto considered one the most accepted models for describing the glottal pulse. Furthermore, the LF model takes advantage of the derivative of the glottal glow to model the labial radiation, producing consequently better results and a more natural speech [45, 61]. The derivative and the glottal flow according to the LF model can be seen in Figure 3.1.

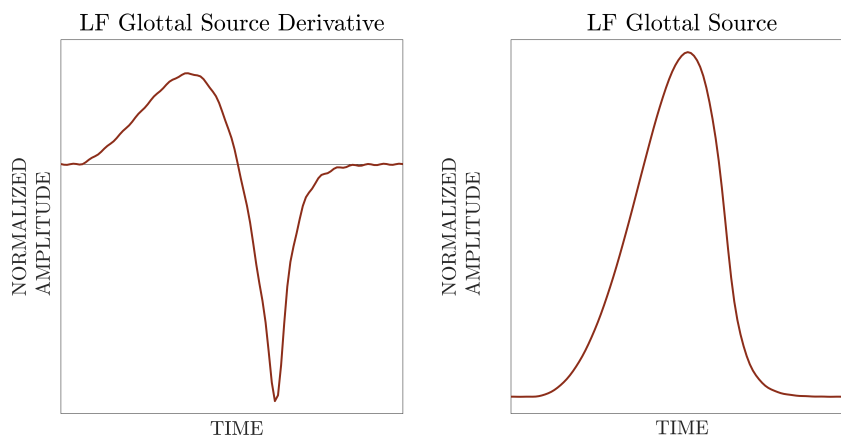


Figure 3.1: Waveshape of the glottal source derivative (left) and the glottal source (right) according to the LF model.

3.2.2 Rosenberg model

The Rosenberg Glottal Model is characterized according to different pulse amplitude, width and skew values [62]. The most popular model is described by:

$$g_R(t) = \begin{cases} \frac{A_v}{2} \left[1 - \cos\left(\frac{\pi t}{t_e}\right) \right] & , 0 \leq t \leq T_e \\ A_v \cos \frac{\pi(t-t_e)}{2t_c} & , t_e \leq t \leq T_c \\ 0 & , t_c \leq t \leq T_0 \end{cases} \quad (3.2)$$

For which, the A_v represents the peak amplitude of the glottal pulse, t_c or T_0 is the time of the time-derivative minimum and t_e is the period interval. The model defines clearly three phases concerning the glottal cycle, the opening, closing and closed phase. This model is considered to be an efficient alternative to the LF model in terms of computation time [42]. The waveshape of the glottal source according to the Rosenberg model can be seen in Figure 3.2.

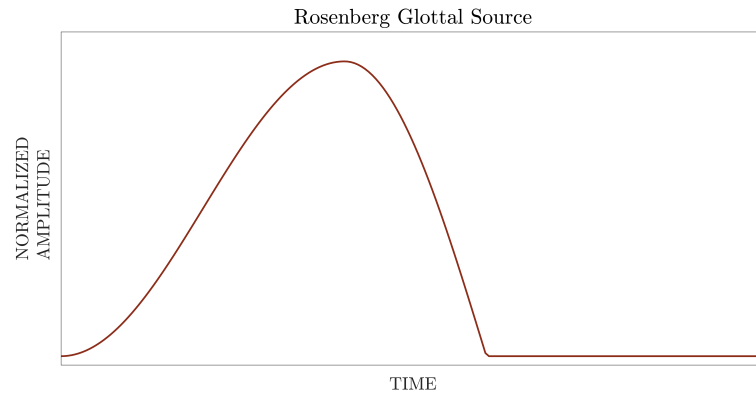


Figure 3.2: Waveshape of the glottal source according to the Rosenberg model.

3.2.3 Other models

The models that have been addressed previously are the most well-accepted by the scientific community and are usually a reference for the development of new models. However, there have been other models created during the last half century. In Table 3.1 summarises these other models. The table shows a brief description of each model, the required number of parameters and their contributions.

Table 3.1: Main theoretical glottal source models with a brief description, the number of parameters required and their improvements [11, 63].

MODEL	DESCRIPTION	# PARAMETERS	IMPROVEMENTS
Rosenberg (1971) [64]	Described by separate trigonometric functions for the opening and the closing phases of the glottal pulse	3	-
Fant (1975) [65]	Described by separate trigonometric functions for the opening and the closing and the closed phases of the glottal pulse	4	Controls the slope of the descending branch
Liljencrants-Fant (1985) [66]	Represents the first derivative of the glottal flow volume velocity pulse Described by the combination of sinusoidal and exponential functions	4	Incorporates labial radiation effects
Fujisaki-Ljungqvist (1986) [62]	Described by polynomial functions	6	Greater detail in modelling the glottal pulse shape
KLGLOTT88 (1987) [67]	Derived from the Rosenberg model	6	Considers turbulence noise generation at the glottis
Rosenberg++ (1998) [68]	Derived from the Rosenberg model, but uses the LF parameters	6	Computationally more efficient and perceptually equivalent
CALM (2003) [57]	Described in the spectral domain	5	Accounts for the mixed causal/anticausal phase behavior of the source
EE1 (2010) [69]	Described by the combination of sinusoidal and exponential functions	5	Ability to adjust the opening and closing phases slopes separately
EE2 (2012) [70]	Redefines EE1 parameters (speed of opening and speed of closing)	6	Lower computational complexity, faster generation and more accurate pulse shape manipulation

Among these is the Rosenberg++ model, which requires less computational time and still maintaining its efficiency in producing good results in perceptual terms, when compared to the LF model [68]. The Fant model was an earlier presented model that worked with independent parameters, three of them related with frequency, amplitude and the exponential growth constant of a sinusoid. Another parameter was added regarding the exponential recovery time constant of the return phase [66]. This model ensures that all found waveforms are fitted with the least amount of parameters and is compliant to meet unusual waveshapes [66]. The Klatt model describes the glottal pulse characteristics with two simple features such as the fundamental frequency or the pulse peak amplitude [67]. The Causal-Anticausal Linear Model (CALM) characterizes the glottal signal in the spectral domain, by using the glottal peak and the spectral slope [57].

3.3 Glottal source parameterization

3.3.1 Time domain features

From the glottal waveform it is possible to extract time-domain features, which characterize its shape [11]. The waveform can be divided into different phases of the glottal cycle, as shown in Figure 3.3, from which relevant instants can be noted, such as the glottal opening instant or the glottal closing instant. These are used for accounting the glottal flow pulse or measuring the duration of each phase [71].

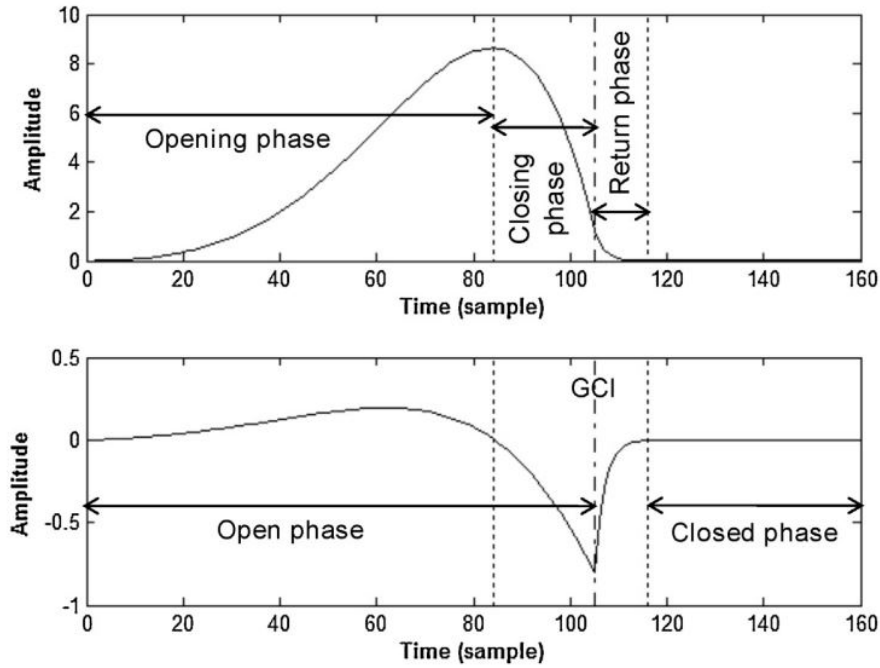


Figure 3.3: Illustrative example of the waveshape of the glottal source (top) and the glottal source derivative (bottom), where are depicted the different phases and events of the glottal cycle [59].

The LF model is created using time-domain parameters, such as the Open Quotient (OQ), $OQ = \frac{T_c}{T_0}$, the Asymmetry coefficient, $\alpha_m = \frac{T_z}{T_c}$, and the Voice Speed Quotient (SQ), $SQ = \frac{T_z}{T_c - T_z}$ [71, 44].

In order to overcome a major difficulty related with locating accurately the relevant instants, the Quasi-Open Quotient (QOQ), $QOQ = \frac{T_2 - T_1}{T_0}$ is used, that describes the relative glottal open phase [71].

Parameters can also be obtained from the amplitude of peaks of the glottal pulse or its derivative [72].

In 1995, a glottal feature was proposed that characterizes the glottal closing phase that correlates with voice quality, the Basic Shape Parameter [73].

Later, the Normalized Amplitude Quotient (NAQ), $NAQ = \frac{A_{max} - A_{min}}{D_{min} \cdot T_0}$, a similar parameter was developed that relates the glottal flow maximum and its derivative minimum [74].

3.3.2 Frequency domain features

Features with relevant information can be extracted from the glottal flow signal spectral content. Only recently, these features have been studied more thoroughly, since it used to be time-consuming, contaminated with artifacts and dependent on other time-domain features [59]. However, these computational challenges have been overcome during the most recent years and several frequency-domain features have been described.

From the spectrum of the LF glottal source model one can distinguish the glottal formant, a low-frequency resonance, that can be described by both frequency and bandwidth [71, 59].

The Parabolic Spectrum Parameter (PSP) is based on a spectral error measure that results from the comparison of a model with the estimated glottal flow [75]. This estimates the spectral slope by fitting a second-order polynomial to the source flow spectrum [76].

The difference between amplitude of the fundamental frequency and the second harmonic from the glottal source magnitude spectrum, $H_1 - H_2$, is also a common parameter used to describe the glottal flow [77].

The Harmonic to Noise Ratio (HNR) and the Harmonic Richness Factor, $HRF = \frac{\sum_{k \geq 2} H_k}{H_1}$ are two parameters used for assessing the amount of harmonics in the glottal source spectrum. The latter, specifically, calculates the number of harmonics present in the glottal source magnitude spectrum and corresponds to the ratio of the sum of the k harmonics and the fundamental frequency amplitudes [78, 79, 80].

3.4 Summary

The review made on glottal source models and its estimation reveals that there are still limitations and challenges that prevent these methods to describe and characterize accurately the real glottal source. From this chapter, we conclude that an empirical model built from real acoustic pressure signals and its characterization would improve the understanding of speech production processes and the performance of speech applications that rely on theoretical models.

The following chapter describes the experimental procedure for the dataset acquisition, as well as preliminary analysis and perceptual tests carried out in order to understand the content carried by the glottal source signals.

Chapter 4

Data Acquisition and Dataset Characterization

THE creation of a reliable dataset is required for further characterizing the glottal source signal and, therefore, a procedure was designed for the acquisition of the real glottal source signal. In this chapter the data acquisition process is explained and a description of the equipment used is given, followed by a characterization of the dataset created.

4.1 Signal acquisition

The acquisition of these signals was performed under controlled conditions in the Otorhinolaryngology (ORL) Department of Centro Hospitalar e Universitário de São João (CHUSJ) by two otorhinolaryngology specialists, Dr. Jorge Spratley (Co-Supervisor) and Dr. Laurentino Mendes. The procedure used the following equipment for the signal acquisition (specifications can be found in Appendix A.1 and Appendix A.2):

- two high quality microphones with reduced dimensions (B6 Omnidirectional Lavalier);
- a 128 kHz USB audio/MIDI interface (Focusrite) with 2 stereo channels;
- two phantom power adaptors (MZA 900 P);
- a flexible rhyno-laryngo fiberscope (ENF-XP OLYMPUS);
- a nasogastric tube (6mm diameter).

The signal acquisition consisted of a normal Fiber Naso-Pharyngo-Laryngoscopy (FNPL) exam with a larger than usual working channel in order to accommodate the internal microphone, which was placed at a distance of approximately 1 cm away from the vocal folds. The external microphone was placed using an earpiece adapter to fixate its location approximately 5 cm away from the mouth. This procedure is illustrated in Figure 4.1 where the internal microphone is represented in blue and the external microphone is represented in orange. A similar approach was already attempted for glottal source signal acquisition in [81]. However, the use of different microphones for the acquisition of each signal lead us to believe that their results are questionable.

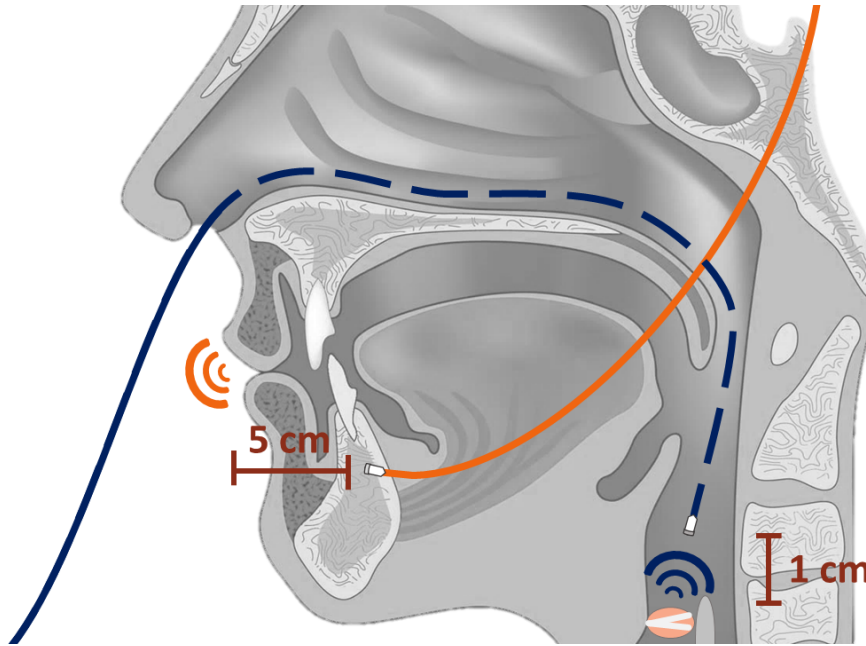


Figure 4.1: Scheme of the positioning for the internal microphone (blue) with an approximated distance to the vocal folds and the external microphone (orange) with an approximated distance to the mouth.

The patients were recruited and treated according to the ethical diligences required by the Comissão de Ética para a Saúde (CES) of CHSJ. A document with the information regarding the procedure was handed out to each participant and an informed consent document filled in by the participants. Both the document with the information regarding the procedure and the approval document signed by CES can be found in Appendix A.3.

The inclusion criteria used for this study regarding the participants were:

- having at least 18 years of age;
- leading a healthy lifestyle, e.g. non-smoker;
- absent history of voice disorders;
- showing viability for the procedure after anterior rhinoscopy inspection.

The choice of the three sustained vowels (/a/, /i/ and /u/) has to do with the fact that these vowels represent the extreme values for the first two formants (F1 and F2) for the acoustic triangle formed by the oral tonic vowels in the standard European Portuguese (EP) [82], as shown in Figure 4.2. This choice is important since the second part of this work consists in the estimation and modelling of the vocal tract filter. Therefore, it is convenient to have a set of vowels covering the associated formant frequency range.

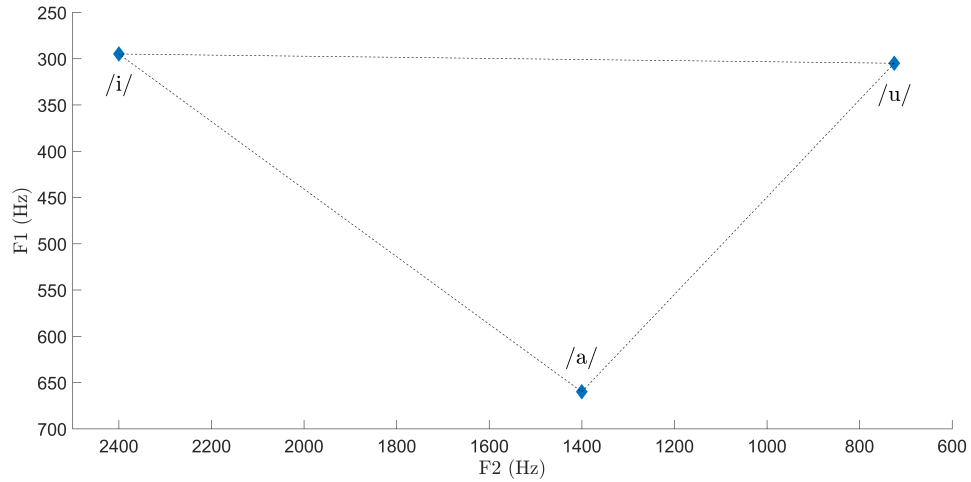


Figure 4.2: Illustration of the acoustic triangle for the oral tonic vowels $/a/$, $/i/$ and $/u/$ of the standard EP. Adapted from [83].

The EGG signals and their time aligned speech signals were collected for future reference and to be used as the ground-truth for events of the glottal cycle. However, in this dissertation these signals were not carefully analysed due to time shortage.

4.2 Dataset description

The database comprises the recordings of six healthy speakers, three males and three females, as characterized in Table 4.1.

Table 4.1: Characterization of the 6 volunteer speakers regarding gender and age.

Speaker	Gender	Age
1	Male	25
2	Female	23
3	Female	19
4	Female	22
5	Male	27
6	Male	22

Each acquisition file contains the recording of two time-aligned signals of three sustained vowels ($/a/$, $/i/$ and $/u/$), as shown in Figure 4.3. The vowels were uttered in the most natural way as possible and two time-aligned signals were acquired: one signal recorded externally, close to the mouth, and one recorded internally, near the vocal folds.

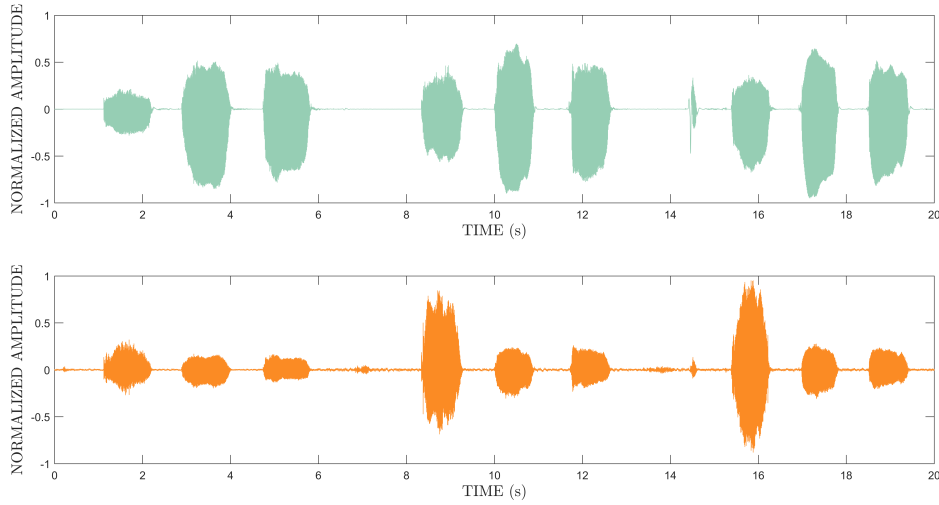


Figure 4.3: Example of the time-aligned stereo recording of speaker 6 and containing three repetitions of the three sustained vowels (/a/, /i/ and /u/). The upper figure corresponds to the signal recorded near the vocal folds and the lower figure corresponds to the signal recorded outside close to the mouth.

The internal and the external microphone signals were recorded, respectively, as the Left and the Right channel of a stereo time-aligned signal and subsequently separated for individual analysis. These individual signals were further segmented to isolate relevant portions of the uttered vowels for detailed analysis regarding time, spectral magnitude, and spectral phase structure. This segmentation resulted in a total of 108 separate files, 54 for the external signals and 54 for the internal signals.

4.3 Preliminary Analysis

The waveshape typically observed for a signal collected from the glottal source and the waveshape of its corresponding speech signal is represented, respectively, in the upper and lower part of Figure 4.4.

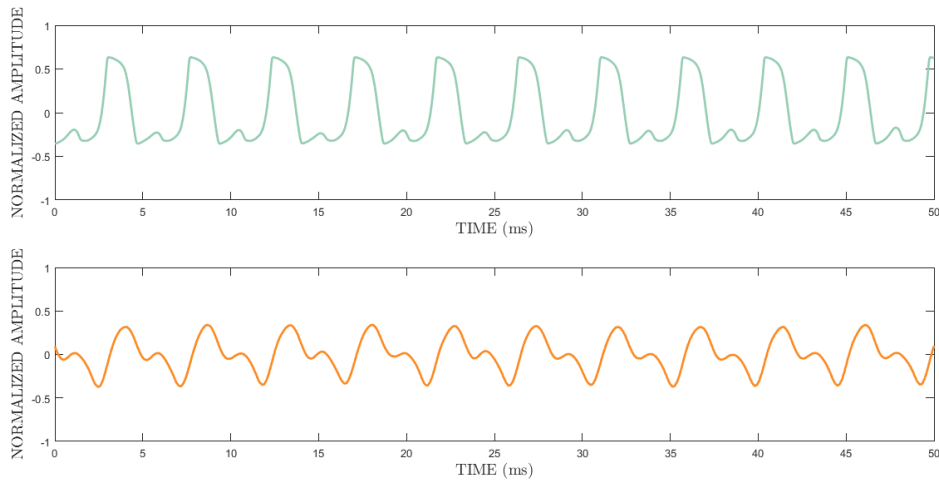


Figure 4.4: Shape of the acoustic pressure signals, the upper signal recorded near the vocal folds and the bottom signal recorded close to the mouth of speaker 3 for the sustained vowel /i/.

A preliminary analysis of the recordings lead to the conclusion that some signals had one or more additional impulses per cycle in the internal signals, such as the one illustrated in the upper part of Figure 4.5. In order to avoid compromising the analysis, only the signals obtained from speakers 2, 3 and 4 were considered in the subsequent analysis. It is important to note, however, that the signals which showed an absence of additional impulses corresponded exclusively to signals collected from female speakers. This phenomenon was first mentioned by Timcke, who describes it as multiphasic patterns of vibration [84]. This most likely occurs due to the asynchronous behaviour of the vocal folds, which leads to an irregular flow passage and results in a combined signal. Since this is observed in most of the cases for the male speakers, it is possible that the length of the vocal folds is related to this behavioural characteristic.

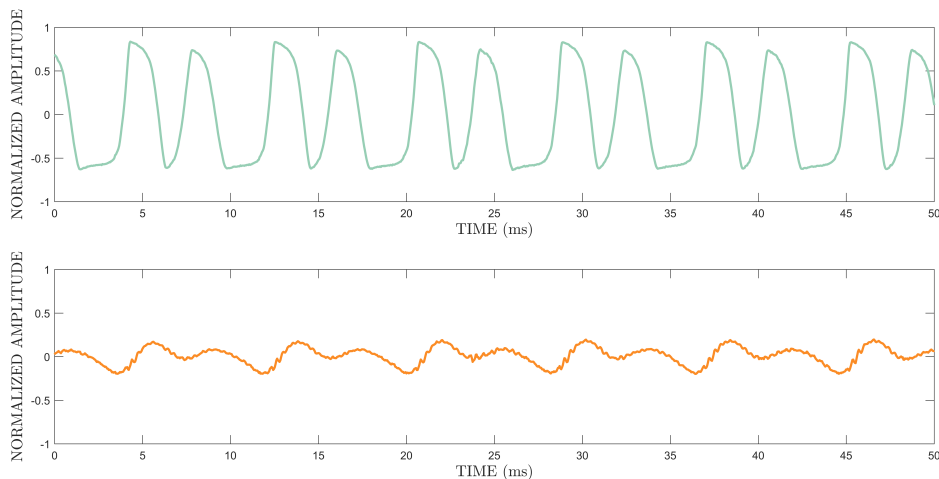


Figure 4.5: Shape of the acoustic pressure signals, the upper signal recorded near the vocal folds showing clear influence of two distinct signals and the bottom signal recorded close to the mouth of speaker 5 (male) for the sustained vowel /i/.

4.4 Perceptual tests

Perceptual tests were carried out in order to obtain a better understanding of the internal signals linguistic content. For this purpose, 29 participants were recruited to perform informal tests, under the recommendation of using headphones, and they were required to identify 18 different samples (6 for each vowel) of the internal acoustic pressure signal recordings collected from 6 different speakers. Specifically, after listening to a recorded internal signal (i.e. near the vocal folds), subjects were asked to identify their correspondence to a known Portuguese vowel. In the first part of the test, the participants were given the chance to choose among all the 9 main oral vowels in EP (/à/, /â/, /e/, /é/, /i/, /ê/, /ó/, /ô/ and /u/). In the second part, the choice was limited to the vowels which were recorded for this study (/a/, /i/ and /u/). The perceptual tests results are displayed in Figure 4.6 according to repetition and divided in the two parts of the test. The 95% Confidence Intervals (CIs) were computed using the Adjusted Wald method for binary data [85].

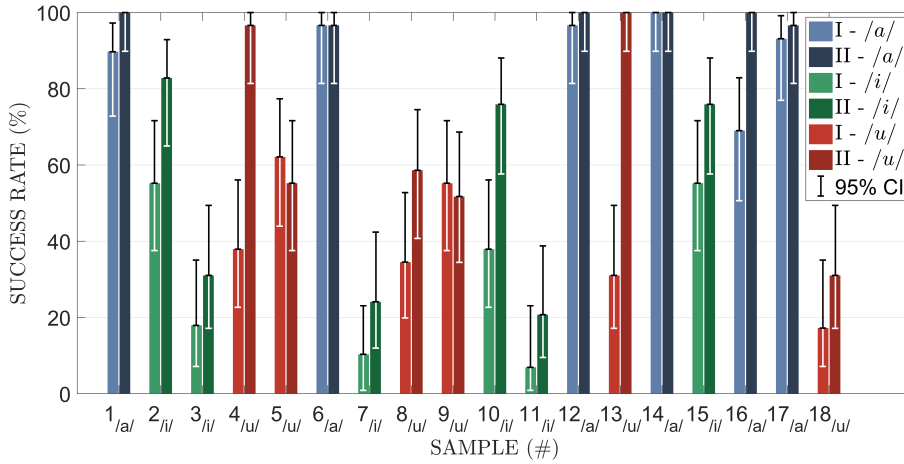


Figure 4.6: Results obtained for the perceptual tests describing the percent correct identification of the signals given. The lighter bars correspond to the first part and the darker to the second part of the perceptual tests.

The first conclusion that emerges from the results in Figure 4.6 and in Table 4.2 is that the success in the vowel identification tests improves from the first to the second test. This is an expected outcome since the diversity of options in the first case is much higher than in the second case (9 versus 3). It can be concluded that between the first and the second part of the test neither the results for the correct identification of the vowel /a/ (from 90.83% to 98.87%) nor the identification of the vowel /u/ (from 39.65% to 65.52%) showed a statistically significant difference with p-values of 0.1581 and 0.1021, respectively. On the other hand, the results obtained for the correct identification of the vowel /i/ (from 30.57% to 51.73%) showed a statistically significant difference with a p-value of 0.0035. The p-values were obtained using the MATLAB function `ttest2()`.

Table 4.2: Results for both parts of the perceptual tests where the values shown correspond to the success rate in identifying the vowel recorded. The p-values obtained regarding the statistical difference between results for Part I and Part II are also depicted.

	<u>/a/</u>	<u>/i/</u>	<u>/u/</u>
PART I	90.83%	30.57%	39.65%
PART II	98.87%	51.73%	65.52%
p-VALUE	<u>0.1581</u>	<u>0.0035</u>	<u>0.1021</u>

Additionally, participants could recognize effectively the vowel /a/, while having more difficulty in recognizing the vowels /i/ and /u/. Specifically, the vowel /i/ was often misidentified as being the vowel /u/, which suggests that the signals recorded for both vowels are very similar in terms of linguistic content. Moreover, this could be explained by the fact that the internal signals for the vowels /i/ and /u/ suffer less influence from the vocal tract filter when compared to the internal signals recorded for the vowel /a/.

4.5 Summary

This chapter outlined the acquisition procedure for glottal source signal. The dataset which will be used throughout this work for the characterization of the glottal source signal was described. This will also be used as the ground truth for the vocal tract filter estimation. Perceptual tests were performed for determining whether the glottal source signals had linguistic content. The results implied that the signals acquired for vowel /a/ suffered the most from the influence of the vocal tract and that the signals acquired for vowel /i/ suffered the least, showing the lowest success rates of **30.57%** and **51.73%**, respectively, for the first and second parts with a statistically significant difference of **0.0035**. The next chapter will explain how these signals were analyzed and the parameters used to characterize them.

Chapter 5

Glottal Source Characterization

THEORETICAL models have provided enough information for speech analysis and synthesis applications hitherto. However, a better understanding of the glottal source could potentially improve the performance of voice technology applications with a more efficient and simple approach [44]. Consequently, an accurate glottal source modelling and its estimation is a crucial task in the field of voice processing. By accessing to the acoustic pressure signal closer to the vocal folds, it is expected a better and more direct study of the glottal source signal. Moreover, conclusions may be drawn about the relationship between the glottal source signal and the corresponding speech signal by comparing the content of the internal signal to its external counterpart. In this chapter, the approach used for analysing the spectral magnitude and phase content is described with the purpose of characterizing the glottal source and to compare it with the theoretical reference models described in Chapter 3.

5.1 Parametric Spectral Analysis

A periodic signal can be decomposed in sinusoidal components that contribute to its harmonic structure [86]. In order to perform a spectral analysis, the time domain signals must first be converted to their frequency domain, so that they can be described according to their magnitude, $|X(e^{j\omega})|$, and phase, $\angle X(e^{j\omega})$ [87]. Since the work is focused on the voiced speech, in particular with sustained vowels, we may assume quasi-stationary conditions. Therefore, by analysing the magnitude spectrum $|X(e^{j\omega})|$ of any given portion of the signal, this portion is expected to be constant to time-shift. However, when analysing the phase spectrum $\angle X(e^{j\omega})$ the opposite scenario is expected. The time-shift influences the frequency components of the signal and, consequently, its absolute phases. Nevertheless, a time-shift independent phase representation would provide valuable information for describing a periodic signal, which is characterized by its unaltered waveshape. For this reason, a phase-related feature was used that allows a time-invariant phase representation.

The block diagram represented in Figure 5.1 describes the DyNaVoiceR framework, where ℓ denotes the harmonic index and the fundamental frequency is denoted by ω_0 . The harmonic magnitude and phase values are represented by A_ℓ and ϕ_ℓ . Each discrete-time signal, $x[n]$, was segmented using a sine window, $h[n]$, with the length of 1024 samples and using a 50% overlap between adjacent frames [88]. This overlap-add analysis takes advantage of the combination of

this discrete-time analysis window with the Odd-Frequency Discrete Fourier Transform (ODFT) and their respective properties for the computation and of the power spectrum [89].

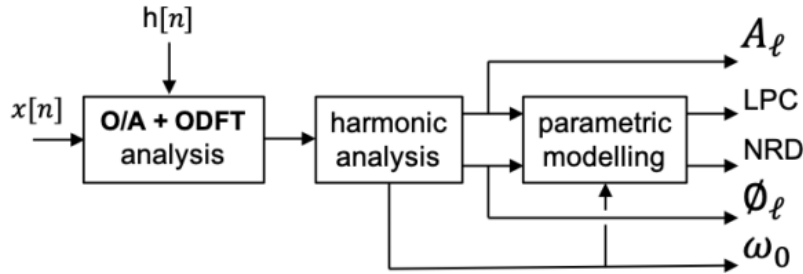


Figure 5.1: Block diagram with the main steps of the analysis and parametric modelling performed by the DyNaVoiceR framework, where the LPC represents the spectral magnitude and the NRD represents the spectral phase of a periodic signal [90].

The harmonic analysis performed by the DyNaVoiceR framework extracts the spectral content from the power spectrum of the selected acoustic signals, as shown in Figure 5.2. From this analysis, accurate values of the respective frequencies, magnitudes and phases of all detected harmonics were estimated [88, 89, 91].

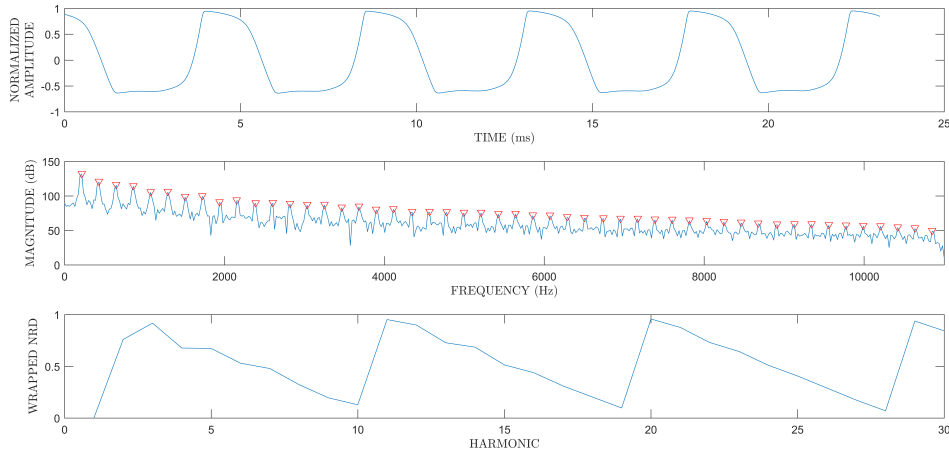


Figure 5.2: Typical waveform of the internal acoustic pressure signal uttered by Speaker 3 for vowel /i/ (top), respective magnitude power spectrum (mid) and the respective harmonic structure alongside with the estimated peaks of the harmonic (red triangles) and wrapped NRD representation (bottom).

5.1.1 Spectral magnitude analysis

The spectral magnitude analysis provides accurate information regarding the peaks of the harmonics, concerning their frequencies and magnitude values.

The magnitude modelling is obtained by using an all-pole (LPC) model of 22nd order, which is appropriate for the diversity of Portuguese oral and nasal vowels and for signals recorded with a sampling frequency of 22050 Hz.

This information is obtained by firstly estimating the Power Spectral Density (PSD) through interpolation of the magnitude values for all the harmonic peaks in a dB scale. Subsequently, the autocorrelation coefficients are calculated according to the Wiener-Khintchine theorem and followed by the computation of the LPC model by employing the Levinson-Durbin recursion [87, 92].

A representation of the magnitude spectrum of a glottal source signal frame for repetition 3 of a sustained /i/ vowel uttered by speaker 3 is shown in Figure 5.3. The peaks of the harmonics are identified by the red triangles and these correspond to their accurately estimated frequency and magnitude values.

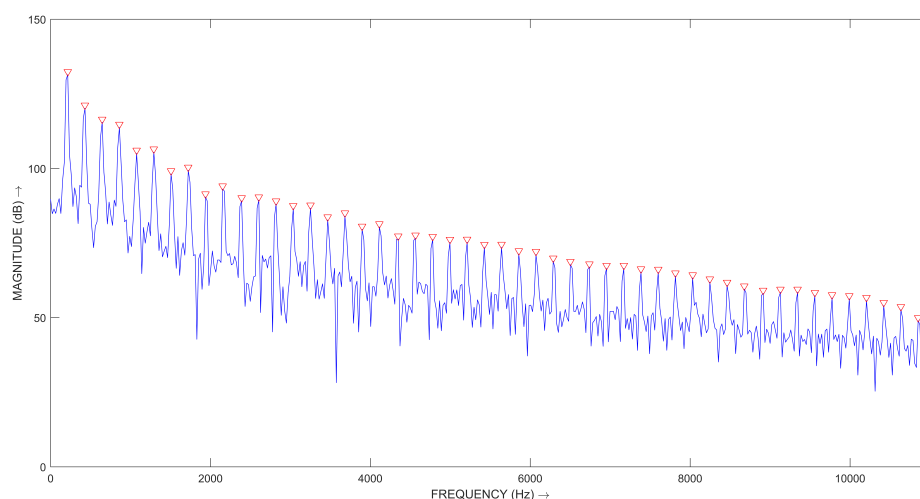


Figure 5.3: Magnitude power spectrum and the respective harmonic structure alongside with the estimated peaks of the harmonic (red triangles) for vowel /i/ from speaker 3.

The natural decay of the glottal pulse spectral magnitudes is used for characterizing the glottal source signal in the frequency domain and can be described by the spectral magnitude slope [93]. For the spectral magnitude analysis, an estimation of the spectral slope was computed for all speakers, considering only the 7 repetitions of the vowel /i/ and 8 repetitions of the vowel /u/. The remaining two repetitions of the vowel /i/ and another one of the vowel /u/ were disregarded from this study, since these segments showed signs of clipping. Conclusions could not be drawn for the vowel /a/, since the effect of vocal tract was evident in the internal signals, as shown in Figure 5.4, and these were disregarded from the spectral magnitude study. Whenever the titles are composed by, *pacX₁_22050_regX₂_X₃*, the figure being analysed corresponds to the file containing the signal recorded using a sampling frequency of 22050 Hz from speaker number *X₁* for the repetition number *X₂* of the uttered *X₃* vowel.

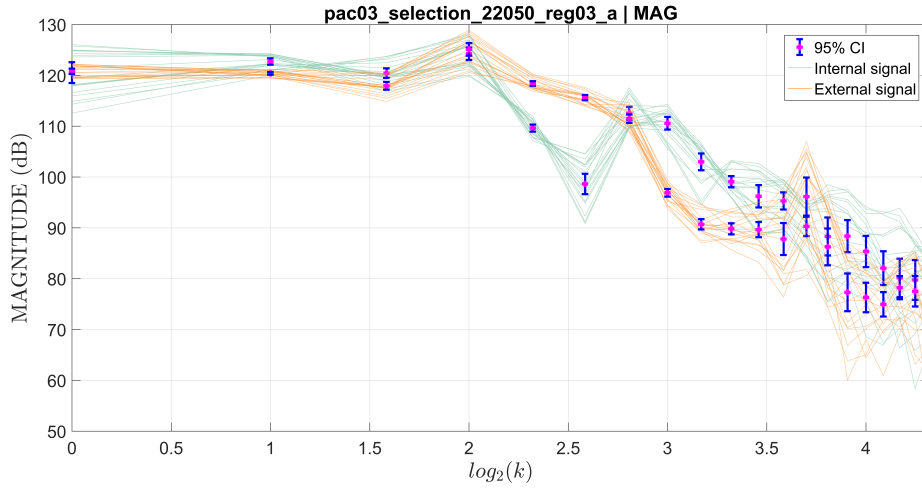


Figure 5.4: Individual frames spectral magnitudes of the internal (blue) and external (orange) signals, and their respective 95% CIs, obtained from repetition 3 of speaker 3 uttering the sustained vowel /a/.

Excluding the first and last seven frames (in order to remove the outliers created by the analysis of the signal's transition regions), this study only considered the frames where a minimum of 20 harmonics were detected. The spectral magnitude slope was computed by fitting a linear model to the mean magnitude values (on a dB scale) of the 19 harmonics above the fundamental frequency of all frames considered for each segment. The linear model was obtained using a logarithmic scale in order to obtain a better fit to the magnitude values, similarly to the spectral magnitude slope values described in the literature [94, 95]. Hence, the spectral magnitude slopes were computed in dB per octave for comparison with the reference values of the theoretical models. For a better visual analysis, figures were generated with a semi-logarithmic scale for all the segments illustrating the magnitude values for each frame, the magnitude means and their respective 95% CIs for each harmonic, as exemplified in Figure 5.5. The figures obtained for the remaining signals are available in Appendix B.1.

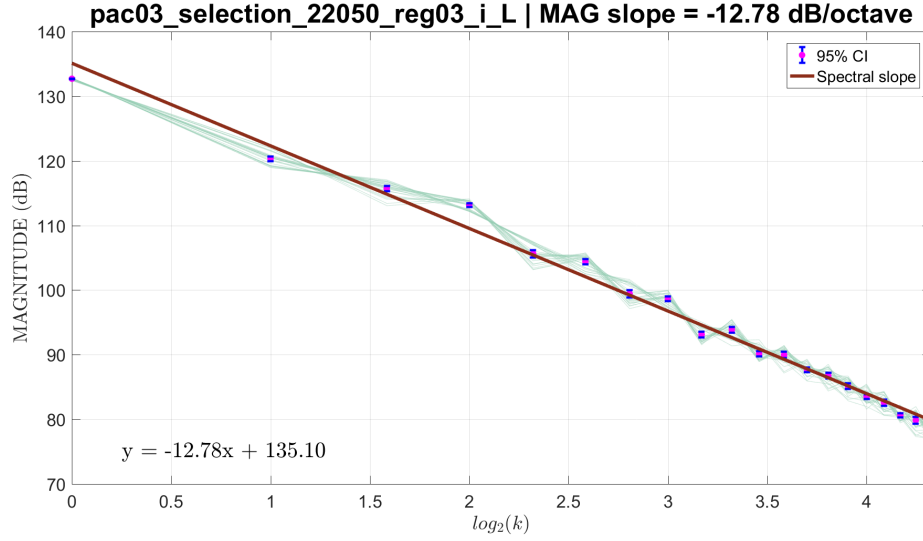


Figure 5.5: Mean spectral magnitude slope (brown) and individual frames spectral magnitudes (blue) and their respective 95% CIs, obtained from repetition 3 of speaker 3 uttering the sustained vowel /i/.

The spectral decay values obtained for each vowel considered in the analysis, /i/ and /u/, are presented in Table 5.1, identified by the respective segment and speaker. Generally, the spectral decay is consistent between repetitions for each specific speaker and vowel.

