UNIVERSIDADE FEDERAL FLUMINENSE
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE TELECOMUNICAÇÕES

LUIS EDUARDO BARRIENTOS SANDOVAL

Modelagem e síntese de vogais cantadas por sopranos hispanoparlantes com estimação dos parâmetros através de predição linear ponderada adaptada

NITERÓI

2023

UNIVERSIDADE FEDERAL FLUMINENSE
ESCOLA DE ENGENHARIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA E DE TELECOMUNICAÇÕES

LUIS EDUARDO BARRIENTOS SANDOVAL

Modelagem e síntese de vogais cantadas por sopranos hispanoparlantes com estimação dos parâmetros através de predição linear ponderada adaptada

Tese de doutorado apresentado ao Programa de Pós-Graduação em Engenharia Elétrica e de Telecomunicações da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Doutor em Engenharia Elétrica e de Telecomunicações. Área de concentração: Sistemas de Telecomunicações .

Orientador:
Prof. D.Sc. EDSON LUIZ CATALDO FERREIRA

NITERÓI

2023

LUIS EDUARDO BARRIENTOS SANDOVAL

Modelagem e síntese de vogais cantadas por sopranos hispanoparlantes com estimação dos parâmetros através de predição linear ponderada adaptada

Tese de doutorado apresentado ao Programa de Pós-Graduação em Engenharia Elétrica e de Telecomunicações da Universidade Federal Fluminense como requisito parcial para a obtenção do título de Doutor em Engenharia Elétrica e de Telecomunicações. Área de concentração: Sistemas de Telecomunicações.

Aprovada em 31 de agosto de 2023.

BANCA EXAMINADORA

Prof. Edson Luiz Cataldo Ferreira, D.Sc. – Orientador, UFF

Prof[a]. Leni Joaquim de Matos, D.Sc. – UFF

Prof[a]. Renata Raposo Del Vecchio, D.Sc. – UFF

Prof. Americo Barbosa da Cunha Junior, D.Sc. – UERJ

Prof. Leonardo Alfredo Forero Mendoza, D.Sc. – UERJ

Niterói

2023

Dedicado a Deus e à minha família.

Agradecimentos

Agradeço a Deus por colocar no meu caminho de vida as bênçãos e faculdades para conquistar este lindo sonho. Toda a glória seja para Ele.

Agradeço à minha namorada Deisy Yurley Basto, pela sua compreensão, respeito, tolerância e por todas as atitudes que a faz merecedora do meu amor.

Agradeço aos meus pais e irmãos, pela confiança na minha capacidade, pelo apoio, motivação e carinho em todos os momento da minha vida.

Agradeço, especialmente, ao professor Edson Cataldo, pela orientação e disposição para a realização deste projeto, pela confiança e apoio nas situações de dificuldade, pela bondade e, sobretudo, sua paciência.

Agradeço à Universidade Federal Fluminense e ao PPGEET, por abrirem as portas e permitirem que me torne Doutor.

Agradeço à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela bolsa de estudos e pelo apoio financeiro de todos os recursos necessários para o desenvolvimento deste trabalho de doutorado.

Resumo

A síntese da voz cantada tem sido um tópico de crescente interesse devido às suas implicações significativas na indústria musical, análise de técnica vocal e aplicações recreativas. Nas últimas décadas, houve inúmeros avanços notáveis que se destacam pela naturalidade e expressividade das sínteses. No entanto, a síntese da voz cantada por sopranos líricos hispanoparlantes apresenta desafios únicos que merecem uma atenção mais aprofundada.

O desafio principal desta tese reside na reprodução precisa das características vocais, incluindo a frequência fundamental (f_o), as frequências dos formantes, o vibrato, o trêmulo e a duração dos fonemas, a fim de alcançar uma síntese natural, expressiva e inteligível. Isso requer consideração das complexidades inerentes à voz cantada por sopranos líricos hispanoparlantes nas suas gamas de tons mais agudos. Para enfrentar esses desafios, o objetivo principal desta tese é desenvolver um sintetizador de vogais cantadas específico para sopranos hispanoparlantes. Isso é alcançado por meio de modelos matemáticos da fonte glotal, frequências formantes estimadas com precisão e a consideração de aspectos como a *Interação Fonte-Filtro* (SFI). Além disso, a adição de efeitos como trêmulo e vibrato que tornem as sínteses mais expressivas e naturais.

O uso inovador dos métodos inéditos desenvolvidos nesta tese, como a *Predição Linear Ponderada adaptada para as Vozes de Canto de Alto Tom* (WLP-HPSV) e o método proposto denominado *Predição Linear Ponderada com Excitação Principal Atenuada Adaptado ao Canto Agudo* (WLP-AME-ADP) para estimar com precisão parâmetros de ressonância, representa um passo significativo em direção a uma síntese vocal autêntica, inteligível e expressiva. A aplicação desses métodos no processo de estimação de frequências formantes aprimora a qualidade e o realismo das sínteses, especialmente ao reproduzir as características únicas das sopranos de língua espanhola.

Este estudo oferece duas abordagens de síntese de vogais cantadas: uma com parâmetros estimados de forma geral e outra com parâmetros estimados com precisão, considerando as características individuais das cantoras. A abordagem refinada demonstra resultados notavelmente superiores em termos de naturalidade e inteligibilidade, destacando a importância da parametrização personalizada. Os resultados apresentados têm o potencial de enriquecer e revolucionar a síntese de vozes cantadas, não apenas para as sopranos líricas de língua espanhola, mas também para outros tipos vocais e gêneros musicais. Ao abordar os desafios específicos do canto agudo, este estudo estabelece as bases para futuros desenvolvimentos na síntese vocal interativa. A integração de efeitos como vibrato e trêmulo aumenta a expressividade das vozes sintetizadas, abrindo possibilidades criativas na música, teatro e produção audiovisual.

Palavras-chave: síntese de voz cantada, modelagem da fonte glotal, estimativa de frequências formantes, interação fonte-filtro, inteligibilidade, naturalidade, sopranos hispanoparlantes.

Abstract

The synthesis of sung voice has been a topic of growing interest due to its significant implications in the music industry, vocal technique analysis, and recreational applications. In recent decades, there have been numerous remarkable advances that stand out for the naturalness and expressiveness of the syntheses. However, the synthesis of sung voice by Spanish-speaking lyric sopranos presents unique challenges that warrant further attention.

The primary challenge of this thesis lies in the precise reproduction of vocal characteristics, including the fundamental frequency (f_o), formant frequencies, vibrato, tremolo, and phoneme duration, in order to achieve a natural, expressive, and intelligible synthesis. This calls for careful consideration of the complexities inherent in the sung voice of Spanish-speaking lyric sopranos in their higher pitch ranges. To address these challenges, the main objective of this thesis is to develop a specific synthesizer for sung vowels by Spanish-speaking sopranos. This is achieved through mathematical models of the glottal source, accurately estimated formant frequencies, and consideration of aspects such as *Source-Filter Interaction* (SFI). Additionally, the addition of effects like tremolo and vibrato to make the syntheses more expressive and natural.

The innovative use of novel methods developed in this thesis, such as *Weighted Linear Prediction adapted to High-Pitched Singing Voices* (WLP-HPSV) and the proposed method called *Weighted Linear Prediction with Attenuated Main Excitation adapted to the High-Pitched Singing* (WLP-AME-ADP) for accurately estimating resonance parameters, represents a significant step towards authentic, intelligible, and expressive vocal synthesis. The application of these methods in the estimation of formant frequencies enhances the quality and realism of the syntheses, especially in reproducing the unique characteristics of Spanish-speaking sopranos.

This study offers two approaches to synthesizing sung vowels: one with parameters estimated in a general manner and another with parameters estimated precisely, considering the individual characteristics of the singers. The refined approach demonstrates notably superior results in terms of naturalness and intelligibility, emphasizing the importance of personalized parameterization. The presented results have the potential to enrich and revolutionize the synthesis of sung voices, not only for Spanish lyrical sopranos but also for other vocal types and musical genres. By addressing the specific challenges of high-pitched singing, this study establishes the foundations for future developments in interactive vocal synthesis. The integration of effects such as vibrato and tremolo enhance the expressiveness of synthesized voices, opening creative possibilities in music, theater, and audiovisual production.

Keywords: singing voice synthesis, glottal source modeling, formant frequency estimation, source-filter interaction, intelligibility, naturalness, spanish-speaking sopranos.

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

RNA	*Redes Neurais Artificiais* 1
SFI	*Interação Fonte-Filtro* iv
WLP	*Predição Linear Ponderada* 4
LPC	*Codificação Preditiva Linear* 38
STE	*Short-Time Energy* 40
AME	*Excitação Principal Atenuada* vi
EGG	*Eletroglotografia* 41
LF	*Liljencrants-Fant* vi
HMM	*Modelos Ocultos de Markov* 24
QCP	*Fase Quase Fechada* 42
RTV	*Resposta do Trato Vocal* 42
OP	*Fase Aberta* 20
CP	*Fase Fechada* 15
GIF	*Filtragem Inversa da Glote* 42
DAW	*Estação de Áudio Digital* 55
SSD	*Solid-State Drive* 56
WAV	*Waveform Audio File* 57
ZFF	*Filtragem de Frequência Zero* 22
SEDREAMS	*Detecção de Eventos de Fala usando Excitação Residual e Sinal Baseado na Média* 22
TFD	*Transformada de Fourier Discreta* 12
FIR	*Finite Impulse Response* 45
GOI	*Instante de Abertura Glótica* 4
GCI	*Instante de Fechamento Glótico* 4

WLP-AME *Predição Linear Ponderada com Excitação Principal Atenuada* 37

WLP-AME-ADP *Predição Linear Ponderada com Excitação Principal Atenuada Adaptado ao Canto Agudo* iv

WLP-HPSV *Predição Linear Ponderada adaptada para as Vozes de Canto de Alto Tom* iv

DYPSA *Dynamic Programming Projected Phase-Slope Algorithm* 41

SVC *Síntese de Voz Cantada* 48

Sumário

Capítulo 1

Introdução

A síntese da voz cantada é um campo de estudo que tem atraído considerável interesse de pesquisa ao longo das últimas décadas [10, 11, 12, 13]. A capacidade de criar vozes cantadas artificialmente tem implicações significativas na indústria musical, desde produções profissionais até aplicações recreativas. Por exemplo, o instrumento musical simulado computacionalmente que permite a síntese em tempo real de vozes cantadas desenvolvido por D'Alessandro et al. [10], ou ainda o trabalho recente sobre um sintetizador de voz para canto baseado em *Redes Neurais Artificiais* (RNA) apresentado por Chandna et al. [13]. No entanto, a síntese da voz cantada, especialmente quando se trata de sopranos, que são conhecidas por suas vozes agudas e brilhantes, em comparação com outros tipos de vozes, apresenta desafios que exigem específicas abordagens para alcançar resultados satisfatórios [14].