Table 5.1: Spectral magnitude slope values obtained for different repetitions of each sustained vowel (/i/ and /u/) according to the speaker. The values below correspond to the mean value ($\bar{x}$) and to the standard deviation (σ). The bottom values correspond to the mean value of all spectral magnitude slope values.

Speaker	Repetition #	/i/	/u/
2	1	-14.52	-13.61
	2	-13.06	-12.66
	3	-13.90	-12.78
3	1	-	-
	2	-12.72	-13.06
	3	-12.78	-12.30
4	1	-	-13.94
	2	-11.99	-13.36
	3	-11.09	-13.40
$\bar{x}$		-12.85	-13.15
σ		± 0.83	± 0.45
$\bar{x}/i/,/u/$		-13.01	

The mean spectral magnitude slope values obtained for the vowel /i/ and the vowel /u/ correspond, respectively, to **-12.85 $\pm$ 0.83 dB/oct** and **-13.15 $\pm$ 0.45 dB/oct**. It is noticeable that the mean slope value for the vowel /u/ is relatively higher than the mean slope value for the vowel /i/ and the standard deviation confirms the consistency of these values. Furthermore, as Table 5.1 shows, both the average slope value for the vowel /i/ and vowel /u/, as well as the

average value of **-13.01** dB/oct for both values, fall in between the reference values of **-12 dB/oct** and **-16 dB/oct**, given by the Rosenberg model and the LF model, respectively.

The experimental values obtained for the spectral magnitude are highly congruent as shown by their respective 95% CIs, depicted in Figure 5.6 and Figure 5.7. The characterization of this empirical model obtained from the spectral magnitude of all the glottal source signals recorded constitutes one of the goals of this work.

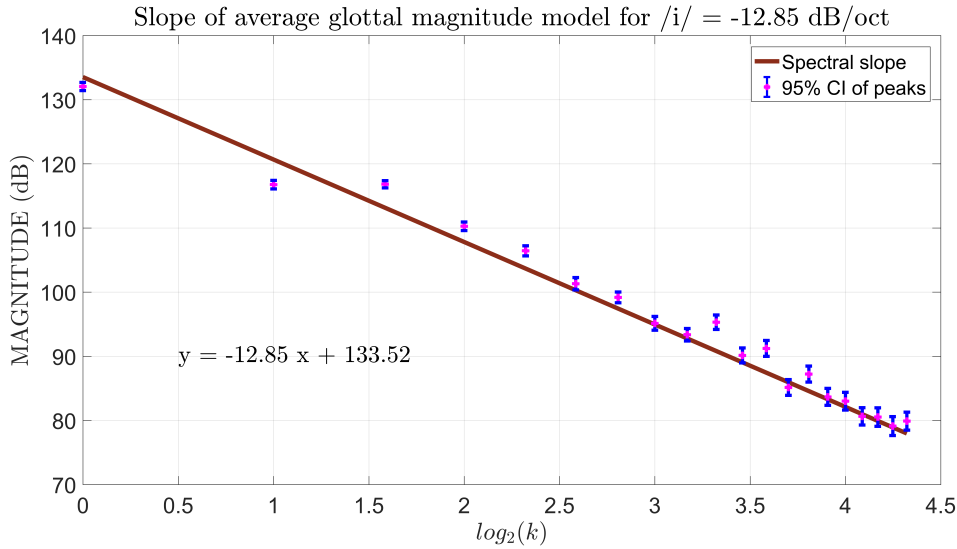


Figure 5.6: Linear regression model that represents the mean spectral magnitude found for all repetitions collected from different speakers uttering the sustained vowel /i/.

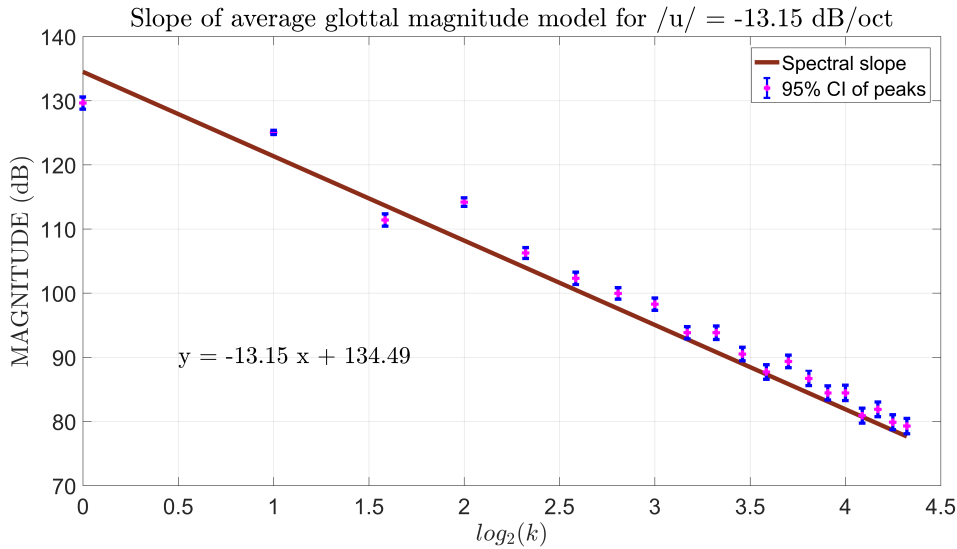


Figure 5.7: Linear regression model that represents the mean spectral magnitude found for all repetitions collected from different speakers uttering the sustained vowel /u/.

The models obtained empirically, as described above, are able to fit the data correctly for any speaker, regardless of the chosen repetition, as illustrated by the example in Figure 5.8 example, corresponding the vowel /i/ of repetition 3 from speaker 3, where the mean magnitude values

of each harmonic are represented with the respective 95% CIs, along with the mean spectral magnitude slope model previously obtained for the vowel /i/.

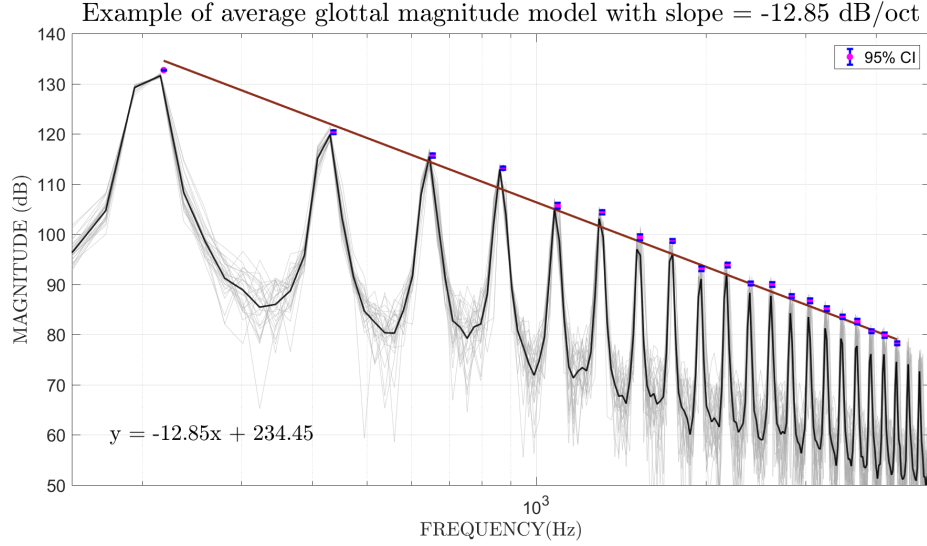


Figure 5.8: Spectral magnitude mean and the representation of the magnitude power spectra for all frames and its corresponding mean for repetition 3 from speaker 3 uttering the sustained vowel /i/.

5.1.2 Spectral phase analysis

In order to describe a periodic signal, it would be necessary a phase-related feature that allowed time-invariant representation. According to Equation 5.1, when describing a periodic signal as an harmonic discrete-time signal, where L corresponds to the number of the magnitude spectrum harmonics below the Nyquist Frequency, the signal phase structure includes L starting (or initial) phase values (ϕ_ℓ).

$$x[n] = \sum_{\ell=0}^{L-1} A_\ell \sin(n\omega_\ell + \phi_\ell) \quad (5.1)$$

However, the phase structure mentioned earlier is shift dependent and, therefore, by shifting the waveform in n samples, the phase values are altered. However, when dealing with harmonic signals, $\omega_\ell = (\ell + 1)\omega_0$, the equation that describes a discrete-time signal can be represented as Equation 5.2, where the phase structure is portrayed as a time-shift invariant feature [96].

$$x[n] = \sum_{\ell=0}^{L-1} A_\ell \sin[(\ell + 1)(n\omega_\ell + \phi_\ell) + 2\pi\text{NRD}_\ell] \quad (5.2)$$

The Normalized Relative Delay (NRD) can be obtained according to Equation 5.3. The NRD feature can be represented in a wrapped or unwrapped format and ranges from 0 to 1 for each harmonic, since it represents the normalized phase. This work focused on the latter in order

to facilitate interpretation, which corresponds to a smooth and more easily interpretable phase-related feature [97].

$$\mathbf{NRD}_\ell = \frac{\phi_\ell - (\ell + 1)\phi_0}{2\pi}, \quad \ell = 0, 1, \dots, L - 1 \quad (5.3)$$

When dealing with a periodic signal, this normalized phase-related feature represents the phase delay for each individual harmonic in relation to the fundamental frequency and represents the phase contribution to its shape invariance [98]. For most cases, the unwrapped NRD vectors were prone to follow a linear tendency. Having this in mind, the behaviour of the unwrapped NRD was approximated by a linear regression model of the 19 harmonics above the fundamental frequency in each ODFT frame of the signal. Similarly to the previous study, the first and last seven frames were excluded (in order to remove the outliers created by the analysis of the fade in and fade out regions) and the frames considered were the ones with a minimum of 20 harmonics detected. Following this procedure, the NRD slope was obtained, both for internal and external signals, among all speakers, for 14 signals regarding vowel /a/, for 14 signals regarding the vowel /i/ and 16 signals regarding vowel /u/. The experimental values that were obtained for the NRD models were plotted alongside with the NRD vectors, their respective mean and 95% CIs for each harmonic regarding each repetition. Two examples are given in Figure 5.9 (internal signal) and Figure 5.10 (external signal). The remaining results are available in Appendix B.2.

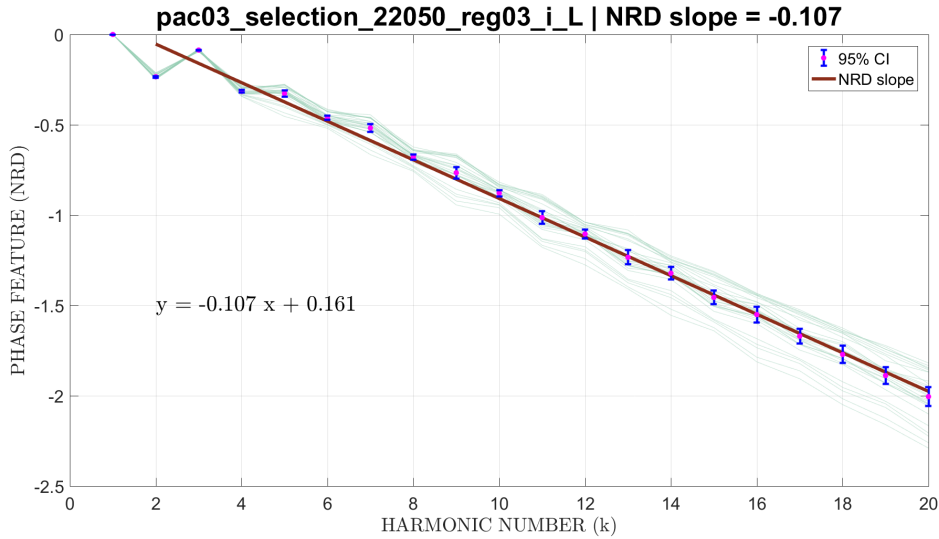


Figure 5.9: Spectral NRD slope mean, the unwrapped NRD values (blue), and the corresponding mean and the 95% CIs of the internal signal for repetition 3 from speaker 3 uttering the sustained vowel /i/.

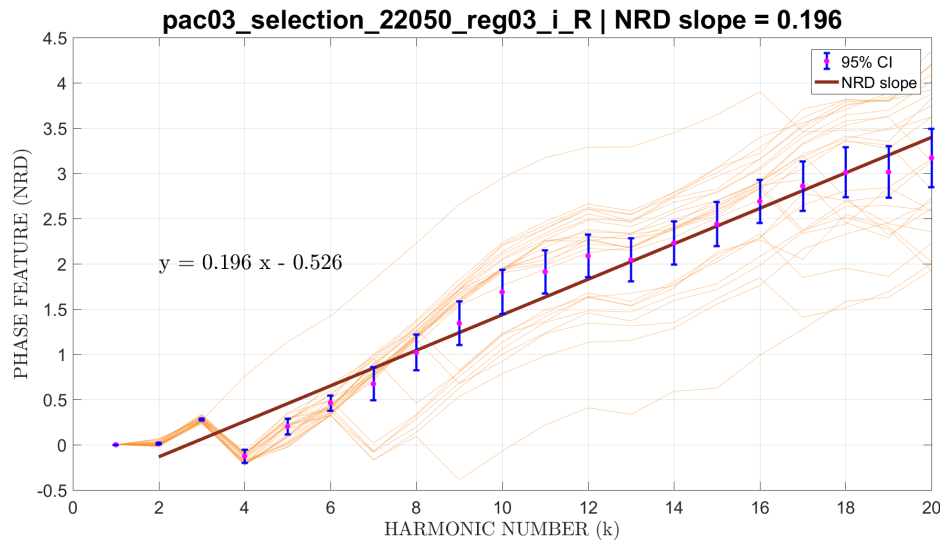


Figure 5.10: Spectral NRD slope mean, the unwrapped NRD values (orange), its corresponding mean and the 95% CIs of the external signal for repetition 3 from speaker 3 uttering the sustained vowel /i/.

These results showed that the external signals regarding all vowels had a positive NRD slope, while only the internal signals regarding the vowels /i/ and /u/ had a negative NRD slope. The internal signals regarding the vowel /a/ showed a positive NRD slope similarly to the external signals. This reassures that the internal signals are not inverted due to its acquisition conditions. Moreover, taking into account the fact that all the internal signals were recorded at the same distance from the vocal folds, the positive polarity of the vowel /a/ suggests that there is an earlier influence of the vocal tract in this case. This hypothesis is supported by the results of the perceptual tests carried out to evaluate the linguistic content of the internal signals, where vowel /a/ had the highest rate for successful identification. Table 5.2 shows the slope values obtained for the NRD models for each vowel considered for this study (/a/, /i/ and /u/) according to the speaker and repetition.

Table 5.2: Spectral NRD slope values for different repetitions of each sustained vowel (/a/, /i/ and /u/) according to the speaker. The mean spectral NRD slope values ($\bar{x}$) and the standard deviation (σ) are also depicted according to vowel and location, alongside with the difference values between the internal and the external NRD means according to vowel.

Speaker	Repetition #	/a/		/i/		/u/	
		L	R	L	R	L	R
2	1	0.084	0.077	-0.084	0.131	-0.119	0.031
	2	0.092	0.067	-0.029	0.132	-0.081	0.046
	3	0.101	0.098	-0.089	0.185	-0.123	0.138
3	1	-	-	-	-	-	-
	2	0.133	0.231	-0.077	0.161	-0.117	0.046
	3	0.106	0.186	-0.107	0.196	-0.147	0.078
4	1	-	-	-	-	-0.029	0.140
	2	0.148	0.169	-0.074	0.199	-0.058	0.124
	3	0.160	0.183	-0.037	0.156	-0.068	0.146
$\bar{x}$		0.118	0.144	-0.071	0.166	-0.093	0.094
σ		± 0.024	± 0.055	± 0.022	± 0.023	± 0.034	± 0.043
$\bar{x}_R - \bar{x}_L$		0.026		0.238		0.187	

The analysis of the results in Table 5.2 indicates that these unwrapped NRD slope values are highly vowel dependent. The internal signals presented values around **0.12** for vowel /a/, **-0.07** for vowel /i/ and **-0.09** for vowel /u/. The external signals seem to have a steeper slope with values around **0.14** for vowel /a/ and **0.17** for vowel /i/, except for the slope value of **0.09** found for vowel /u/.

It should be noted that the average NRD slope value found for the internal signal of vowel /a/ of **0.12** has a similar value to the NRD slope reference value of **0.09** given by the LF glottal source model [42]. Additionally, the average NRD slope values found for the external signals of vowels /a/ and /i/ of, respectively, **0.14** and **0.17**, replicate approximately the NRD slope obtained from experimental human data of approximately **0.15** [96].

The NRD values and models were reproduced graphically for both signals in Figure 5.11 for a better visual perception of the NRD relation between internal and external signals according to the respective repetition. The remaining illustrations are available in Appendix B.2.

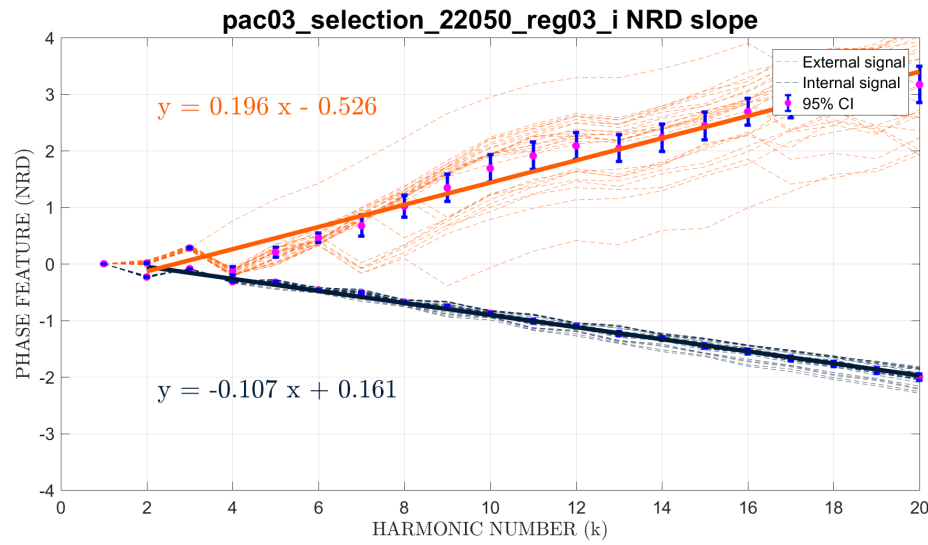


Figure 5.11: Spectral NRD slope mean found for the internal signal (blue) and the external signal (orange). The unwrapped NRD for the internal signal (blue) and the external signal (orange) are depicted, its corresponding mean and the 95% CIs for repetition 3 from speaker 3 uttering the sustained vowel /i/.

A relation was obtained for the difference between the external NRD mean vectors and the internal NRD mean vectors for time-aligned signals. For that reason, the NRD means were represented alongside with the difference between them and the equation of the linear regression model that describes its development, as shown in Figure 5.12. The remaining illustrations can be found in Appendix B.2.

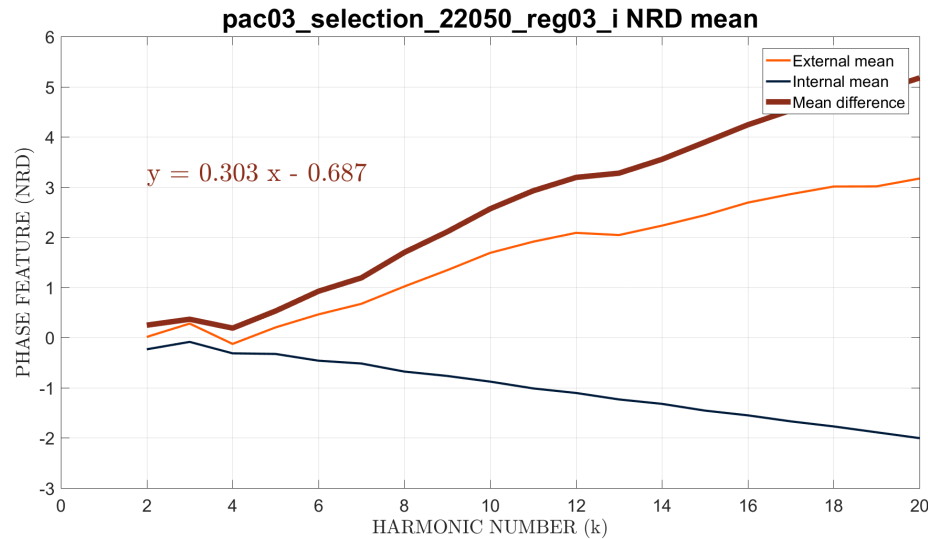


Figure 5.12: Difference between the external NRD mean (orange) and the internal NRD mean (blue) for repetition 3 of speaker 3 uttering the sustained vowel /i/. The equation that corresponds to the linear regression model of the NRD difference is represented in brown.

After determining these results, it was important to study this behaviour in more detail. In this perspective, a linear regression was used to study the difference between the internal and

external signals NRD means of all the signals according to vowel. Therefore the differences were plotted and the equations of the linear regression models for the mean difference were depicted according to vowel, as shown in Figure 5.13. The goal was to verify the existence and study the relation between inner and outer spectral content to infer the behaviour of the glottal source signal from the information available in the external speech signal.

From this study, it could be concluded that the slope of the NRD mean difference model for the vowel /i/ is the highest, with a value of **0.246**, closely followed by the slope of the NRD mean difference model for the vowel /u/, with a value of **0.186**, and, finally, followed by the slope of the NRD mean difference model for the vowel /a/, with a value of **0.032**.

Moreover, it can be observed in Figure 5.13 that the behaviour of the NRD mean difference for the first 10 harmonics is quite similar for both /i/ and /u/ vowels.

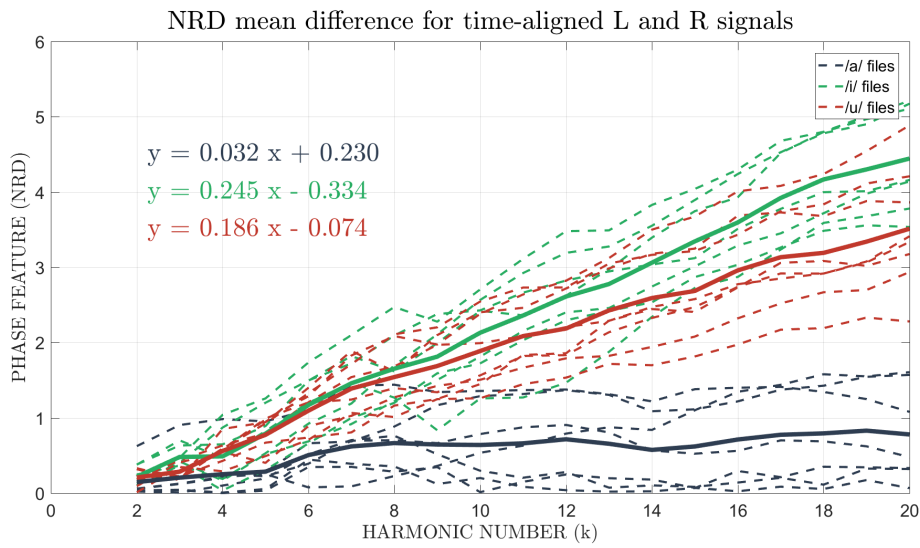


Figure 5.13: Differences between the external and internal NRD means for all signals. Also depicted the equations that correspond to the linear regression model of the NRD difference for each vowel.

The characterization of these empirical NRD models obtained from the spectral phase (NRD) values of all the glottal source signals recorded corresponds to another goal of this work.

5.2 Statistical analysis

In this section, a descriptive statistical tool is used for the representation of the empirical data distribution of the magnitude and phase (NRD) values estimated for the first 20 harmonics of each recorded sample. The purpose of this analysis is to verify the quality of the dataset, namely the congruence of the spectral magnitude and phase estimated data. A comparison is made between the values obtained for the internal and external signals for the same vowels. The boxplots regarding the spectral magnitude values estimated for the first 20 harmonics of the internal signals for repetition 3 from speaker 3 uttering vowel /i/ and vowel /u/ are illustrated in Figure 5.14 and in Figure 5.15, respectively.

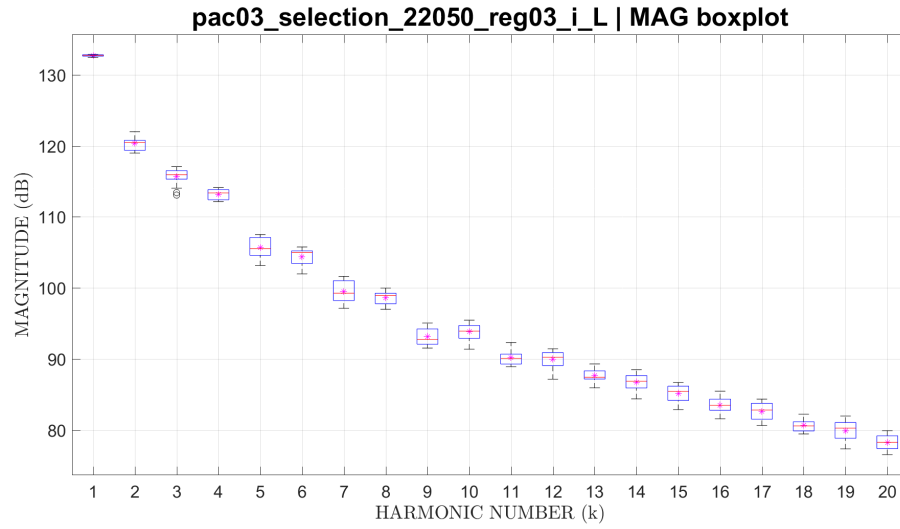


Figure 5.14: Boxplots of all the spectral magnitude values for the first 20 harmonics of the internal signal for repetition 3 from speaker 3 uttering vowel /i/.

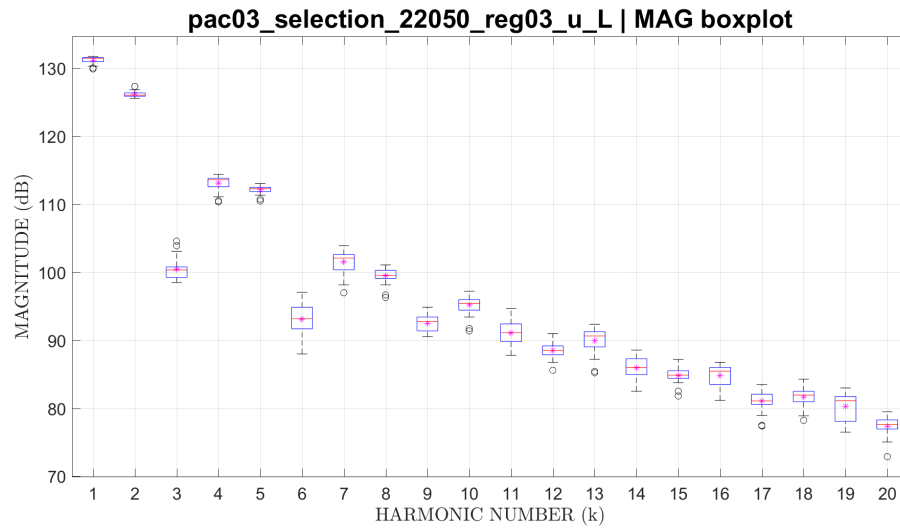


Figure 5.15: Boxplots of all the spectral magnitude values for the first 20 harmonics of the internal signal for repetition 3 from speaker 3 uttering vowel /u/.

After analysing the boxplot figures provided in Appendix C.1, some significant conclusions can be drawn. The first relates to the spectral magnitude values which present condense boxplots with low variability, as can be seen in Figure 5.14 and Figure 5.15. Secondly, that some harmonics present noticeable deviations from the logarithmic relation previously studied, as shown in Figure 5.15.

The boxplots for the estimated NRD values are depicted from Figure 5.16 to Figure 5.21.

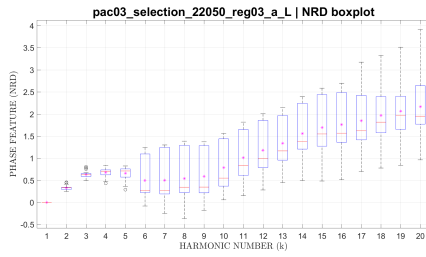


Figure 5.16: Boxplots of all the unwrapped NRD values for the first 20 harmonics of the internal signal for repetition 3 from speaker 3 uttering vowel /a/.

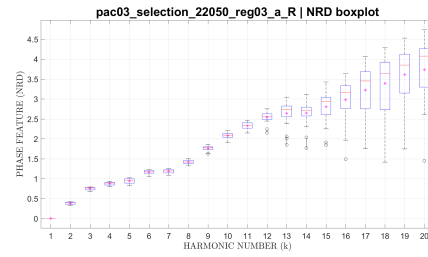


Figure 5.17: Boxplots of all the unwrapped NRD values for the first 20 harmonics of the external signal for repetition 3 from speaker 3 uttering vowel /a/.

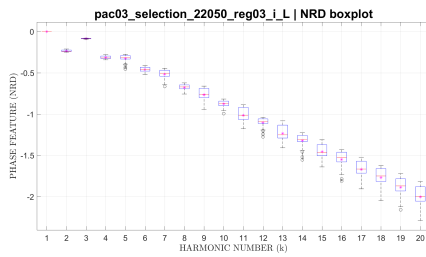


Figure 5.18: Boxplots of all the unwrapped NRD values for the first 20 harmonics of the internal signal for repetition 3 from speaker 3 uttering vowel /i/.

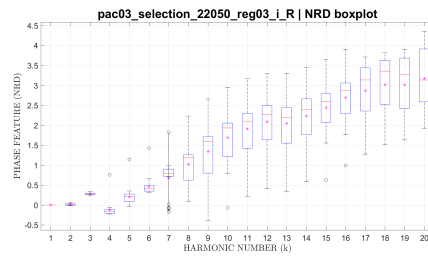


Figure 5.19: Boxplots of all the unwrapped NRD values for the first 20 harmonics of the external signal for repetition 3 from speaker 3 uttering vowel /i/.

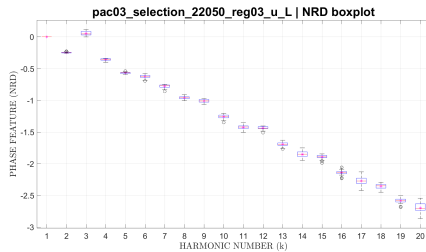


Figure 5.20: Boxplots of all the unwrapped NRD values for the first 20 harmonics of the internal signal for repetition 3 from speaker 3 uttering vowel /u/.

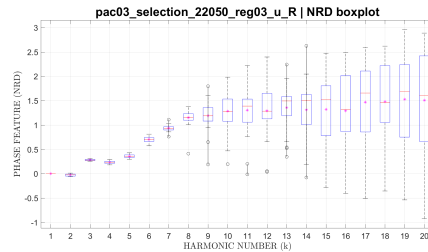


Figure 5.21: Boxplots of all the unwrapped NRD values for the first 20 harmonics of the external signal for repetition 3 from speaker 3 uttering vowel /u/.

Regarding the NRD values, these show higher variability for the internal signals when compared to the external signals in the specific case of vowel /a/, illustrated respectively in Figure 5.16 and Figure 5.17. The opposite was verified for vowels /i/ and /u/, where the boxplots show less variability for the internal signals, as can be seen in Figure 5.18 and Figure 5.20 for the internal signals and in Figure 5.19 and Figure 5.21 for the external signals. The figures of the remaining repetitions and speakers depict similar distributions and can be found in Appendix C.2.

A conclusion that can be drawn from these boxplots and the ones available in Appendix C is the fact that the first harmonics show substantially less variability when compared to the higher

harmonics, this is reassuring since the first harmonics are the most important for defining a periodic signal shape invariance.

5.3 Empirical model of the glottal source

Each speaker produces a unique glottal source signal, carrying idiosyncratic information. When that is combined with their own vocal tract characteristics results in a recognizable speech signal. One of the main goals for accurately modelling the real glottal source is to improve speech synthesis in terms of speech quality and naturalness. However, many theoretical models fail to perceptually fit natural voices due to the complex associations between the physical and the psychoacoustic events [63]. Having this in mind, an empirical glottal source model is presented based on real data collected directly from and quite near the vocal folds.

It is known that it is possible to define the shape invariance of a given periodic waveform using both independent spectral magnitude and spectral phase (NRD) models and reconstruct it in the time and frequency domain [99, 100, 96]. Considering that the periodic component of a speech signal may be decomposed on a series of harmonic related sine waves, such signals require at least three independent parameters to be generated: the fundamental frequency $\omega_0 = \frac{2\pi}{T_0}$; the spectral magnitude values obtained from the signal harmonics, A_ℓ ; and the coefficients from the shift-invariant phase-related model, NRD_ℓ . Therefore, the glottal source signal $gs(t)$ can be defined according to Equation 5.4, where L represents the total number of harmonics and ℓ represents the index number of each harmonic.

$$gs(t) = \sum_{\ell=1}^L A_\ell \sin\left(\frac{2\pi}{T_0} \ell t + 2\pi NRD_\ell\right) \quad (5.4)$$

The glottal source derivative signal $dgs(t)$ can be obtained by differentiating Equation 5.4. According to Fourier Theory, the derivative of $x(t)$ may be computed by multiplying $X(j\omega)^*(j\omega)$ on the frequency domain, where $X(j\omega)$ corresponds to the Fourier Transform of $x(t)$. Therefore, this corresponds to multiplying the magnitude of the Fourier transform by the frequency and adding $\pi/2$ to the phase [96] and, thus, except for a constant scaling factor, we obtain Equation 5.5.

$$dgs(t) = \sum_{\ell=1}^L \ell A_\ell \sin\left(\frac{2\pi}{T_0} \ell t + 2\pi NRD_\ell + \pi/2\right) \quad (5.5)$$

Since the glottal source carries speaker-dependent characteristics, these empirical models were generated for each vowel and speaker. The resulting models show, however, similar wave-shapes as Figure 5.22 shows. The fundamental frequency used for this representation was 220 Hz and the sampling frequency 22050 Hz. These empirical models were synthesized using the spectral magnitude and phase information regarding the first 20 harmonics. Similarly, the idealized glottal source signal was synthesized using the spectral magnitude and phase information for the first 20 harmonics of the LF model, obtained through the `v_glottf()` function from VOICEBOX, which is a speech toolbox available for MATLAB [101].

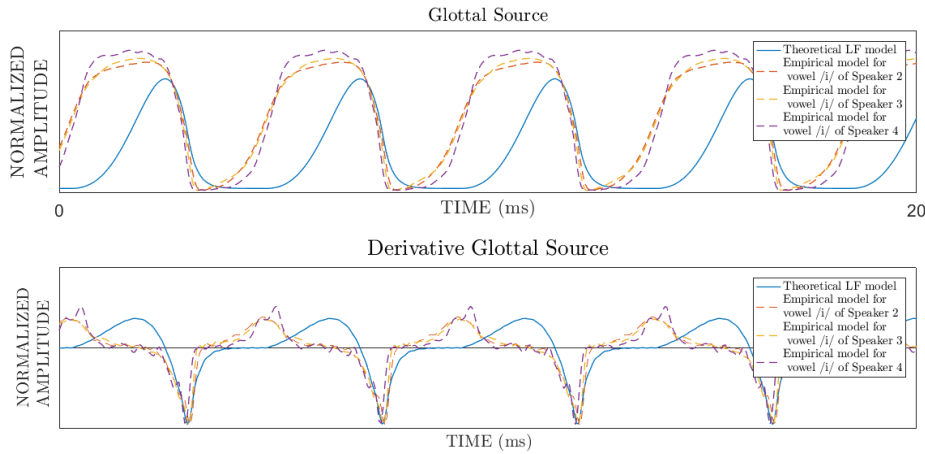


Figure 5.22: Waveform of the empirical model of the glottal source for vowel /i/ obtained for Speaker 2 (orange), for Speaker 3 (yellow) and for Speaker 4 (purple), alongside with the waveform of theoretical model LF (blue) and their derivatives.

According to Figure 5.22, the empirical glottal signals present a longer opening phase and a shorter return and closed phase than the LF glottal source signal. The empirical glottal source signal derivatives show an early maximum when compared to the LF glottal source derivative signal. Moreover, it depicts a stationary period (null glottal source derivative) during the opening phase and before the closing phase, besides the expected stationary period during the closed phase.

These empirical models intend to be a proof of concept and, even though these may differ from the well-accepted theoretical models, they show high similarity between the synthesized glottal source signals for different speakers and to the real glottal source signals recorded for each speaker. There are certainly studies that can follow up this unexplored approach for a better understanding of this topic.

5.4 Summary

This chapter described the estimation of the magnitude and phase (NRD) features for each vowel. These spectral features were then analysed and characterized. For the spectral magnitude analysis, the natural harmonic decay was estimated through the spectral slope of the empirical spectral magnitude model of both sustained vowels /i/ and /u/ of **-13 dB/oct**. The spectral phase (NRD) analysis of the internal signals resulted in slope values of **0.118** for the vowel /a/, **-0.071** for the vowel /i/ and **-0.093** for the vowel /u/. The analysis of the slope values for the mean differences between the external and internal signals showed values of **0.032** for the vowel /a/, **0.246** for the vowel /i/ and **0.186** for the vowel /u/. Using these features, a synthesized signal of the glottal source was obtained and compared to the theoretical LF model.

The focus of the next chapter is the vocal tract filter estimation and its characterization in terms of magnitude and phase.

Chapter 6

Vocal Tract Characterization

ACCORDING to the Source-Filter model, previously explained in Chapter 2, the speech production system can be divided into: the source, corresponding to the glottal excitation signal; and the filter, corresponding to the effect of the vocal tract structure. The Vocal Tract Filter (VTF) conveys the linguistic content by modulating the source signal in time and frequency. This filter reproduces the effect of the resonances and anti-resonances formed in the oral and nasal cavities, which results, in particular, in the presence of prominences on the magnitude spectrum, known as formant frequencies in the speech signal. In this chapter, the procedures followed for estimating the VTF according to vowel and speaker will be described, as well as the results obtained for the perceptual tests carried out using the synthetic signals generated with the estimated VTF.

6.1 Estimation of the vocal tract filter

6.1.1 Deconvolution approach

A first approach was attempted by using the concept of deconvolution, considering the Source-Filter Theory, which presupposes the simplified speech production described as a convolution between the excitation signal $g(t)$ and the impulse response of the vocal tract $v(t)$, resulting in a speech signal $s(t)$ in the time domain, as shown in Equation 6.1.

$$s(t) = g(t)v(t) \quad (6.1)$$

When expressed in the frequency domain, the speech $S(\omega)$ can be modelled as the product between the source $G(\omega)$ and the VTF $V(\omega)$, as described in Equation 6.2

$$S(\omega) = G(\omega)V(\omega) \quad (6.2)$$

The VTF transfer function $V(\omega)$ is estimated by dividing the speech signal $S(\omega)$ by the glottal source signal $G(\omega)$, in the frequency domain, as described in Equation 6.3.

$$V(\omega) = \frac{S(\omega)}{G(\omega)} \quad (6.3)$$

Therefore, by computing the inverse Fourier transform of the transfer function, $V(\omega)$, it is possible to obtain the impulse response of the VTF in the time domain, $v(t)$.

Although this approach seems rather straightforward, the experimental results were inconclusive due to the high variability of the glottal source signal in comparison with the VTF, which constrains the validity of the solution of Equation 6.3.

6.1.2 Adaptive filtering approach

Adaptive filtering was another approach attempted to estimate the VTF, which is based on iterative adjustments of the coefficients of an adaptive filter. This opportunity exists because both input and output signals are known. This technique consists in finding a filter matching the impulse response of the unknown filter according to a statistical criteria that minimizes the error signal, typically the minimization of the root mean square error. In the present case, we are dealing with a system identification problem where the goal is to design an adaptive filter that provides an approximation model for the unknown system [102]. The processing structure of the adaptive filter modelling can be represented by the block diagram shown in Figure 6.1.

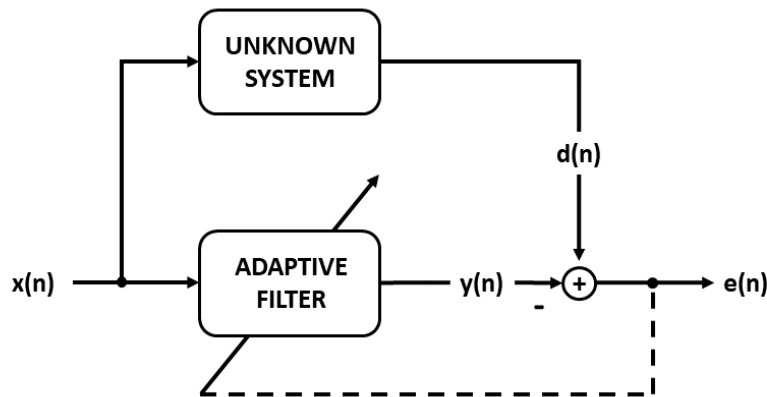


Figure 6.1: Block diagram representing the structure of a general adaptive filtering processing for system identification.

For the error minimization procedure, a gradient algorithm known as Least Mean Squares (LMS) was used, a common technique used for this purpose [103]. However, the conditions for convergence in order to obtain an optimal estimate of the unknown system were not achieved. In fact, even when the adaptive filter order was varied between 10 and 500, convergence was not obtained for a single well defined filter. For this reason, it was not possible to guarantee that the adaptive filter learns towards a global minimum, rather than towards a local minimum of the error function. Once more, this is related to the fact that the source signal has its energy restricted in several limited bands that correspond to its harmonic structure, resulting in multiple local minima of the error function and, therefore, it is not possible to ensure that this iterative process delivers a useful solution.

6.1.3 Holistic filter design approach

Finally, a holistic approach was carried out to overcome the limitations of the attempts previously described in Subsection 6.1.1 and Subsection 6.1.2 and accurately estimate the VTF, by taking advantage of the spectral envelopes computed for the internal and external signals to design digital filters that approximate the transfer function of a useful VTF. In order to evaluate the relevance of phase information, two filters were computed for all the repetitions of each vowel from different speakers. For this reason, a linear-phase all-zero Finite Impulse Response (FIR) filter and an all-pole Infinite Impulse Response (IIR) filter were obtained with equal magnitude response $|V(\omega)|$ for each pair of concomitant files (corresponding to the internal and external recordings). Using the harmonic analysis described in Chapter 5, we take advantage of the spectral envelope of the all-pole (LPC) model obtained from the Power Spectral Density (PSD) of both internal and external signals. Using these two spectral envelopes, the frequency response of a prototype filter was computed by performing the difference, on a dB scale, between the computed speech and glottal source spectral envelopes, as shown in Figure 6.2.

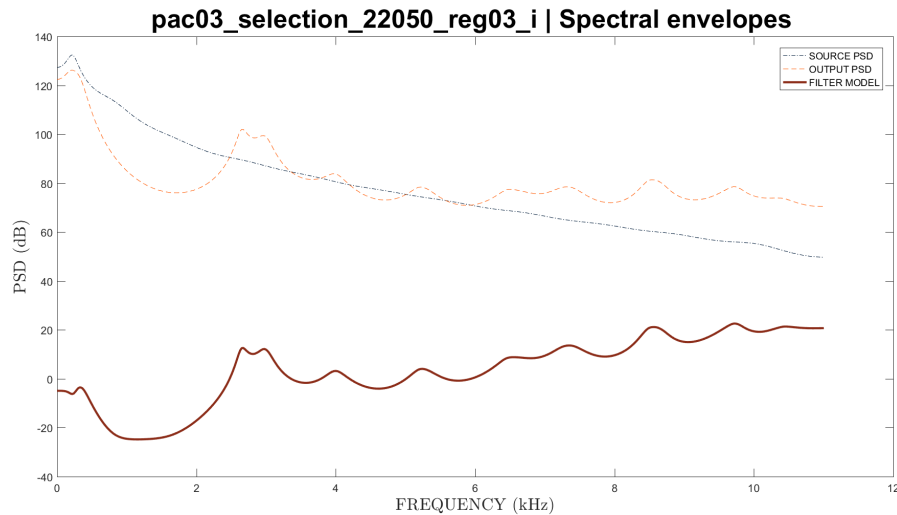


Figure 6.2: Spectral envelopes of the internal (blue) and external (orange) harmonic structure and frequency response of the prototype filter (red).

The IIR filter was obtained using the magnitude frequency response of this prototype. From the magnitude squared of the prototype, the autocorrelation coefficients were firstly obtained by using the Wiener-Khintchine theorem. Then, the parameters of the 22nd order all-pole model were obtained through the Levinson-Durbin recursion [92]. This was followed by the design of a single-band FIR filter by using the IIR filter magnitude frequency response. The linear-phase property of the FIR filter was ensured by the use of a single band Parks-McClellan optimal equiripple design of order 500, independently of the desired magnitude frequency response [90]. Both IIR and FIR filters are depicted alongside the prototype in Figure 6.3.

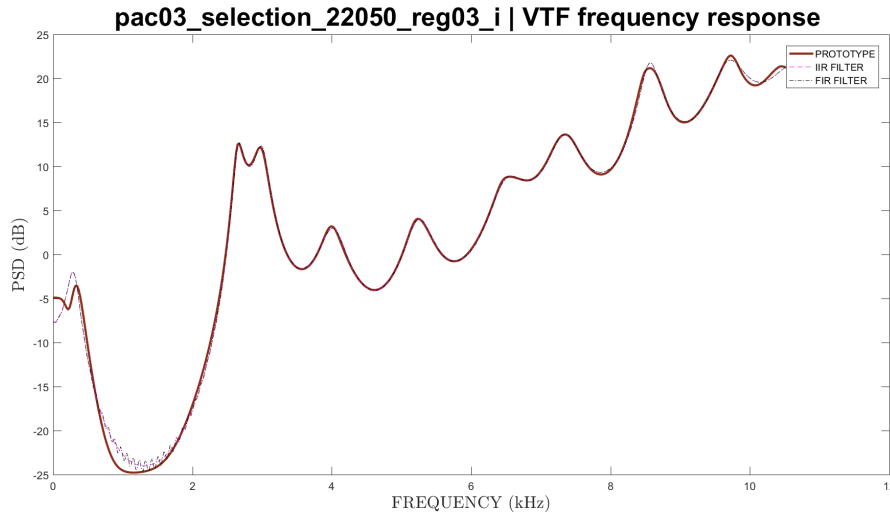


Figure 6.3: Frequency response of the prototype filter and the corresponding frequency response of the IIR and FIR filters computed for the repetition 3 of the vowel /i/ from speaker 3.

Overall, the magnitude frequency responses for both IIR and FIR filters were very similar, as shown by the dashed lines in Figure 6.3. The figures for other filters can be found in Appendix D, where a small ripple effect may be observed for some FIR filters computed for the vowels /i/ and /u/ from speaker 4.

The estimated filters for each vowel were then compared for different repetitions for each speaker. When observing Figure 6.4 it is possible to conclude that the estimated VTFs have a very similar frequency response for all the three repetitions of the vowel /i/ from speaker 2. The same conclusion can be drawn from the figures regarding other vowels and speakers available in Appendix D.

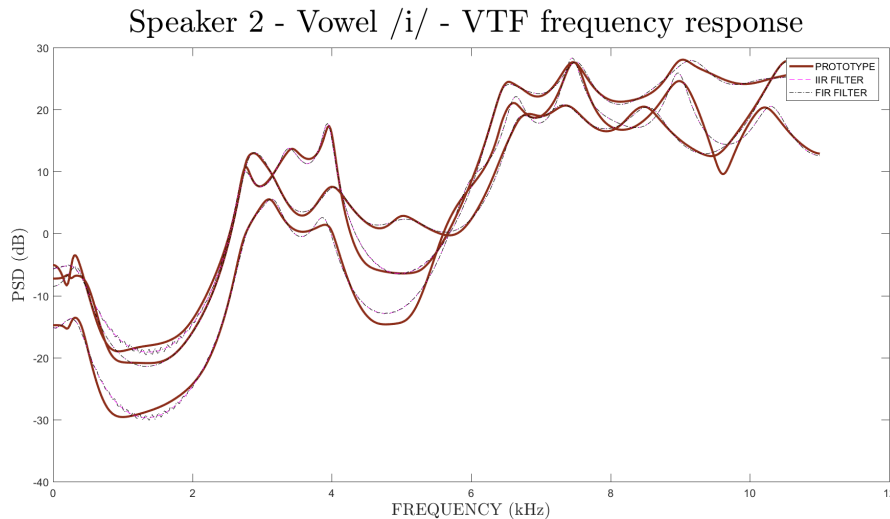


Figure 6.4: Frequency response of the prototype filters and the corresponding frequency response of the IIR and FIR filters obtained for all three repetitions of vowel /i/ from speaker 2.

Figure 6.5 represents the frequency responses of the different filters (and their prototype) obtained for different speakers for the vowel /i/. It is known that vowels are defined mostly by the location the first two formants [10]. In fact, as observed in Figure 6.5, the low variability shown in the lower frequencies indicates that these frequencies, where the formant frequencies are located, are related to the linguistic content while the greater variability at higher frequencies should be related to the speaker idiosyncratic characteristics. A similar conclusion can be inferred from the figures for the other two vowels (/a/ and /u/) available in Appendix D.

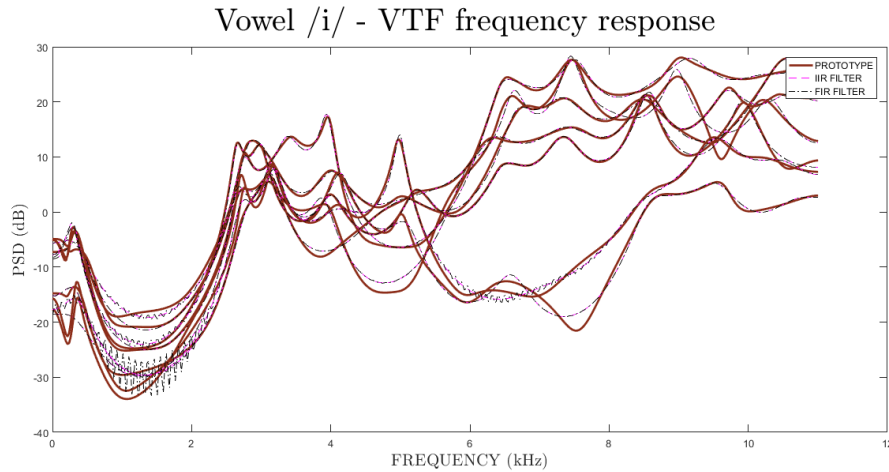


Figure 6.5: Frequency response of the prototype filters and the corresponding frequency response of the IIR and FIR filters obtained for all the repetitions of vowel /i/ from all speakers.

Subsequently, synthetic signals were generated for each vowel and speaker using as the source, the signal recorded by the internal microphone and as the filter, the estimated IIR and FIR filters representing the vocal tract approximation. The synthesized signals obtained from each filter version (IIR and FIR) were compared by the DyNaVoiceR Project team and no significant perceptual differences were found. Having this in mind, it was concluded that the phase differences do not play a major role in the perceptual impact of the synthetic vowels studied. However, even though the results show that phase contribution is not relevant for perceptual differences in short duration sustained vowels, phase contribution should not be disregarded in interconnected speech [90].

6.2 Perceptual tests

The synthetic signals obtained with the FIR filter corresponding to the estimated VTF were used to conduct perceptual tests to compare the generated synthetic speech signals with its corresponding speech signal recorded by the external microphone, in order to evaluate whether the VTF was estimated correctly and if it carries idiosyncratic characteristics that allow the speaker identification.

Twenty five volunteer listeners participated in the perceptual tests under informal conditions, though the participants were recommended the use of headphones. The participants were given an original reference sample of a vowel utterance and three synthetic samples, one from the same speaker and two from other speakers uttering the same vowel. The participants were asked to

identify the synthetic sample corresponding to the given reference sample. Additionally, the participants were asked to grade the similarity level between the chosen synthetic signal and the reference signal, using for this purpose a scale ranging from 1 (low similarity) to 5 (high similarity). Hence, a reference sample was used for each vowel (/a/, /i/ and /u/) of each speaker, which amounts to a total of 9 real audio references and 27 synthetic samples.

The results obtained with these perceptual tests can be observed in Figure 6.6 and Figure 6.7, where each female speaker is identified by a different color. The 95% CIs obtained for the success rate were computed using the Adjusted Wald method for the binary data [85] and the 95% CIs obtained for the similarity were computed using the MATLAB function *ttest2()* for continuous data.

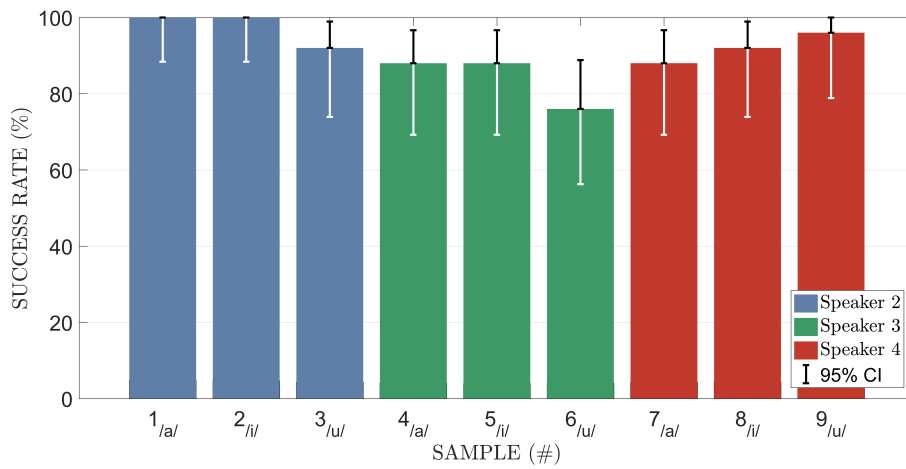


Figure 6.6: Results obtained with the perceptual tests regarding the accuracy in identifying the correct speaker. The blue bars correspond to Speaker 2, the green bars correspond to Speaker 3 and the red bars correspond to Speaker 4.

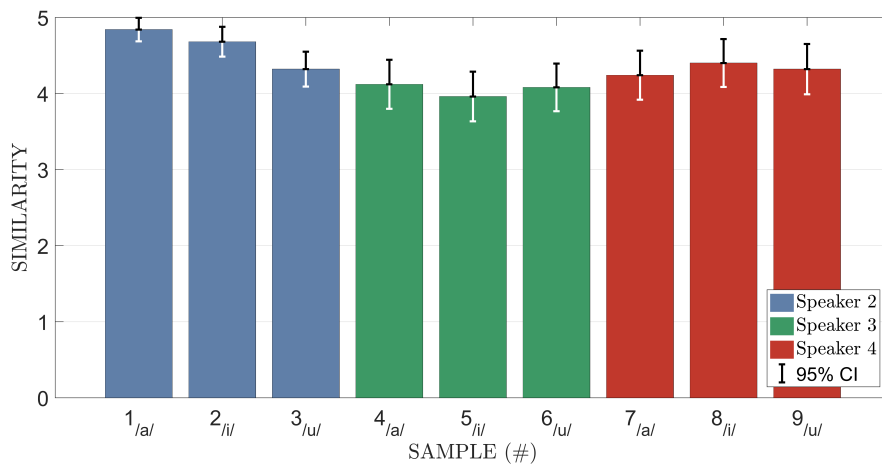


Figure 6.7: Results obtained with the perceptual tests regarding the degree of similarity that the participants gave to the chosen sample when compared to the reference sample. The blue bars correspond to Speaker 2, the green bars correspond to Speaker 3 and the red bars correspond to Speaker 4.

Figure 6.6 represents the percentage of cases for which the speaker was correctly identified and it can be concluded that the majority of the participants identified correctly the speakers given as a reference. According to the results, speaker 2 (blue) was the most accurately identified, with 100.00% success rate for both vowels /a/ and /i/, and speaker 3 (green) the least, with 76.00% for vowel /u/. The vowels that obtained a higher success rate for speakers 2 and 3 were vowels /a/ and /i/, while for speaker 4 was vowel /u/.

Based on the results depicted in Figure 6.7 it can be stated that the synthetic signals resemble the natural signals, since every sample achieved a mean score above 4 out of 5, which according to the scale corresponds to very similar. A statistical analysis was performed regarding the similarity results obtained for a significance level of 5% using the MATLAB function *ttest()* for continuous data.

Table 6.1: *p*-values obtained for the statistical analysis between the degree of similarity for different vowels for the same speaker.

	S2	S3	S4
/a/ vs /i/	0.1931	0.4748	0.4669
/a/ vs /u/	0.0004	0.8551	0.7221
/i/ vs /u/	0.0178	0.5864	0.7196

Table 6.2: *p*-values obtained for the statistical analysis between the degree of similarity for different speakers for the same vowels.

	/a/	/i/	/u/
S2 vs S3	0.0002	0.0004	0.2092
S2 vs S4	0.0014	0.1277	1
S3 vs S4	0.5890	0.0509	0.2827

In Table 6.1, it can be observed a statistically significant difference for Speaker 2 between the degree of similarity for vowels /a/ and /u/, with *p*-values of **0.0004**. Table 6.2 shows a statistically significant difference between the degree of similarity for vowel /a/ between speakers 2 and 3 and speakers 2 and 4, with *p*-values of **0.0002** and **0.0014**, respectively, and for vowel /i/ between speakers 2 and 3, with *p*-values of **0.0004**.

Table 6.3: Results of the perceptual tests where the values shown correspond to the average success rate in identifying the reference and average similarity to the reference according to speaker. The last column corresponds to the mean value ($\bar{x}$) for the three speakers.

	Speaker 2	Speaker 3	Speaker 4	$\bar{x}$
SUCCESS RATE	97.33%	84.00%	92.00%	91.11%
SIMILARITY	4.61	4.05	4.32	4.33

It can be seen in Table 6.3 that both success rate and similarity show good agreement, proving that the participants gave a higher similarity grade when they were more convinced of their choice. A conclusion to be drawn is the fact that speaker 2 showed the highest success rate and degree of similarity with 97.33% and 4.61, respectively, followed by speaker 4 with 92.00% and 4.32. Finally, speaker 3 showed the lowest success rate and degree of similarity with 84.00% and 4.05.

Overall, the results indicate that participants were able to select the correct speaker in most instances and that the synthetic signal was found to be highly similar to the reference signal with a mean success rate of **91.11%** and mean degree of similarity of **4.33**. Therefore, we conclude that the attempt to estimate the VTF was successfully achieved using this last approach and that the filter seems to carry idiosyncratic characteristics associated to each speaker. It should also be stated that the synthesized data seemed to have better quality than the corresponding recorded

real speech data. This is probably due to the surrounding noise recorded by the external microphone, which was not captured by the internal microphone, nor did it manifest its effect in the synthesized data.

6.3 Summary

The deconvolution and the adaptive filtering techniques were not well succeeded, since the obtained results did not produce useful vocal tract filters. However, by using the holistic design filter approach for the estimation of the VTF, it was possible to obtain synthesized data with a mean success rate of **91.11%** and a mean degree of similarity of **4.33**. The synthetic signals seemed very similar to the natural signals, as shown by the results of the perceptual tests, which indicates that this holistic approach allowed to obtain a good approximation of the VTF. In the next chapter, the major conclusions of this research work will be presented, as well as future trends.

Chapter 7

Conclusions and Future Work

SPEECH is an ability taken for granted. It is such a natural part of our life, since the moment we are born until the moment we breath our last breath, and we do not value it enough. We only realise its importance, when restrained from using it.

Firstly, it is clear that the waveshape of the internal signals shows discrepancies when compared to the glottal source waveshape of theoretical models. In fact, it even differs when compared between speakers, probably due to slight variations in the acquisition conditions and possibly due to the different phonation modes [11]. Nonetheless, to the best of our knowledge, there is no evidence in the literature of recent references reporting research work dealing with the real, i.e. physiological, glottal waveshape. In fact, in real case scenarios, the LF model has been stated to fall short when manipulated in the time domain [104]. For this reason, a spectral approach was outlined for accurately modelling the glottal source according to the spectral content extracted from real data opposing to the theoretical models. Thereby, a more accurate spectral model gives more flexibility to overcome shape constrictions imposed by the idealized models and improve the naturalness of synthesized speech signals.

The segments regarding the internal recordings of the sustained vowel /a/ show a more significant influence of the vocal tract filter than the observed in the vowels /i/ and /u/. In fact, for the vowel /a/, the power spectra of the internal recordings are very similar to the power spectra of the external recordings. This may be explained by the articulation of the vocal tract for this vowel, which consists in fewer constrictions when compared to the vowels /i/ or /u/, enabling the contributions of the supralaryngeal cavities to contaminate the internal signal through echoed signals.

In terms of spectral magnitude analysis, the value obtained for the spectral slope of the empirical spectral magnitude model of both sustained vowels /i/ and /u/ of **-13 dB/oct** approximates more to the Rosenberg reference value of **-12 dB/oct** [64], rather than to the **-16 dB/oct** reference from the LF glottal model [35].

In terms of spectral phase analysis, the values obtained for the empirical spectral NRD models vary according to the vowel. On one hand, the analysis of the internal signals resulted in slope values of **0.118** for the vowel /a/, **-0.071** for the vowel /i/ and **-0.093** for the vowel /u/. On the other, the analysis of the external signals resulted in slope values of **0.144** for the vowel /a/, **0.166** for the vowel /i/ and **0.094** for the vowel /u/. The slope value obtained for the mean differences between the external and internal signals for the vowel /a/ was **0.032**, for the vowel

/i/ was **0.246** and, lastly, for the vowel /u/ was **0.186**. This last analysis relates the NRD values obtained for the internal glottal source signal with the NRD values obtained for the external speech signal and shows a larger difference for the recordings obtained for vowels /i/ and /a/, when compared to vowel /a/.

The perceptual tests that were carried out support the hypothesis that the internal recordings of vowel /a/ show a larger effect of the vocal tract filter when comparing to the internal recordings of vowels /i/ and /u/. Another important conclusion from these tests is the fact that the internal recordings of vowels /i/ and /u/ show very similar linguistic content since these were often misidentified as the opposite.

The estimation of VTF was possible using the holistic design filter approach. The filters obtained showed relevant similarities not only among the same speaker, but also for different speakers according to vowel. The perceptual tests carried to with the synthesized data, showed that the filters were correctly estimated for the signals recorded with a mean success rate of **91,11%** and a mean degree of similarity of **4,33**. Additionally, it was concluded that phase contribution is not substantial in terms of perceptual impact for vowel utterance.

In conclusion, a characterization of both spectral magnitude and phase was performed in order to describe the glottal source signal as accurately as possible, which fulfills the purpose of this dissertation. Additionally, empiric glottal source models are described according to speaker in order to preserve the idiosyncratic information. Finally, the estimation of the vocal tract filter for a given vowel was successful according to each speaker and shown to replicate faithfully the signals recorded externally.

This study has only scratched the surface of this uncharted topic. Future directions are given on the continuity of this dissertation:

- Improve the dataset and record for a larger variety of speakers and different types of phonation (e.g. whispering);
- Estimate the glottal source using different state-of-the-art techniques developed in more recent studies and compare it with the glottal source empirical models obtained, in order to validate that the recorded signal obtained corresponds to the real glottal source signal;
- Compare the empiric glottal source model obtained and the theoretical glottal source models described in the literature, regarding the relation between the glottal source derivative behaviour and the physiological events;
- Study the cause of the difference in NRD slope polarity between the signals captured internally and externally, in the case of /i/ and /u/ vowels, which is probably related the acoustic radiation effects.

References

- [1] Nelson R. Williams. Occupational groups at risk of voice disorders: a review of the literature. *Occupational medicine*, 53(7):456–460, 2003. Cited on page 1.
- [2] Lorraine O. Ramig and Katherine Verdolini. Treatment efficacy: voice disorders. *Journal of Speech, Language, and Hearing Research*, 41(1):S101–S116, 1998. Cited on page 1.
- [3] Nelson Roy, Ray M. Merrill, Steven D. Gray, and Elaine M. Smith. Voice disorders in the general population: prevalence, risk factors, and occupational impact. *The Laryngoscope*, 115(11):1988–1995, 2005. Cited on page 1.
- [4] Zhi Tao, Xue-Dan Tan, Tao Han, Ji-Hua Gu, Yi-Shen Xu, and He-Ming Zhao. Reconstruction of normal speech from whispered speech based on RBF neural network. In *2010 Third International Symposium on Intelligent Information Technology and Security Informatics*, pages 374–377. IEEE, 2010. Cited on page 1.
- [5] Tomoki Toda, Mikihiro Nakagiri, and Kiyohiro Shikano. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(9):2505–2517, 2012. Cited on page 1.
- [6] Hideaki Konno, Mineichi Kudo, Hideyuki Imai, and Masanori Sugimoto. Whisper to normal speech conversion using pitch estimated from spectrum. *Speech Communication*, 83:10–20, 2016. Cited on page 1.
- [7] James L. Flanagan. *Speech analysis synthesis and perception*, volume 3. Springer Science & Business Media, 2013. Cited on page 5.
- [8] Henry Gray. *Anatomy Descriptive and Surgical: Gray’s Anatomy*. Running Press, 1974. Cited on pages 6 and 7.
- [9] Kenneth C. Jones and Anthony J. Gaudin. *Introdução à Biologia – 3ª edição*. Fundação Calouste Gulbenkian, Lisboa, 2000. Cited on page 6.
- [10] Johan Sundberg and Thomas D. Rossing. The science of singing voice. *Acoustical Society of America*, 1990. Cited on pages 6, 8, and 51.
- [11] Thomas Drugman, Paavo Alku, Abeer Alwan, and Bayya Yegnanarayana. Glottal source processing: From analysis to applications. *Computer Speech & Language*, 28(5):1117–1138, 2014. Cited on pages 7, 9, 18, 20, and 55.
- [12] Kenneth N. Stevens. *Acoustic phonetics*, volume 30. MIT press, 2000. Cited on pages 7 and 8.
- [13] Maria H. M. Mateus, Isabel Falé, and Maria J. Freitas. *Fonética e fonologia do português*. Universidade Aberta, Lisboa, 2016. Cited on pages 7 and 8.
- [14] Minoru Hirano, Yuki Kakita, Koichi Ohmaru, and Shigejiro Kurita. Structure and mechanical properties of the vocal fold. In *Speech and language*, volume 7, pages 271–297. Elsevier, 1982. Cited on page 8.

- [15] Katharine Murphy. *Digital signal processing techniques for application in the analysis of pathological voice and normophonic singing voice*. PhD thesis, Polytechnic University of Madrid, 2008. Cited on pages 8, 12, 13, and 14.
- [16] Thomas F. Quatieri. *Discrete-time speech signal processing: principles and practice*. Pearson Education India, 2006. Cited on page 8.
- [17] Aníbal Ferreira. Implantation of voicing on whispered speech using frequency-domain parametric modelling of source and filter information. In *2016 International Symposium on Signal, Image, Video and Communications (ISIVC)*, pages 159–166. IEEE, 2016. Cited on page 8.
- [18] Janwillem Van den Berg. Myoelastic-aerodynamic theory of voice production. *Journal of speech and hearing research*, 1(3):227–244, 1958. Cited on page 9.
- [19] Jody Kreiman and Diana Sidtis. *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons, 2011. Cited on page 9.
- [20] Ronald J. Baken. Electroglottography. *Journal of Voice*, 6(2):98–110, 1992. Cited on page 9.
- [21] Ingo R. Titze, Brad H. Story, Gregory C. Burnett, John F. Holzrichter, Lawrence C. Ng, and Wayne A. Lea. Comparison between electroglottography and electromagnetic glottography. *The Journal of the Acoustical Society of America*, 107(1):581–588, 2000. Cited on page 9.
- [22] Bertil Sonesson. A method for studying the vibratory movements of the vocal cords. *The Journal of Laryngology & Otology*, 73(11):732–737, 1959. Cited on page 9.
- [23] Paavo Alku. Glottal inverse filtering analysis of human voice production—a review of estimation and parameterization methods of the glottal excitation and their applications. *Sadhana*, 36(5):623–650, 2011. Cited on pages 9 and 13.
- [24] Nathalie Henrich, Christophe d’Alessandro, Boris Doval, and Michèle Castellengo. On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *The Journal of the Acoustical Society of America*, 115(3):1321–1332, 2004. Cited on page 9.
- [25] Sevasti-Zoi Karakozoglou. Glottal source analysis: a combinatory study using high-speed videoendoscopy and electroglottography. *Master’s thesis. Université Paris-Sud XI, University of Crete*, 2010. Cited on pages 9, 10, and 12.
- [26] Donald G. Childers, D. M. Hicks, G. P. Moore, L. Eskenazi, and A. L. Lalwani. Electroglottography and vocal fold physiology. *Journal of Speech, Language, and Hearing Research*, 33(2):245–254, 1990. Cited on page 9.
- [27] Marie-Agnes Faure and Andre Muller. Stroboscopy. *Journal of Voice*, 6(2):139–148, 1992. Cited on page 10.
- [28] M. Kiritani, T. Yoshiie, S. Kojima, Y. Satoh, and K. Hamada. Fission-fusion correlation by fission reactor irradiation with improved control. *Journal of nuclear materials*, 174(2-3):327–351, 1990. Cited on page 10.
- [29] Vinicius F. Guimarães. Estabilização de imagens para laringoscopia. *Master’s Thesis. University of São Paulo*, 2008. Cited on pages 10 and 11.
- [30] Jan G. Švec and Harm K. Schutte. Videokymography: high-speed line scanning of vocal fold vibration. *Journal of Voice*, 10(2):201–205, 1996. Cited on page 10.
- [31] Habib Moukalled, Dimitar D. Deliyski, Raphael R. Schwarz, and Song Wang. Segmentation of laryngeal high-speed videoendoscopy in temporal domain using paired active contours. *Models and analysis of vocal emissions for biomedical applications : 6th international workshop*. Cited on page 10.

- [32] Dimitar D. Deliyski, Pencho P. Petrushev, Heather S. Bonilha, Terri T. Gerlach, Bonnie Martin-Harris, and Robert E. Hillman. Clinical implementation of laryngeal high-speed videoendoscopy: challenges and evolution. *Folia Phoniatrica et Logopaedica*, 60(1):33–44, 2008. Cited on page 12.
- [33] Louis K. Pitman. Nasolaryngoscope, October 24 1961. US Patent 3,005,452. Cited on page 12.
- [34] M. H. Mateus, I. Falé, and M. J. Freitas. *Fonética e Fonologia do português*. Universidade Aberta, 2005. Cited on page 12.
- [35] Gunnar Fant. *Acoustic theory of speech production*. Number 2. Walter de Gruyter, 1970. Cited on pages 12 and 55.
- [36] Hannu Pulakka. Analysis of human voice production using inverse filtering, high-speed imaging, and electroglottography. *Master's thesis. Aalto University*, 2005. Cited on page 12.
- [37] Johan Sundberg. The acoustics of the singing voice. *Scientific American*, 236(3):82–91, 1977. Cited on page 12.
- [38] George Kafentzis. On the inverse filtering of speech. *Master's thesis. University of Crete*, 2010. Cited on page 12.
- [39] Gilles Degottex. *Glottal source and vocal-tract separation: estimation of glottal parameters, voice transformation and synthesis using a glottal model*. PhD thesis, Paris 6, 2010. Cited on page 12.
- [40] H. R. Javkin, Norma Antónanzas-Barroso, and Ian Maddieson. Digital inverse filtering for linguistic research. *Journal of Speech, Language, and Hearing Research*, 30(1):122–129, 1987. Cited on pages 13, 14, 16, and 17.
- [41] Ahmet M. Kondo. *Digital speech: coding for low bit rate communication systems*. John Wiley & Sons, 2005. Cited on page 13.
- [42] Sandra O. Dias. Estimation of the glottal pulse from speech or singing voice. *Master's thesis. University of Porto*, 2012. Cited on pages 14, 18, 19, and 40.
- [43] M. A. Nwachuku. Inverse filtering techniques in speech analysis. *Nigerian Journal of Technology*, 1(1), 1975. Cited on page 15.
- [44] Thomas Drugman. Advances in glottal analysis and its applications. *University of Mons, Belgium*, 2011. Cited on pages 15, 21, and 31.
- [45] Jacqueline Walker and Peter Murphy. A review of glottal waveform analysis. In *Progress in nonlinear speech processing*, pages 1–21. Springer, 2007. Cited on pages 15 and 18.
- [46] D. Wong, J. Markel, and A. Gray. Least squares glottal inverse filtering from the acoustic speech waveform. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(4):350–355, 1979. Cited on page 15.
- [47] D. Veeneman and S. BeMent. Automatic glottal inverse filtering from speech and electroglottographic signals. *IEEE transactions on acoustics, speech, and signal processing*, 33(2):369–377, 1985. Cited on page 16.
- [48] M. D. Plumpe, T. F. Quatieri, and D. A. Reynolds. Modeling of the glottal flow derivative waveform with application to speaker identification. *IEEE Transactions on Speech and Audio Processing*, 7(5):569–586, 1999. Cited on page 16.
- [49] Paavo Alku, Carlo Magi, Santeri Yrttiaho, Tom Bäckström, and Brad Story. Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering. *the Journal of the Acoustical Society of America*, 125(5):3289–3305, 2009. Cited on pages 16 and 17.
- [50] D. M. Brookes and D. S. F. Chan. Speaker characteristics from a glottal airflow model using robust inverse filtering. 16:501–501, 1994. Cited on page 16.

- [51] Huiqun Deng, Rabab K. Ward, Michael P. Beddoes, and Murray Hodgson. A new method for obtaining accurate estimates of vocal-tract filters and glottal waves from vowel sounds. *IEEE Transactions on Audio, Speech, and Language Processing*, 14(2):445–455, 2006. Cited on page 16.
- [52] Elliot Moore and Mark Clements. Algorithm for automatic glottal waveform estimation without the reliance on precise glottal closure information. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages I–101. IEEE, 2004. Cited on page 16.
- [53] Damien Vincent, Olivier Rosenc, and Thierry Chonavel. Estimation of glottal source parameters based on an arx model. In *Ninth European Conference on Speech Communication and Technology*, 2005. Cited on page 16.
- [54] Paavo Alku. Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech communication*, 11(2-3):109–118, 1992. Cited on page 16.
- [55] Paavo Alku and Erkki Vilkmann. Estimation of the glottal pulseform based on discrete all-pole modeling. In *Third International Conference on Spoken Language Processing*, 1994. Cited on page 16.
- [56] Baris Bozkurt and Thierry Dutoit. Mixed-phase speech modeling and formant estimation, using differential phase spectrums. In *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003. Cited on page 17.
- [57] Boris Doval, Christophe d’Alessandro, and Nathalie Henrich Bernardoni. The voice source as a causal/anticausal linear filter. 2003. Cited on pages 17 and 20.
- [58] Baris Bozkurt, Boris Doval, Christophe d’Alessandro, and Thierry Dutoit. Zeros of Z-transform representation with application to source-filter separation in speech. *IEEE signal processing letters*, 12(4):344–347, 2005. Cited on page 17.
- [59] Thomas Drugman, Baris Bozkurt, and Thierry Dutoit. Complex cepstrum-based decomposition of speech for glottal source estimation. *arXiv preprint arXiv:1912.12602*, 2019. Cited on pages 17 and 21.
- [60] Thomas Drugman, Thomas Dubuisson, Alexis Moinet, and Thierry Dutoit. Glottal source estimation robustness a comparison of sensitivity of voice source estimation techniques. 2008. Cited on page 18.
- [61] Alan Ó Cinnéide. PhD transfer report. *Institute of Technology, Dublin*, 2008. Cited on page 18.
- [62] Hiroya Fujisaki and Mats Ljungqvist. Proposal and evaluation of models for the glottal source waveform. In *ICASSP’86. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 11, pages 1605–1608. IEEE, 1986. Cited on pages 19 and 20.
- [63] Jody Kreiman, Marc Garellek, Gang Chen, Abeer Alwan, and Bruce R. Gerratt. Perceptual evaluation of voice source models. *The Journal of the Acoustical Society of America*, 138(1):1–10, 2015. Cited on pages 20 and 45.
- [64] Aaron 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. Cited on pages 20 and 55.
- [65] Gunnar Fant. Vocal source analysis- a progress report. *STL-QPSR*, 20(3-4):31–53, 1979. Cited on page 20.
- [66] Gunnar Fant, Johan Liljencrants, and Qi-guang Lin. A four-parameter model of glottal flow. *STL-QPSR*, 4(1985):1–13, 1985. Cited on page 20.
- [67] Dennis H. Klatt. Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America*, 82(3):737–793, 1987. Cited on page 20.

- [68] Raymond Veldhuis. A computationally efficient alternative for the Liljencrants–Fant model and its perceptual evaluation. *The Journal of the Acoustical Society of America*, 103(1):566–571, 1998. Cited on page 20.
- [69] Yen-Liang Shue and Abeer Alwan. A new voice source model based on high-speed imaging and its application to voice source estimation. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5134–5137. IEEE, 2010. Cited on page 20.
- [70] Gang Chen, Yen-Liang Shue, Jody Kreiman, and Abeer Alwan. Estimating the voice source in noise. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012. Cited on page 20.
- [71] Boris Doval, Christophe d’Alessandro, and Nathalie Henrich. The spectrum of glottal flow models. *Acta acustica united with acustica*, 92(6):1026–1046, 2006. Cited on pages 20 and 21.
- [72] Christer Gobl and Ailbhe Ní Chasaide. Amplitude-based source parameters for measuring voice quality. In *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*, 2003. Cited on page 21.
- [73] Gunnar Fant. The LF-model revisited. transformations and frequency domain analysis. *Speech Trans. Lab. Q. Rep., Royal Inst. of Tech. Stockholm*, 2(3):40, 1995. Cited on page 21.
- [74] Paavo Alku, Tom Bäckström, and Erkki Vilkmán. Normalized amplitude quotient for parametrization of the glottal flow. *the Journal of the Acoustical Society of America*, 112(2):701–710, 2002. Cited on page 21.
- [75] Paavo Alku, Helmer Strik, and Erkki Vilkmán. Parabolic spectral parameter—a new method for quantification of the glottal flow. *Speech Communication*, 22(1):67–79, 1997. Cited on page 22.
- [76] Matti Airas. TKK Aparat: An environment for voice inverse filtering and parameterization. *Logopedics Phoniatrics Vocology*, 33(1):49–64, 2008. Cited on page 22.
- [77] Helen M. Hanson. Individual variations in glottal characteristics of female speakers. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 772–775. IEEE, 1995. Cited on page 22.
- [78] Peter J. Murphy and Olatunji O. Akande. Quantification of glottal and voiced speech harmonics-to-noise ratios using cepstral-based estimation. In *ISCA Tutorial and Research Workshop (ITRW) on Non-Linear Speech Processing*, 2005. Cited on page 22.
- [79] Donald G. Childers and Chih K. Lee. Vocal quality factors: Analysis, synthesis, and perception. *the Journal of the Acoustical Society of America*, 90(5):2394–2410, 1991. Cited on page 22.
- [80] Donald G. Childers and Jose A. Diaz. Speech processing and synthesis toolboxes, 2000. Cited on page 22.
- [81] Erica E. Fukuyama. Análise acústica da voz captada na faringe próximo à fonte glótica através de microfone acoplado ao fibrolaringoscópio. *Revista Brasileira de Otorrinolaringologia*, 67(6):776–786, 2001. Cited on page 23.
- [82] Marco A. M. Oliveira. Modelização de filtro de trato vocal para reconstrução de voz disfônica. *Master’s thesis. University of Porto*, 2020. Cited on page 24.
- [83] Maria R. Delgado Martins. Análise acústica das vogais tónicas em português. *Boletim de Filologia*, 22(3):303–314, 1973. Cited on page 25.
- [84] Rolf Timcke, Hans von Leden, and Paul Moore. Laryngeal vibrations: Measurements of the glottic wave: Part II— Physiologic variations. *AMA archives of otolaryngology*, 69(4):438–444, 1959. Cited on page 27.

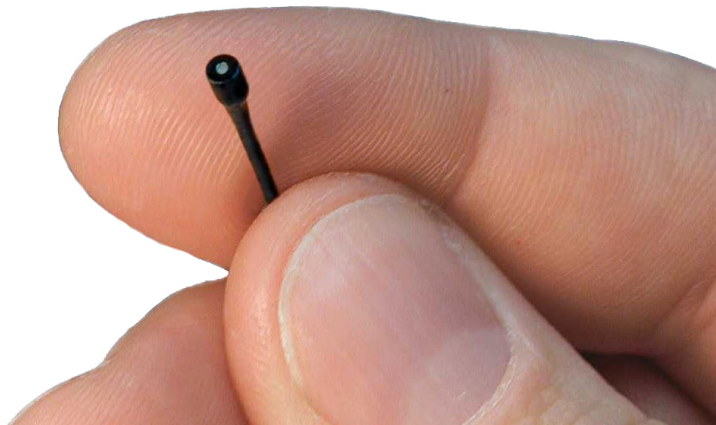
- [85] Jeff Sauro and James R. Lewis. *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann, 2016. Cited on pages 28 and 52.
- [86] Alan V. Oppenheim, Alan S. Willsky, and S. Hamid. *Signals and systems, processing series*, 1997. Cited on page 31.
- [87] Alan V Oppenheim, John R. Buck, and Ronald W. Schafer. *Discrete-time signal processing*. Vol. 2. Upper Saddle River, NJ: Prentice Hall, 2001. Cited on pages 31 and 33.
- [88] Anibal J. S. Ferreira. Combined spectral envelope normalization and subtraction of sinusoidal components in the ODFT and MDCT frequency domains. In *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics (Cat. No. 01TH8575)*, pages 51–54. IEEE, 2001. Cited on pages 31 and 32.
- [89] Aníbal Ferreira and Deepen Sinha. Accurate and robust frequency estimation in the ODFT domain. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, 2005., pages 203–206. IEEE, 2005. Cited on page 32.
- [90] Aníbal Ferreira. On the physiological validity and perceptual impact of the group delay response of all-pole vocal tract models. *Submitted for the Journal of the Audio Engineering Society (AES)*, 2020. Cited on pages 32, 49, and 51.
- [91] Aníbal Ferreira and Ricardo Sousa. DFT-based frequency estimation under harmonic interference. In *2010 4th International Symposium on Communications, Control and Signal Processing (ISCCSP)*, pages 1–6. IEEE, 2010. Cited on page 32.
- [92] Monson H. Hayes. *Statistical digital signal processing and modeling*. John Wiley & Sons, 2009. Cited on pages 33 and 49.
- [93] Michel Jackson, Peter Ladefoged, Marie Huffman, and Norma Antoñanzas-Barroso. Measures of spectral tilt. *The Journal of the Acoustical Society of America*, 77(S1):S86–S86, 1985. Cited on page 33.
- [94] Richard L. Miller. Nature of the vocal cord wave. *The Journal of the Acoustical Society of America*, 31(6):667–677, 1959. Cited on page 34.
- [95] James L. Flanagan. Some properties of the glottal sound source. *Journal of Speech and Hearing Research*, 1(2):99–116, 1958. Cited on page 34.
- [96] Aníbal J Ferreira and José M Tribolet. A holistic glottal phase-related feature. In *21st International Conference on Digital Audio Effects (DAFx-18). Aveiro, Portugal*, 2018. Cited on pages 37, 40, and 45.
- [97] José Tribolet. A new phase unwrapping algorithm. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 25(2):170–177, 1977. Cited on page 38.
- [98] Ricardo Sousa and Aníbal Ferreira. Importance of the relative delay of glottal source harmonics. In *Audio Engineering Society Conference: 39th International Conference: Audio Forensics: Practices and Challenges*. Audio Engineering Society, 2010. Cited on page 38.
- [99] Thomas F. Quatieri and Robert J. McAulay. Shape invariant time-scale and pitch modification of speech. *IEEE Transactions on Signal Processing*, 40(3):497–510, 1992. Cited on page 45.
- [100] Michael Portnoff. Time-scale modification of speech based on short-time fourier analysis. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(3):374–390, 1981. Cited on page 45.
- [101] Mike Brookes. Voicebox: Speech processing toolbox for matlab. *Software, available [Mar. 2011] from www.ee.ic.ac.uk/hp/staff/dmb/voicebox/voicebox.html*, 47, 1997. Cited on page 45.