A síntese da voz cantada por sopranos hispanoparlantes é um tópico relevante, porém menos explorado em comparação com outras línguas [14, 12, 11, 15]. A predominância do inglês na indústria musical global direcionou grande parte da pesquisa e desenvolvimento em síntese vocal para contextos de língua inglesa. Além disso, a diversidade de dialetos e sotaques do espanhol, em diferentes regiões, apresenta desafios adicionais na busca por um sistema de síntese vocal que capture com precisão as nuances e características de cada variante. As implicações dessa falta de atenção são notáveis. Primeiramente, limita-se à disponibilidade de ferramentas e recursos de alta qualidade em síntese da voz cantada para treinadores vocais, cantores, músicos, compositores e produtores musicais de língua espanhola. Além disso, a escassez de pesquisas nesse campo dificulta o avanço da tecnologia de síntese da voz cantada voltada para a comunidade hispanoparlante [16]. Diante desses desafios, é essencial explorar e desenvolver técnicas de síntese da voz cantada que atendam às necessidades específicas das sopranos hispanoparlantes, buscando alcançar

resultados mais satisfatórios nesse contexto.

Um dos principais desafios na síntese da voz cantada por sopranos é a reprodução precisa das características vocais, como a frequência fundamental (f_o), frequências dos formantes, projeção vocal, vibrato, trêmulo, duração dos fonemas e a dinâmica [9, 17, 18]. Esses aspectos são essenciais para obter uma síntese de voz cantada natural e de alta qualidade, pois influenciam diretamente a percepção tonal, a ressonância vocal, a inteligibilidade, a expressividade e a naturalidade da voz sintetizada [19].

A estimativa das frequências dos formantes na voz cantada apresenta desafios significativos. Historicamente, essa tarefa tem sido complexa devido à natureza aguda do canto e às complexidades inerentes ao trato vocal. Métodos indiretos, como o uso de vibradores externos ou fontes acústicas de banda larga, foram tentados, mas essas abordagens possuem limitações e imprecisões. Tradicionalmente, obter estimativas precisas de formantes a partir de sinais agudos tem sido quase impossível [20]. No entanto, avanços recentes em técnicas de análise espectral e predição linear têm mostrado promessa em superar essas dificuldades, permitindo medições precisas das frequências dos formantes mesmo no contexto do canto agudo [21, 22, 23]. Esse progresso abre novas possibilidades para explorar características da voz cantada, aprimorar a identificação de cantores, e compreender técnicas vocais.

A projeção vocal é fundamental para garantir que a voz cantada seja ouvida de forma clara e alta, mesmo em longas distâncias, espaços amplos, ou em meio ao som produzido por uma orquestra [24]. No canto clássico, as sopranos têm a capacidade de cantar em frequências acima do $D4$ (240 Hz) e, em seu registro mais agudo, podem até ultrapassar o $D6$ (1047 Hz). No entanto, para que o som seja amplificado e percebido com um volume alto, elas precisam usar a técnica de ajuste de formantes, permitindo que elas se destaquem sem muito esforço, ou seja, com eficiência vocal.

Nos registros agudos do canto feminino, a f_o é mais alta do que a frequência do primeiro formante (F_1), especialmente nas vogais /*e*/, /*i*/, /*o*/ e /*u*/. Isso resulta em um som aparentemente mais *fraco*, uma vez que F_1 não aumenta a amplitude de f_o. Para contornar esse problema, as cantoras sopranos aumentam involuntariamente a abertura da boca, elevando a frequência F_1 e fazendo-a coincidir ligeiramente acima de f_o, permitindo que o primeiro formante reforce a amplitude de f_o. Esse fenômeno foi observado por Sundberg e chamado de ajuste de formantes [25, 26]. No entanto, esse ajuste de formantes pode afetar a inteligibilidade das vogais e apresentar um desafio na síntese da voz cantada [17, 27].

Por outro lado, efeitos comuns à voz humana, como dinâmica, vibrato, trêmulo e duração dos fonemas, podem adicionar expressividade e emoção à voz cantada. A dinâmica refere-se à variação na intensidade ou volume da voz durante uma interpretação [28]. O vibrato consiste em uma variação periódica da f_o para adicionar expressividade [9]. O trêmulo é uma variação rápida da intensidade vocal. Já a duração dos fonemas se refere ao tempo de pronúncia dos sons individuais. Ao recriar esses efeitos na voz de canto sintetizada, é possível melhorar a naturalidade e a qualidade vocal da síntese, proporcionando uma experiência mais realista e envolvente [7].

A interação fonte-filtro (SFI, source-filter interaction) refere-se à influência do trato vocal na fonte glótica e desempenha um papel fundamental na produção da voz cantada [29, 30]. Essa interação tem sido amplamente estudada na área de síntese vocal. Embora também ocorra na fala, a voz cantada exibe um grau mais significativo de SFI. Além disso, a voz cantada apresenta uma ampla variedade de tons, variações controladas de tom, duração das frases, prosódia e um maior alcance dinâmico, o que torna o processamento da voz cantada um desafio [31]. Portanto, é crucial que um sintetizador de voz cantada seja capaz de modelar com precisão a interação fonte-filtro, garantindo uma síntese de voz mais natural e realista.

1.1 Motivação

A motivação subjacente à realização deste estudo emerge da necessidade de aprimorar as técnicas de síntese de voz cantada, especialmente nos registros agudos das sopranos líricos hispanoparlantes. A complexidade intrínseca da voz cantada, com seus fenômenos como vibrato e trêmulo, demanda abordagens refinadas para capturar com fidelidade sua expressividade. Além disso, a interação complexa entre a fonte (glote) e o filtro (trato vocal) apresenta desafios consideráveis na estimativa de frequências formantes e suas larguras de banda. O presente trabalho surge para superar esses obstáculos, alavancando a inovação no desenvolvimento de métodos precisos e incorporando essas descobertas em um processo de síntese de alta qualidade. O resultado é uma abordagem mais sofisticada e confiável na criação de sínteses vocais realistas, ampliando os horizontes da pesquisa em síntese de voz e sua aplicação em diversas áreas, como música, entretenimento e tecnologias assistivas.

1.2 Objetivos

Para enfrentar os desafios da síntese da voz cantada por sopranos, esta tese tem como objetivo principal desenvolver um sintetizador de vogais cantadas específico para sopranos hispanoparlantes, com base em modelos matemáticos da fonte glotal, frequências formantes estimadas pelo método de *Predição Linear Ponderada* (WLP) (Weigthing Linear Prediction ou predição linear ponderada) e consideração de aspectos como interação fonte-filtro e variações de trêmulo e vibrato, a fim de produzir sinais sintetizados que se destacam pela naturalidade e inteligibilidade.

Para atingir o objetivo principal desta tese, os seguintes objetivos específicos foram definidos:

- Realizar uma revisão abrangente da literatura sobre síntese de voz cantada, modelos da fonte glotal, estimação das frequências formantes e predição linear ponderada.
- Construir um banco de dados contendo gravações de vogais cantadas por sopranos hispanoparlantes, considerando variações de estilo e expressão vocal.
- Desenvolver um sintetizador por formantes que permita controlar características como intensidade, variação de tom, duração dos fonemas e projeção vocal para a síntese de voz cantada.
- Implementar modelos matemáticos da fonte glotal para a geração do sinal glotal, considerando parâmetros como *Instante de Fechamento Glótico* (GCI), *Instante de Abertura Glótica* (GOI) e período fundamental.
- Estimar as frequências formantes das vogais cantadas utilizando o método de WLP, adaptado especificamente para a voz cantada. Além da síntese propriamente dita, esse método é uma grande contribuição dessa tese.
- Avaliar a qualidade, naturalidade e inteligibilidade dos sinais de voz sintetizados pelo sistema proposto, comparando-os com gravações reais de sopranos hispanoparlantes.
- Realizar análises perceptuais e objetivas para verificar a eficácia do sintetizador de voz cantada, considerando parâmetros como similaridade, clareza e expressividade vocal.

Ao superar esses objetivos, a síntese de voz cantada por sopranos hispanoparlantes pode abrir novas possibilidades criativas, tanto na música quanto em aplicações relacionadas,

como a produção de áudio, cinema e jogos. Além disso, contribuirá para o avanço do conhecimento no campo da síntese da voz cantada, oferecendo novas perspectivas e abordagens para aprimorar a qualidade e realismo das vozes sintetizadas.

A estratégia metodológica adotada para esta pesquisa se concentrou na síntese de vogais cantadas por sopranos líricos hispanoparlantes em seus registros mais agudos, por meio de um sintetizador por formantes. Nessa abordagem, houve uma integração de inovações, incluindo a incorporação de características singulares da voz cantada, como vibrato, trêmulo e duração dos fonemas. Além disso, foram desenvolvidos dois métodos inéditos para estimar com precisão as frequências formantes e suas larguras de banda, lidando com a complexa interação fonte-filtro inerente à voz cantada. A pesquisa também promoveu melhorias no sintetizador por meio de abordagens interativas, fundamentadas no modelo fonte-filtro. Adicionalmente, a construção de uma abrangente base de vozes de diferentes cantoras sopranos líricos contribuiu substancialmente para o desenvolvimento desta pesquisa.

1.3 Estrutura da tese

Esta tese foi organizada em oito capítulos distintos, começando com uma introdução que estabelece o contexto e a importância da pesquisa. O segundo capítulo oferece uma revisão bibliográfica abrangente, explorando conceitos essenciais, teorias e técnicas relevantes para a síntese da voz cantada. Em seguida, o terceiro capítulo se concentra na estimação das frequências formantes da voz cantada, apresentando métodos avançados de predição linear ponderada. O quarto capítulo aborda a síntese da voz cantada por sopranos, destacando abordagens não interativas e interativas. O quinto capítulo detalha a metodologia da pesquisa, incluindo a construção da base de vozes e os procedimentos de análise e síntese. Os resultados da pesquisa são apresentados no sexto capítulo, com uma análise detalhada das sínteses realizadas. O sétimo capítulo contém as conclusões principais deste estudo, enquanto o oitavo capítulo explora possíveis direções para trabalhos futuros na área. Finalmente, a tese é complementada por uma seção de referências e um apêndice que contém os algoritmos cruciais desenvolvidos durante esta pesquisa.