- [102] Garima Malik and Amandeep S. Sappal. Adaptive equalization algorithms: an overview. *International Journal of Advanced Computer Science and Applications*, 2(3), 2011. Cited on page 48.
- [103] Simon S. Haykin. *Adaptive filter theory*. Pearson Education India, 2005. Cited on page 48.
- [104] Jody Kreiman, Bruce R. Gerratt, and Norma Antonanzas-Barroso. Analysis and synthesis of pathological voice quality. *Unpublished users' manual, retrieved February, 22:2006*, 2006. Cited on page 55.

Appendix A

Data Acquisition and Dataset Characterization

A.1 Microphone Specifications



COUNTRYMAN ASSOCIATES INC

B6 OMNIDIRECTIONAL LAVALIER MICROPHONE

Only one tenth of an inch in diameter, the B6 is the smallest lavalier in the world and outperforms microphones many times its size.

The swappable protective caps provide moisture resistance and color options and let you shape the frequency response to suit different applications or to match other microphones.

With exceptionally low handling noise and rugged construction, the B6 is the ideal choice for theater, broadcast, churches, and general lavalier applications.

Unobtrusive

Smaller than the *cable* of other lavaliers, the B6 is easily hidden in hair or on costumes, or taped to a performer's face. The B6 comes in five colors to match clothing, hair, and skin tones, or use a felt tip marker to color the white caps for near-perfect concealability. Never suffer the hassle and degraded sound quality of under-clothing miking again.

Rugged and Reliable

The B6 is exceptionally resistant to makeup, sweat, and moisture when used with the supplied protective caps, and is well-suited to use in hair or on the body. The protective caps are easily removed for cleaning or replacement, and the Aramid-reinforced cable gives it world-class survivability.

Exceptional Sound Quality

The highest-quality audio available in a lavalier mic. Low distortion at SPL up to 140 dB on 48 V Phantom Power. The Aramid cable and ultra-thin diaphragm combine to set a new standard for low handling noise. The tiny size and natural sound pickup make the B6 easy to position for ambient noise and feedback rejection.

Versatile

Swappable protective caps let you shape the frequency response for different situations or to match other microphones. Versions available for different speaking or singing styles, with up to 140 dB SPL capability.



Supplied with carrying case, black and white cable clips, wind screen, and 3 protective caps.

Frequency Response : 20 Hz to 20 kHz

Operating Current : 500 μ A

Operating Voltage : 1 to 2 Volts

Power Supply Voltage :

+3 V with 2.7 k Ω load

+5 V with 6.8 k Ω load

+9 V with 15 k Ω load

The B6 Lavalier is available in three sensitivities:

B6W4

standard (gray band) for most uses

Sensitivity: 16.0 mV/Pascal

Equivalent Acoustic Noise: 24 dBA SPL

Overload Sound Level: 120 dB SPL

B6W5

low sensitivity (red band): head mic for theater

Sensitivity: 7.0 mV/Pascal

Equivalent Acoustic Noise: 29 dBA SPL

Overload Sound Level: 130 dB SPL

B6W6

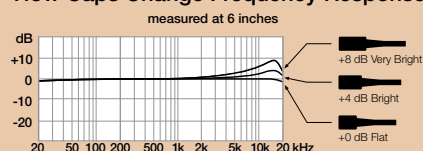
very low sensitivity (blue band): instrument / near mouth

Sensitivity: 2 mV/Pascal

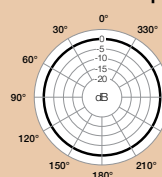
Equivalent Acoustic Noise: 39 dBA SPL

Overload Sound Level: 140 dB SPL

How Caps Change Frequency Response



1 kHz Polar Response



B6 Lavalier: Frequently Asked Questions



How do I choose the right color for my skin tone?

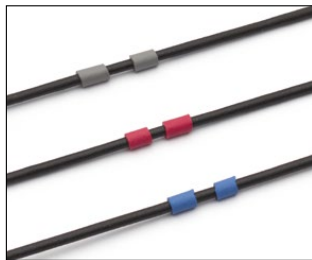
Tan works very well for Caucasian skin tones as well as olive complexions. Light beige works well for light and pink tones. Cocoa is the ideal choice for very light brown to chocolate tones. Choose black for deep brown skin, or on other tones when you want the mic to be visible. When in doubt, choose the darker option. That's because a mic that's too light can resemble a scar or blemish, while a mic that's slightly darker than the background resembles a shadow and draws less attention.



Which cap should I use?

The B6 Lavalier should always be used with a protective cap in place to keep sweat, makeup, and other foreign material out of the microphone. The three omni caps each have a different high-frequency response characteristic that controls the amount of "crispness" or "sibilance" (response at 15 kHz). To identify caps, compare size to the drawings.

The omni ships with the +4 dB protective cap fitted to the mic. This will boost the perceived amount of presence in your sound, while leaving the lower frequencies unchanged. If you experience problems with high-frequency feedback, you should switch to the 0 dB cap.



Which sensitivity should I choose?

Making a microphone more sensitive to catch soft sounds means it will overload sooner for loud sounds. Because sound pressure levels vary between individuals and applications, we provide three sensitivities with three overload or clipping characteristics.

- The most sensitive (W4, gray band) is for general speaking, such as presentations or sermons, where the mic is positioned on the chest or lapel
- The middle sensitivity (W5, red band) is ideal for use as a head mic in theater
- The least sensitive mic (W6, blue band) with the highest overload sound level is a good choice for instrument applications, opera, or where the microphone will be positioned very near the mouth.

What are other popular placement tips for the B6?

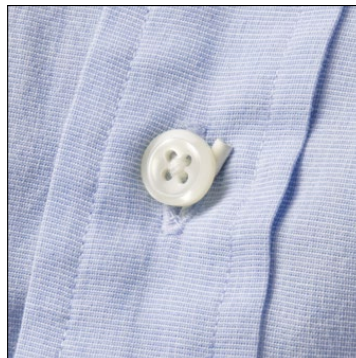
Hair Mic

The B6 is very popular as an ultraminiature hair microphone. This placement provides good gain before feedback and natural sound that doesn't change when the head moves. (Microphone extended here for visibility)



Behind a Button

Positioned in the knot of a tie or behind a shirt button, the B6 delivers flawless audio and hides in plain sight. When placed completely under clothing use the +4 dB or +8 dB protective cap to boost high frequencies.



Attached to Eyeglasses

The B6 is also popular on stage attached to glasses. Placement on eyeglasses with O-rings or tape is secure, discreet, and a convenient alternative to hair miking.



A.2 Rhino-Laryngo Fiberscope Specifications

OLYMPUS[®]

Your Vision, Our Future

RHINO-LARYNGO FIBERSCOPE

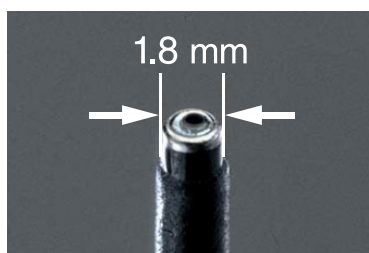
ENF-XP

The Optimal Choice for Observation of Narrow Upper Airway Passages



Ultra-slim 2.2 mm Insertion Tube — Perfectly Sized for Narrow Nasal Cavities.

Featuring an ultra-slim insertion tube diameter of 2.2 mm as well as high resolution optics and brightness, the Olympus ENF-XP is perfectly suited for observing the nasal cavity, larynx and pharynx. Optimally sized for narrow nasal cavities, it also makes observation of adult paranasal sinuses easier than ever, while its angulation range of 130° both up and down enables trachea insertion maneuverability. For minimally invasive insertion and clear, reliable observation, depend on the Olympus ENF-XP Rhino-Laryngofiberscope.



Ultra-slim 2.2 mm Insertion Tube Diameter

Optimally sized for smaller nasal cavities. This endoscope also makes observation of adult paranasal sinuses easier than ever.



Wide Angulation Range — 130° Up and Down

The endoscope wide bending range of 130° up/130° down allows trouble-free insertion, even through the trachea, allowing you to focus more on observation.

RHINO-LARYNGO FIBERSCOPE ENF TYPE XP Specifications

Field of view	75°
Depth of field	2.5 - 50 mm
Range of tip bending	Up 130°/ Down 130°
Distal end outer diameter	1.8 mm
Insertion tube diameter	2.2 mm
Working length	300 mm
Total length	530 mm

Olympus Rhino-Laryngofiberscope Lineup

Versatile, durable and reliable Olympus rhino-laryngofiberscopes ensure smooth operation. A full range of rhino-laryngo fiberscopes for your procedural needs. Olympus rhino-laryngofiberscope — the right choice for superior performance.



ENF-P4



ENF-T3



ENF-GP

Specifications, design and accessories are subject to change without any notice or obligation on the part of the manufacturer.

OLYMPUS®

OLYMPUS MEDICAL SYSTEMS CORP.
Shinjuku Monolith, 2-3-1 Nishi-Shinjuku, Shinjuku-ku, Tokyo 163-0914, Japan

For a complete listing of
sales and distribution locations visit
www.olympus.com

Printed in Japan R0025E2-122011

A.3 Ethical Approval

Unidade de Investigação

Tomei conhecimento. Nada a opor. À DC.

21 de Maio de 2020

A Coordenadora da Unidade de Investigação

(Prof.ª Doutora Ana Azevedo)



n.º 82 / 2020

DIRECÇÃO CLÍNICA
2020, 5, 22

PEDIDO DE AUTORIZAÇÃO

Realização de Investigação

Exmo. Senhor Presidente do Conselho de Administração
do Centro Hospitalar de São João

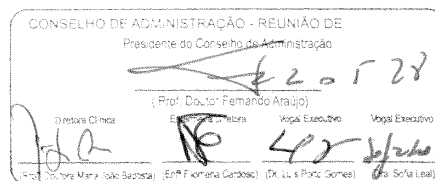
21.5.2020

Nome do Investigador Principal:

Prof. Aníbal João de Sousa Ferreira

Título da Investigação:

Dysphonic to Natural Voice Reconstruction (DyNaVoiceR) | Accurate
glottal source estimation and modelling



Pretendo realizar no(s) Serviço(s) de:

Otorrinolaringologia

a investigação em epígrafe, solicito a V. Exa., na qualidade de Investigador/Promotor, autorização para a sua efetivação.

Para o efeito, anexo toda a documentação referida no dossier da Comissão de Ética do Centro Hospitalar de São João/Faculdade de Medicina da Universidade do Porto respeitante à investigação, à qual enderecei pedido de apreciação e parecer.

Com os melhores cumprimentos.

O Investigador/Promotor

Porto, 13 de Fevereiro de 2020 .

Aníbal João de Sousa Ferreira
assinatura

• Centro Hospitalar São João •
Centro de Epidemiologia Hospitalar

14.5.2020

Parecer da Comissão de Ética do

Centro Hospitalar Universitário de São João / Faculdade de Medicina da Universidade do Porto

Título do Projeto: Dysphonic to natural voice reconstruction (DyNaVoiceR) - Accurate glottal source estimation and modelling

Nome do Investigador Principal: Prof. Doutor Aníbal João de Sousa Ferreira

Onde decorre o Estudo: No Serviço de Otorrinolaringologia do CHUSJ. Apresentou declaração da Dra. Margarida Santos. O Prof. Doutor Jorge Spratley é o profissional de ligação.

Objetivos do Estudo:

Caraterizar, com precisão, a verdadeira excitação acústica da glote, principalmente a vibração das pregas vocais, sob condições naturais de interação entre laringe e trato vocal.

Conceção e Pertinência do estudo:

Estudo que pretende recolher conhecimento crucial para o desenvolvimento de procedimentos inovadores e confiáveis que permitam sintetizar e implantar, de forma não invasiva, componentes de fala normal em fala sussurrada. Pretende ainda obter modelos mais fidedignos da fonte glótica que permita estimar modelos do filtro do trato vocal com mais precisão.

Os voluntários saudáveis serão submetidos a uma nasofibrolaringoscopia normal, mas com a inserção de um microfone especial no canal de trabalho para poder proceder à recolha do sinal glótico o mais próximo das pregas vocais. Outro microfone idêntico será colocado no exterior próximo da boca do voluntário de forma a recolher dois sinais temporalmente alinhados para posterior análise e comparação.

Estão definidos critérios de inclusão (maiores de 18 anos; indivíduos saudáveis) e de exclusão (fumadores; indivíduos com histórico de distúrbios de voz; indivíduos que não apresentem viabilidade após inspeção por rinoscopia anterior).

Benefício/risco:

Desconforto e possível irritação temporária causados pelo procedimento (nasofibrolaringoscopia). O incómodo de uma deslocação ao Serviço de Otorrinolaringologia do CHUSJ para uma consulta única.

Confidencialidade dos dados: É garantida a confidencialidade, ainda que não seja especificado o modo como esta é alcançada.

Respeito pela liberdade e autonomia do sujeito de ensaio:

Dispõe de uma adequada informação ao participante e de modelo de Consentimento Informado do CHUSJ.

Curriculum do investigador: Adequado à investigação.

Data previsível da conclusão do estudo: fevereiro de 2021

Conclusão: Proponho um parecer favorável à realização deste projeto de investigação. Solicitando, porém, informação sobre:

- a) O modo como serão recrutados os voluntários;
- b) Como é garantida a confidencialidade dos dados, desde logo o registo de som e imagem.

Porto, 19 de março de 2020

91

O Relator da CE, Prof. Doutor Rui Nunes

Peder Brito

Todas as questões foram esclarecidas
e confirmados pelo relator.

11/05/2020

Peder Brito



Questionário para submissão de Investigação

Exmo. Sr. Presidente da Comissão de Ética do Centro Hospitalar de São João/
Faculdade de Medicina da Universidade do Porto,

Pretendo realizar a investigação infracitada, solicito a V. Exa., na qualidade de Investigador, a sua apreciação e a elaboração do respetivo parecer. Para o efeito, anexo toda a documentação requerida.

IDENTIFICAÇÃO DO ESTUDO

Título da investigação: Dysphonic to Natural Voice Reconstruction (DyNaVoiceR) | Accurate glottal source estimation and modelling

Nome do investigador: Prof. Doutor Aníbal João de Sousa Ferreira

Endereço eletrónico: ajf@fe.up.pt

Contacto telefónico: 962631521

Caracterização da investigação:

- ☐ Estudo retrospectivo ☐ Estudo observacional ☐ Estudo prospetivo
☐ Inquérito ☒ Outro. Qual? Experimental Transversal

Tipo de investigação:

- ☒ Com intervenção ☐ Sem intervenção

Formação do investigador em boas práticas clínicas (GCP): ☐ Sim ☒ Não

Promotor (se aplicável): Prof. Doutor Jorge Spratley

Nome do orientador de dissertação/tese (se aplicável): Prof. Doutor Aníbal João de Sousa Ferreira

Endereço eletrónico: ajf@fe.up.pt

Local/locais onde se realiza a investigação: CHUSJ

Data prevista para início: ____ / ____ / ____

Data prevista para o término: ____ / ____ / ____

PROTOCOLO DO ESTUDO

Síntese dos objetivos:

O estudo visa caracterizar com precisão a verdadeira excitação acústica da glote, principalmente a vibração das pregas vocais, sob condições naturais de interação entre laringe e trato vocal.

Fundamentação ética (ganhos em conhecimento/inação; ponderação benefícios/riscos):

Irão ser recolhidos conhecimentos cruciais para o desenvolvimento de procedimentos inovadores e confiáveis que permitem sintetizar e implantar de forma não invasiva componentes de fala normal em fala sussurrada. Além disso, permitirá a obtenção de modelos mais fidedignos da fonte glótica que, por sua vez, permitirá estimar modelos do filtro do trato vocal com mais precisão.

CONFIDENCIALIDADE

De que forma é garantida a anonimização dos dados recolhidos de toda a informação?

É garantida a privacidade e confidencialidade.

O investigador necessita ter acesso a dados do processo clínico? ☐ Sim ☒ Não

Está previsto o registo de imagem ou som dos participantes? ☒ Sim ☐ Não

Se sim, está prevista a destruição deste registo após o sua utilização? ☐ Sim ☒ Não

CONSENTIMENTO

O estudo implica recrutamento de:

Doentes: ☐ Sim ☒ Não Voluntários saudáveis: ☒ Sim ☐ Não

Menores de 18 anos: ☐ Sim ☒ Não

Outras pessoas sem capacidade do exercício de autonomia: ☐ Sim ☒ Não

A investigação prevê a obtenção de Consentimento Informado: ☒ Sim ☐ Não

Se não, referir qual o fundamento para a isenção:

Existe informação escrita aos participantes: ☒ Sim ☐ Não

PROPRIEDADE DOS DADOS

A investigação e os seus resultados são propriedade intelectual de:

☐ Investigador ☐ Promotor ☒ Ambos ☒ Serviço onde é realizado

☐ Não aplicável Outro: _____

BENEFÍCIOS, RISCOS E CONTRAPARTIDAS PARA OS PARTICIPANTES

Benefícios previsíveis:

Nenhum benefício aplicável.

Riscos/incómodos previsíveis:

Desconforto e possível irritação temporária causados pelo procedimento (nasofibrolaringoscopia).

São dadas contrapartidas aos participantes:

· pela participação ☐ Sim ☒ Não ☐ Não aplicável

· pelas deslocações ☐ Sim ☒ Não ☐ Não aplicável

· pelas faltas ao emprego ☐ Sim ☒ Não ☐ Não aplicável

· por outras perdas e danos ☐ Sim ☒ Não ☐ Não aplicável

CUSTOS / PLANO FINANCEIRO

Os custos da investigação são suportados por:

☒ Investigador ☒ Promotor ☒ Serviço onde é realizado

☐ Não aplicável Outro: 'DyNaVoiceR' (PTDC/EMD-EDM/29308/2017)

Existe protocolo financeiro? ☒ Sim ☐ Não

LISTA DE DOCUMENTOS ANEXOS

- ☒ Pedido de autorização ao Presidente do Conselho de Administração do Centro Hospitalar de São João (se aplicável)
- ☒ Pedido de autorização à Diretora da Faculdade de Medicina da Universidade do Porto (se aplicável)
- ☒ Protocolo do estudo
- ☒ Declaração do Diretor de Serviço onde decorre o estudo
(sendo um estudo na área de enfermagem deve anexar também a concordância da chefia de enfermagem)
- ☒ Profissional de ligação
- ☒ Informação dos orientadores
- ☒ Informação ao participante
- ☒ Modelo de consentimento
- ☐ Instrumentos a utilizar (inquéritos, questionários, escalas, p.ex.): _____
- ☒ Curriculum Vitae abreviado (máx. 3 páginas)
- ☐ Protocolo financeiro
- ☐ Outros:

COMPROMISSO DE HONRA E DECLARAÇÃO DE INTERESSES

Declaro por minha honra que as informações prestadas neste questionário são verdadeiras. Mais declaro que, durante o estudo, serão respeitadas as recomendações constantes da Declaração de Helsínquia (1960 e respetivas emendas), e da Organização Mundial da Saúde, Convenção de Oviedo e das "Boas Práticas Clínicas" (GCP/ICH) no que se refere à experimentação que envolve seres humanos. Aceito, também, a recomendação da CES de que o recrutamento para este estudo se fará junto de doentes que não tenham participado em outro estudo, nos últimos três meses. Comprometo-me a entregar à CES o relatório final da investigação, assim que concluído.

Porto, 13 de fevereiro de 2020
Nome legível: Aníbal João de Sousa Ferreira

Assinado por: ANÍBAL JOÃO DE SOUSA FERREIRA
Num. de Identificação: BI05956694
Data: 2020.02.20 20:46:22 +0000

assinatura

Parecer da Comissão de Ética do Centro Hospitalar de São João/FMUP



CHAVE MÓVEL

Ária da CE de 19 / 03 / 2020

Aguarda esclarecimentos.

Prof. Doutor Filipe Almeida
Presidente da Comissão de Ética

Centro Hospitalar São João

CONSIDERADOS QUE FORAM COMO SATISFATÓRIOS OS
ESCLARECIMENTOS PRESTADOS PELO(A)
INVESTIGADOR(A), A CES APROVA POR UNANIMIDADE O
PARECER DO RELATOR, PELO QUE NADA TEM A OPOR À
REALIZAÇÃO DESTA PROJETO DE INVESTIGAÇÃO.

Prof. Doutor Filipe Almeida
Presidente da Comissão de Ética

11 / 05 / 2020

Appendix B

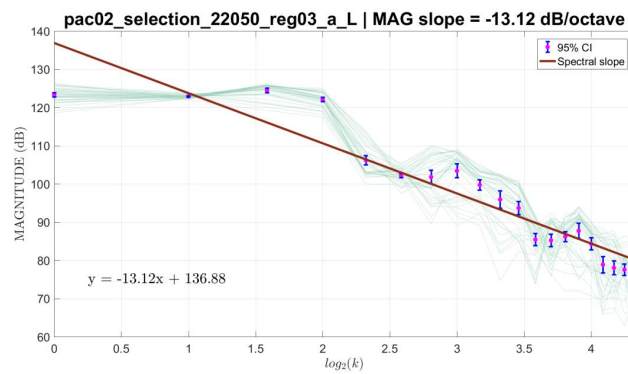
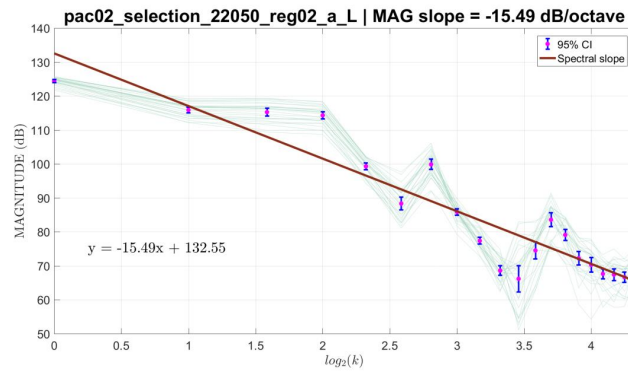
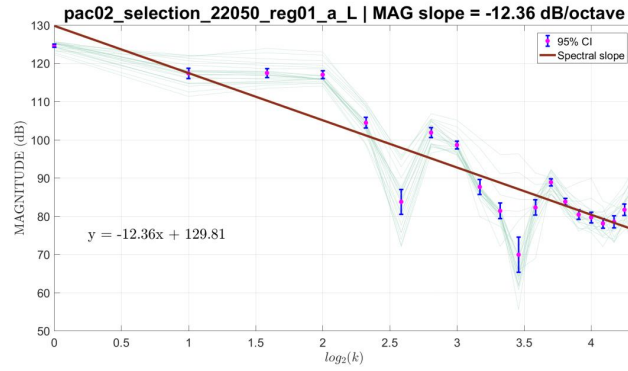
Glottal Source Characterization

B.1 Spectral Magnitude Analysis

SPECTRAL MAGNITUDE SLOPE PER FILE

SPEAKER 2
VOWEL /a/

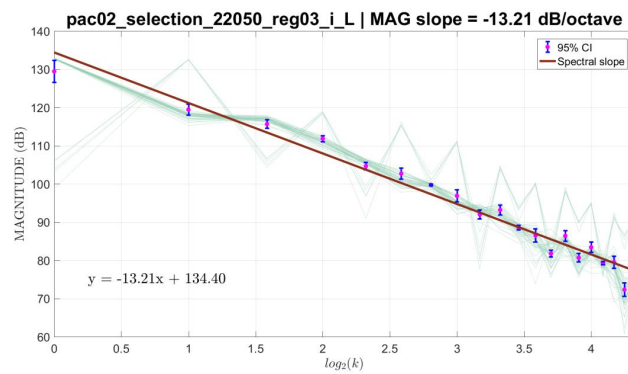
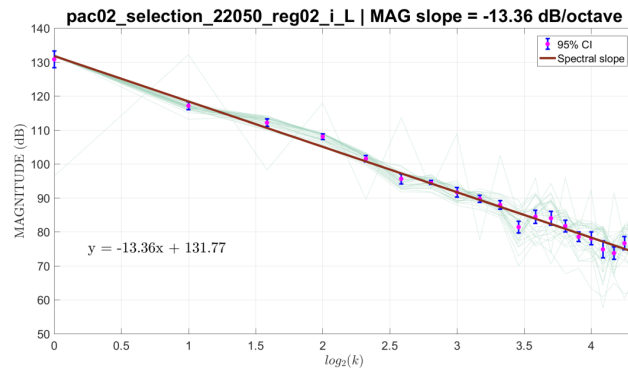
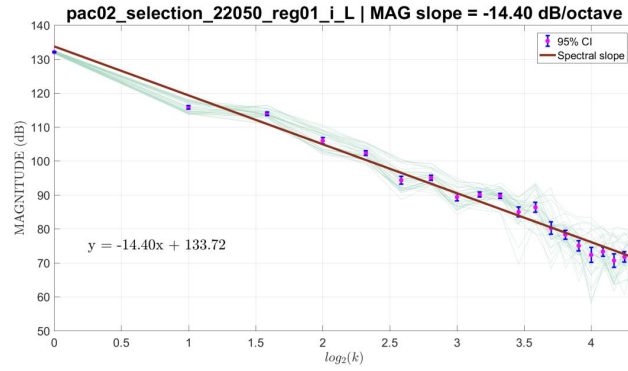
L
(INTERNAL)



SPECTRAL MAGNITUDE SLOPE PER FILE

SPEAKER 2
VOWEL /i/

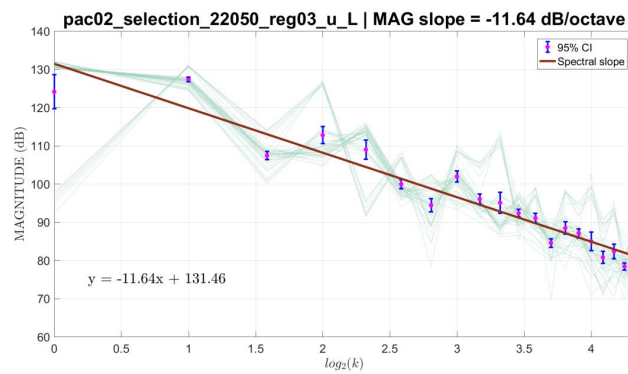
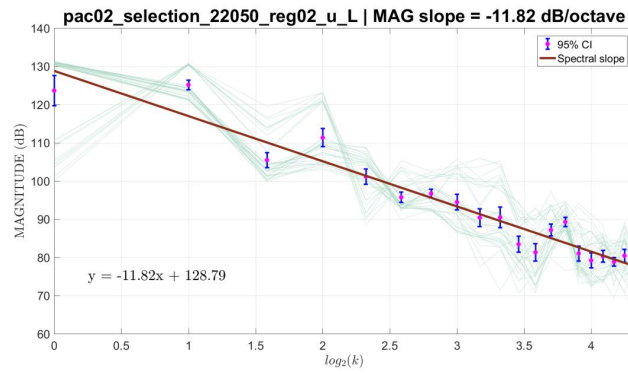
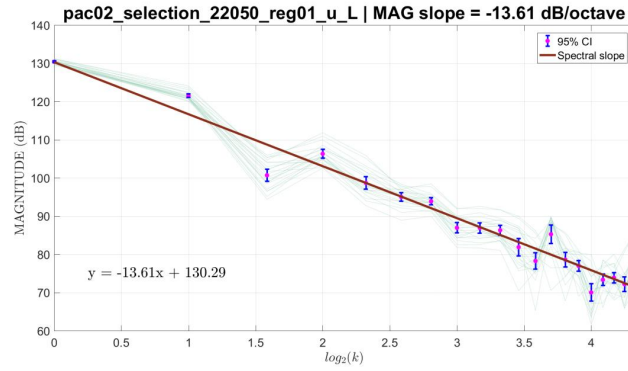
L
(INTERNAL)



SPECTRAL MAGNITUDE SLOPE PER FILE

SPEAKER 2
VOWEL /u/

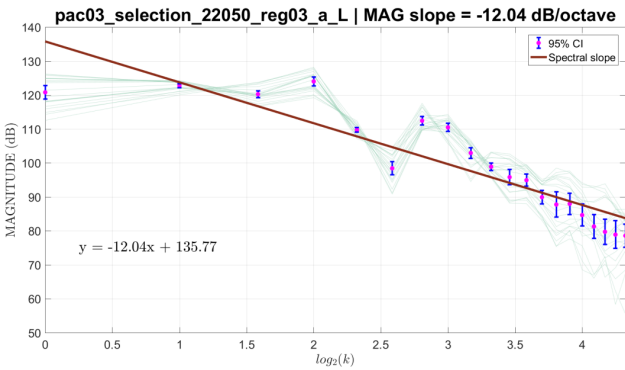
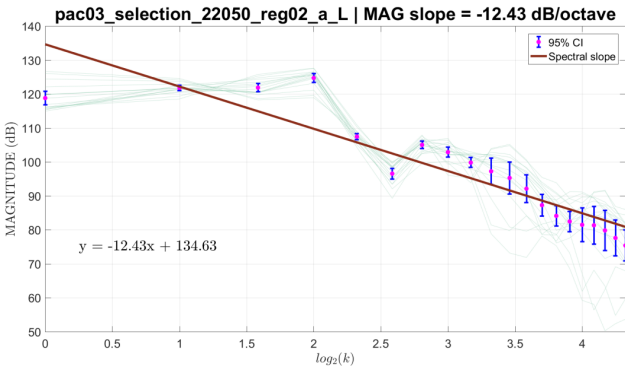
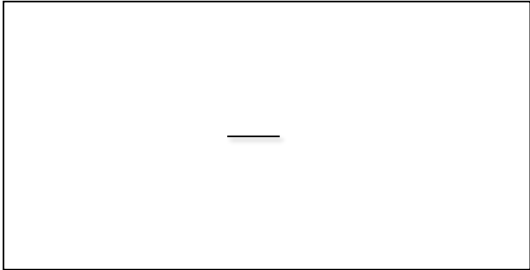
L
(INTERNAL)



SPECTRAL MAGNITUDE SLOPE PER FILE

SPEAKER 3
VOWEL /a/

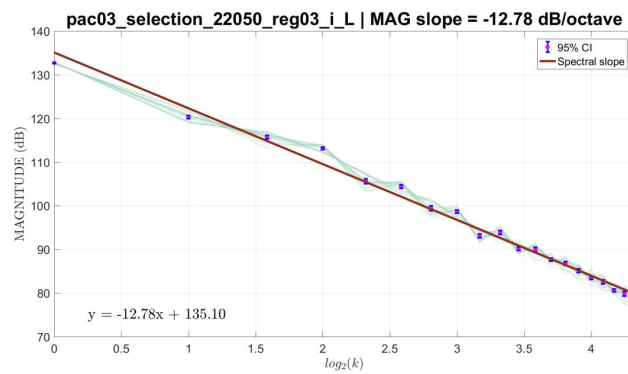
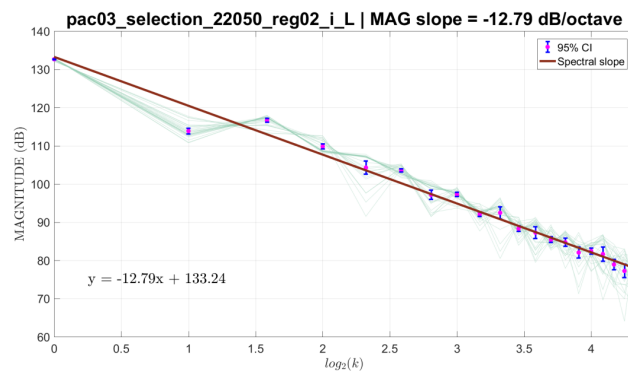
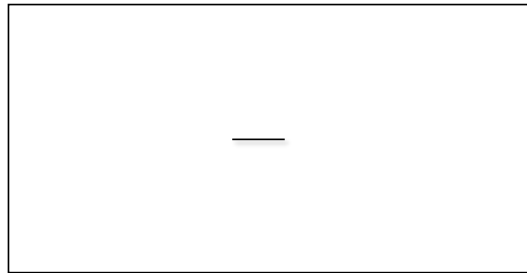
L
(INTERNAL)



SPECTRAL MAGNITUDE SLOPE PER FILE

SPEAKER 3
VOWEL /i/

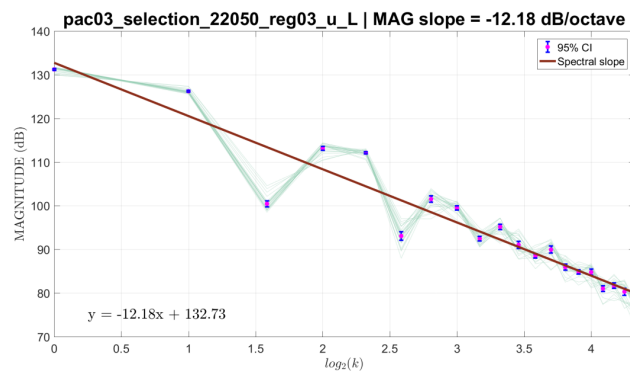
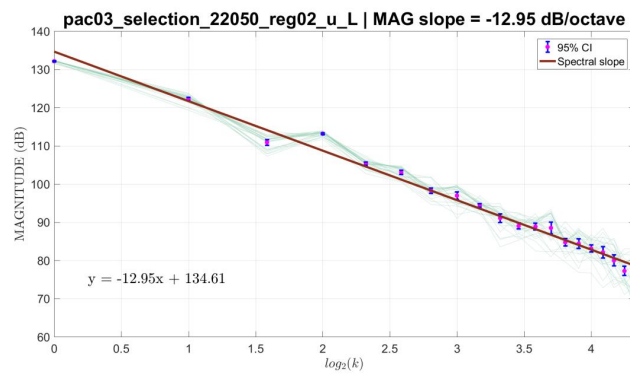
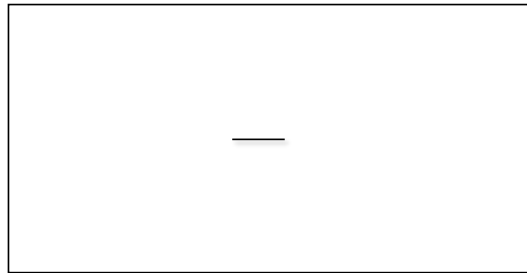
L
(INTERNAL)



SPECTRAL MAGNITUDE SLOPE PER FILE

SPEAKER 3
VOWEL /u/

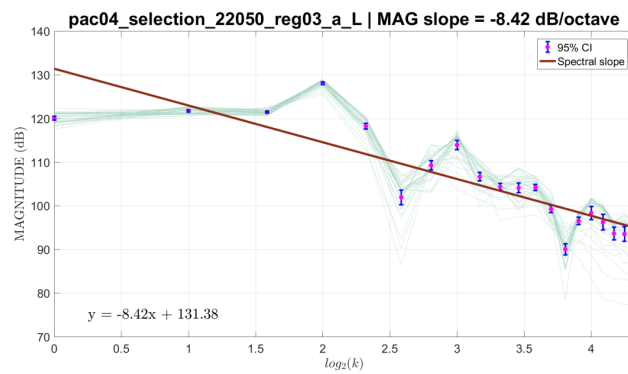
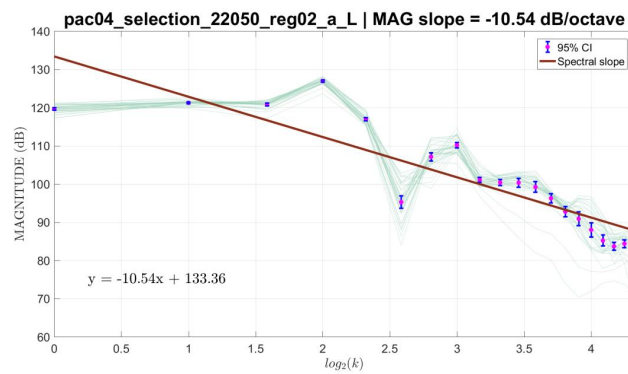
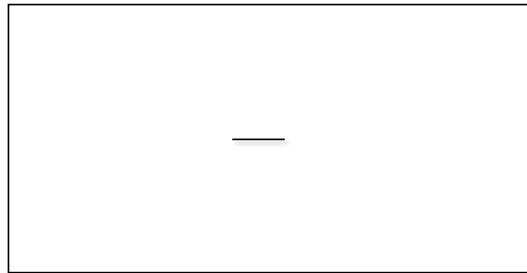
L
(INTERNAL)



SPECTRAL MAGNITUDE SLOPE PER FILE

SPEAKER 4
VOWEL /a/

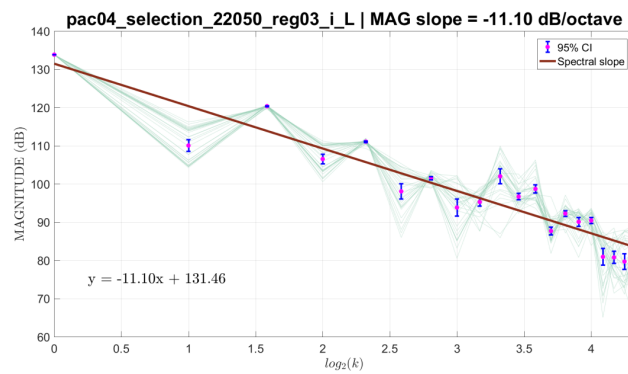
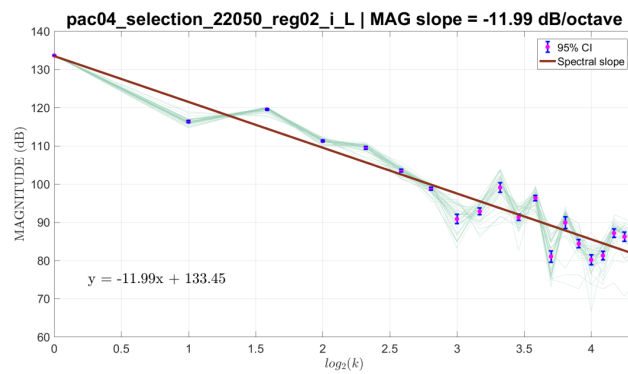
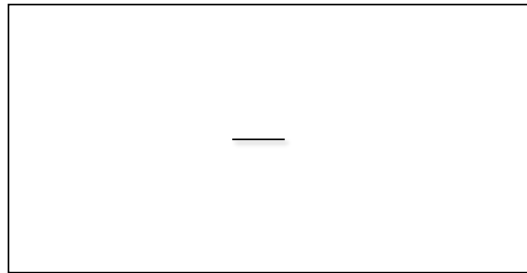
L
(INTERNAL)



SPECTRAL MAGNITUDE SLOPE PER FILE

SPEAKER 4
VOWEL /i/

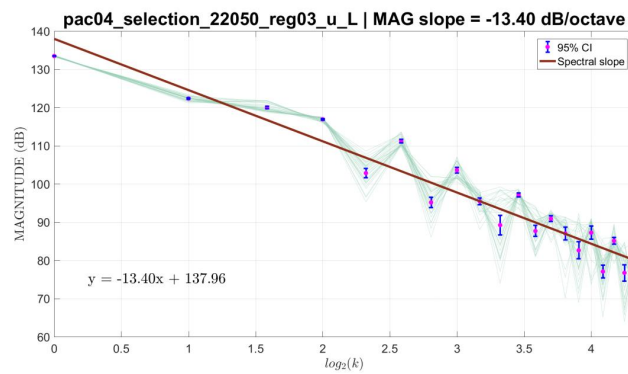
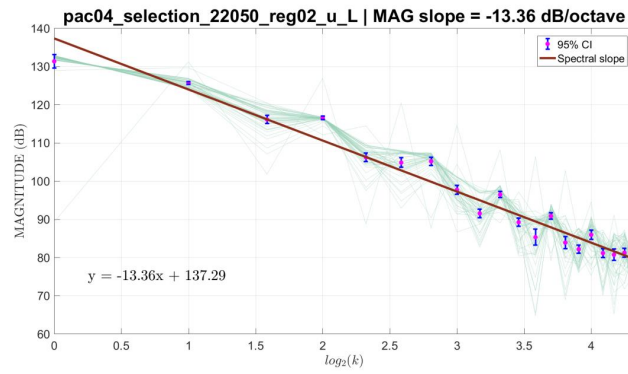
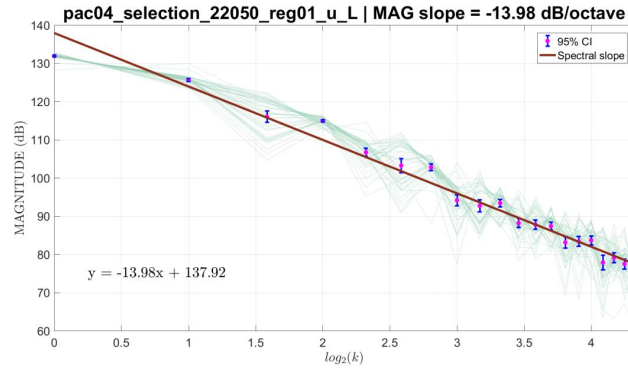
L
(INTERNAL)



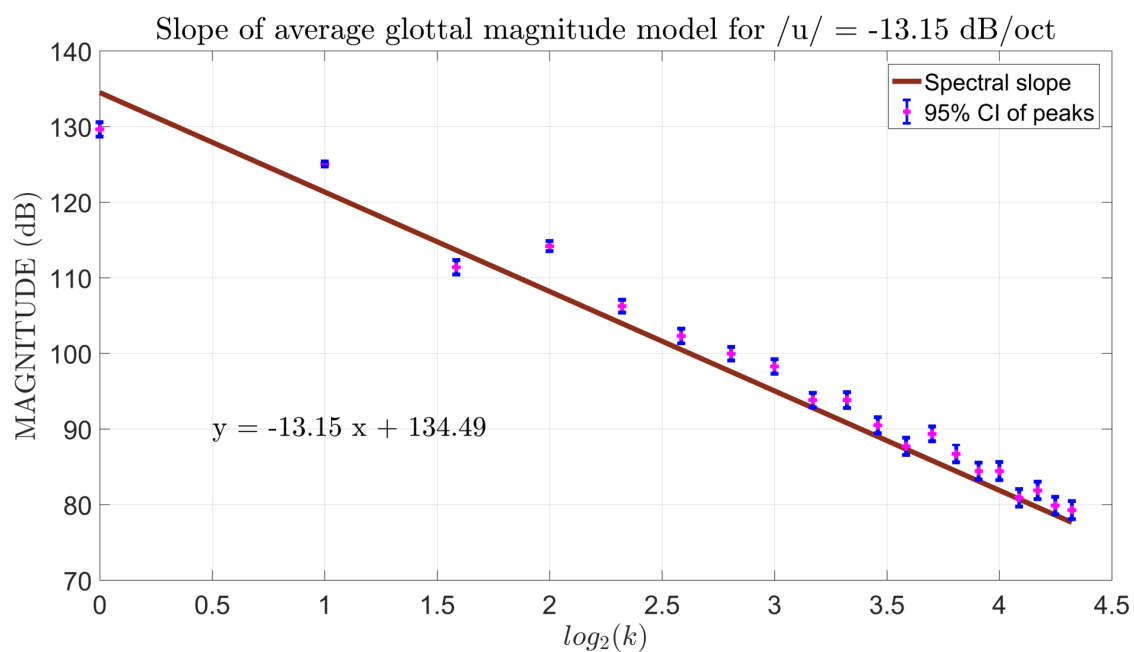
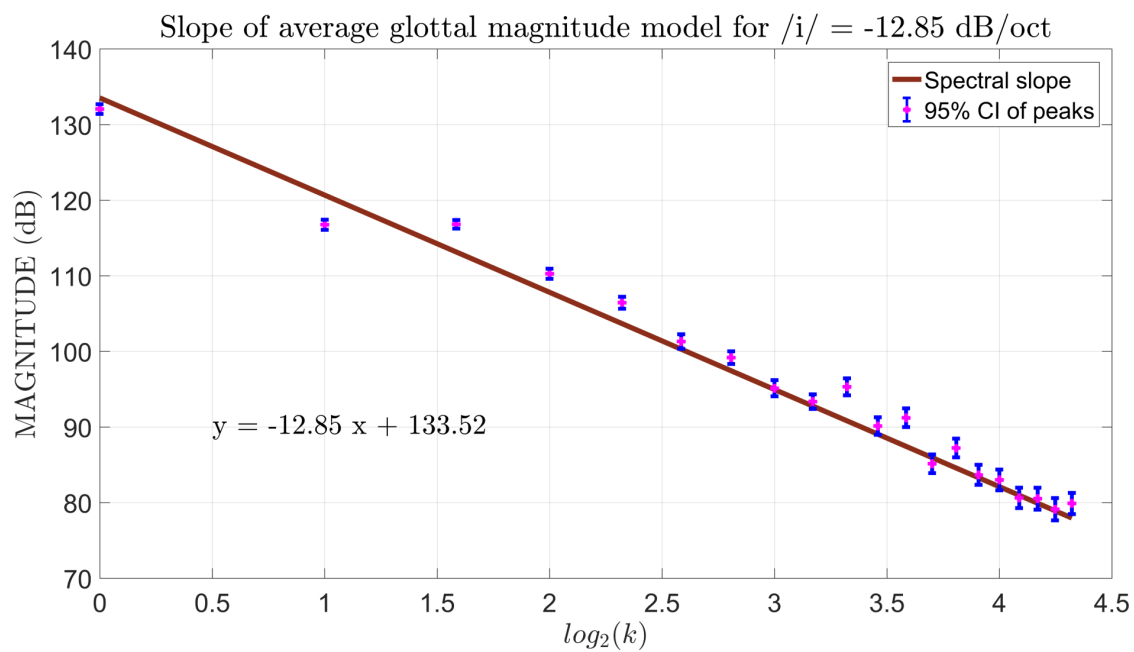
SPECTRAL MAGNITUDE SLOPE PER FILE

SPEAKER 4
VOWEL /u/

L
(INTERNAL)



MEAN SPECTRAL MAGNITUDE SLOPE PER VOWEL



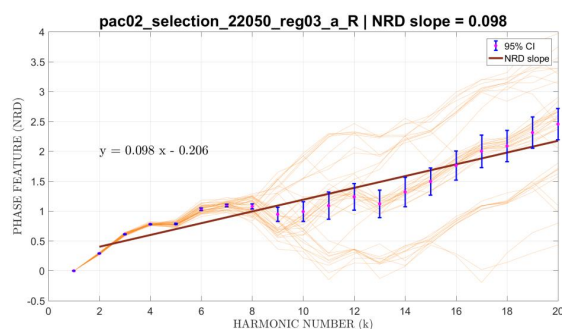
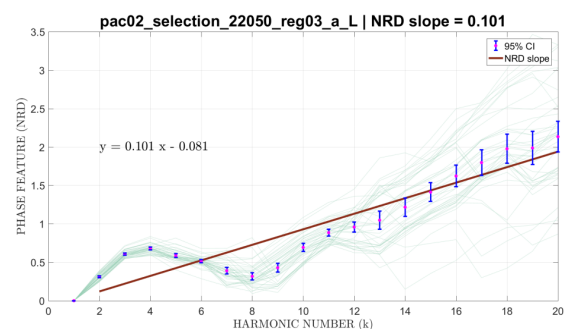
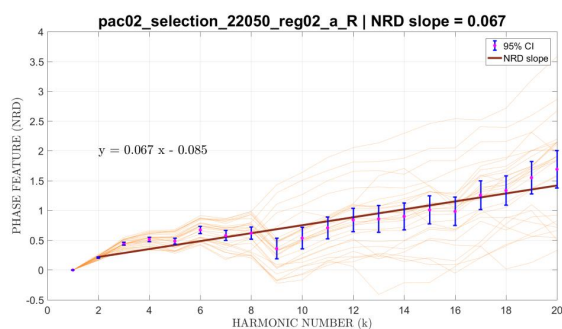
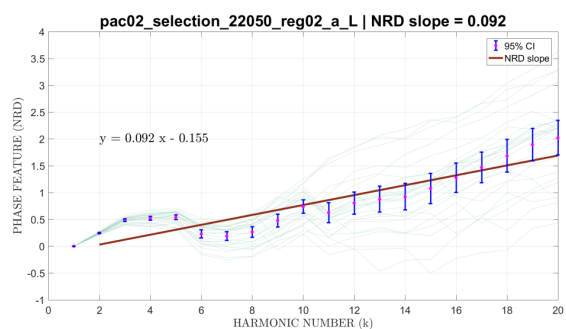
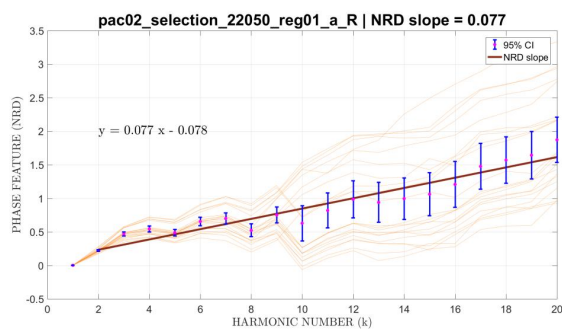
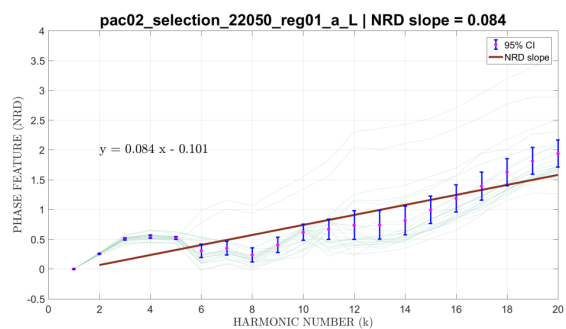
B.2 Spectral Phase Analysis

SPECTRAL NRD SLOPE PER FILE

SPEAKER 2
VOWEL /a/

L
(INTERNAL)

R
(EXTERNAL)

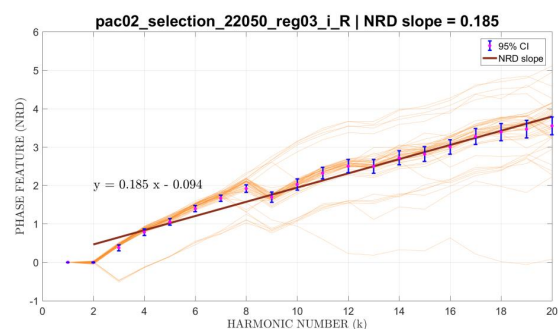
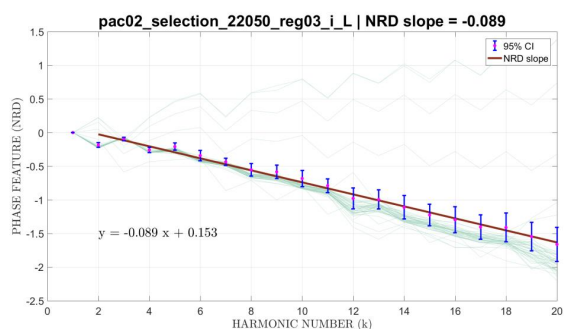
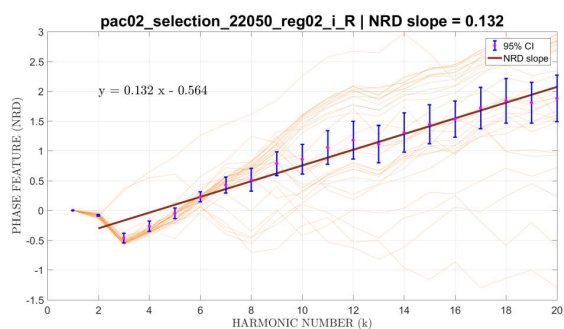
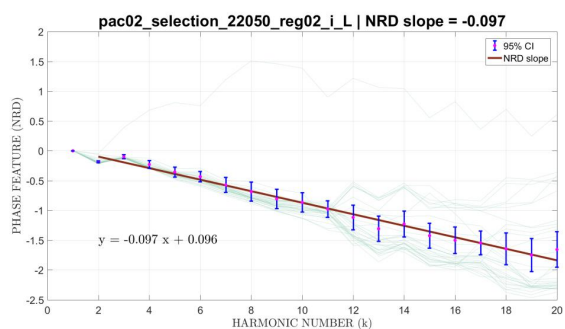
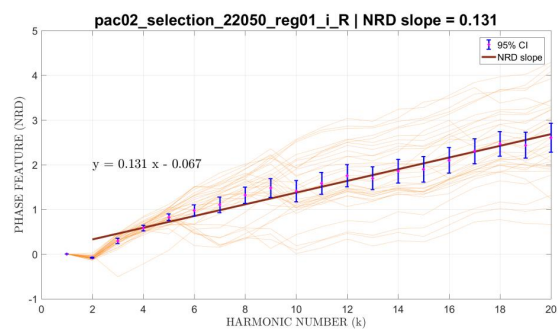
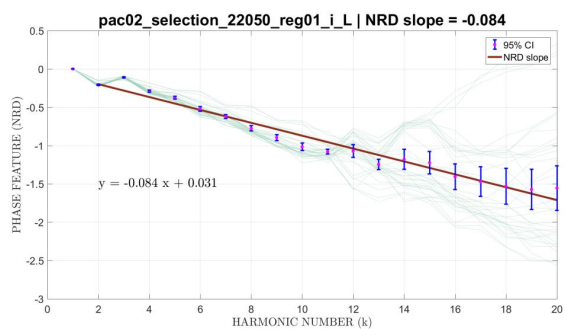


SPECTRAL NRD SLOPE PER FILE

SPEAKER 2
VOWEL /i/

L
(INTERNAL)

R
(EXTERNAL)

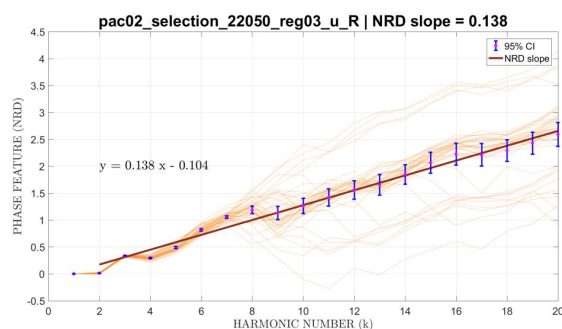
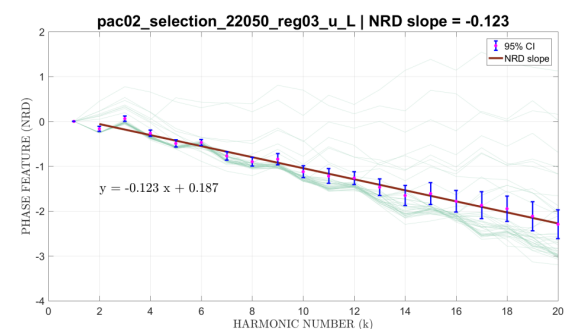
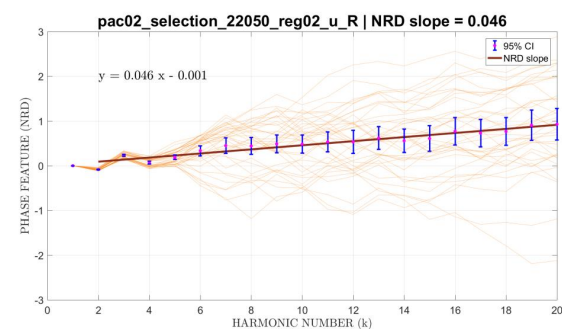
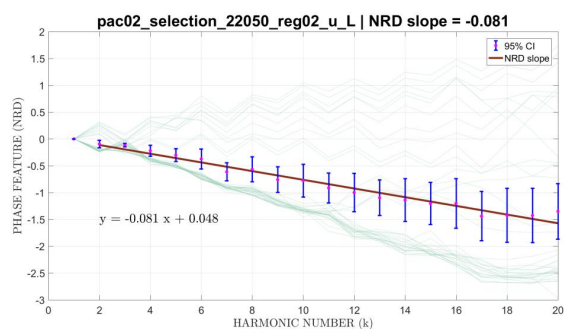
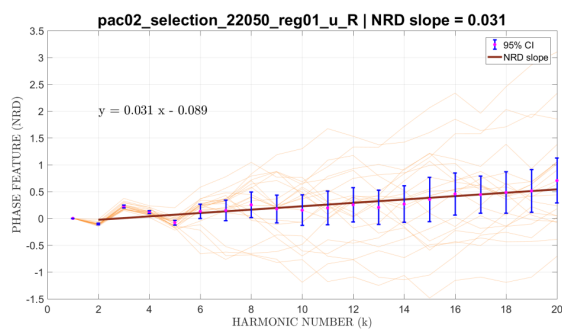
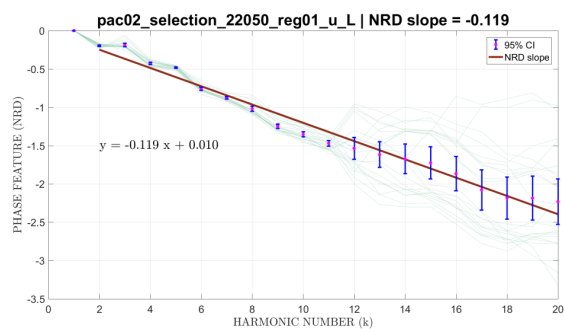


SPECTRAL NRD SLOPE PER FILE

SPEAKER 2
VOWEL /u/

L
(INTERNAL)

R
(EXTERNAL)

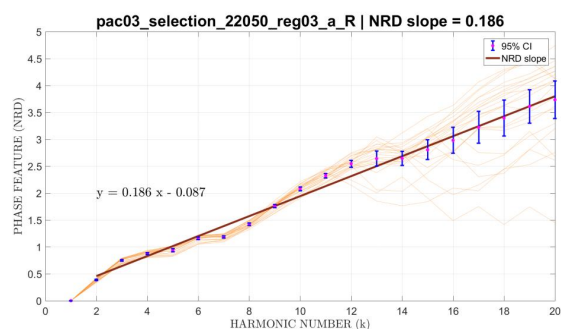
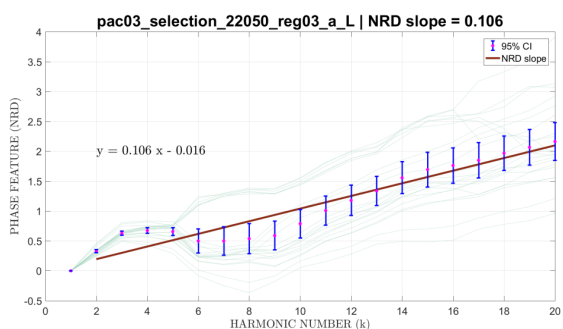
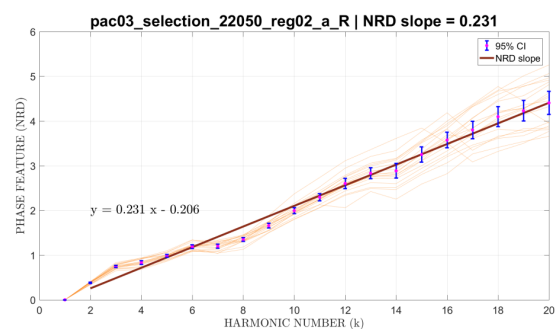
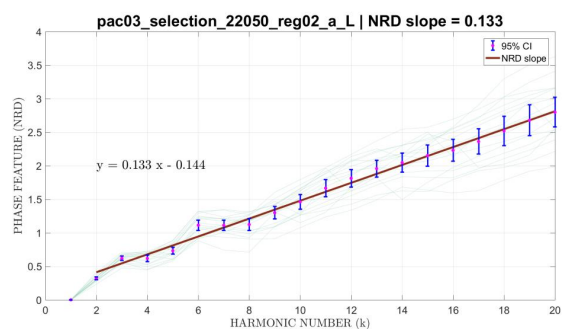
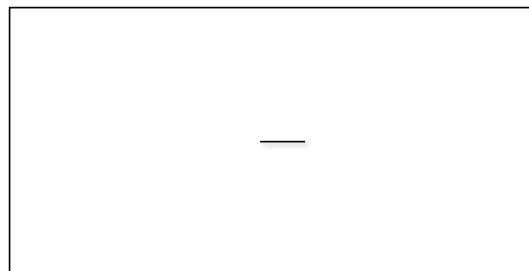
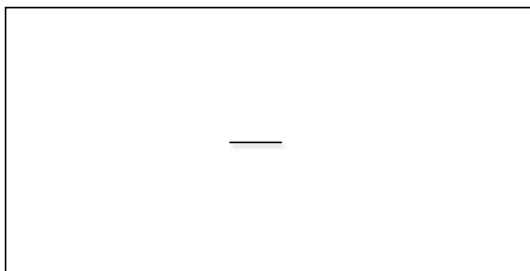


SPECTRAL NRD SLOPE PER FILE

SPEAKER 3
VOWEL /a/

L
(INTERNAL)

R
(EXTERNAL)

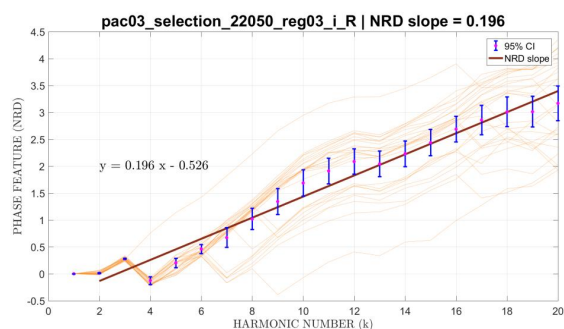
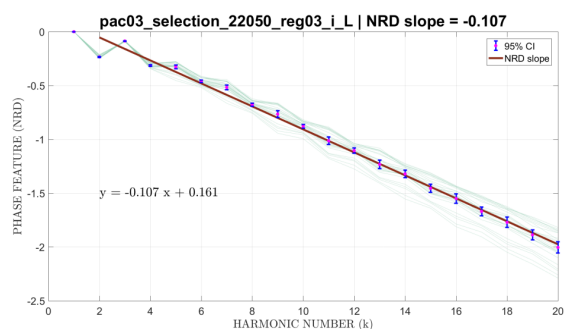
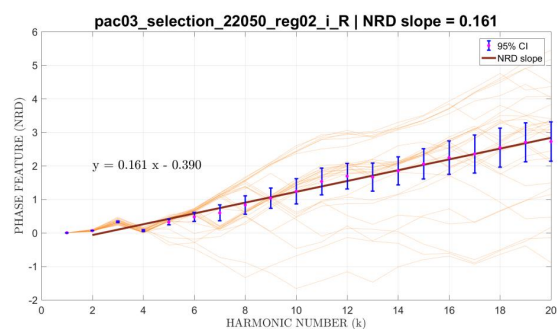
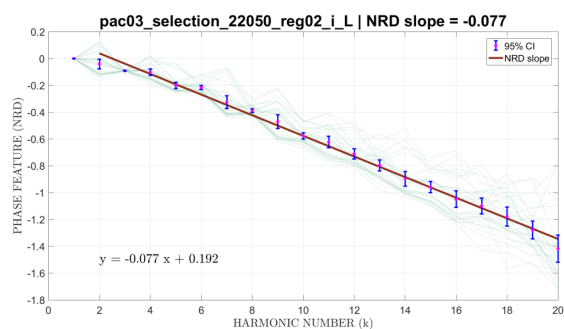
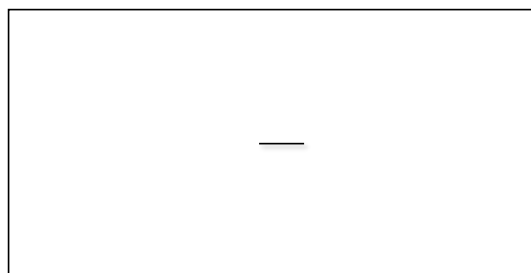
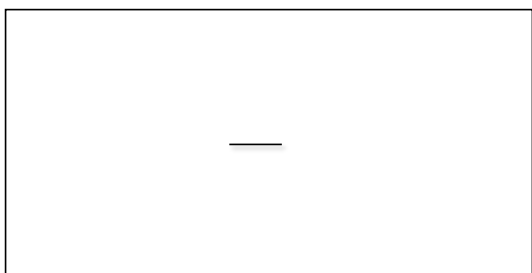


SPECTRAL NRD SLOPE PER FILE

SPEAKER 3
VOWEL /i/

L
(INTERNAL)

R
(EXTERNAL)

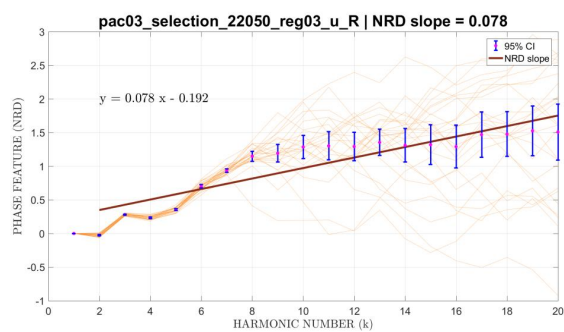
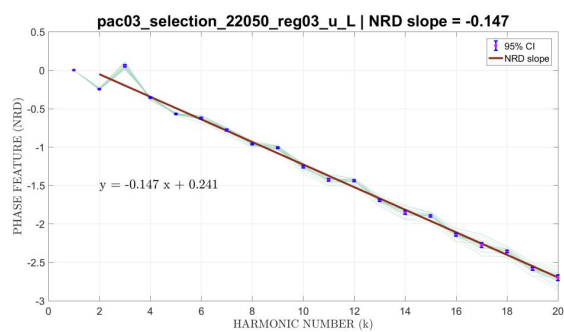
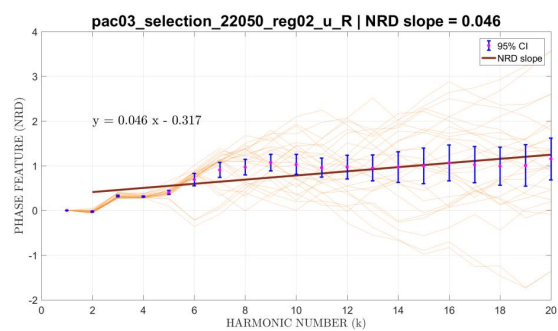
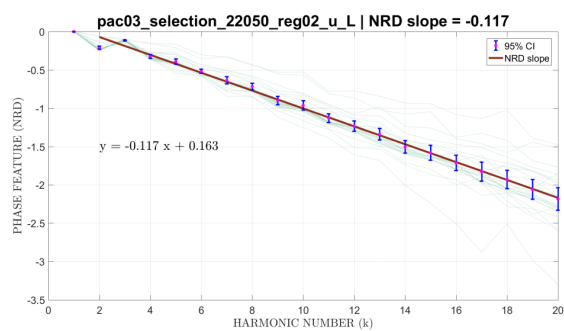
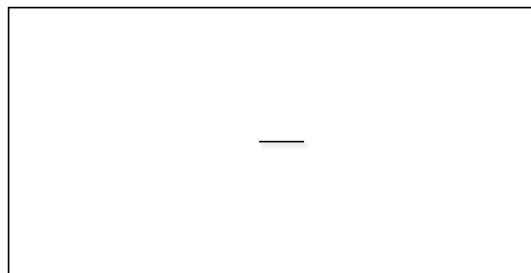
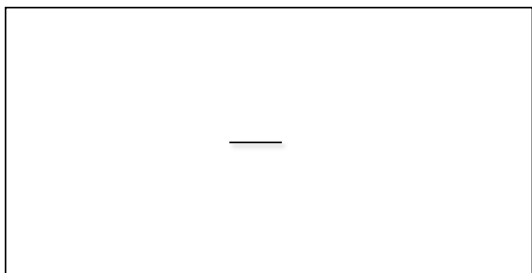


SPECTRAL NRD SLOPE PER FILE

SPEAKER 3
VOWEL /u/

L
(INTERNAL)

R
(EXTERNAL)

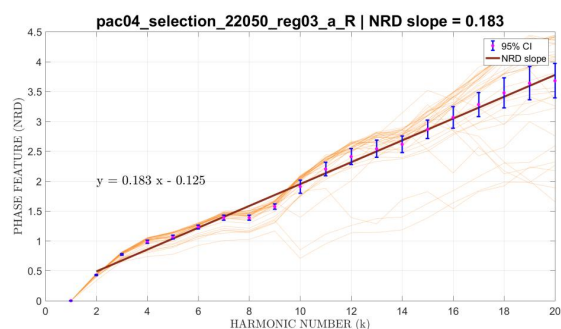
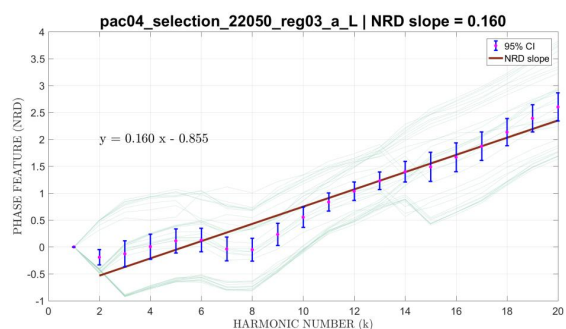
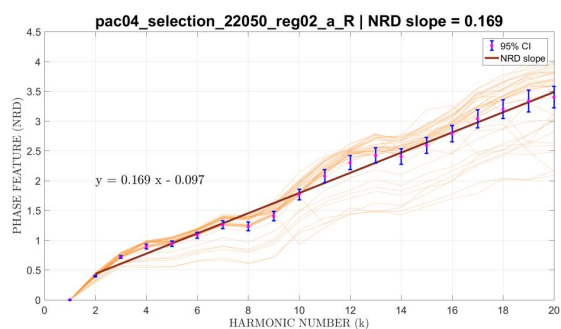
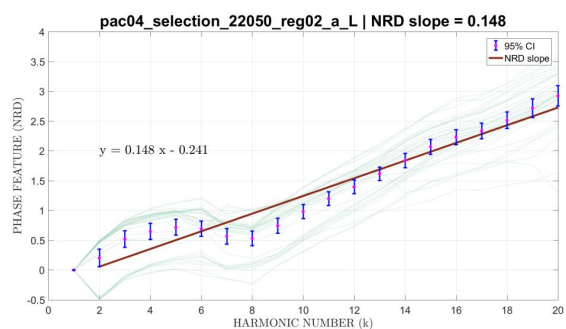
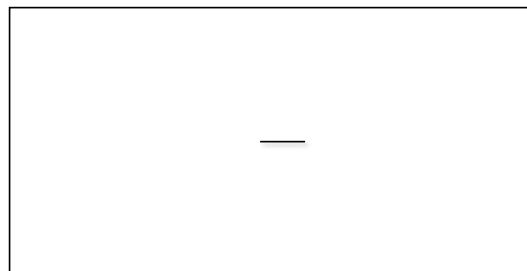
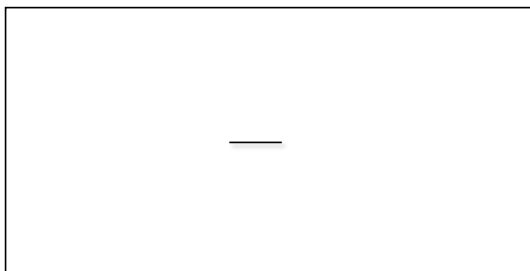


SPECTRAL NRD SLOPE PER FILE

SPEAKER 4
VOWEL /a/

L
(INTERNAL)

R
(EXTERNAL)

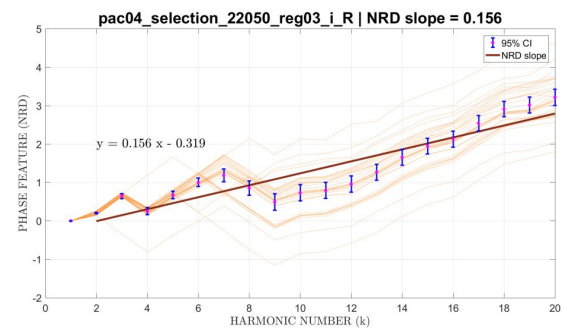
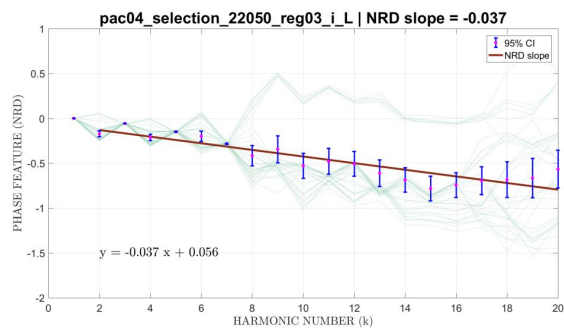
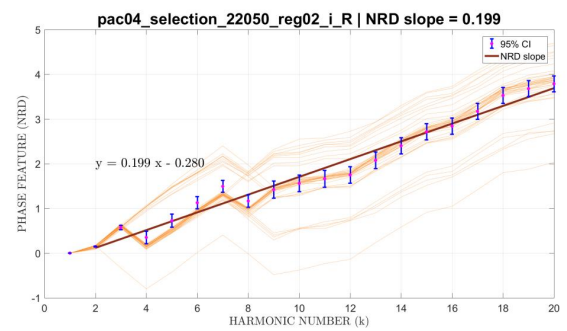
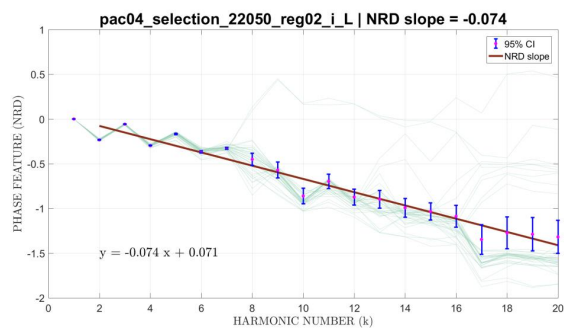
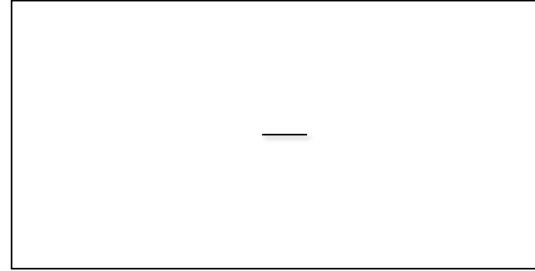
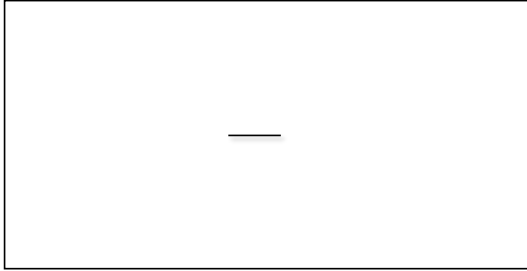


SPECTRAL NRD SLOPE PER FILE

SPEAKER 4
VOWEL /i/

L
(INTERNAL)

R
(EXTERNAL)

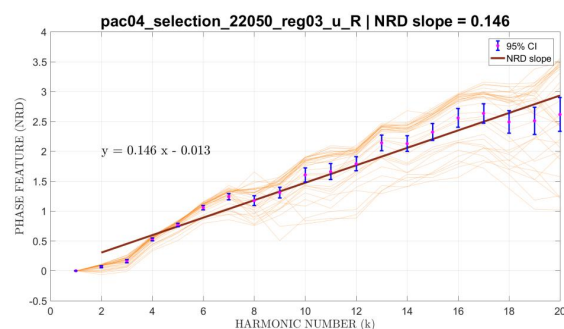
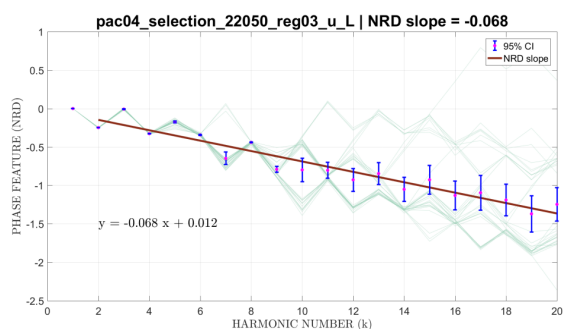
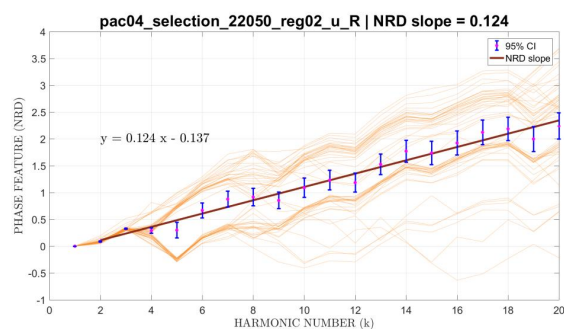
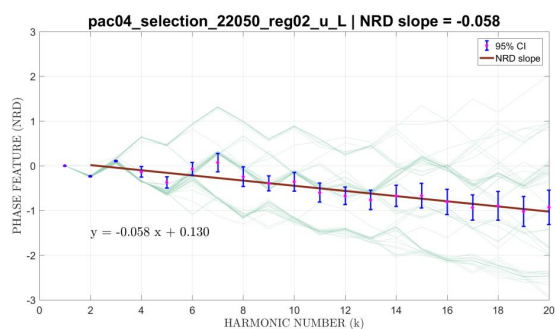
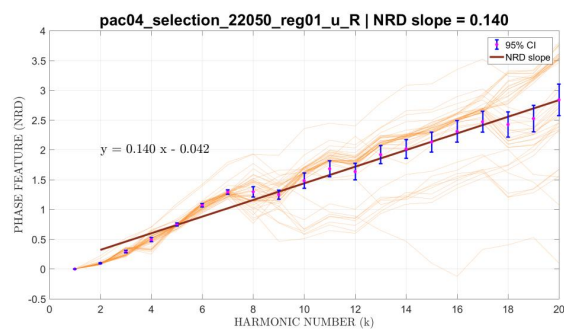
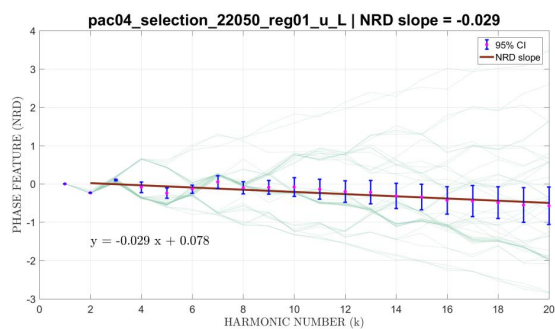


SPECTRAL NRD SLOPE PER FILE

SPEAKER 4
VOWEL /u/

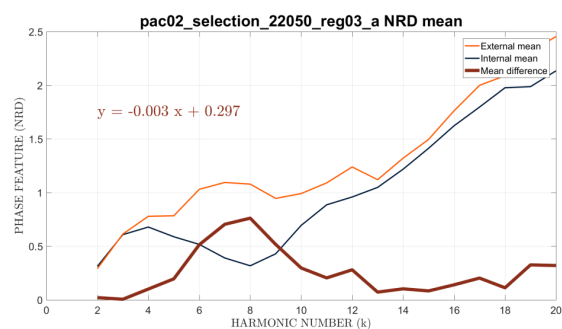
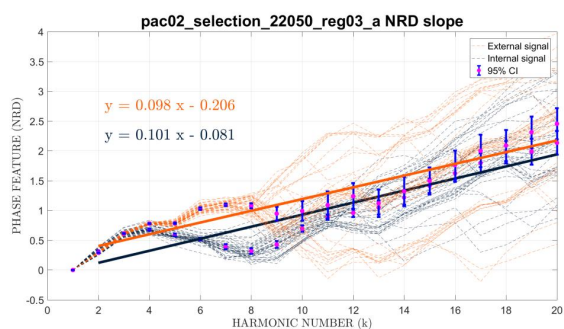
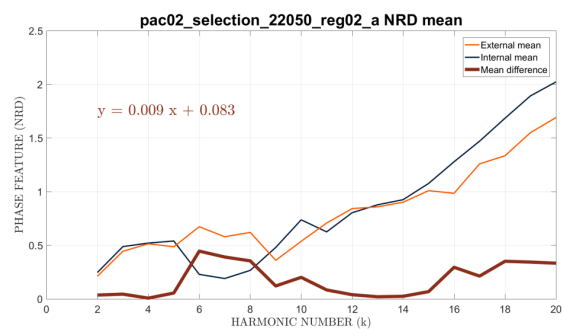
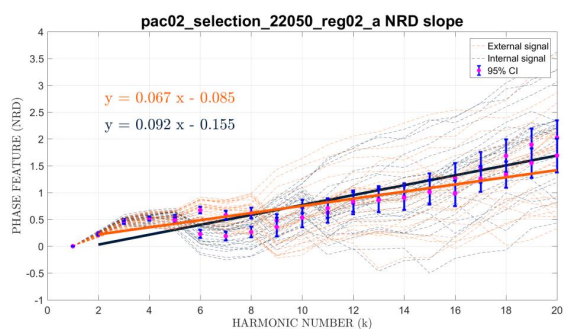
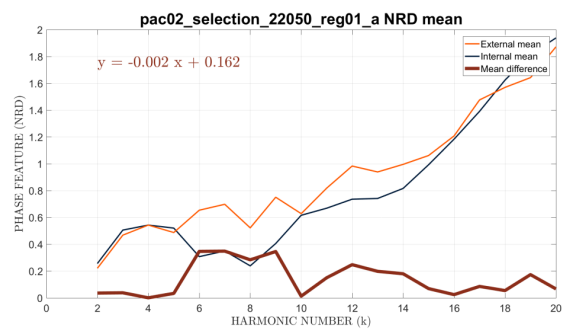
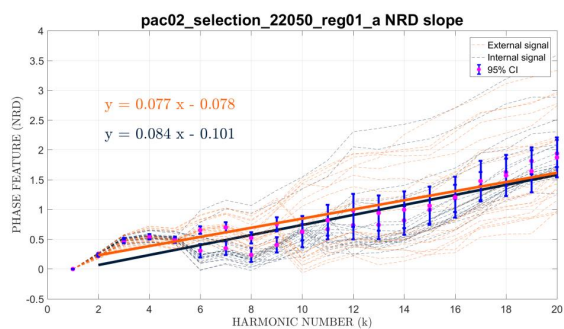
L
(INTERNAL)

R
(EXTERNAL)