Capítulo 2

Revisão Bibliográfica

2.1 Fisiologia e produção da voz humana

A voz é a base da comunicação humana, através da qual transmitimos informações e expressamos sentimentos e ideias. Sua excepcional importância levou pesquisadores e pensadores ao longo da história da humanidade a estudar sua fisiologia. Inúmeros experimentos foram realizados com o objetivo de construir teorias e entender as características morfológicas da laringe, determinando assim a forma como a voz humana é produzida.

A produção da voz não ocorre somente na laringe, mas sim pela ação coordenada de várias estruturas do corpo humano. Essas estruturas formam o sistema fonador, como ilustrado na Figura. 2.1, e incluem músculos de diferentes regiões do corpo, elementos do aparelho respiratório e digestivo, conforme mencionado por [32]. A produção efetiva da voz baseia-se na coordenação de três ações realizadas pelo sistema fonador: a respiração, a fonação e a ressonância.

A respiração é uma das funções mais vitais do corpo humano, desempenhando um papel fundamental na produção da voz. Sem a intervenção do sistema pulmonar, a voz não poderia ser gerada. De acordo com Titze [33], o tórax e o abdômen, juntamente com os pulmões, atuam como um fole para criar um fluxo de ar, enquanto a glote funciona como uma válvula reguladora desse fluxo. Durante a inspiração (inalação), os pulmões expandem-se, permitindo que o ar flua da boca em direção a eles, com a glote relativamente aberta. Já na expiração (exalação), os pulmões se contraem, empurrando o ar de volta à boca.

No aparelho fonador, a laringe é responsável por produzir a principal fonte de som da voz humana. Localizada entre a faringe e a traqueia, a laringe é principalmente composta

por cartilagem, com apenas um osso (hioide) que flutua em relação ao esqueleto.

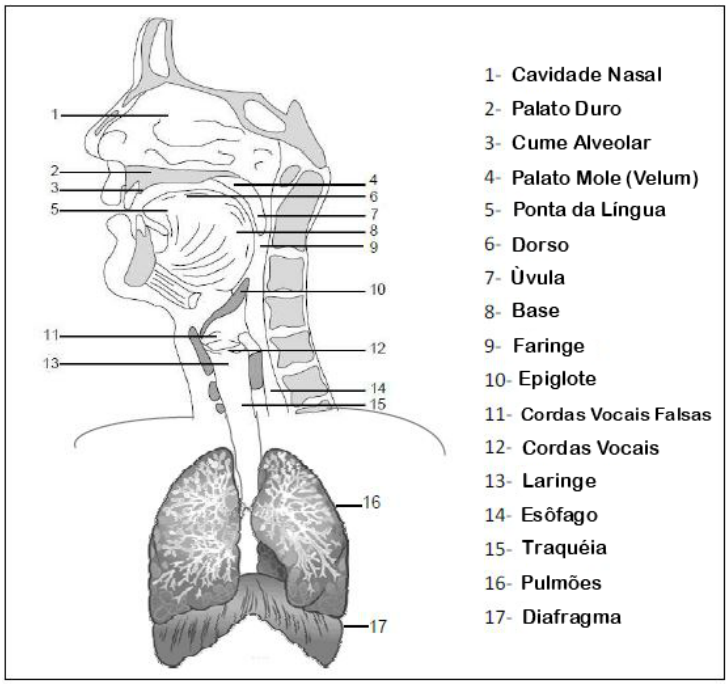


Figura 2.1: Sistema da produção da voz humana [1].

As cordas vocais são os componentes funcionais mais relevantes da laringe, atuando como geradores de som através de sua oscilação com a passagem do ar [1]. Elas consistem em três camadas distintas: o músculo tireoaritenoideo, a camada de ligamento vocal e uma membrana mucosa, trabalhando em conjunto para promover a ondulação durante a produção do som. Na Figura. 2.2, é possível observar a glote, uma abertura de forma triangular entre as cordas vocais.

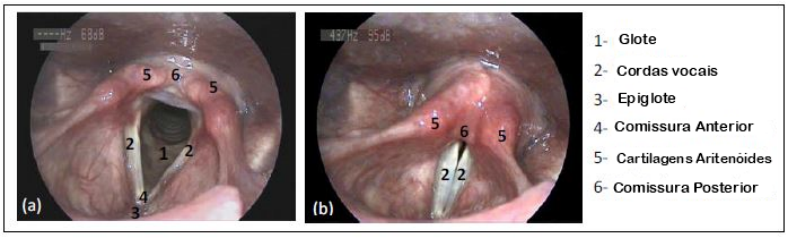


Figura 2.2: Vista das cordas vocais da laringe (a) Cordas vocais abertas, (b) Cordas vocais fechadas [1].

No processo de respiração, a glote se abre, permitindo a livre passagem do ar, sem ocorrer a oscilação das cordas vocais e, portanto, sem produção sonora. Para a fonação acontecer, a glote se fecha (adução), e uma coluna de ar, expelida dos pulmões, quebra a resistência criada pelas cordas vocais para a passagem do ar, resultando em um som ou zumbido. Esse processo é conhecido como ciclo glotal, conforme ilustrado na Figura. 2.3.

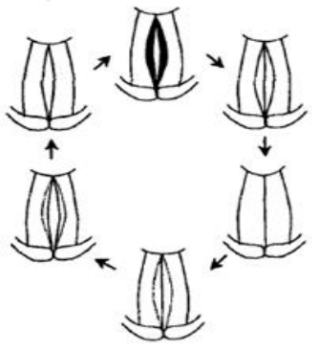


Figura 2.3: Representação de um ciclo glotal [2].

Quando as pregas vocais se encontram em adução, a pressão do ar abaixo da glote (pressão subglótica) se torna maior do que a pressão do ar acima dela (pressão supraglótica). Essa força gradualmente separa as cordas vocais, começando pela parte inferior até que reste apenas o contato na parte superior. Em seguida, as cordas vocais se fecham novamente por baixo, devido ao efeito Bernoulli, que cria uma sucção e permite o fechamento completo delas. A teoria mais aceita para explicar a fonação é a aerodinâmica mioelástica (Van den Berg, 1958), complementada pela teoria Muco-ondulatória de Perelló (1962), que combina a configuração das cordas (mioelástica) e o efeito do ar (aerodinâmica) [2].

Os sons gerados pela oscilação das cordas vocais são chamados de vozeados ou sonoros, enquanto aqueles produzidos sem oscilação, com as cordas vocais abertas e excitação glotal barulhenta, são referidos como não vozeados [1]. As vogais incluem-se nos sons vozeados.

As cordas vocais variam em comprimento de acordo com o gênero e a idade do indivíduo. Em recém-nascidos, elas são mais curtas e finas, com cerca de 3mm, enquanto em adultos, o comprimento varia de 13mm a 17mm no sexo feminino e de 17mm a 24mm no sexo masculino. A amplitude da abertura glótica é de aproximadamente 3mm [3].

A oscilação das cordas vocais é responsável pela produção de uma variedade de sons com uma f_o específica, que tende a ser mais alta em crianças e mulheres e está inversamente relacionada ao comprimento das cordas vocais. Quanto mais curtas as cordas, maior será a f_o. Essa característica é uma das razões pelas quais as vozes de crianças e mulheres geralmente soam mais agudas em comparação com as vozes de homens adultos. A compreensão dessa relação entre o comprimento das cordas vocais e a f_o é fundamental para a síntese da voz cantada, especialmente quando se trata de reproduzir vozes femininas e infantis com precisão e naturalidade.

A f_o das oscilações laríngeas nos tons sustentados varia em torno de 60 Hz a 1500 Hz,

abrangendo uma ampla faixa de frequências. Para cobrir essa extensão, o aparelho fonador possui vários modos vibratórios conhecidos como mecanismos laríngeos. A Figura. 2.4 mostra a extensão vocal média em voz falada e cantada.

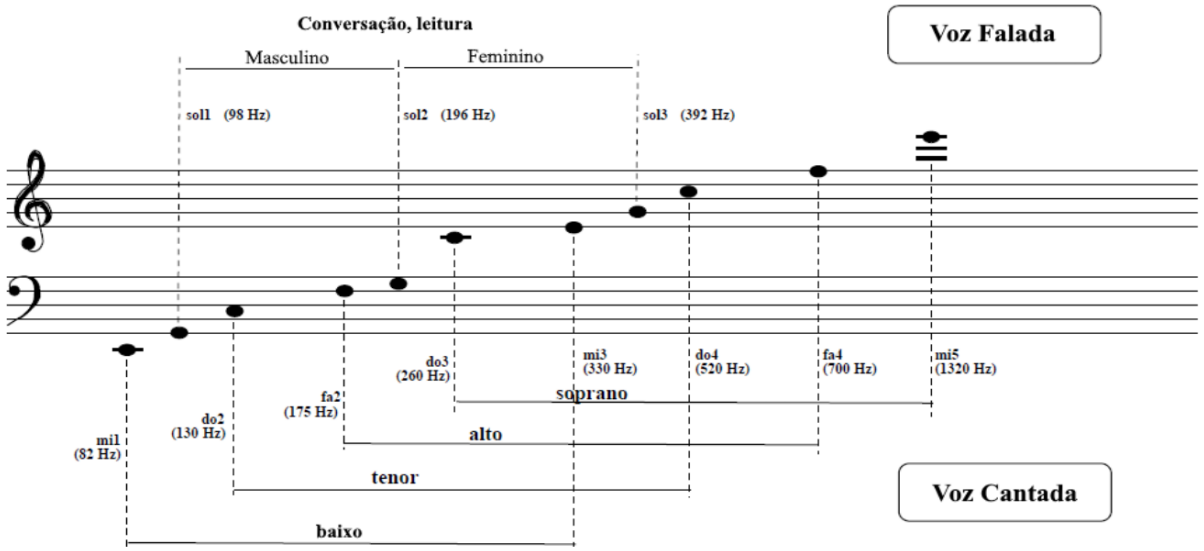


Figura 2.4: Extensão vocal média em voz falada e cantada [3].