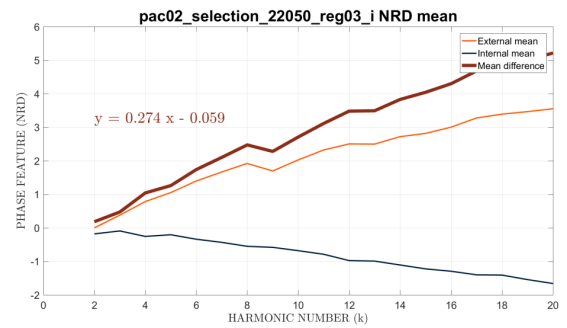
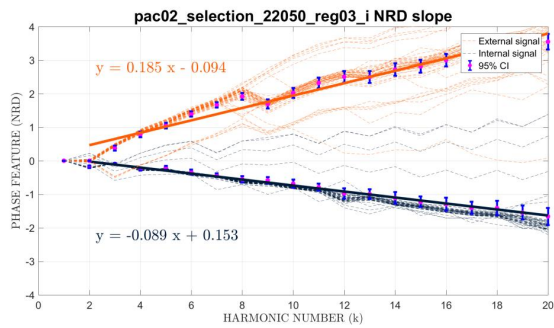
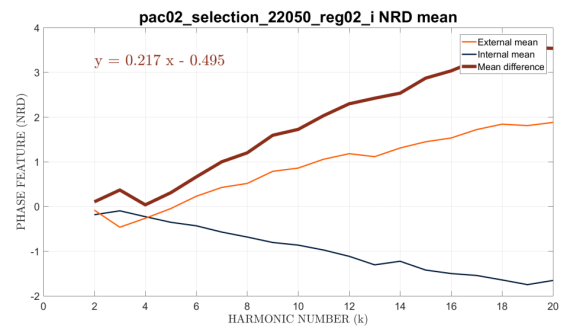
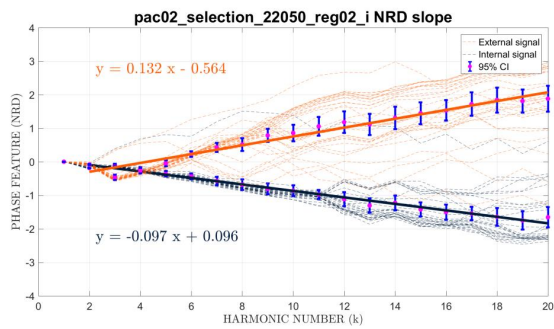
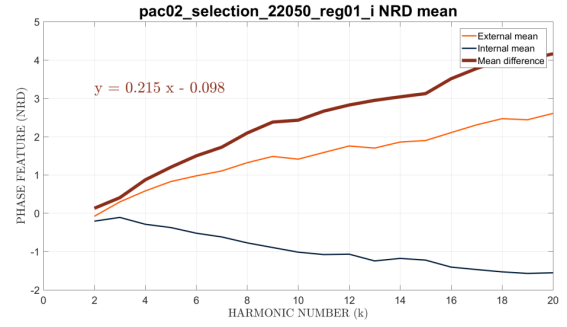
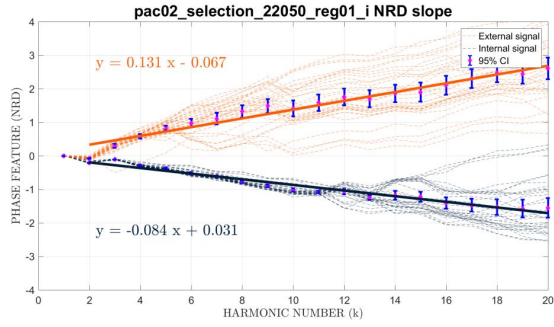
SPECTRAL NRD SLOPE AND NRD MEAN DIFFERENCE

SPEAKER 2
VOWEL /a/



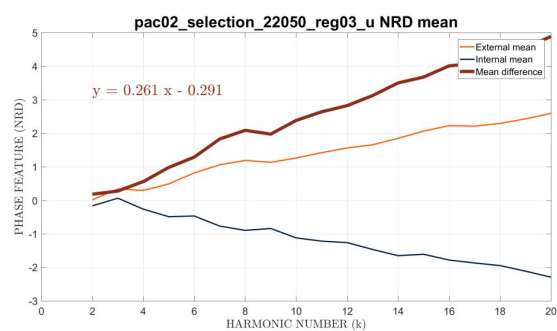
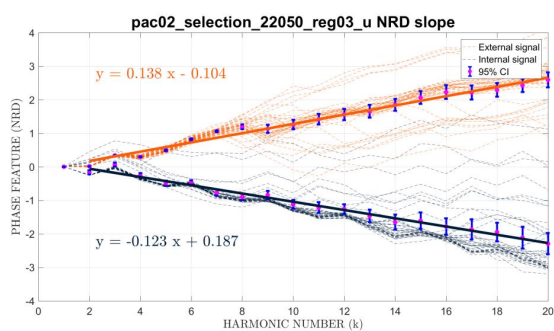
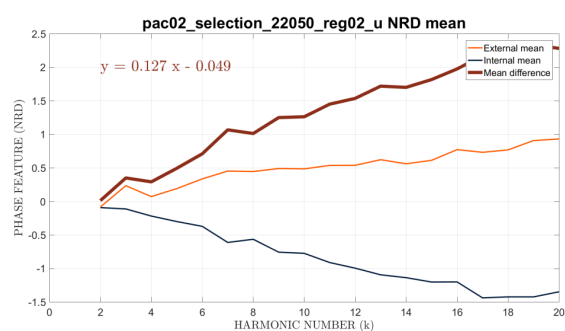
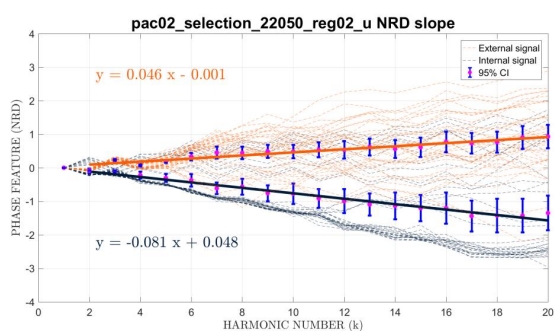
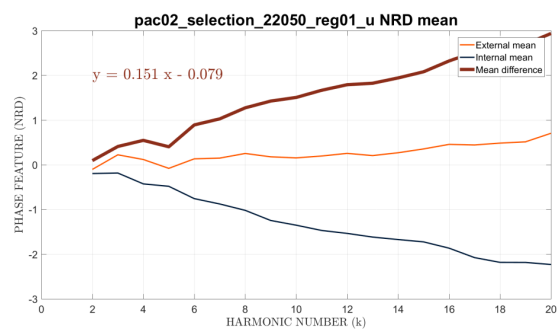
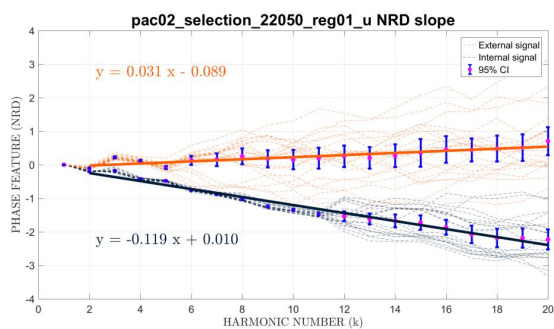
SPECTRAL NRD SLOPE AND NRD MEAN DIFFERENCE

SPEAKER 2
VOWEL /i/



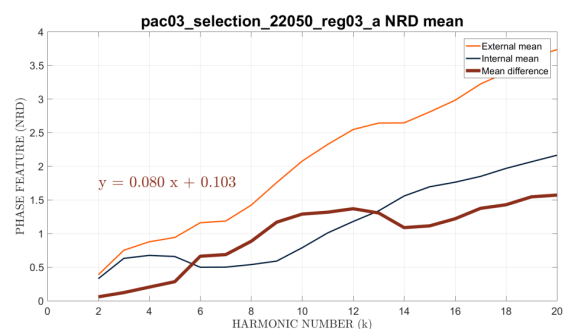
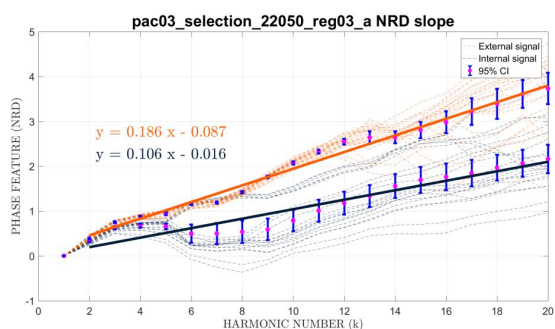
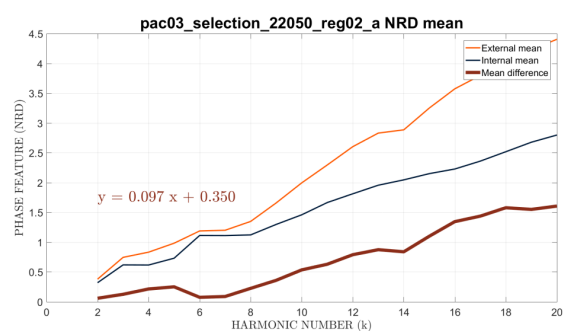
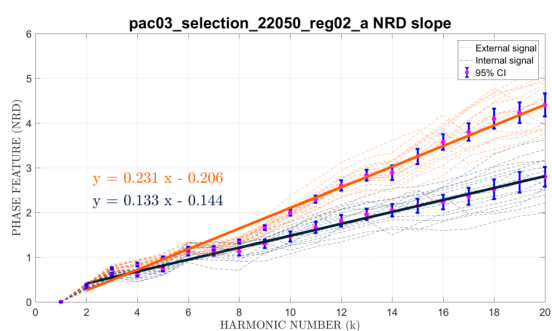
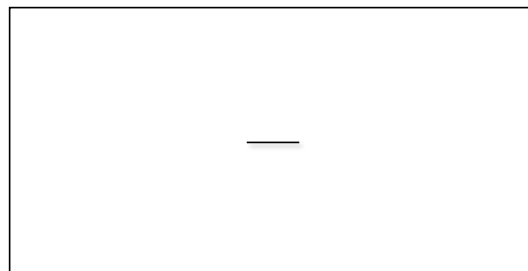
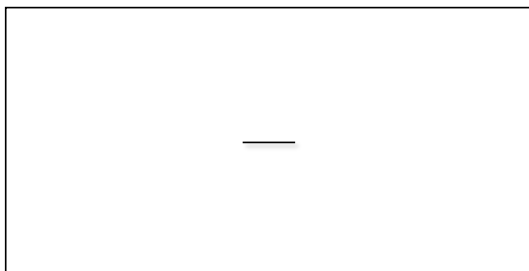
SPECTRAL NRD SLOPE AND NRD MEAN DIFFERENCE

SPEAKER 2
VOWEL /u/



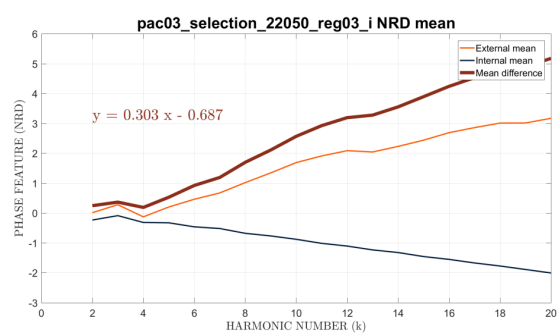
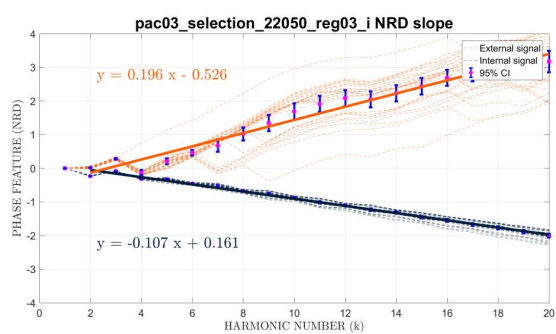
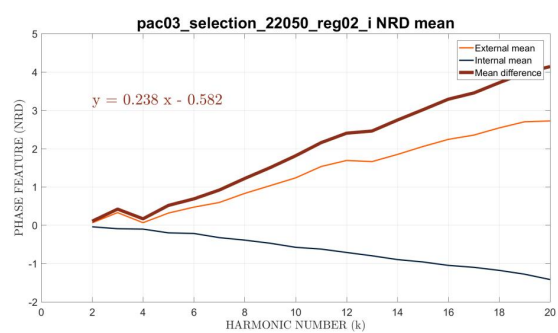
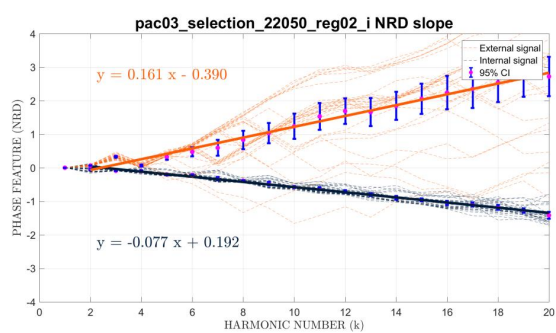
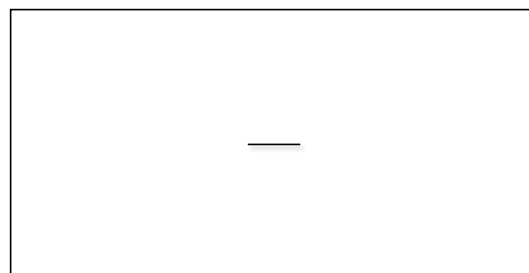
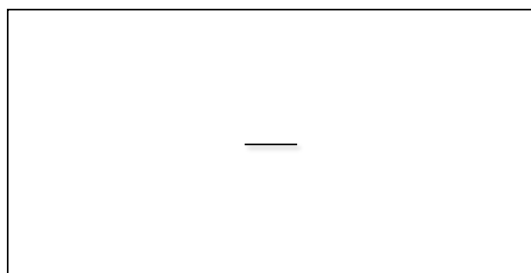
SPECTRAL NRD SLOPE AND NRD MEAN DIFFERENCE

SPEAKER 3
VOWEL /a/



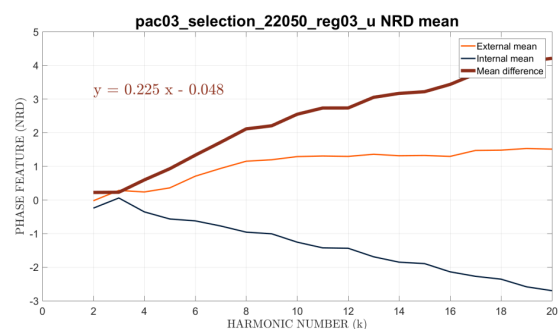
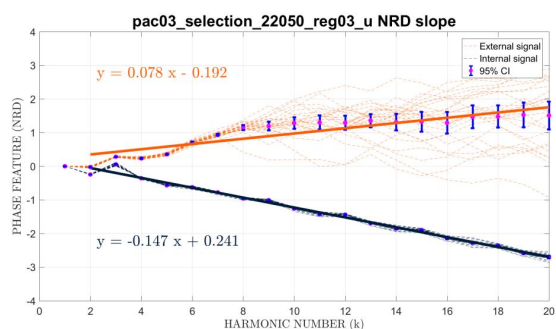
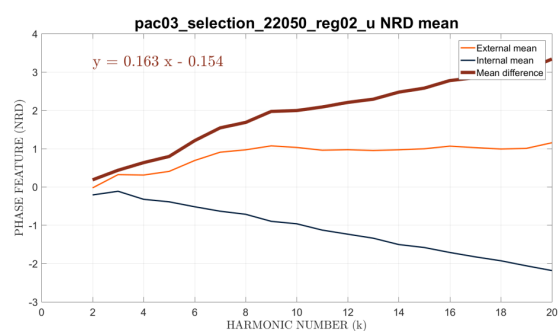
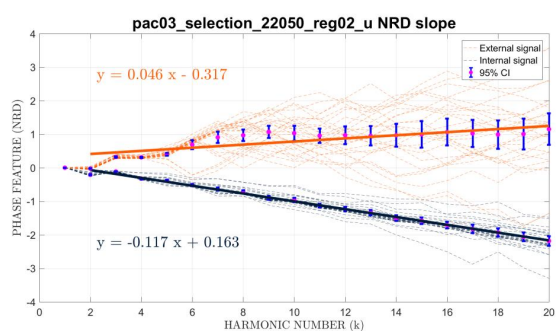
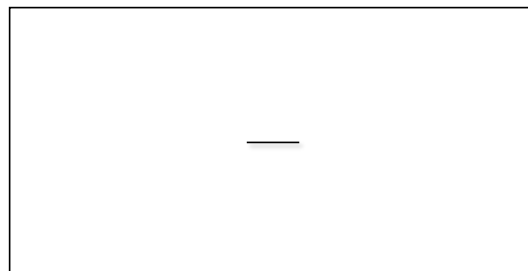
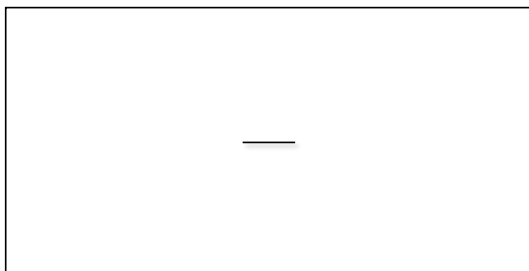
SPECTRAL NRD SLOPE AND NRD MEAN DIFFERENCE

SPEAKER 3
VOWEL /i/



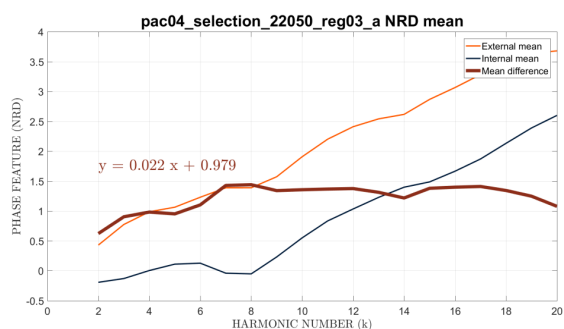
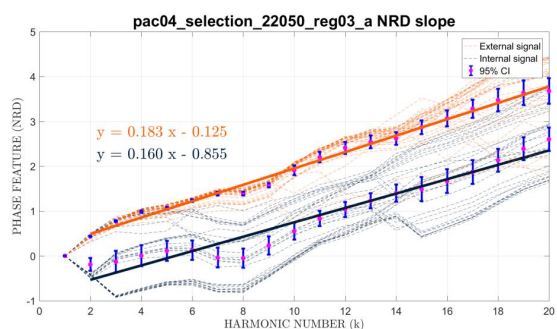
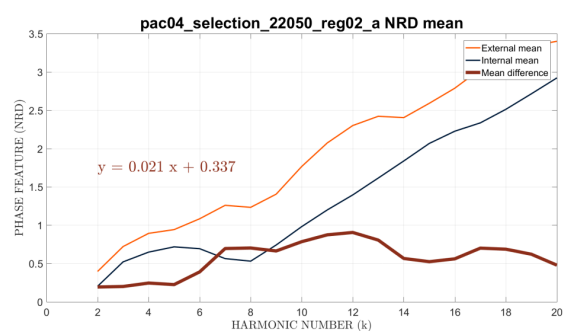
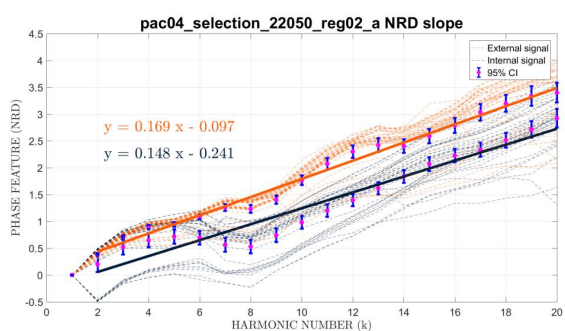
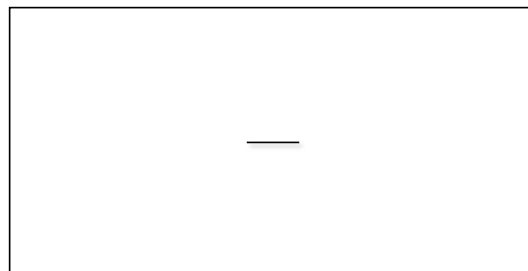
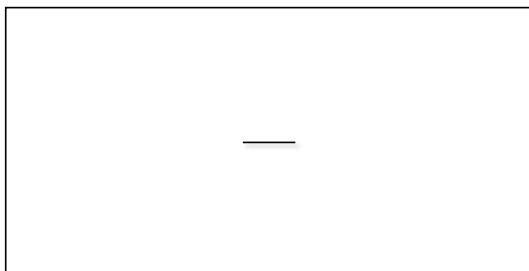
SPECTRAL NRD SLOPE AND NRD MEAN DIFFERENCE

SPEAKER 3
VOWEL /u/



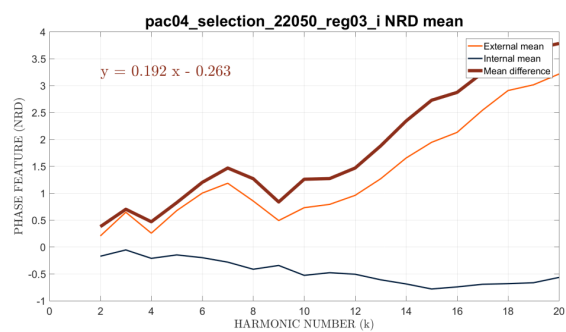
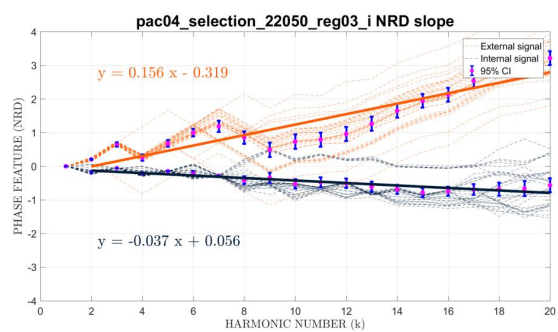
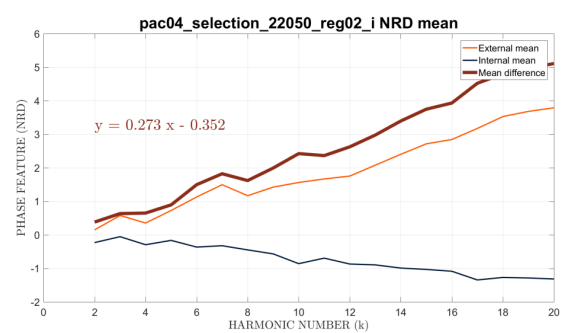
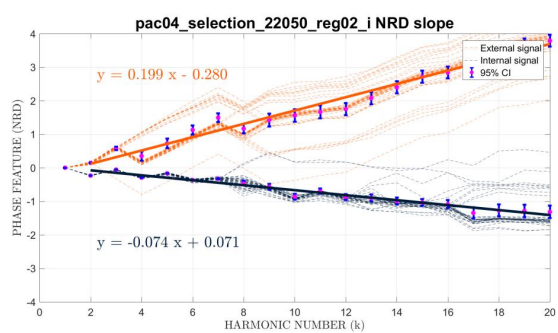
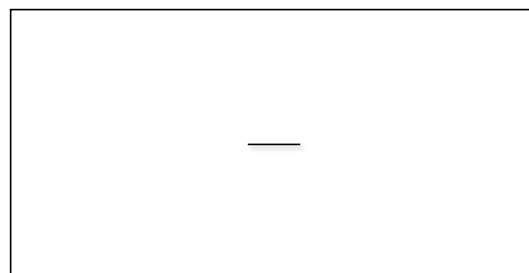
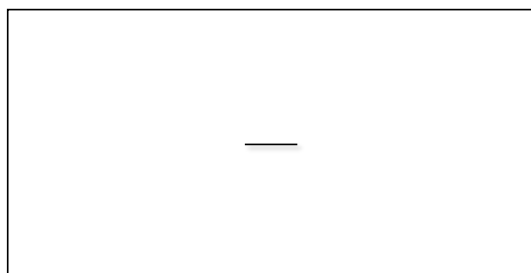
SPECTRAL NRD SLOPE AND NRD MEAN DIFFERENCE

SPEAKER 4
VOWEL /a/



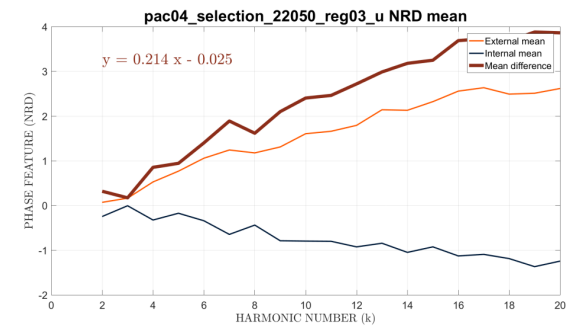
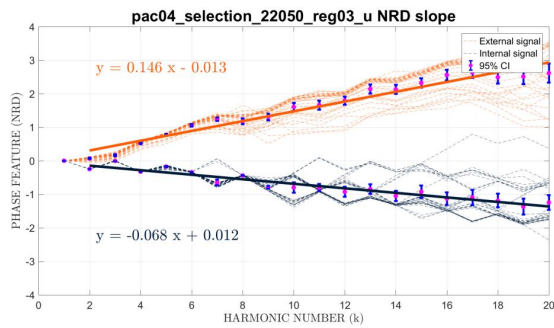
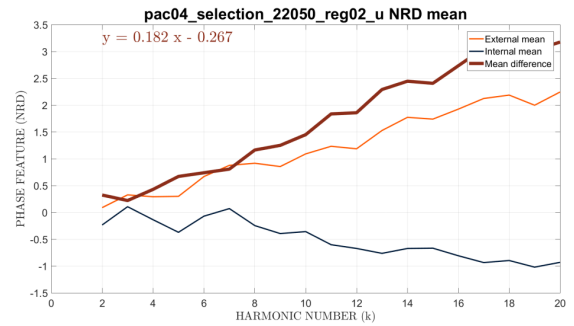
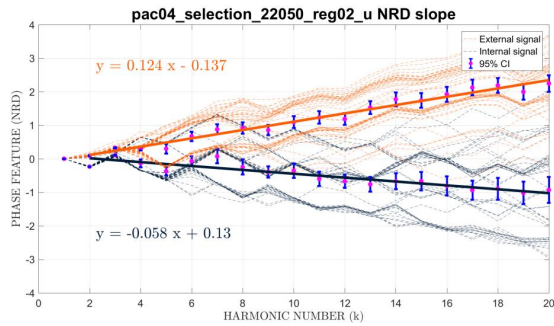
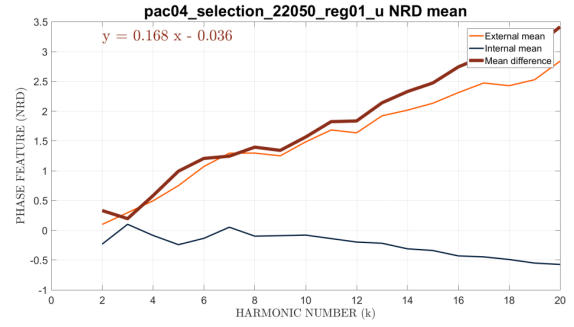
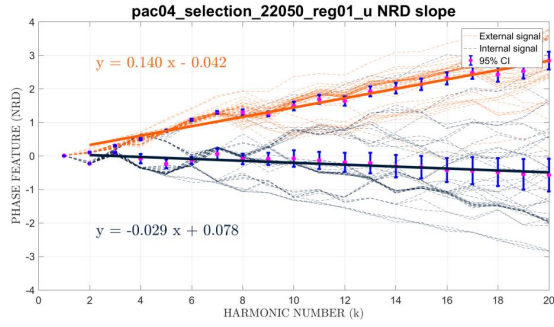
SPECTRAL NRD SLOPE AND NRD MEAN DIFFERENCE

SPEAKER 4
VOWEL /i/

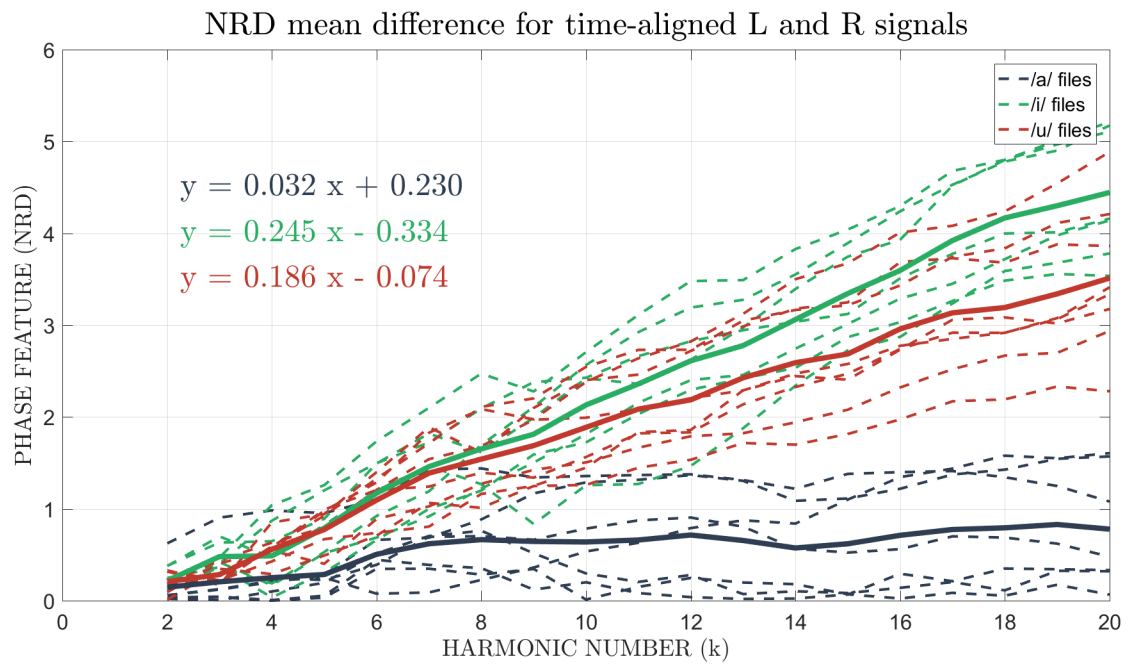


SPECTRAL NRD SLOPE AND NRD MEAN DIFFERENCE

SPEAKER 4
VOWEL /u/



SPECTRAL NRD MEAN DIFFERENCE FOR ALL FILES



Appendix C

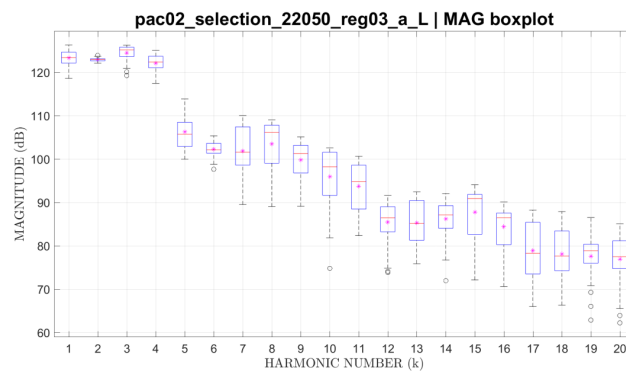
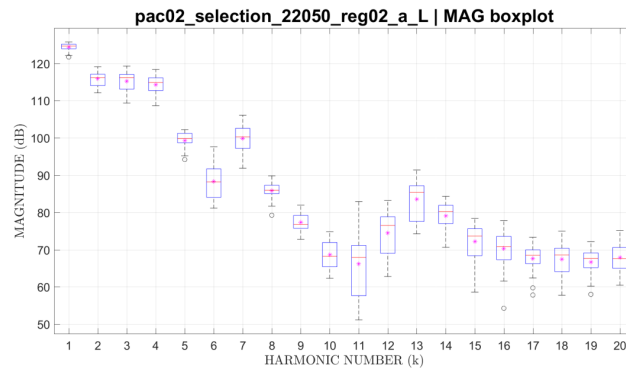
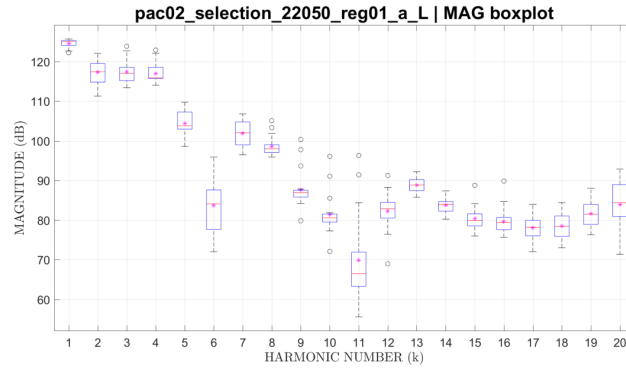
Statistical Analysis

C.1 Spectral Magnitude Boxplots

SPECTRAL MAGNITUDE BOXPLOTS PER FILE

SPEAKER 2
VOWEL /a/

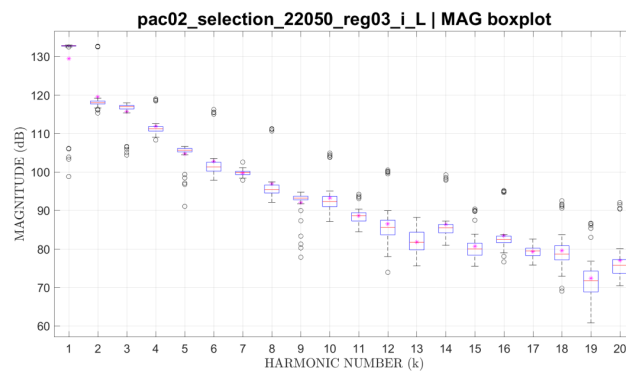
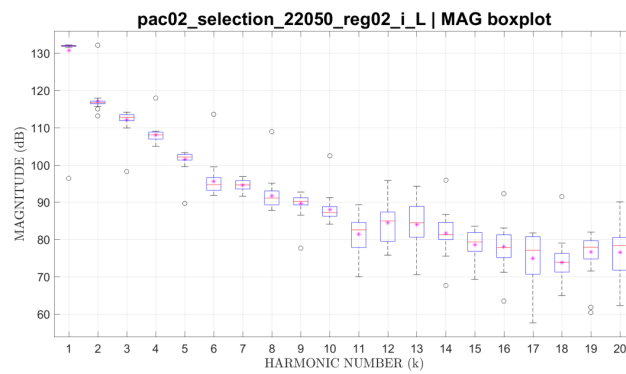
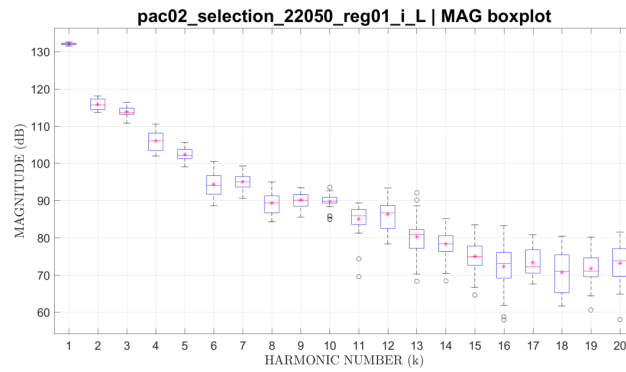
L
(INTERNAL)



SPECTRAL MAGNITUDE BOXPLOTS PER FILE

SPEAKER 2
VOWEL /i/

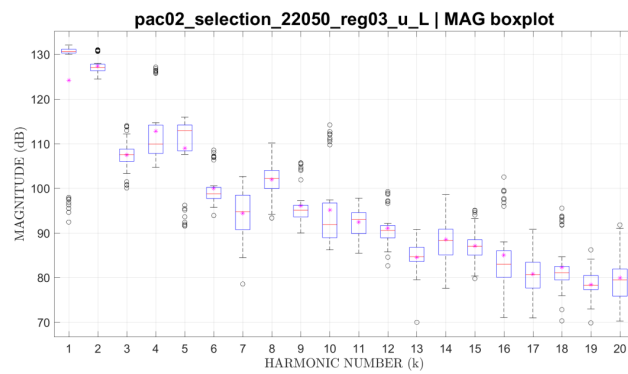
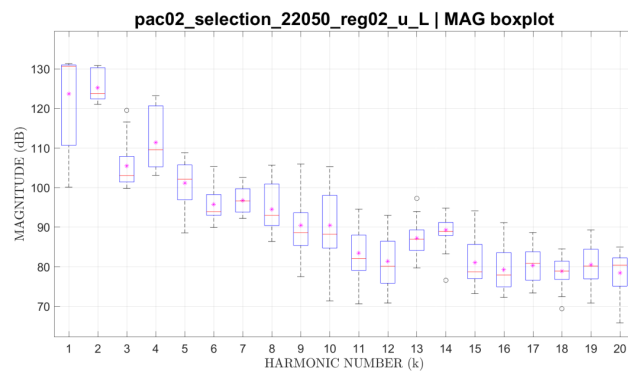
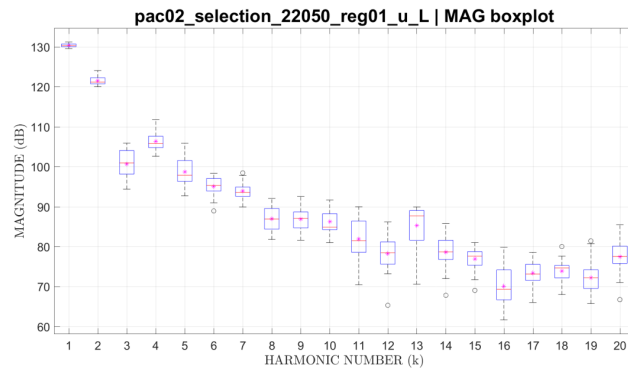
L
(INTERNAL)



SPECTRAL MAGNITUDE BOXPLOTS PER FILE

SPEAKER 2
VOWEL /u/

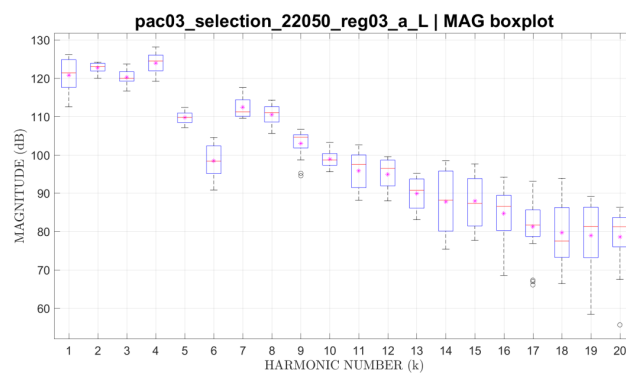
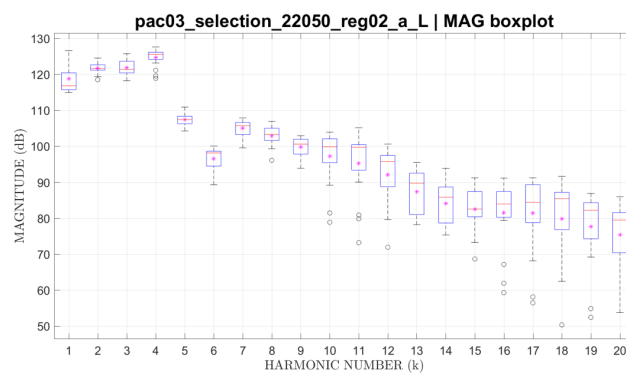
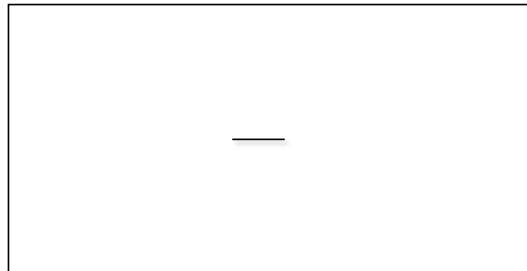
L
(INTERNAL)



SPECTRAL MAGNITUDE BOXPLOTS PER FILE

SPEAKER 3
VOWEL /a/

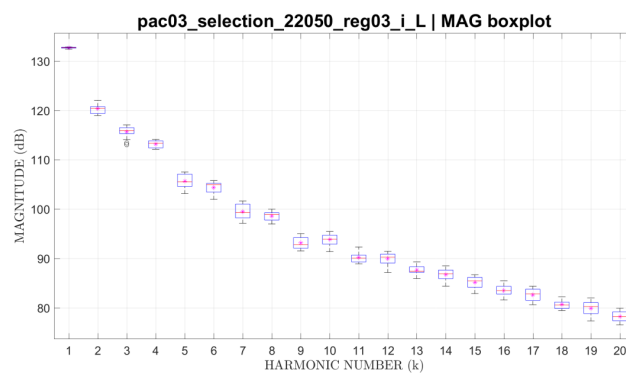
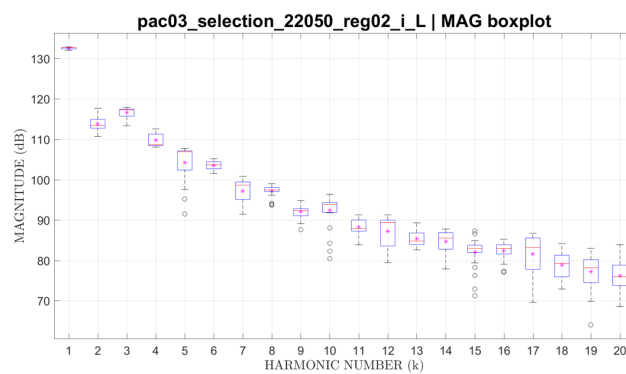
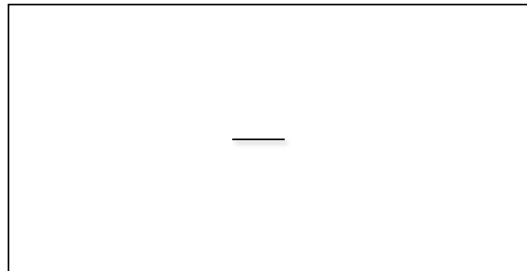
L
(INTERNAL)



SPECTRAL MAGNITUDE BOXPLOTS PER FILE

SPEAKER 3
VOWEL /i/

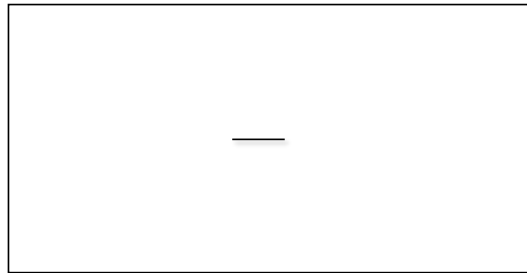
L
(INTERNAL)



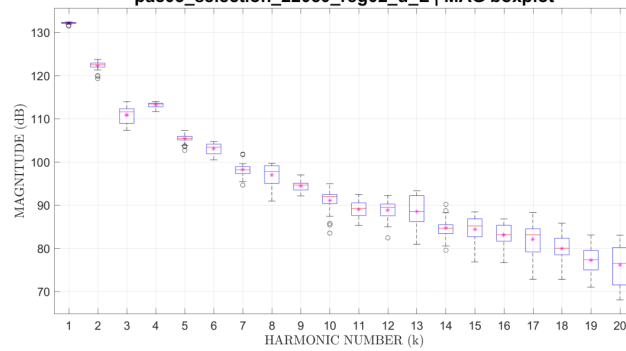
SPECTRAL MAGNITUDE BOXPLOTS PER FILE

SPEAKER 3
VOWEL /u/

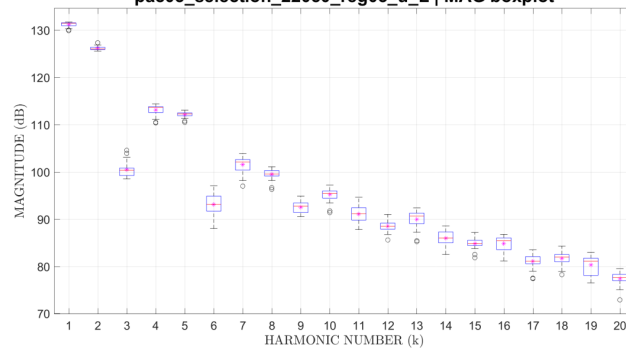
L
(INTERNAL)



pac03_selection_22050_reg02_u_L | MAG boxplot



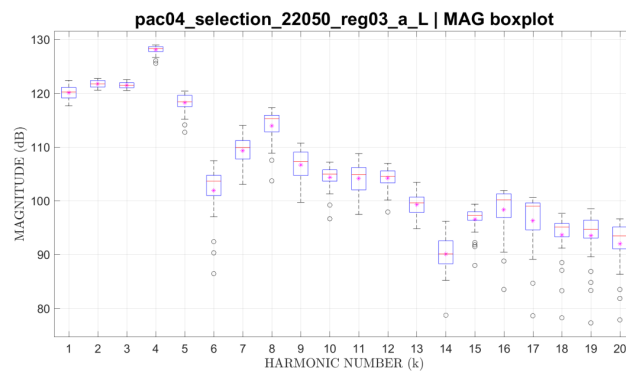
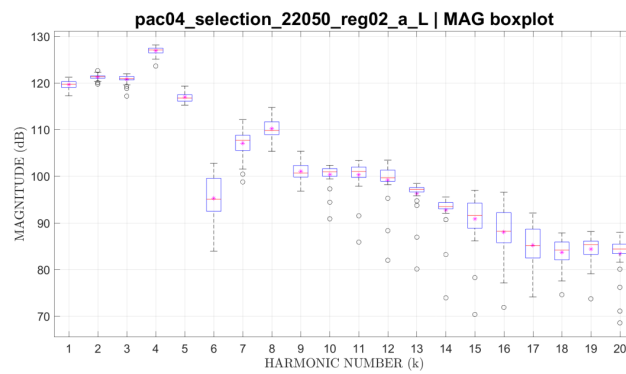
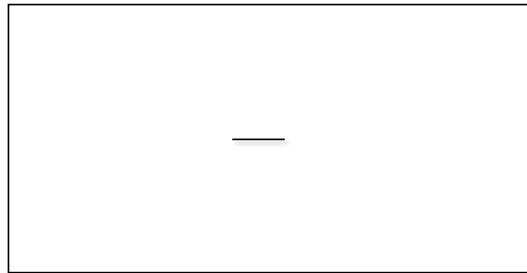
pac03_selection_22050_reg03_u_L | MAG boxplot



SPECTRAL MAGNITUDE BOXPLOTS PER FILE

SPEAKER 4
VOWEL /a/

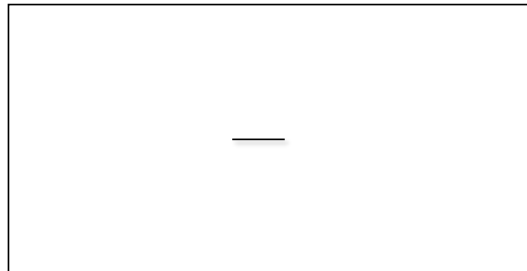
L
(INTERNAL)



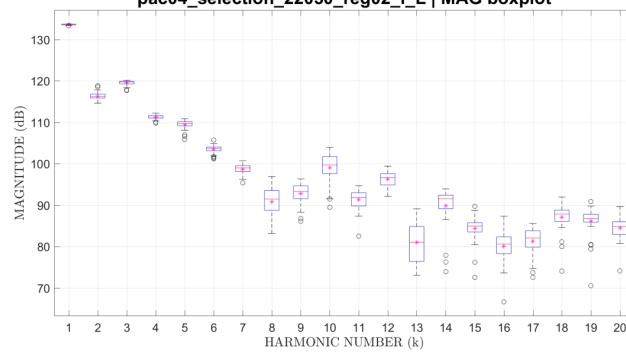
SPECTRAL MAGNITUDE BOXPLOTS PER FILE

SPEAKER 4
VOWEL /i/

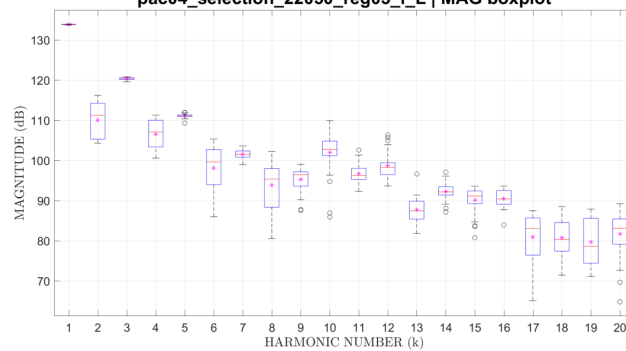
L
(INTERNAL)



pac04_selection_22050_reg02_i_L | MAG boxplot



pac04_selection_22050_reg03_i_L | MAG boxplot

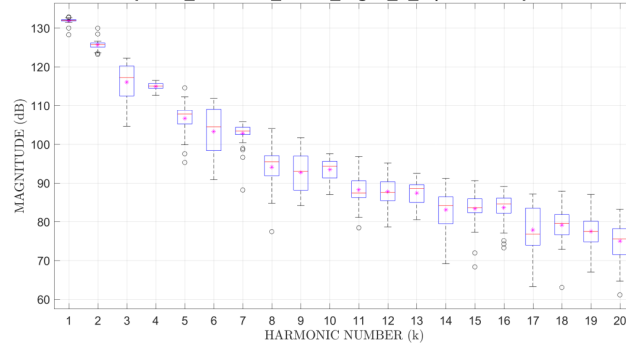


SPECTRAL MAGNITUDE BOXPLOTS PER FILE

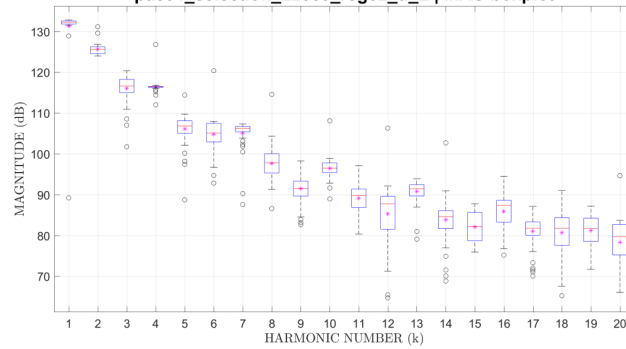
SPEAKER 4
VOWEL /u/

L
(INTERNAL)

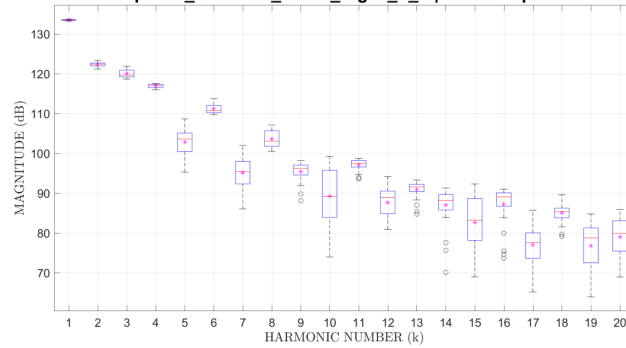
pac04_selection_22050_reg01_u_L | MAG boxplot



pac04_selection_22050_reg02_u_L | MAG boxplot



pac04_selection_22050_reg03_u_L | MAG boxplot



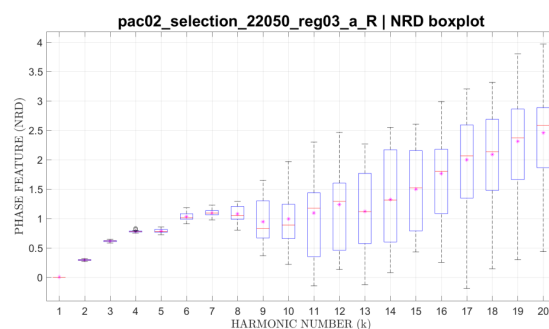
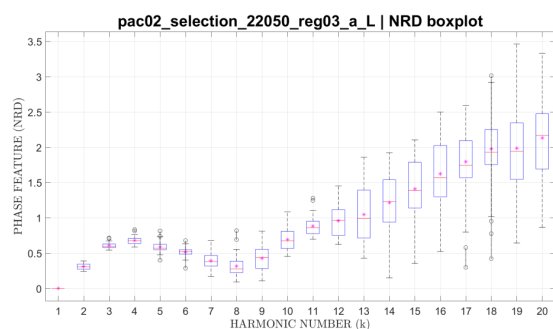
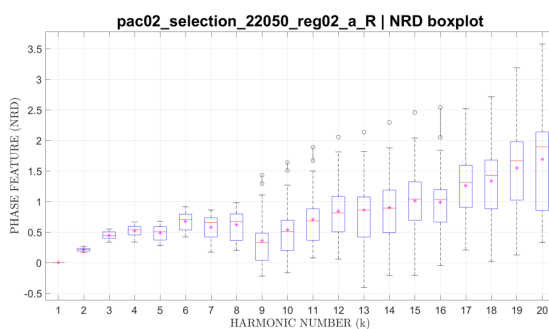
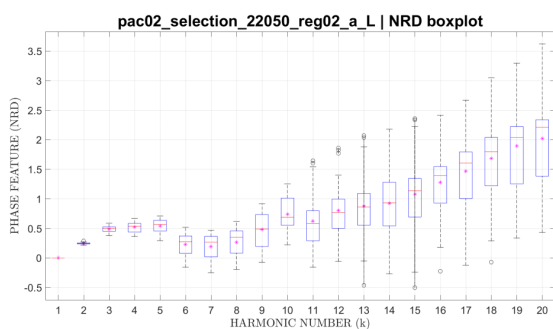
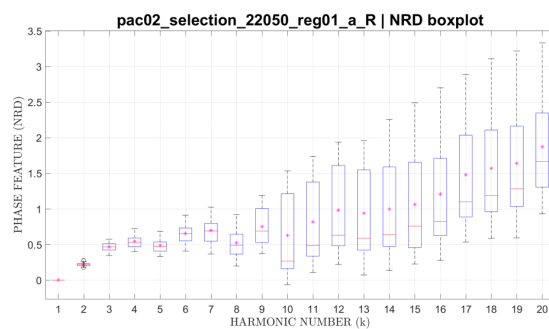
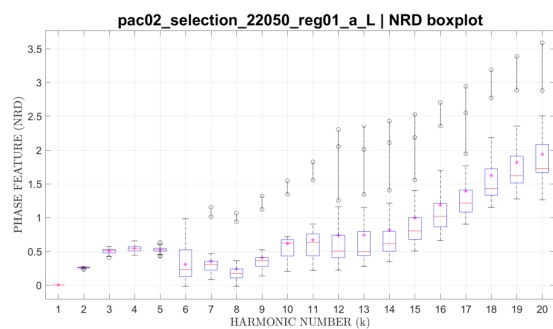
C.2 Spectral Phase Boxplots

SPECTRAL NRD BOXPLOTS PER FILE

SPEAKER 2
VOWEL /a/

L
(INTERNAL)

R
(EXTERNAL)

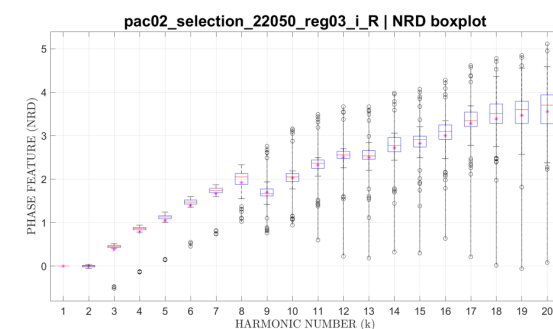
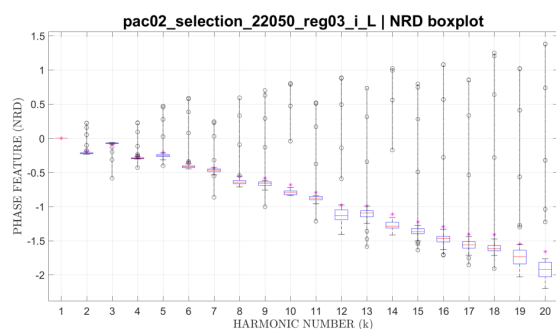
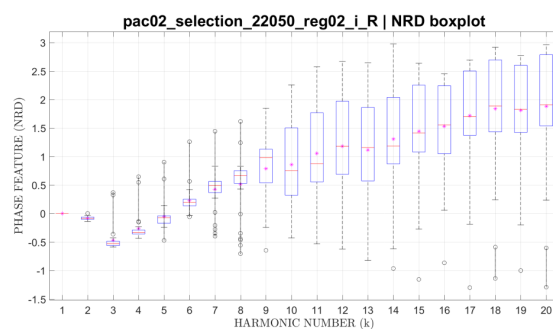
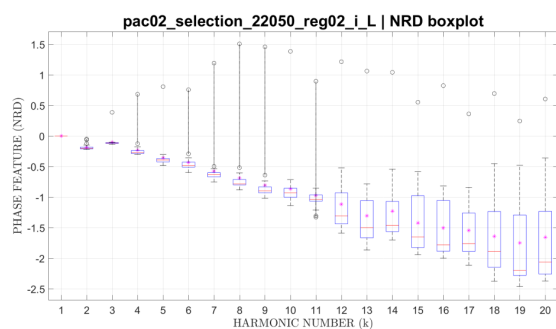
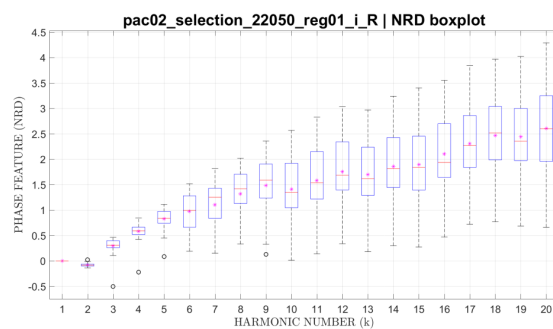
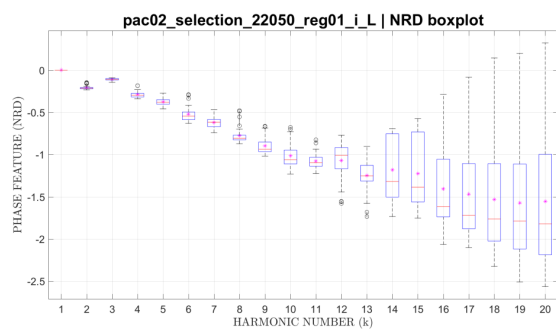


SPECTRAL NRD BOXPLOTS PER FILE

SPEAKER 2
VOWEL /i/

L
(INTERNAL)

R
(EXTERNAL)

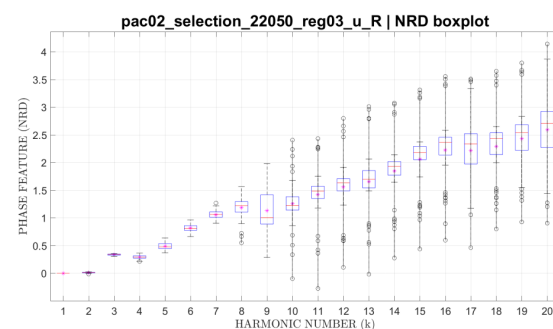
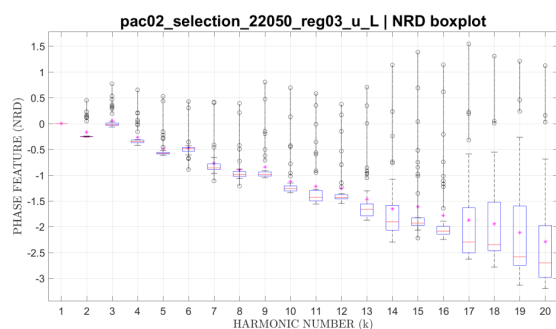
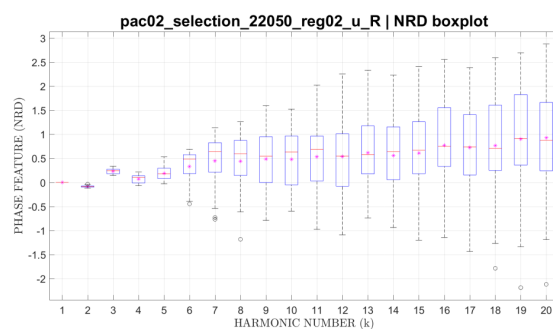
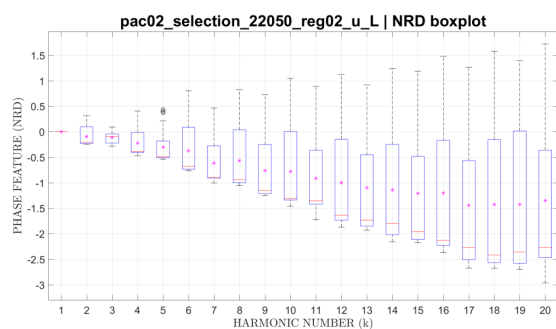
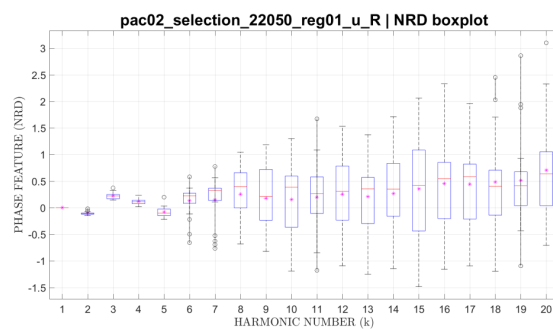
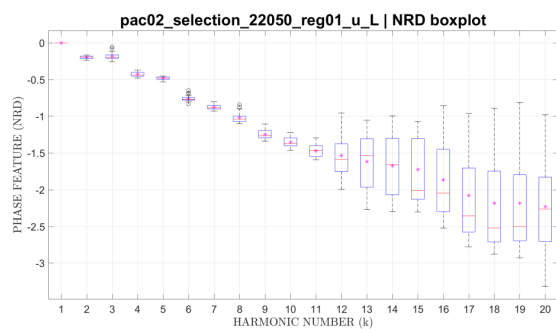


SPECTRAL NRD BOXPLOTS PER FILE

SPEAKER 2
VOWEL /u/

L
(INTERNAL)

R
(EXTERNAL)

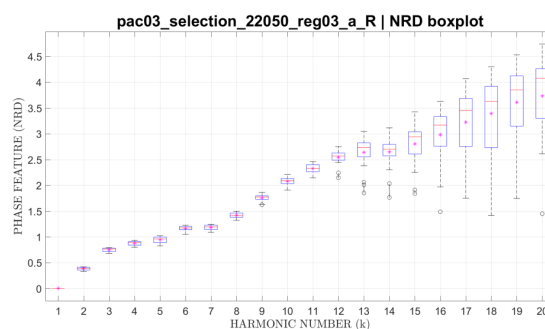
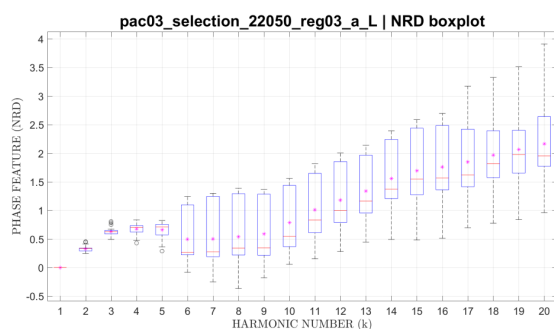
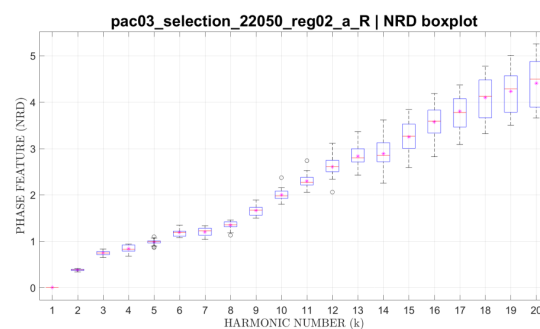
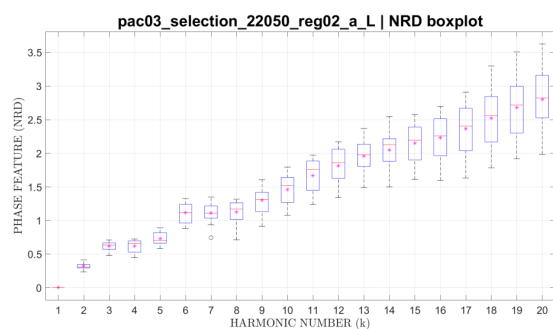
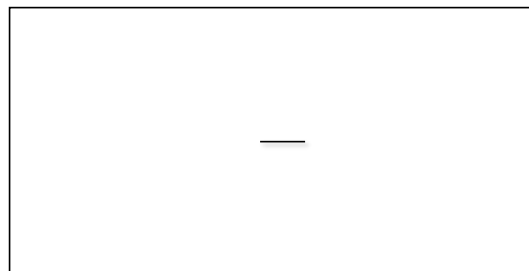
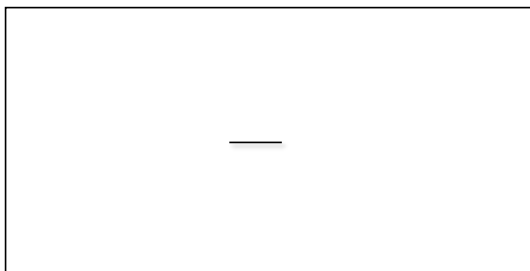


SPECTRAL NRD BOXPLOTS PER FILE

SPEAKER 3
VOWEL /a/

L
(INTERNAL)

R
(EXTERNAL)

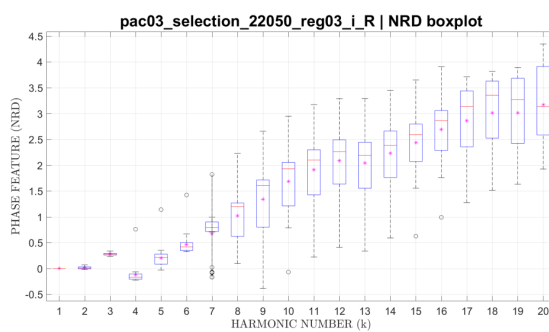
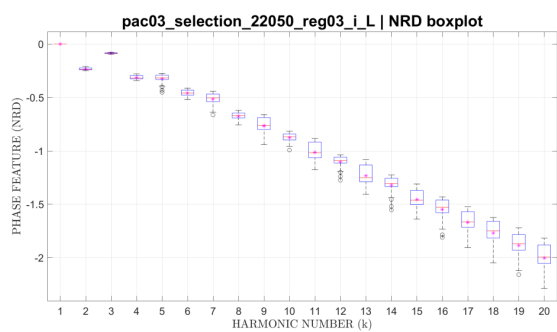
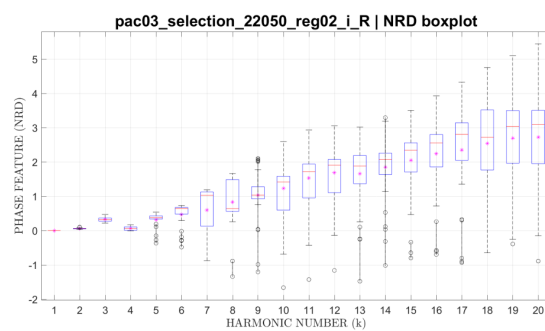
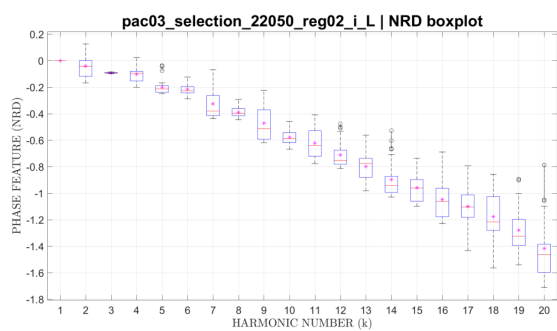
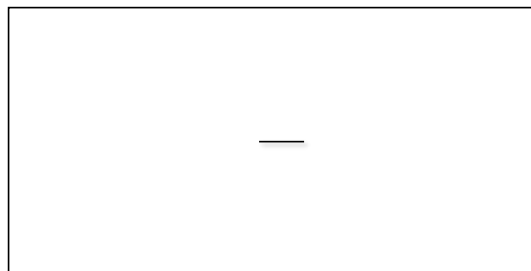
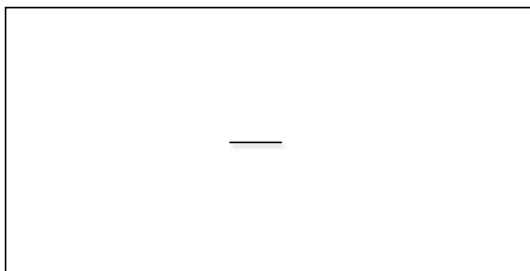


SPECTRAL NRD BOXPLOTS PER FILE

SPEAKER 3
VOWEL /i/

L
(INTERNAL)

R
(EXTERNAL)

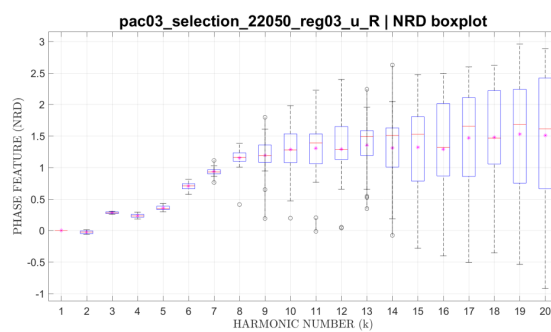
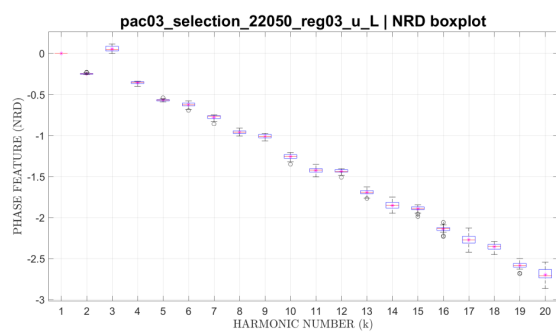
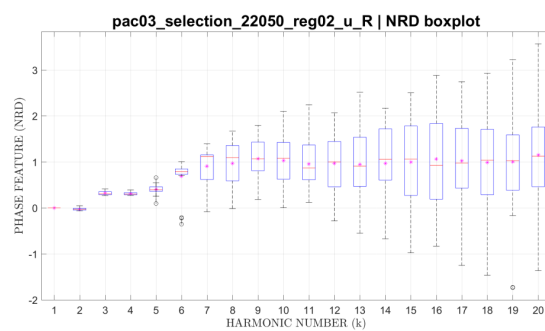
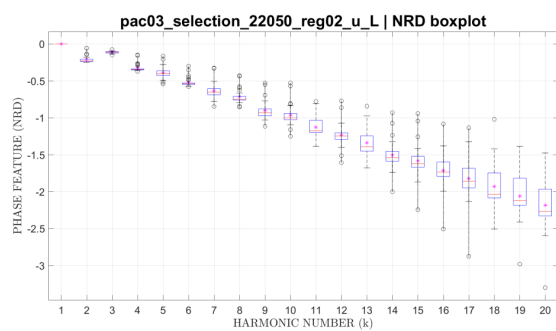
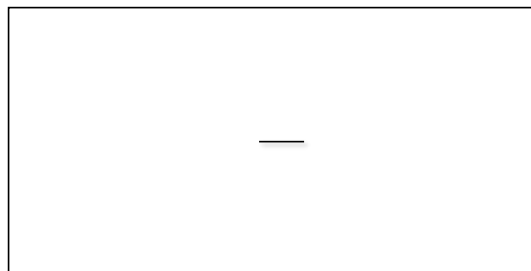
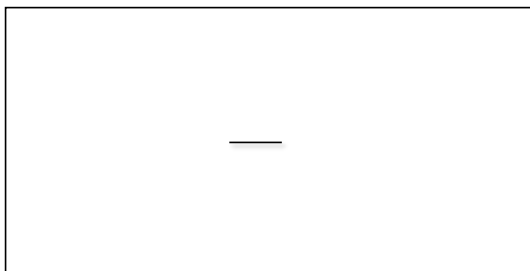


SPECTRAL NRD BOXPLOTS PER FILE

SPEAKER 3
VOWEL /u/

L
(INTERNAL)

R
(EXTERNAL)

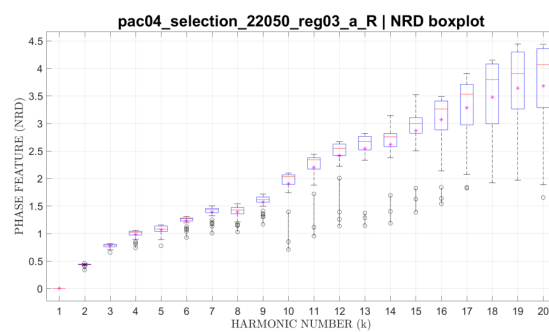
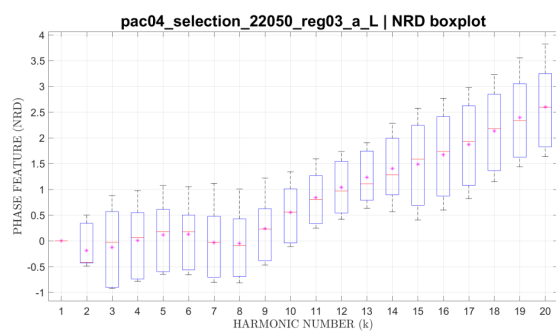
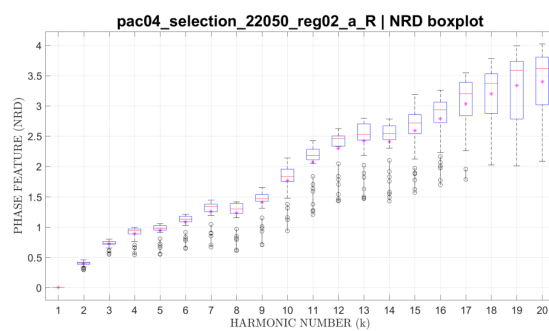
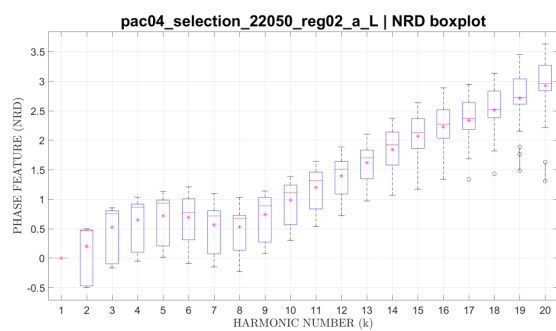
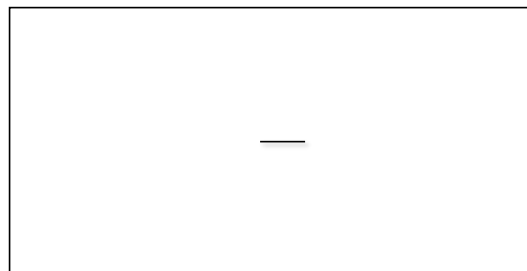
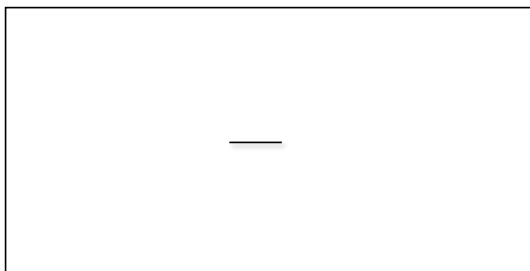


SPECTRAL NRD BOXPLOTS PER FILE

SPEAKER 4
VOWEL /a/

L
(INTERNAL)

R
(EXTERNAL)

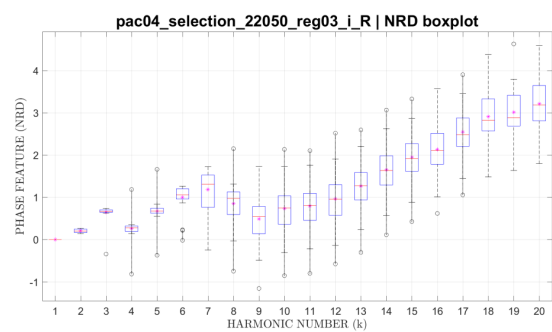
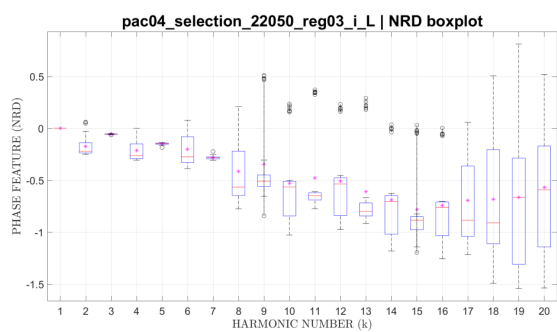
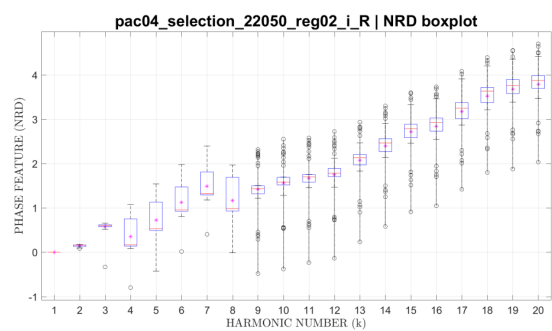
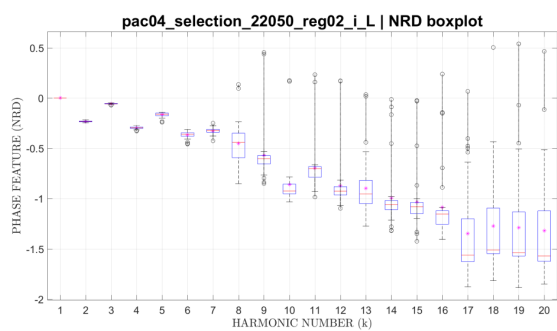
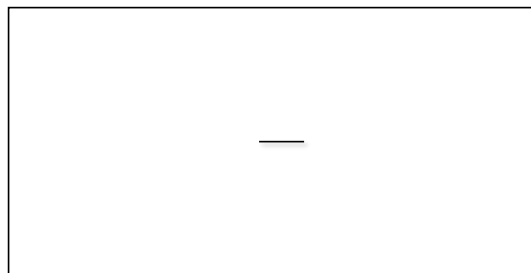
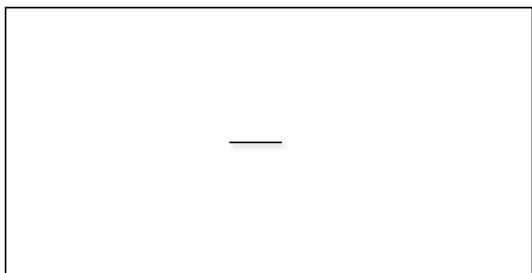


SPECTRAL NRD BOXPLOTS PER FILE

SPEAKER 4
VOWEL /i/

L
(INTERNAL)

R
(EXTERNAL)

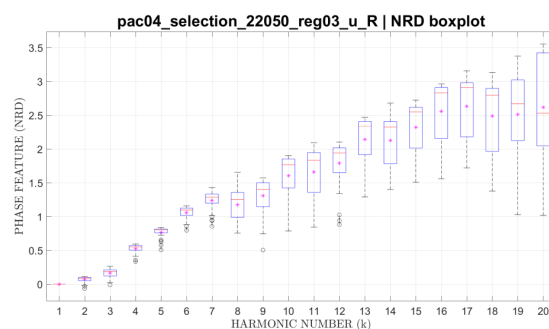
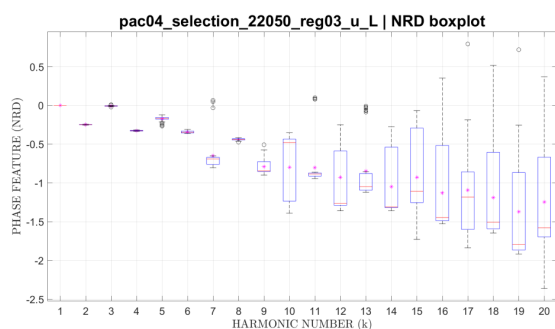
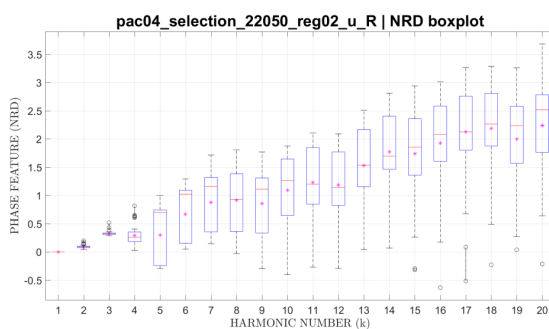
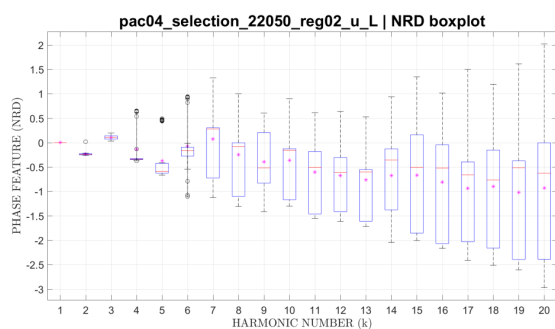
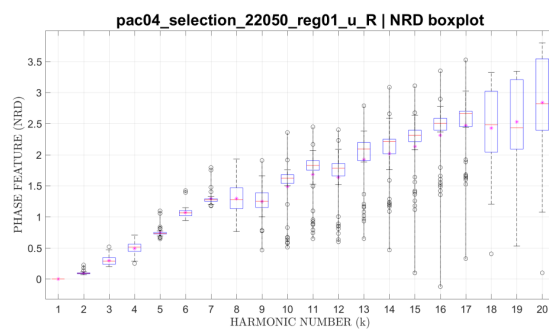
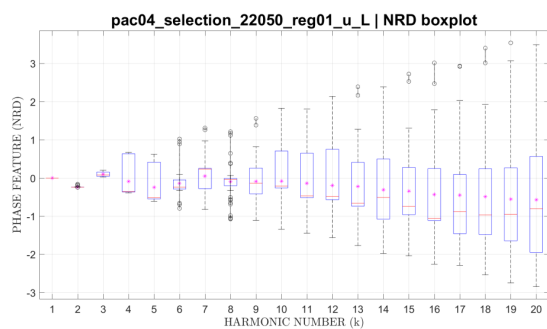


SPECTRAL NRD BOXPLOTS PER FILE

SPEAKER 4
VOWEL /u/

L
(INTERNAL)

R
(EXTERNAL)