O aparelho ressonador, também conhecido como trato vocal, é composto pela caixa de ressonância inferior, que inclui a faringe, traqueia, brônquios e pulmões, e pela caixa de ressonância superior, compreendendo a cavidade bucal e nasal. Essa configuração pode ser observada na Figura. 2.5.

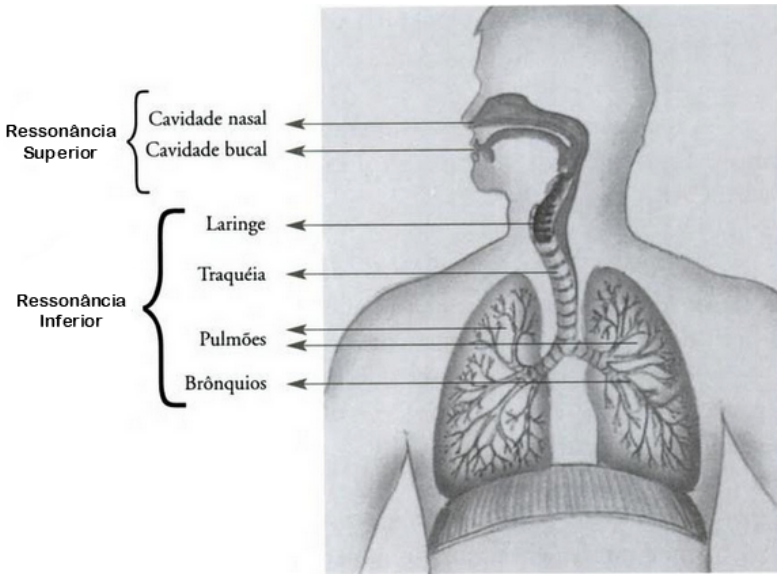


Figura 2.5: Aparelho ressonador ou trato vocal [4].

Graças aos articuladores, o som proveniente das pregas vocais é dirigido para o trato

vocal e as manobras feitas para modificar a forma do trato vocal durante a fonação é denominada articulação [34]. Os principais articuladores usados pelos cantores para moldar o som são: a língua, a mandíbula, o vélum e os lábios. Em menor grau, alguns músculos da faringe também podem ser usados para aplicar algumas constrições em locais diferentes. O trato vocal pode então ser modelado como um tubo simples com diâmetro variável determinado pela posição desses articuladores. Dependendo dessa forma, o trato vocal ressoa em algumas frequências específicas. Essas ressonâncias do trato vocal são chamadas de *formantes* [7].

Tanto na voz cantada como na falada, a ressonância vocal consiste na modificação do som produzido pelas cordas vocais, resultando em efeitos de atenuação e amplificação do sinal sonoro em regiões específicas de frequência [35]. Os ressonadores, além de amplificar ou atenuar o som gerado pelas pregas vocais, também conferem brilho e facilitam a transição entre notas graves e agudas.

No canto, a ressonância geralmente se concentra na parte superior do trato vocal, resultando em uma elevação da energia sonora nessa região. O objetivo de alcançar uma ressonância equilibrada é aliviar a sobrecarga muscular da laringe. Muitos cantores utilizam um certo grau de nasalidade para dissipar a energia sonora sem sobrecarregar a laringe.

Todas as vogais são sonoras, exigindo a oscilação das cordas vocais, e podem ser produzidas pela ação do trato vocal. As diferentes vogais do espanhol (/*a*/, /*e*/, /*i*/, /*o*/, /*u*/) são produzidas por uma articulação apropriada: os dois primeiros formantes estão relacionados à vogal produzida, sendo o primeiro formante principalmente influenciado pela abertura da mandíbula e o segundo formante pela posição da língua. Os próximos três formantes estão mais relacionados ao timbre e à identidade da voz, sendo o terceiro formante particularmente influenciado pela posição da ponta da língua e o quarto pelas dimensões da laringe [7].

As frequências ressonantes do trato vocal são chamadas de frequências formantes e são fundamentais na produção dos sons sonoros. Os dois primeiros formantes são os mais importantes, pois determinam a maior parte do colorido da voz, enquanto o terceiro, o quarto e o quinto formantes contribuem para o timbre da voz [26]

A análise acústica das vogais espanholas é realizada usando os dois primeiros formantes, de acordo com a seguinte distribuição de posicionamento, ilustrada na Figura. 2.6.

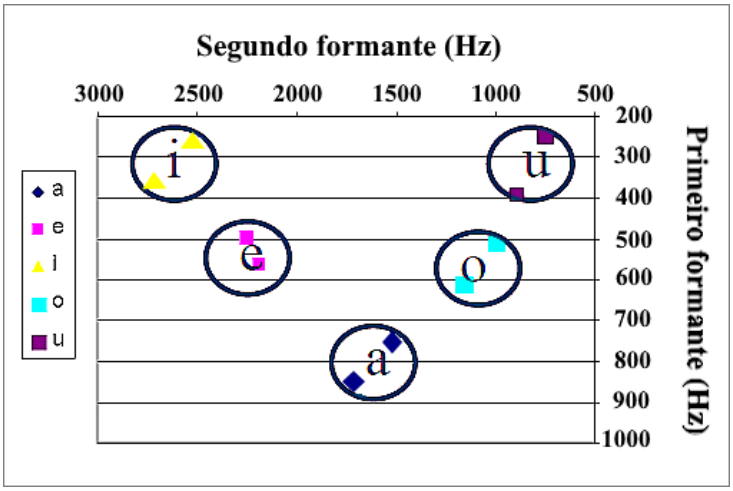


Figura 2.6: Triângulo articulatório de fonemas vocálicos da língua espanhola [5].

As vogais /i/ e /u/ ocupam as áreas limítrofes (canto superior direito e esquerdo) do triângulo articulatório, enquanto /a/ ocupa a parte central inferior, estendendo-se entre /i/ e /u/. As demais vogais espanholas estão posicionadas em pontos intermediários entre /i/-/a/ e /u/-/a/, formando um triângulo articulatório que representa suas respectivas posições. Nesse contexto, /a/, /i/ e /u/ estão nas pontas do triângulo articulatório, e o primeiro formante, F_1, é representado no eixo vertical, enquanto o segundo formante, F_2, é representado no eixo horizontal. Esses valores das frequências do primeiro e segundo formantes são fundamentais para a análise acústica das vogais e têm um papel significativo na caracterização e diferenciação dos sons vocálicos espanhóis.

2.2 A teoria fonte-filtro

A teoria proposta por Gunnar Fant [36, 37], em 1960, aborda o mecanismo de produção da voz, modelando-o como a convolução entre uma fonte de excitação e um sistema de filtros lineares conectados em série. Nesse contexto, a fonte de excitação é representada pelo fluxo de ar que faz vibrar as cordas vocais, enquanto a ressonância vocal, responsável por modificar o som (onda sonora) ao longo do tempo, é representada por um filtro, assim como os lábios que irradiam a voz. No entanto, essa teoria é baseada em duas suposições fundamentais.

1. A fonte e o filtro são considerados independentes entre si.

2. No domínio do tempo, o processo da produção da fala pode ser representado pela convolução dos elementos envolvidos, como a fonte glotal, o filtro correspondente ao

trato vocal e o filtro correspondente à radiação dos lábios e narinas.

A primeira suposição implica que a fonte glotal é igual ao fluxo glotal, o que, na realidade, não é perfeitamente válido devido à interação entre fonte e filtro. O fluxo glotal depende, em certo grau, das variações da impedância do trato vocal [1]. Muitos algoritmos e técnicas utilizados na área de processamento da fala são baseados neste modelo bastante simples. Esse modelo é simplificado para mostrar a relação entre a fonte glotal, o trato vocal e a irradiação pelos lábios/narinas, como ilustrado na Figura. 2.7.

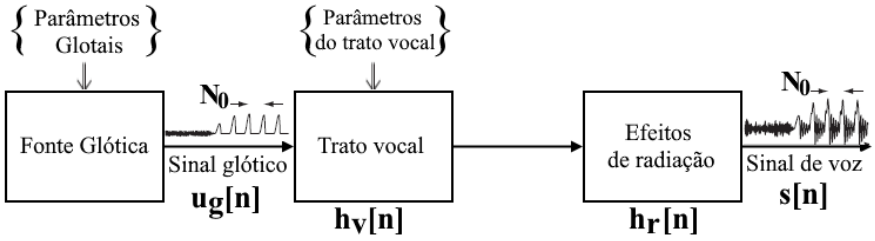


Figura 2.7: Modelo fonte-filtro de Fant.

O sinal de voz gerado, $s[n]$, é resultado da convolução entre o sinal glótico $u_g[n]$, a resposta ao impulso correspondente $h_v[n]$ do filtro que modela o trato vocal, e a radiação pela boca, para a qual a função de resposta ao impulso é $h_r[n]$. Essa convolução é representada pela equação (2.1).

$$s[n] = h_r[n] * h_v[n] * u_g[n]\,, \tag{2.1}$$

ou, no domínio da frequência,

$$S(k) = H_R(k)\, H_V(k)\, U_G(k)\,, \tag{2.2}$$

onde $S(k)$ significa a *Transformada de Fourier Discreta* (TFD) do sinal de voz gerado.

Com essa formulação, não há acoplamento entre a fonte glotal e o trato vocal, simplificando o modelo.

2.2.1 Modelagem da fonte glotal

Qualquer tecido no aparelho fonador capaz de vibrar pode ser considerado uma possível fonte acústica, mas a vibração das cordas vocais é a fonte predominante para a produção

dos sons vozeados, e é essa vibração que influencia a qualidade da voz [38]. Essa vibração é quase periódica, mantida por um mecanismo auto-sustentado.

A teoria Fonte-Filtro simplifica as características da fonte acústica, ou seja, a fonte glotal, tornando mais fácil modelá-la como um trem de pulsos glotais, ou sinal glótico, com características semelhantes às produzidas na glote. Embora na síntese da voz o sinal glótico possa ser gerado por qualquer trem de pulsos, para obter uma voz mais natural é recomendável que o trem de pulsos seja o mais próximo possível do que é produzido na glote, como ilustrado na Figura. 2.8.

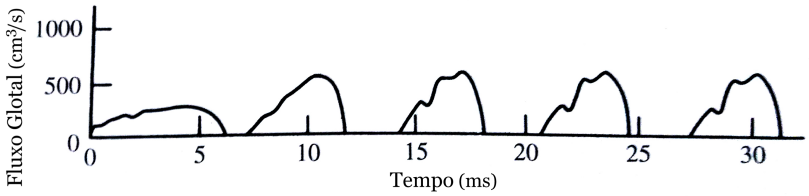


Figura 2.8: Sinal glótico real da voz [6].

Na forma de onda da Figura. 2.8, podemos observar que a inclinação dos pulsos glotais para a direita é causada por uma ascensão lenta e uma queda abrupta do fluxo glotal, o que indica que o acúmulo de fluxo de ar está atrasado em relação ao deslocamento das cordas vocais.

Para abordar o sinal glótico, foram desenvolvidos modelos que procuram explicá-lo e reproduzí-lo com precisão. Ele é representado pela forma de onda de sua velocidade, que, em um ciclo glotal completo, é denominado pulso glotal.

O pulso glotal representa um ciclo glotal completo, composto por fases que estão relacionadas aos aspectos fisiológicos da fonação, ou seja, ao movimento vibratório das cordas vocais. Esse pulso é modelado sob certas suposições sobre o seu comportamento, e as fases do movimento glótico são ilustradas na Figura. 2.9.

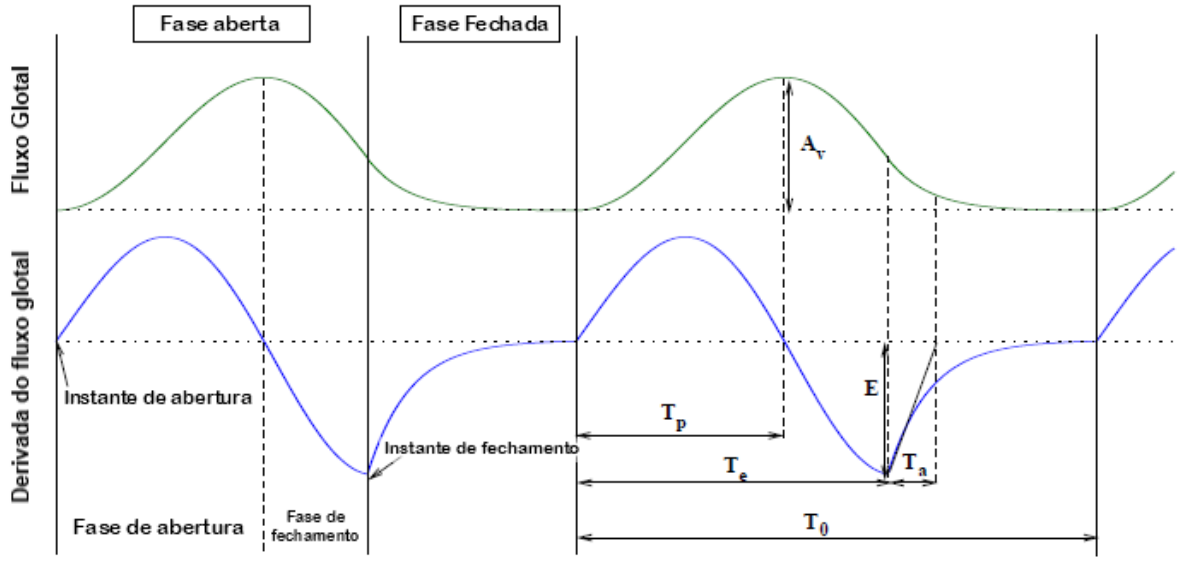


Figura 2.9: Terminologia usada para descrever a fonte glótica [3].

Na parte superior da Figura. 2.9, podemos observar a forma de onda do fluxo glotal e, na parte inferior, a forma de onda de sua derivada. T_0 representa o período fundamental de um pulso glótico, T_e é o tempo de duração da fase de abertura, T_p é o tempo de duração da fase de abertura, T_a é o tempo de duração efetiva da fase de retorno, E corresponde à velocidade de fechamento e A_v é a amplitude de vozeamento.

Antes de analisar as diferentes fases do movimento glótico, é importante compreender os momentos em que ocorrem a abertura e o fechamento glótico, bem como o instante em que ocorre o máximo fluxo glotal, do ponto de vista fisiológico.

O GOI é definido como o instante em que o fluxo glotal aumenta em relação ao seu valor mínimo. Fisiologicamente, corresponde ao momento em que as cordas vocais começam a separar-se desde a sua parte superior. O GCI, por sua vez, está associado ao máximo de excitação glótica, ou seja, o momento em que a derivada do fluxo glótico atinge seu valor mínimo. Fisiologicamente, esse momento corresponde ao instante em que as cordas vocais se unem desde a parte inferior.

O termo *fechamento abrupto* é utilizado quando a derivada do fluxo glótico é descontínua no GCI [3].

- **Fase de abertura:** Definida entre o instante de abertura glótica e o instante de máximo fluxo.
- **Fase de fechamento:** Definida entre o instante de máximo fluxo e o instante de

fechamento glótico.

- **Fase aberta** (OP): Compreende as fases de abertura e fechamento, ou seja, é definida entre o instante de abertura e o instante de fechamento glótico.
- **Fase fechada** (CP): Definida entre os instantes de fechamento e de abertura glótica do ciclo seguinte.
- **Fase de retorno:** Definida entre o momento de máximo da excitação e o momento em que o fluxo atinge seu valor mínimo.

Devido a um vazamento glótico durante a *Fase Fechada* (CP), não é possível considerar o fluxo glótico como nulo durante esta fase [1].

2.2.1.1 Modelo glotal de Rosenberg

A.E. Rosenberg [39] realizou experimentos para determinar o efeito da variação dos tempos relativos de abertura e fechamento glotal na qualidade vocal, criando um modelo básico do pulso glótico. Ele utilizou a técnica de filtragem inversa para extrair três parâmetros da forma de onda glotal e, a partir deles, modelou seis formas diferentes do pulso glótico através de seis expressões matemáticas distintas. No tempo contínuo, a forma mais conhecida do pulso glótico de Rosenberg é expressa por:

$$g(t) = \begin{cases} 0.5A_v[1 - cos(\pi t/(T_P))] & , \ 0 \leq t \leq T_P \\ A_v cos(\pi(t - T_P)/(2T_N)) & , \ T_P < t \leq T_P + T_N \\ 0 & , \ T_P + T_N < t \leq T_0 \, , \end{cases} \tag{2.3}$$

este modelo trigonométrico apresenta duas funções para representar as fases de abertura e fechamento do fluxo glótico. Na expressão, A_v é uma constante que está relacionada à amplitude do pulso glotal, enquanto T_P (tempo de abertura) e T_N (tempo de fechamento) são parâmetros que controlam as porções do pulso com inclinação positiva e negativa, respectivamente. A Figura. 2.10 ilustra a forma de onda do pulso glótico e sua derivada.

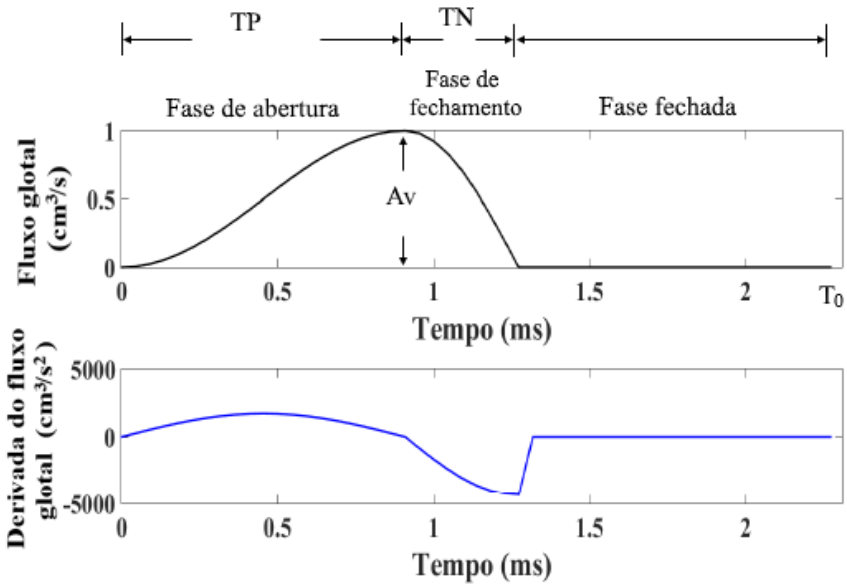


Figura 2.10: Pulso glotal do modelo de Rosenberg e sua derivada.

O parâmetro $\alpha_1 = T_P/T_0$ corresponde ao tempo relativo de abertura, enquanto $\alpha_2 = T_N/T_0$ representa o tempo relativo de fechamento. Notavelmente, esse modelo é altamente flexível, permitindo a obtenção de diferentes formas do pulso glótico por meio de outras expressões matemáticas do mesmo modelo.

A forma do pulso glótico influencia várias características que percebemos na voz do falante; por exemplo, um tom mais grave ou mais agudo pode ser obtido alterando a frequência do pulso glótico. As variações na forma de onda que representam a velocidade volumétrica glótica resultam em diferentes tipos de voz bem identificados. Essas modificações na forma do pulso glótico desempenham um papel crucial na produção de vozes distintas e auxiliam na modelagem e síntese de vozes cantadas com maior precisão e expressividade.

2.2.1.2 Modelo glotal de Liljencrants-Fant

Liljencrants e Fant propuseram, em 1985, um modelo para a derivada do fluxo glótico, conhecido como modelo LF [40]. Esse modelo, com cinco parâmetros no domínio do tempo, em conjunto com o comprimento do ciclo glotal, determina unicamente a forma do pulso. A Figura. 2.11 ilustra a forma de onda do pulso glótico e sua derivada no tempo contínuo, conforme descrito por esse modelo.

O modelo LF é amplamente utilizado para representar o comportamento da voz, permitindo a síntese de vozes com características específicas e contribuindo para a criação

de vozes artificiais de alta qualidade, especialmente em aplicações de síntese vocal. Ele oferece uma abordagem detalhada e precisa para modelar a variação temporal do fluxo glótico e, consequentemente, da fonte glotal na produção da voz. Sua aplicação é essencial para a síntese realista de vozes cantadas, onde a expressividade e a naturalidade são fundamentais.

Os parâmetros descritos são cruciais para o entendimento do modelo LF e a representação precisa do pulso glótico:

- $\boldsymbol{E_e}$: Refere-se ao valor mínimo da derivada do fluxo glótico, o qual corresponde ao máximo de excitação glótica, ou seja, o momento em que as cordas vocais estão mais afastadas.
- $\boldsymbol{T_0}$: Representa o período fundamental, ou seja, a duração completa de um ciclo glótico, incluindo todas as fases desde a abertura até o próximo fechamento glótico.
- $\boldsymbol{T_e}$: Refere-se ao instante de excitação máxima, indicando o momento em que a taxa de aumento do fluxo glótico atinge seu valor máximo.
- $\boldsymbol{T_p}$: Representa o instante do máximo do fluxo glotal, indicando o ponto de pico do movimento das cordas vocais durante o ciclo glótico.
- $\boldsymbol{T_a}$: Corresponde à constante de tempo da fase de retorno, ou seja, o tempo de duração efetiva da fase em que as cordas vocais retornam à sua posição de fechamento para o próximo ciclo glótico.

Esses parâmetros são fundamentais para modelar a forma e o comportamento do pulso glótico, permitindo representar a excitação glótica de maneira detalhada e realista, possibilitando, assim, a síntese de vozes com diferentes características e expressividade.

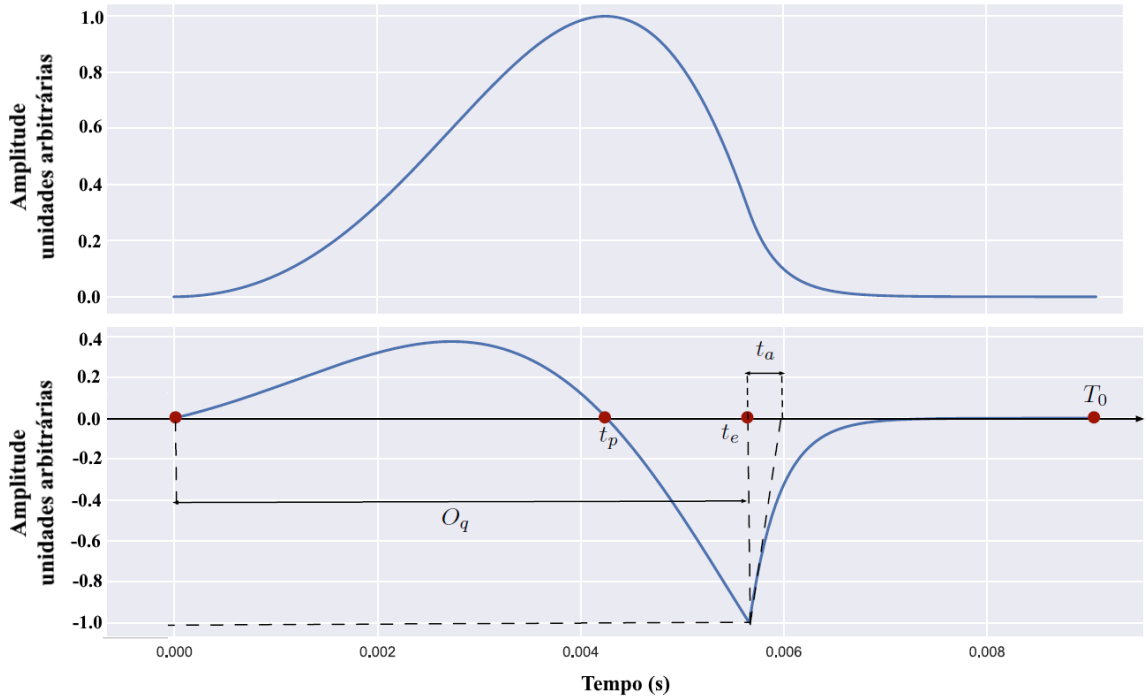


Figura 2.11: Pulso glotal do modelo de LF e sua derivada [7].

O modelo LF é construído por uma parte senoidal modulada por uma exponencial crescente no intervalo de tempo entre 0 e T_e, representando a fase de excitação máxima das cordas vocais. Em seguida, há uma fase de retorno exponencial decrescente entre T_e e T_0, que retrata o movimento das cordas vocais retornando à posição de fechamento. Esse modelo matemático detalhado permite uma representação precisa da forma de onda do pulso glótico, capturando a dinâmica da excitação glótica e suas mudanças ao longo do ciclo glótico [41].

$$g'(t) = \begin{cases} -E_e\, e^{a(t-T_e)} \dfrac{sin(\frac{\pi t}{T_p})}{sin(\frac{\pi T_e}{T_p})} & , \ 0 \leq t \leq T_e \\ \dfrac{-E_e}{\varepsilon T_a}(e^{-\varepsilon(t-T_e)} - e^{-\varepsilon(T_0-T_e)}) & , \ T_e \leq t \leq T_0\,, \end{cases} \tag{2.4}$$

o parâmetro ε é calculado com base na condição de continuidade da derivada do fluxo glótico, $g'(t)$, no instante de fechamento glótico. Essa continuidade é essencial para garantir uma transição suave entre a fase de fechamento e a fase de retorno do ciclo glótico, permitindo que o modelo LF reproduza de forma mais precisa a forma de onda do pulso glótico e suas características acústicas [3].

$$g'(t = T_e^-) = g'(t = T_e^+)\,, \tag{2.5}$$

que dá como resultado a seguinte equação implícita:

$$\varepsilon T_a = 1 - e^{-\varepsilon(T_0 - T_e)}\,, \tag{2.6}$$

o parâmetro a é calculado com base na condição de continuidade do fluxo glótico, $g(t)$, no instante de fechamento glótico.

$$g(t = T_e^-) = g(t = T_e^+)\,, \tag{2.7}$$

a partir desta condição de continuidade, surge a seguinte equação implícita:

$$\frac{1}{a^2 + (\frac{\pi}{T_p})^2}(e^{-aT_e}\frac{\pi/T_p}{sin(\pi T_e/T_p)} + a - \frac{\pi}{T_p}cotg(\pi T_e/T_p)) = \frac{T_0 - T_e}{e^{\varepsilon(T_0 - T_e)} - 1} - \frac{1}{\varepsilon} \tag{2.8}$$

2.2.2 Unificação dos modelos temporais da fonte glotal

Na busca por uma análise mais comparativa entre os modelos de fluxo glótico de domínio do tempo, de forma a entender suas vantagens e desvantagens, Boris Doval e Christophe d'Alessandro [42] propuseram uma estrutura unificada. Eles conseguiram demonstrar que todos os modelos podem ser representados por um conjunto comum de cinco parâmetros no domínio do tempo. Esses parâmetros foram classificados da seguinte maneira: três parâmetros de escala (período fundamental, amplitude de vozeamento, quociente de abertura), um parâmetro de forma (coeficiente de assimetria) e um parâmetro de continuidade de fechamento (constante de tempo da fase de retorno). Essa abordagem permite reescrever os parâmetros dos diferentes modelos de fluxo glótico em termos desses cinco parâmetros fundamentais, facilitando a comparação e análise entre eles, proporcionando uma visão mais integrada dos aspectos acústicos e perceptivos da voz [42]. Esses parâmetros são os seguintes:

- **Amplitude de vozeamento** (A_v): Define-se como a diferença entre o valor mínimo e máximo do fluxo glótico, representando a amplitude do pulso glótico.
- **Velocidade de Fechamento** (E_e): É o valor mínimo da derivada do fluxo glótico,

correspondendo à velocidade do fluxo no instante de fechamento glótico.

- **Quociente de abertura** (O_q): Relação entre a duração da *Fase Aberta* (OP) e o período fundamental do ciclo glotal. Pode variar teoricamente entre 0 (sem abertura) e 1 (sem fechamento ou fechamento incompleto), mas na prática, costuma situar-se entre 0.3 e 0.98. Pode ser calculado como $O_q = \frac{T_e}{T_0}$.
- **Coeficiente de assimetria** (α_m): Número adimensional que define o instante de máximo da onda de fluxo glótico relativo aos parâmetros O_q e T_0. Varia entre 0.5 (forma simétrica) e 1 (caso limite de uma forma muito assimétrica). É expresso como $\alpha_m = \frac{T_p}{O_q T_0}$.
- **Constante de tempo de fase de retorno** (T_a): Representa a diferença entre o instante do máximo de excitação e o instante em que o fluxo atinge seu valor mínimo. Seu valor é nulo em casos de fechamento abrupto, e indica a velocidade de fechamento das pregas vocais.

Todos os parâmetros anteriormente descritos são ilustrados na Figura. 2.12.

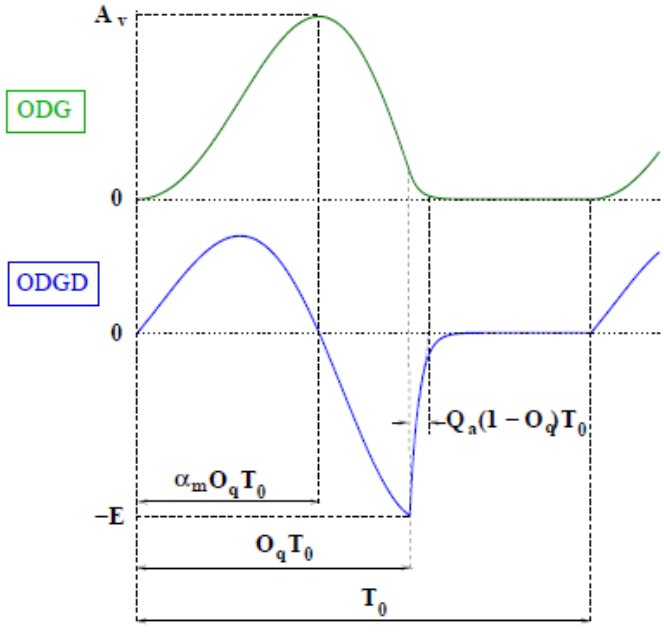


Figura 2.12: Parâmetros unificados da fonte glótica [3].

2.2.2.1 Modelo glotal de Rosenberg com parâmetros unificados

A expressão da Eq. 2.9 corresponde ao modelo de pulso glotal de Rosenberg com parâmetros unificados equivalente à expressão da Eq. 2.3. Os parâmetros no domínio do tempo são

obtidos a partir dos parâmetros originais como: $O_q = (T_P + T_N)/T_0$ e $\alpha_m = T_p/(T_P + T_N)$.

$$g(t) = \begin{cases} 0.5A_v[1 - cos(\pi \dfrac{t}{\alpha_m O_q T_0})] & , \ 0 \leq t \leq \alpha_m O_q T_0 \\ A_v cos(\dfrac{\pi(t - \alpha_m O_q T_0)}{2O_q T_0(1 - \alpha_m)}) & , \ \alpha_m O_q T_0 < t \leq O_q T_0 \\ 0 & , \ O_q T_0 < t \leq T_0 \end{cases} \tag{2.9}$$

2.2.2.2 Modelo glotal de Liljencrants-Fant com parâmetros unificados

Este modelo é composto por uma parte senoidal modulada por uma exponencial ascendente (entre 0 e $O_q T_0$), seguida de uma fase de retorno exponencial decrescente (entre $O_q T_0$ e T_0). Essa versatilidade torna o modelo altamente eficiente para criar a forma de onda do pulso glotal com grande naturalidade.

$$g(t) = \begin{cases} -\dfrac{E_e e^{-aO_q T_0}}{sin(\frac{\pi}{\alpha_m})(a^2 + (\frac{\pi}{\alpha_m O_q T_0})^2)}(\dfrac{\pi}{\alpha_m O_q T_0} + ae^{at} sin(\dfrac{\pi}{\alpha_m O_q T_0} t) - \dfrac{\pi}{\alpha_m O_q T_0} e^{at} cos(\dfrac{\pi}{\alpha_m O_q T_0} t)) & , \ 0 \leq t \leq O_q T_0 \\ -E_e(\dfrac{1}{\varepsilon Q_a (1 - O_q) T_0} - 1)(T_0 - t + \dfrac{1 - e^{\varepsilon(T_0 - t)}}{\varepsilon}) & , \ O_q T_0 \leq t \leq T0\,. \end{cases} \tag{2.10}$$

A expressão temporal da Eq. 2.10 corresponde ao modelo do pulso glotal $g(t)$ de LF com parâmetros unificados, aplicável ao caso de um fechamento não abrupto ($Q_a > 0$). Neste modelo, o parâmetro Q_a está relacionado à duração efetiva da fase de retorno através de $T_a = Q_a(1 - O_q T_0)$ e é conhecido como coeficiente da fase de retorno. Essa característica é uma das diferenças entre o modelo de Rosenberg e o modelo LF.

Os outros parâmetros são definidos da seguinte forma: O_q é o quociente de abertura, que representa o instante de fechamento glótico ($t = O_q T_0$) em relação a T_0. O período fundamental é representado por T_0, e α_m é o coeficiente de assimetria, definido como o instante de máximo do fluxo glótico ($T_m = \alpha_m O_q T_0$) em relação a T_0 e O_q. O coeficiente de assimetria varia entre 0.5 e 1, refletindo a diferença entre a fase de abertura glótica, sempre maior que a fase de fechamento glótico.

Modelagem e simulações desses modelos estão descritos e desenvolvidos em [43].

2.2.3 Métodos usados para estimar GCIs da voz cantada

O instante de excitação principal do sistema do trato vocal durante a fala vocalizada é conhecido como época e ocorre próximo ao instante de fechamento glótico, causado pelo rápido fechamento das pregas vocais [31]. A identificação precisa desse instante de fechamento glótico é de grande importância devido à sua utilização em uma variedade de métodos de análise, modelagem, apresentação e síntese de fala desenvolvidos nos últimos anos. No entanto, na voz cantada, as particularidades como a alta SFI, a presença de tons variados, variações abruptas de frequência entre trechos vocalizados e a variedade de estilos de canto e técnicas, tornam os métodos tradicionais de detecção de GCIs desenvolvidos para a fala não aplicáveis [44]. Nesse contexto, são necessárias abordagens específicas para a voz cantada. Entre os métodos frequentemente empregados para estimar os GCIs na voz cantada, destacam-se:

- **SEDREAMS**: O método de *Detecção de Eventos de Fala usando Excitação Residual e Sinal Baseado na Média* (SEDREAMS) opera identificando os intervalos de curto prazo em que é provável que ocorram os fechamentos glóticos, com auxílio de um sinal baseado na média. As posições exatas dos GCIs são ajustadas ao selecionar as descontinuidades mais acentuadas no sinal residual de Predição Linear, dentro dos intervalos estimados [45].
- **ZFF**: O método de *Filtragem de Frequência Zero* (ZFF) aproveita o fato de que o efeito do fechamento glótico produz uma excitação similar a um impulso. Essa informação está presente em todas as frequências, inclusive na frequência zero. Para capturar a informação presente na região de frequência zero, o sinal é filtrado através de uma cascata de 2 ressonadores, cuja frequência central está localizada em 0 Hz. Esse processo de filtragem minimiza o efeito das ressonâncias do trato vocal [44].
- **ZFF Modificado**: O Método de ZFF modificado é uma versão aprimorada do método ZFF para a estimativa de GCIs. Em lugar de empregar um comprimento de janela fixo, a remoção de tendência na saída dos ressonadores de frequência zero é realizada por meio de janelas de curtos comprimentos, e é aplicada uma sequência de três ressonadores digitais ideais, com polos localizados em 0 Hz [31].

2.2.4 Modelagem do trato vocal

Os efeitos do trato vocal são modelados utilizando um filtro *all-pole filter*, conforme descrito em [6]. Esse modelo considera as ressonâncias (formantes) correspondentes aos polos

da função de transferência do sistema discreto do trato vocal, $V_k(z)$, como ilustrado na Eq. 2.11:

$$V_k(z) = \frac{1 - 2|z_k|\cos(2\pi F_k T) + |z_k|^2}{1 - 2|z_k|\cos(2\pi F_k T)z^{-1} + |z_k|^2 z^{-2}} \tag{2.11}$$

onde z_k representa os polos de ordem k do filtro digital, F_k é a frequência de ressonância de ordem k, e T é o período de amostragem.

Nas larguras de banda de tempo contínuo (analógico), as frequências ressonantes são cerca de $2\sigma_k$, e a frequência central é $2\pi F_k$. No plano complexo, o raio da origem ao polo determina a largura de banda, conforme a Eq. 2.12:

$$|z_k| = e^{-\sigma_k T}\,, \tag{2.12}$$

onde $|z_k| < 1$, considerando que todos os polos correspondentes à função de transferência do sistema discreto do trato vocal, $V_k(z)$, devem estar dentro do círculo unitário, garantindo, assim, a estabilidade do sistema.

Da mesma forma, o ângulo $\theta_k = \angle z_k$ está relacionado com a frequência central, conforme a Eq. 2.13:

$$\theta_k = 2\pi F_k T\,, \tag{2.13}$$

2.3 Modelagem da radiação vocal

A pressão acústica na boca (incluindo os efeitos da radiação) é modelada usando a transformada z, como na Eq. 2.14, tal como sugerido em [6].

$$P_L(z) = R(z)U_L(z)\,. \tag{2.14}$$

sendo:

$P_L(z)$: pressão nos lábios;
$R(z)$: efeitos de radiação;
$U_L(z)$: fluxo acústico nos lábios.

Uma aproximação empregada nesta tese, para modelar os efeitos da radiação vocal, é dada pela Eq. 2.15.

$$R(z) = 1 - 0.95z^{-1}\,. \tag{2.15}$$

2.4 Técnicas de síntese de voz cantada

As técnicas de síntese de voz cantada desempenham um papel crucial na criação de vozes artificiais que se assemelham à fonação humana. No geral, têm um papel vital na ampliação das possibilidades musicais, no desenvolvimento de ferramentas de treinamento vocal, na melhoria da acessibilidade da comunicação e na contínua pesquisa sobre a voz humana. Nesta seção, são revisados os principais métodos usados para sintetizar a voz cantada, os quais podem ser categorizados em quatro amplas abordagens: síntese de formantes, síntese de modelagem física, síntese por unidades concatenadas e síntese paramétrica estatística (incluindo abordagens baseadas em *Modelos Ocultos de Markov* (HMM) e RNA).

2.4.1 Síntese por formantes

Os métodos de síntese por formantes são aqueles baseados no modelo fonte-filtro, utilizando um modelo de fonte gerado geralmente como uma mistura de componentes sonoros e ruído. O sinal da fonte é filtrado por um conjunto de filtros que simulam a função de transferência do trato vocal, com ressonâncias específicas correspondentes aos formantes parametrizados de acordo com cada fonema. Muitos sistemas de síntese têm utilizado essa abordagem desde a década de 1960 até os dias de hoje. Entre eles, destacam-se sistemas de síntese nos quais o filtro é modelado tanto por uma conexão em cascata quanto por uma conexão paralela de ressonadores digitais, implementados em software [46]. Também podem ser mencionados sintetizadores de canto baseados em formantes, como o instrumento de síntese de vogais cantadas conhecido como *Cantor Digitalis* [47].

Embora essa abordagem não seja altamente versátil para situações de síntese Texto-Canto em geral, ela é adequada para a realização de síntese em tempo real. Além disso, mantém sua utilidade ao conduzir pesquisas dentro de um paradigma de análise por síntese. Isso se torna crucial para verificar hipóteses específicas que podem ser desafiadoras de validar somente por meio de análises diretas em gravações. Portanto, foi selecionada como o modelo para o sistema de síntese desenvolvido ao longo desta tese, como será detalhado no capítulo 4.

2.4.2 Síntese por modelagem física

A síntese por modelagem física, às vezes também chamada de síntese articulatória, baseia-se em modelos numéricos da produção vocal e na resolução das equações físicas subjacentes. Isso difere de modelos espectrais, como a síntese de formantes, que se concentram mais nos princípios da percepção sonora, modelando diretamente o espectro sonoro resultante da voz em vez do próprio processo de produção. Geralmente, modelos de tubo acústico são empregados para modelar o trato vocal, enquanto a fonte é derivada de um modelo de sinal paramétrico, uma tabela de formas de onda ou um modelo mecânico, como um modelo de duas massas [48, 49, 50]. Na modelagem física, uma alteração nos parâmetros tem uma relação direta com uma modificação no mecanismo de produção vocal, ao passo que em modelos espectrais, uma mudança nos parâmetros está mais relacionada a uma alteração na percepção [7].

2.4.3 Síntese por unidades concatenadas

A síntese por unidades concatenadas consiste em selecionar amostras de voz em um banco de dados pré-gravado e pré-anotado, de acordo com um texto de entrada fornecido, e concatená-las para recriar novas palavras e frases. Em termos de inteligibilidade e naturalidade, a principal vantagem dessa técnica é preservar um timbre o mais próximo possível do sinal de fala original, especialmente para variações de timbre que ocorrem naturalmente entre fonemas devido aos efeitos de coarticulação [7]. Para obter os melhores resultados na síntese de voz cantada, podem ser utilizados comprimentos de unidades não uniformes, selecionados em grandes bancos de dados que abrangem uma ampla variedade de contextos, de modo a aplicar o mínimo de processamento a essas unidades. Em tais sistemas, a duração das unidades pode variar de pares de fonemas consecutivos curtos até palavras completas, ou até mesmo grupos de frases que abrangem várias palavras [47]. O sistema deve, então, encontrar o melhor equilíbrio entre selecionar as unidades que melhor correspondem ao contexto-alvo, de acordo com o texto de entrada, e selecionar as unidades que podem ser concatenadas da melhor forma possível. Um exemplo notável desta técnica de síntese é o software comercial Vocaloid [51], que tem obtido considerável sucesso entre o público em geral ao empregar a síntese por concatenação. Isso possibilita a composição e síntese de trechos vocais ao adquirir a base de unidades de um cantor específico.

2.4.4 Síntese baseada em Modelos Ocultos de Markov (HMM)

A síntese baseada em HMM utiliza uma abordagem de fonte-filtro, semelhante à usada na síntese por formantes, mas com a diferença de que os parâmetros da fonte e do filtro são obtidos por meio de estatísticas em vez de regras predefinidas. Durante a etapa de treinamento, características espectrais e parâmetros de excitação são extraídos de um banco de dados de gravações de canto, juntamente com informações contextuais. A partir dos valores extraídos, estatísticas são geradas para cada característica. Durante a síntese, essas estatísticas são usadas para selecionar os melhores valores de cada característica de acordo com o contexto-alvo, sendo agrupados com o auxílio de árvores de decisão. Com essa abordagem, os parâmetros do filtro e da fonte são modelados simultaneamente por meio de HMMs dependentes do contexto [7, 11].

2.5 Particularidades da voz cantada

A voz falada e a voz cantada têm muitas coisas em comum; portanto, muitas técnicas desenvolvidas inicialmente para a síntese da fala foram adaptadas com sucesso para o canto. No entanto, saber falar não significa necessariamente saber cantar bem. Da mesma forma, sintetizar vozes cantadas de boa qualidade não é tão simples como adaptar um sistema de síntese de voz falada para seguir uma melodia dada por uma partitura. Portanto, é essencial considerar as particularidades da voz cantada.

Falar é um processo relativamente espontâneo em que nos concentramos principalmente nas palavras para expressar nossos pensamentos, sem pensar conscientemente na maneira como pronunciamos cada palavra. Não ponderamos sobre a entonação ou o acento de sílabas específicas. Esse aspecto é conhecido como *prosódia*, e é o resultado do conhecimento implícito que temos da língua, pelo menos no caso de nossa língua materna. Também não precisamos pensar na quantidade de pressão que aplicamos nos pulmões ou na posição da laringe. Como aprendemos a falar desde a infância, todos nós nos tornamos especialistas naturais nisso. No entanto, cantar é um processo muito menos natural, exigindo a aprendizagem de muitas técnicas para alcançar um alto nível de habilidade, o que requer anos de treinamento [7].

Em geral, o objetivo principal da fala é transmitir a mensagem e garantir a inteligibilidade das palavras pronunciadas. Em certa medida, algumas emoções ou intenções também são comunicadas, para as quais as palavras não são suficientes. A prosódia ajuda a adicionar informações, moldando o tom e a dinâmica da voz para codificar essas men-

sagens implícitas. No entanto, essa moldagem prosódica ocorre de forma implícita e bem definida, sem resultar de escolhas estéticas do falante. Por exemplo, é uma convenção levantar o tom no final de uma pergunta. Já no canto, além da inteligibilidade, o foco principal geralmente está nas qualidades estéticas da voz, que dependem principalmente do timbre e das diversas entonações produzidas. Isso requer muito mais perícia e controle por parte do cantor em comparação com a fala.

2.6 Características da voz cantada

2.6.1 Estilos de canto

O canto é uma forma universal de expressão, que apresenta características muito diversas ao longo dos países, culturas, gêneros musicais e contextos sociais. Essa diversidade é refletida nos numerosos estilos de canto encontrados no mundo que usam o aparelho vocal de maneiras muito diferentes, explorando vários timbres e expressões da voz humana [52].

Os estilos de canto mais conhecidos são: clássico (ópera ocidental e barroca), popular (música comercial contemporânea, rock, jazz, pop, etc.) e tradicional (músicas folclóricas). O estilo de canto clássico tem sido objeto de investigação nos séculos passados, desde uma perspectiva física e fisiológica e, portanto, esse conhecimento forneceu a base para entender os estilos de canto tradicionais e populares. As características do estilo de canto popular foram comparadas às de cantores clássicos, encontrando diferenças consistentes no nível da pressão sonora, comportamento glótico e controle da pressão subglótica [52].

Exemplos dos estilos de canto clássico e popular são aqui apresentados, indicando as técnicas de canto utilizadas e suas especificidades.

- **Soprano de ópera**: As cantoras cantam em tons muito altos e usam a voz de falsete, modificando o trato vocal para sintonizar suas formantes e assim maximizar a homogeneidade do timbre da voz.
- **Metal**: Os cantores de metal usam uma técnica muito específica para produzir timbres bastante ásperos, envolvendo também algumas estruturas supraglóticas como fontes de vibração, além de suas cordas vocais [7].
- **Pop**: Os cantores de pop utilizam um início soprado, um vocal frito, um corpo de frase com boa qualidade tímbrica e uma frase final onde sobressai o vibrato ou um pequeno sopro. Eles usam falsete para a emissão do registro mais alto e

fazem mudanças na qualidade do timbre progressiva ou abruptamente para fornecer climas diferentes na execução musical. Também empregam a técnica de *Belting*, caracterizada pelo uso de uma laringe elevada, o uso do registro de tórax e notas de alta intensidade. Os cantores pop não cantam com o formante do cantor, em torno de 2800 Hz, usando um pico espectral mais alto, em torno de 3200 a 3600 Hz [53].

2.6.2 Extensão vocal

A extensão vocal corresponde à gama de notas musicais que um cantor pode cantar e, dependendo do mecanismo laríngeo utilizado, a faixa de frequências das notas cantadas pode variar desde alguns poucos Hz até 2000 Hz. A média das vozes masculinas pode variar entre $R\#2$ ($f_o = 78$ Hz) e $F\#4$ ($f_o = 370$ Hz) usando o mecanismo I, e entre $Mi3$ ($f_o = 165$ Hz) e $Mi5$ ($f_o = 660$ Hz) no mecanismo II. As vozes femininas podem variar entre $R3$ ($f_o = 147$ Hz) e $Sol4$ ($f_o = 392$ Hz) em mecanismo I, e entre $Sol3$ ($f_o = 196$ Hz) e $D6$ ($f_o = 1046$ Hz) usando o mecanismo II [3, 52]. Por outro lado, a tessitura vocal não pode ser confundida com a extensão vocal, sendo a tessitura a gama de notas que um cantor pode emitir com conforto. Cantar em volumes ou em intervalos de frequência não adequados às capacidades naturais da voz não só apresenta um risco dramático ou artístico, mas também traz um risco vocal [54].

2.6.3 Registro vocal

As qualidades vocais podem ser classificadas de acordo com os registros de voz, ou seja, cada registro corresponde a uma região da extensão vocal composta por uma série de tons consecutivos produzidos com qualidade vocal semelhante [55]. Os registros de vozes mais conhecidos no canto masculino e feminino são: voz de peito/modal, voz média, de cabeça e apito.

Do ponto de vista fisiológico, a voz humana é produzida através de quatro mecanismos laríngeos (M0, M1, M2, M3), cada um associado a uma configuração diferente das pregas vocais, indo do mais grave, M0, ao mais agudo, M3. O mecanismo laríngeo M1 é usado para produzir voz modal, de tórax e de cabeça masculina. A voz feminina principal é produzida em M2, enquanto para atingir as notas mais altas na gama superior das sopranos ligeiros (registro de apito), pode-se usar M3. Em particular, a primeira transição na voz de soprano pode ocorrer em torno de $E4-F4$. Esse ponto de transição corresponde à mudança M1-M2 no mecanismo laríngeo [52, 15].

O cantor pode alcançar os tons mais agudos de seu registro de voz através de duas técnicas de canto vocal:

1. Aumentando a pressão subglótica e, consequentemente, a tensão muscular nas pregas vocais. Este mecanismo é usado por cantores populares.

2. Abaixando a laringe e dilatando a faringe inferior, gerando alongamento das pregas vocais com menor variação da pressão subglótica. Esse mecanismo, chamado de cobertura de tons agudos, é usado por cantores líricos masculinos.

2.6.4 Classificação vocal

No canto, as vozes podem ser classificadas em diferentes categorias vocais, que são determinadas pelas características físicas e fisiológicas do aparelho vocal de cada cantor. Essa classificação é fundamental para a seleção adequada de repertório e para a compreensão das particularidades de cada voz na execução musical. As vozes femininas são geralmente classificadas como: contralto, mezzo-soprano e soprano, enquanto as vozes masculinas são classificadas como: baixo, barítono e tenor [56].

- **Contralto**: A voz contralto é a voz mais grave entre as vozes femininas e apresenta um timbre rico e profundo. As contraltos têm a capacidade de alcançar notas baixas com facilidade e podem ter uma extensão vocal que varia geralmente do $D3$ ao $Sol5$.

- **Mezzo-soprano**: A mezzo-soprano possui uma extensão vocal intermediária, situando-se entre o contralto e o soprano. Essa voz é versátil e pode ser caracterizada por um timbre mais encorpado e expressivo. A extensão vocal típica de uma mezzo-soprano varia do $L3$