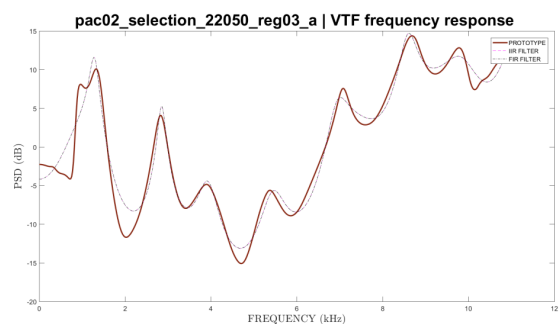
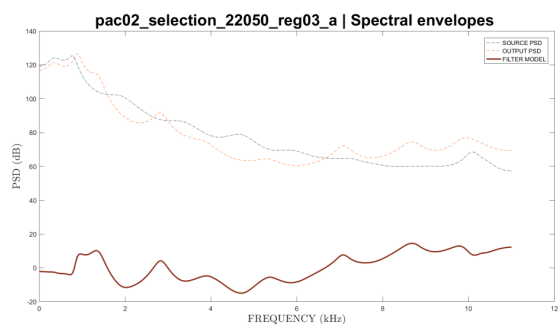
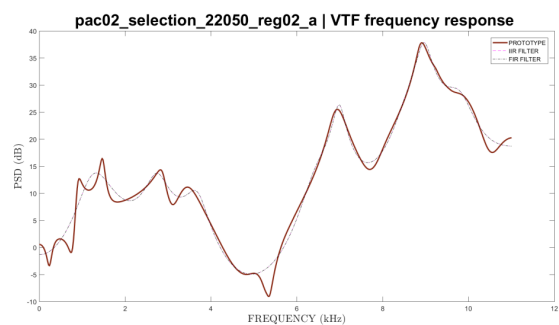
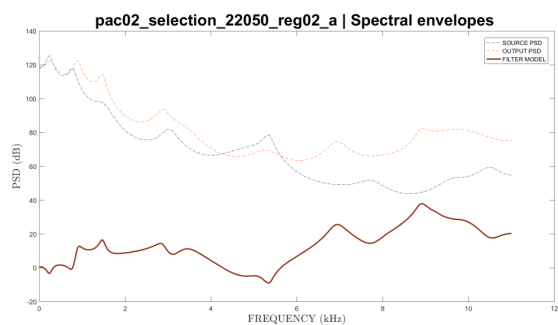
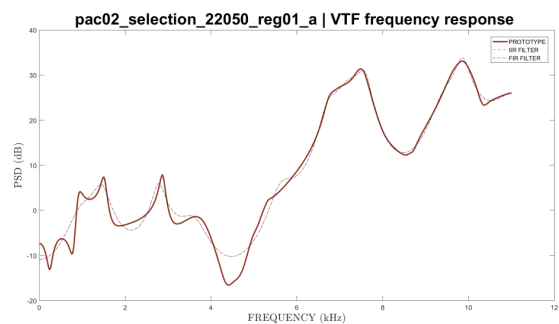
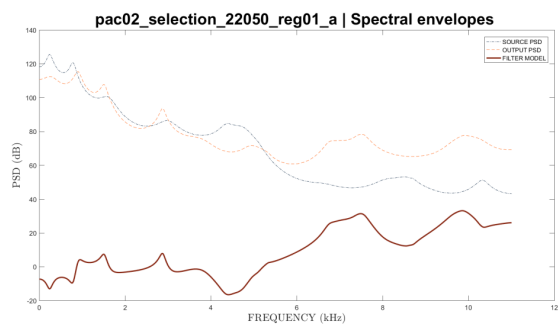


Appendix D

Vocal Tract Characterization

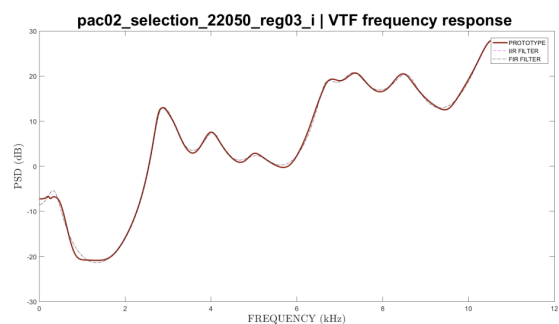
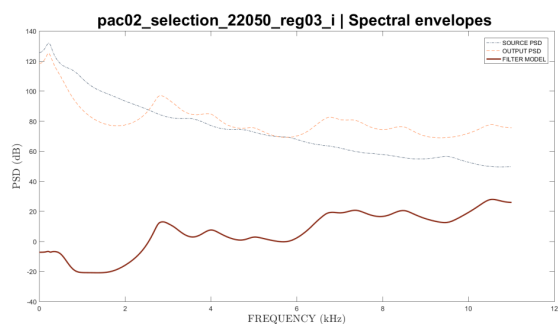
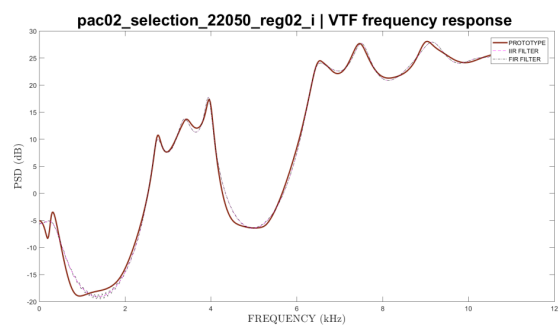
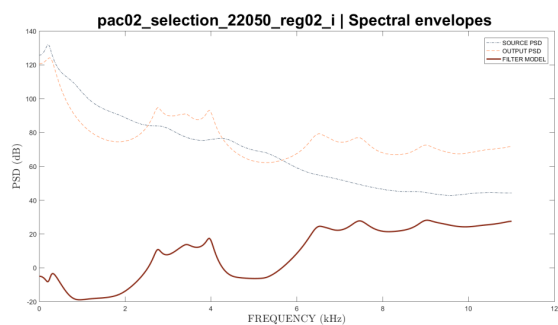
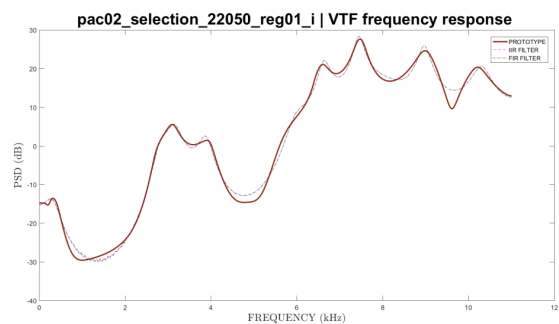
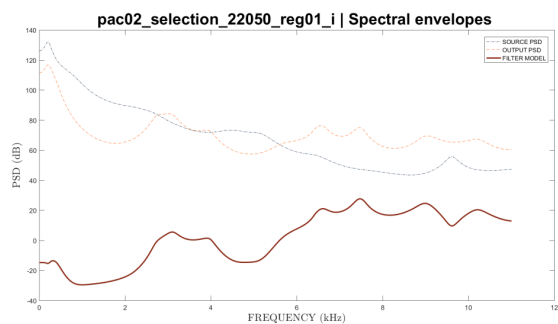
FILTER FREQUENCY RESPONSE

SPEAKER 2
VOWEL /a/



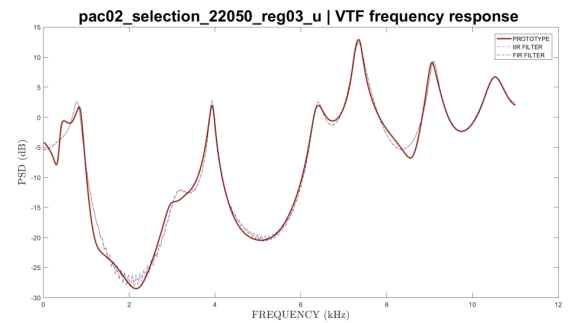
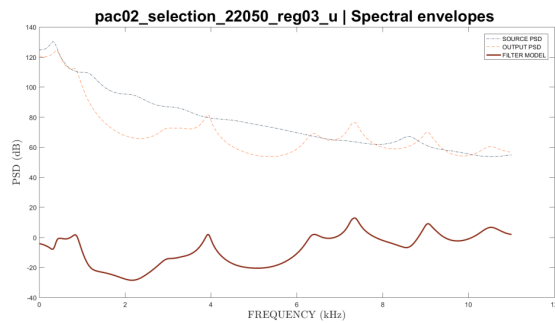
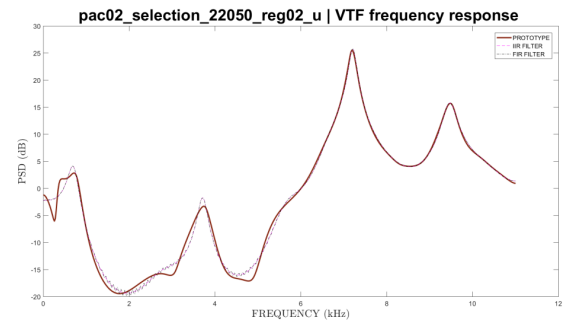
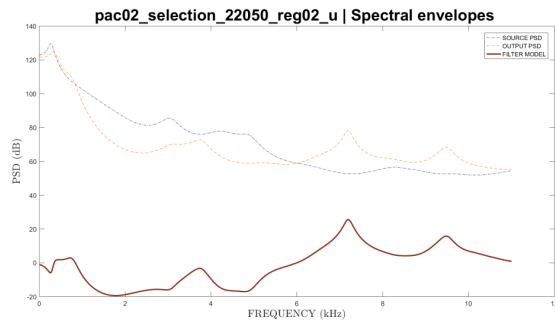
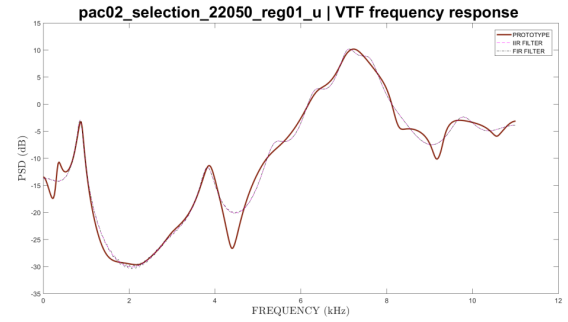
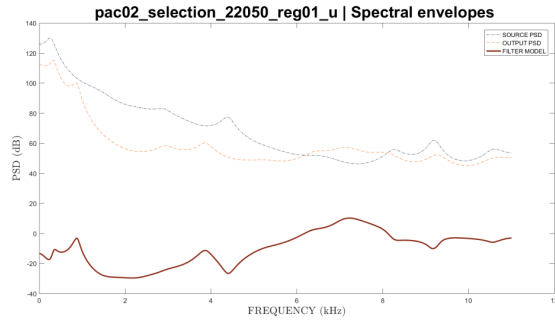
FILTER FREQUENCY RESPONSE

SPEAKER 2
VOWEL /i/



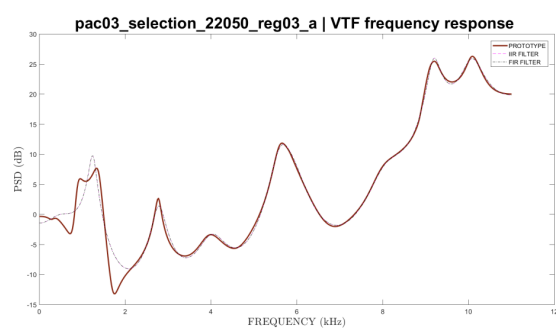
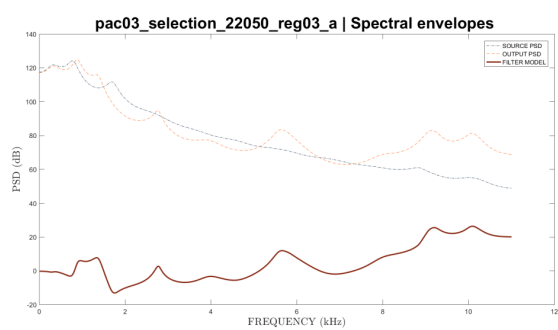
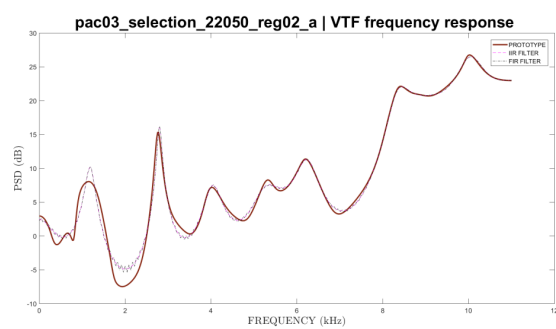
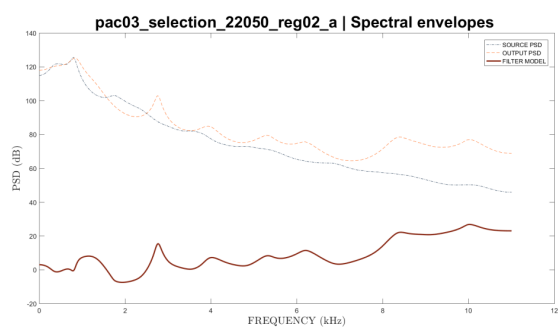
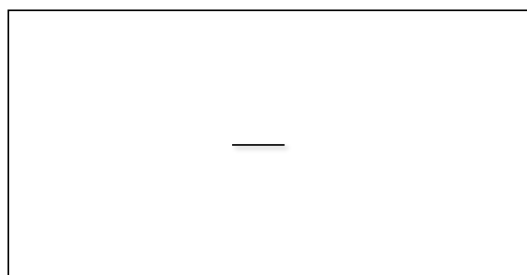
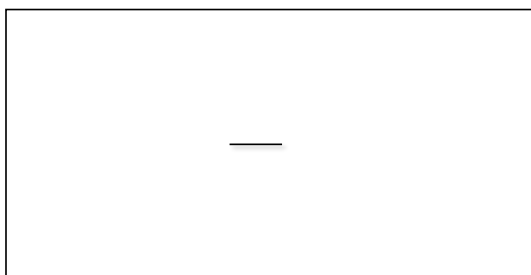
FILTER FREQUENCY RESPONSE

SPEAKER 2
VOWEL /u/



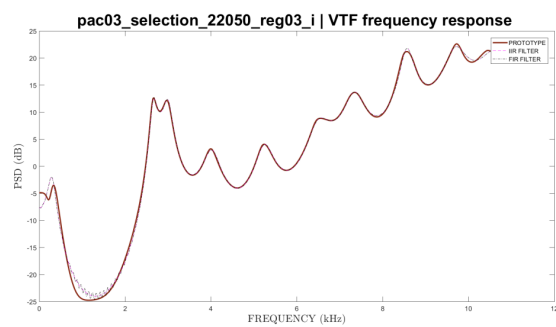
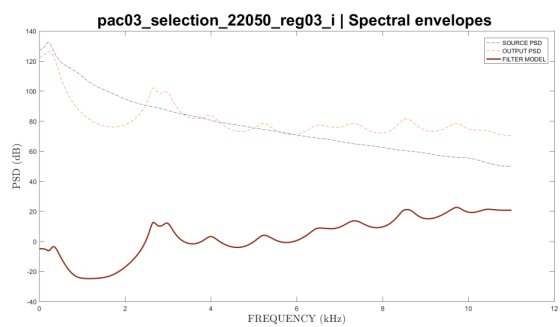
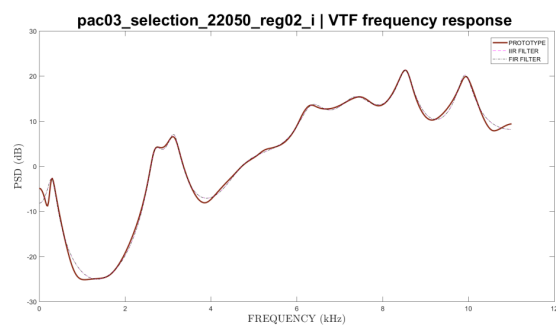
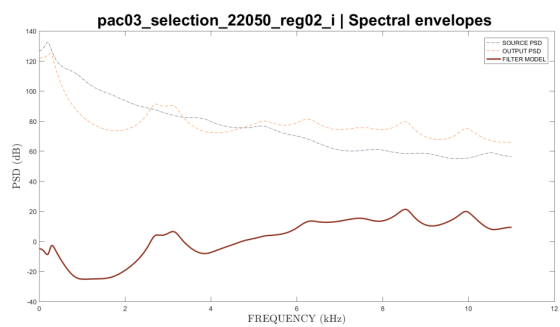
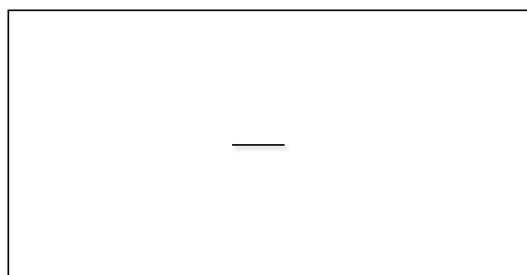
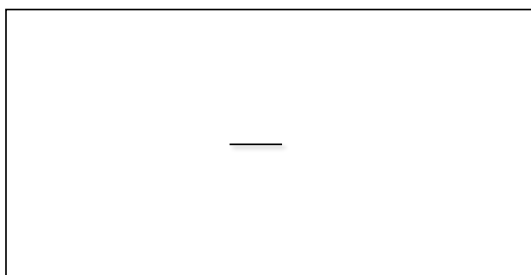
FILTER FREQUENCY RESPONSE

SPEAKER 3
VOWEL /a/



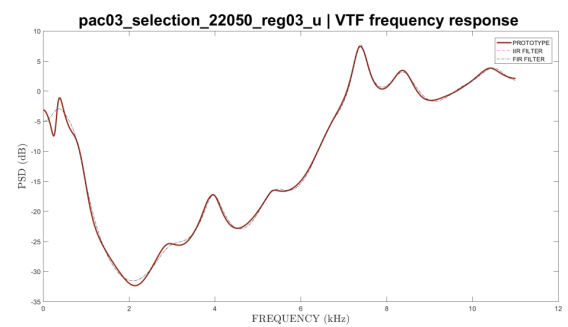
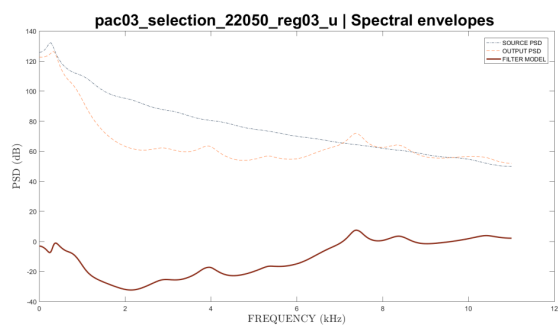
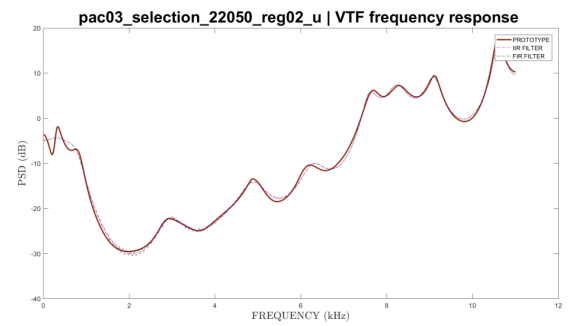
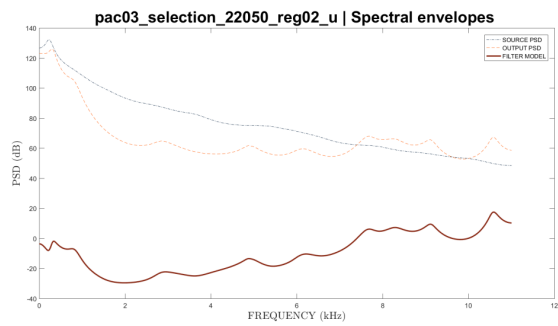
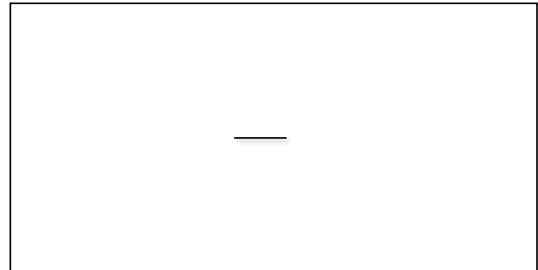
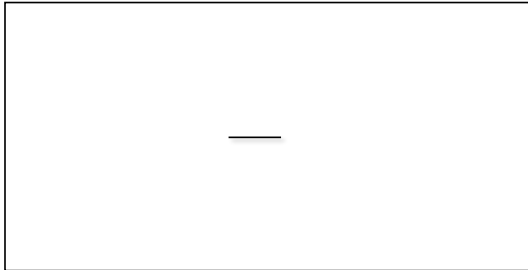
FILTER FREQUENCY RESPONSE

SPEAKER 3
VOWEL /i/



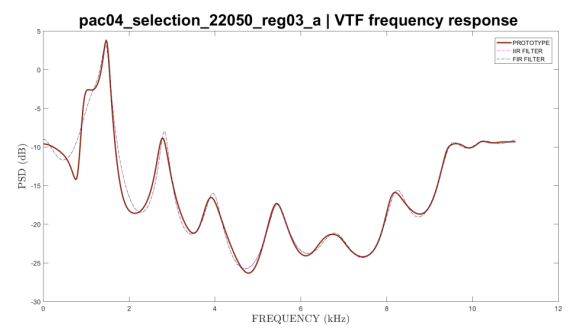
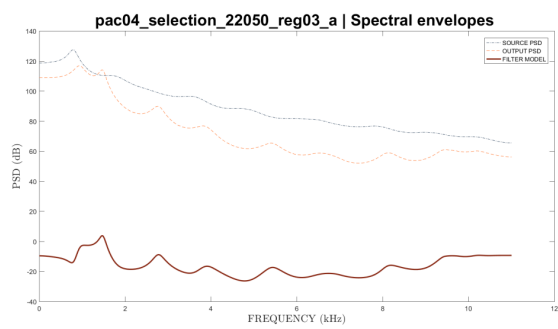
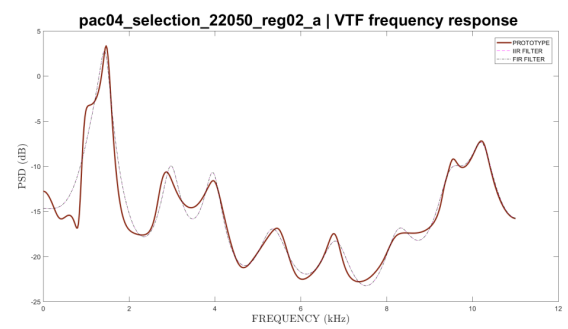
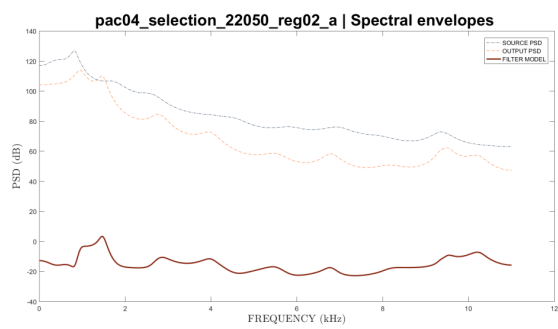
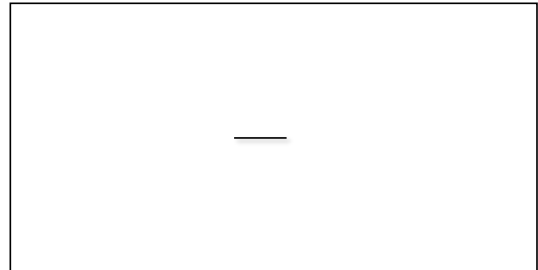
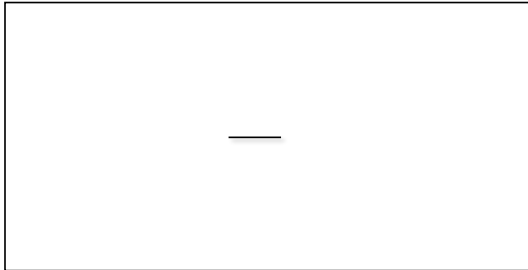
FILTER FREQUENCY RESPONSE

SPEAKER 3
VOWEL /u/



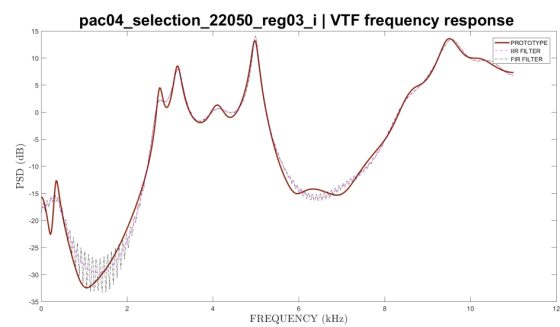
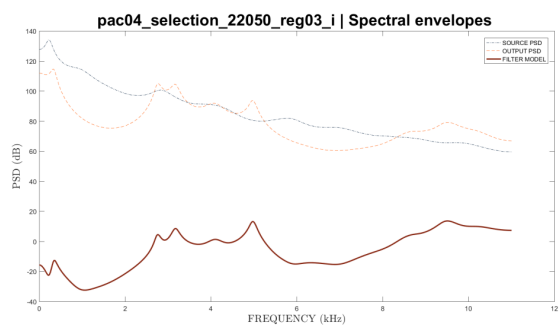
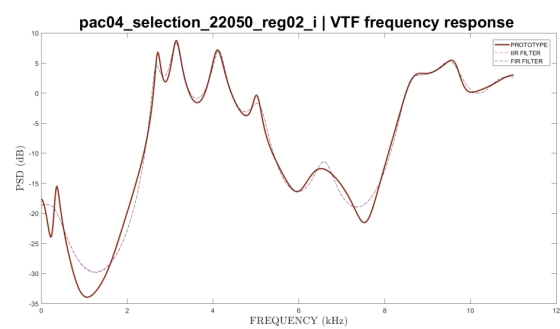
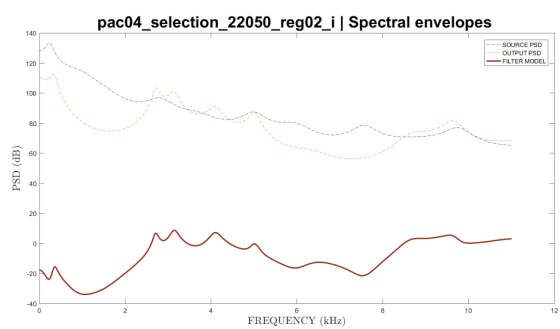
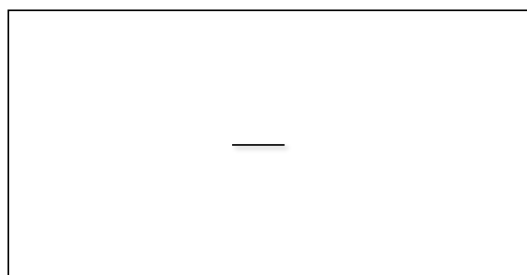
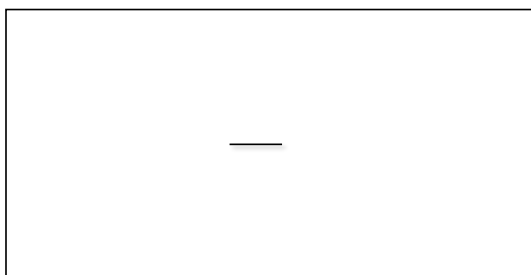
FILTER FREQUENCY RESPONSE

SPEAKER 4
VOWEL /a/



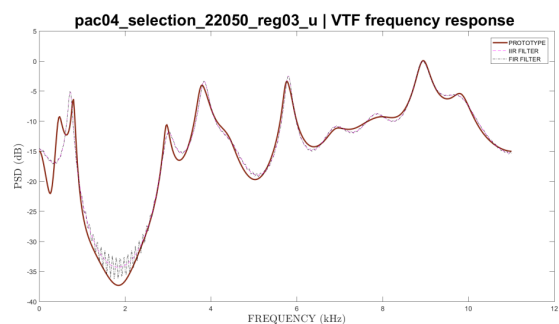
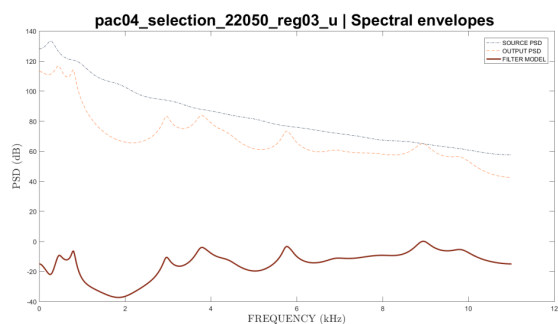
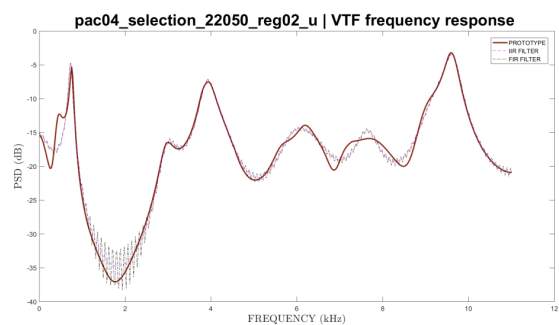
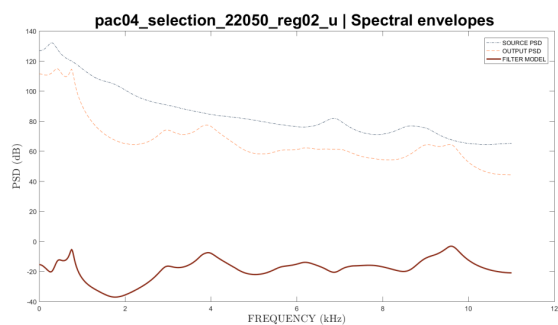
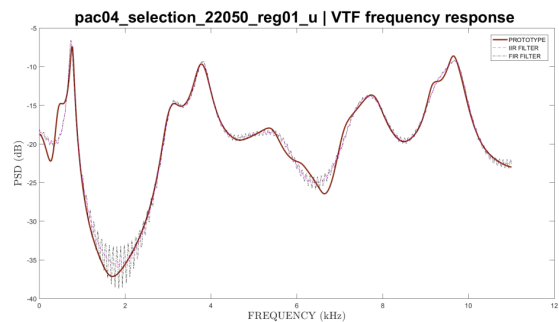
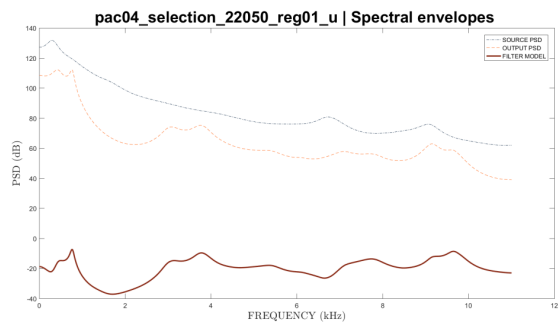
FILTER FREQUENCY RESPONSE

SPEAKER 4
VOWEL /i/



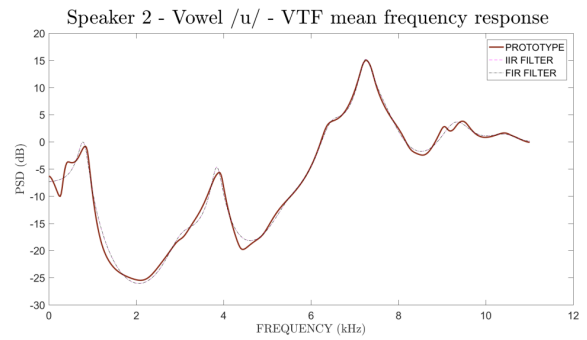
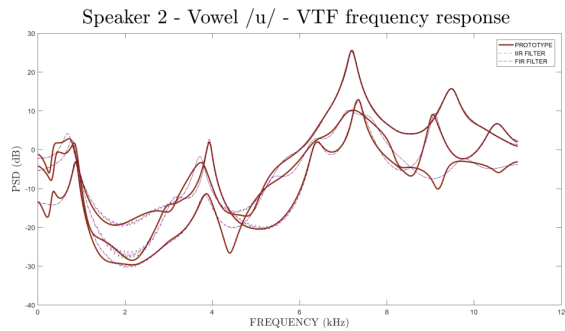
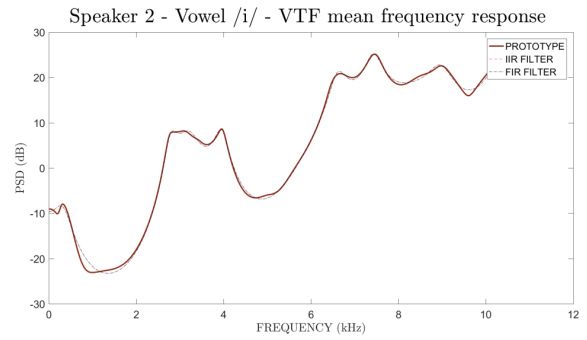
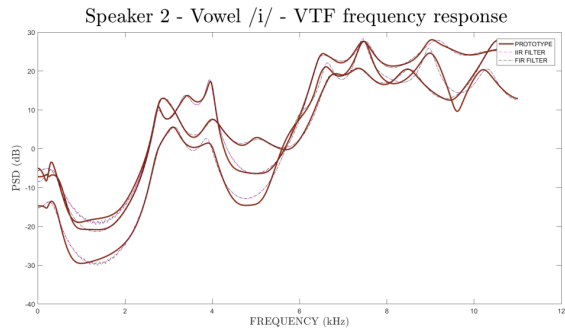
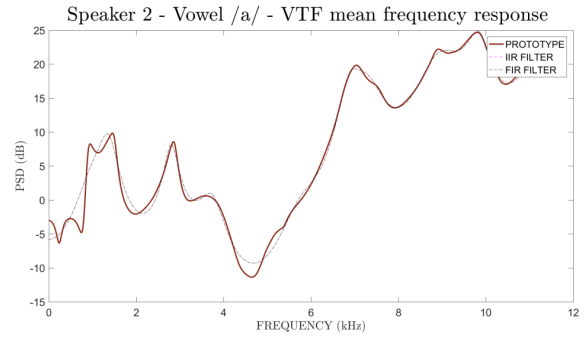
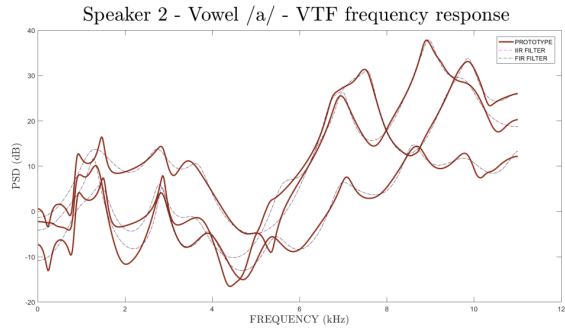
FILTER FREQUENCY RESPONSE

SPEAKER 4
VOWEL /u/



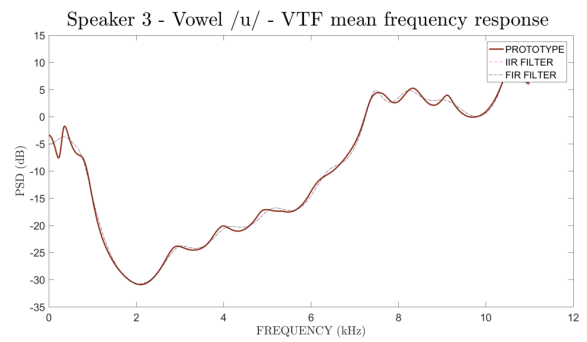
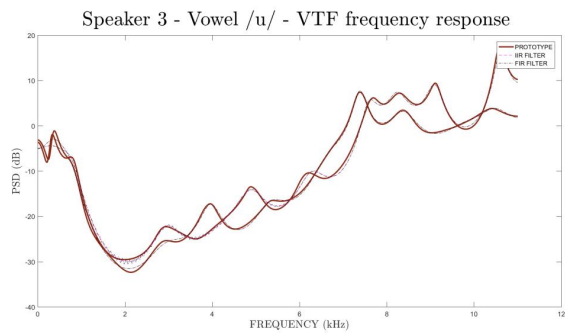
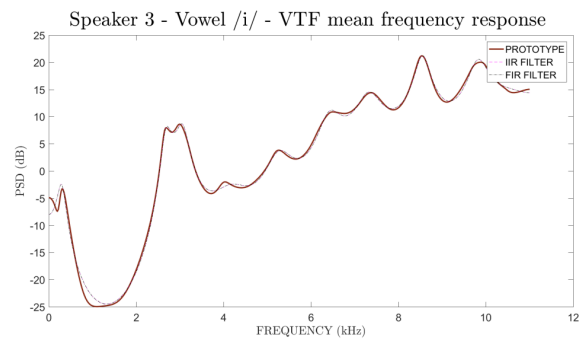
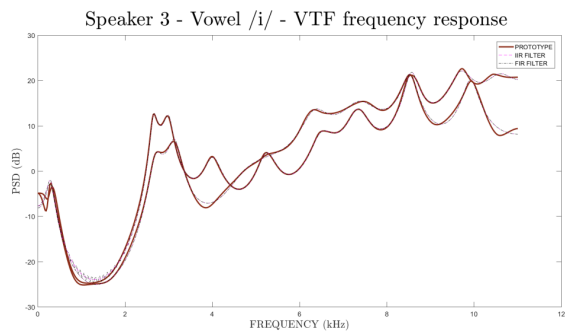
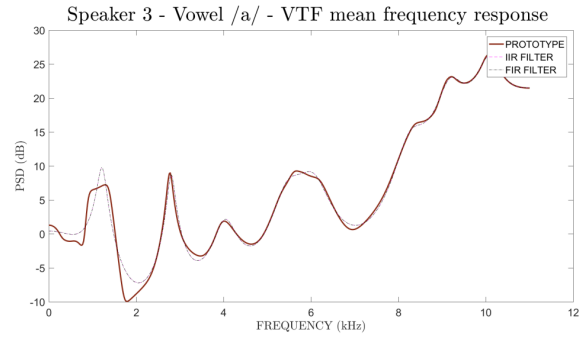
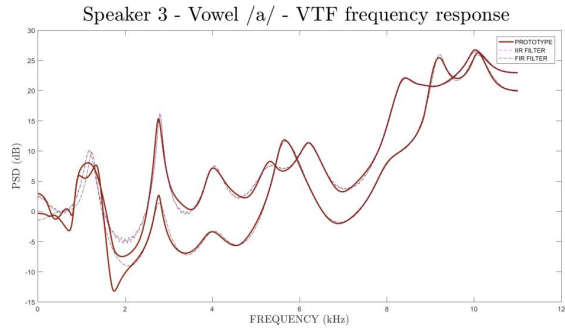
VOCAL TRACT FILTER FREQUENCY RESPONSE PER SPEAKER AND VOWEL

SPEAKER 2



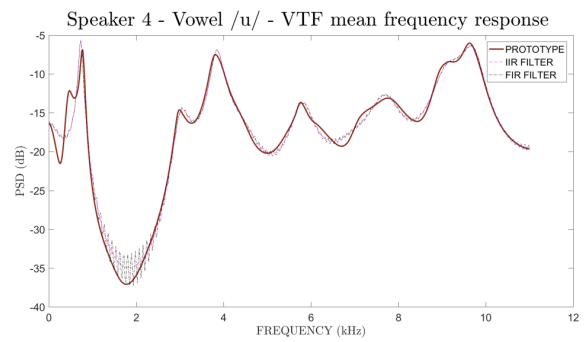
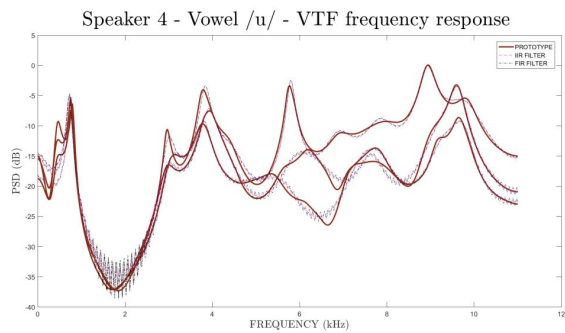
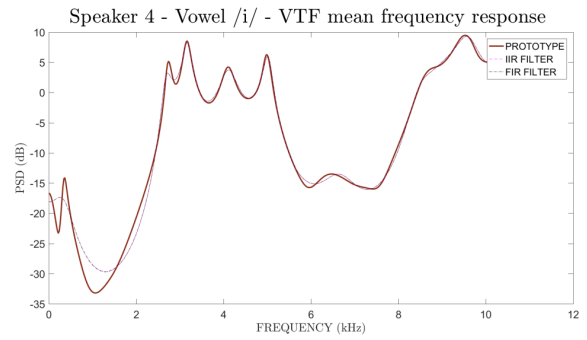
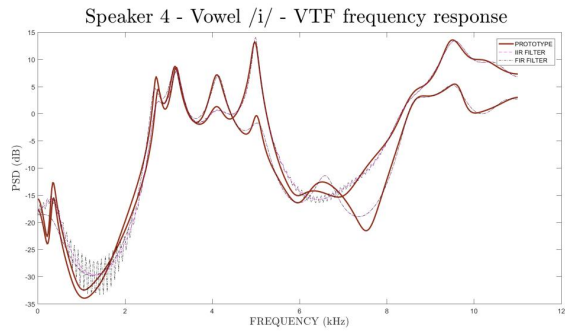
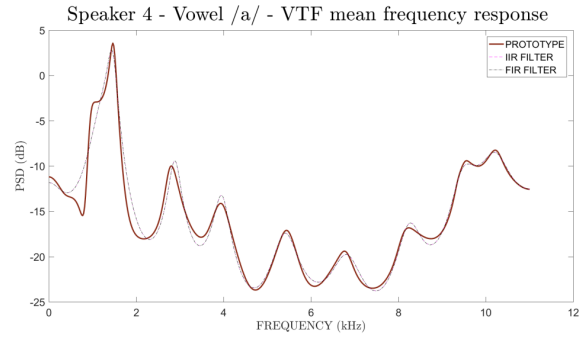
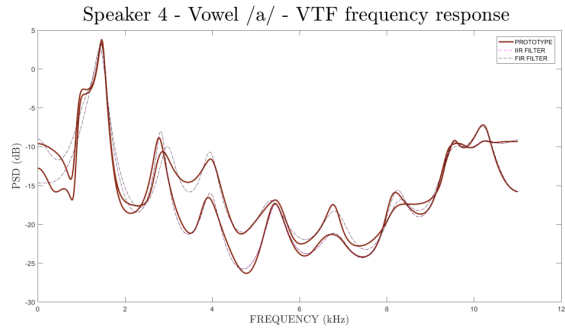
VOCAL TRACT FILTER FREQUENCY RESPONSE PER SPEAKER AND VOWEL

SPEAKER 3



VOCAL TRACT FILTER FREQUENCY RESPONSE PER SPEAKER AND VOWEL

SPEAKER 4



VOCAL TRACT FILTER FREQUENCY RESPONSE PER VOWEL

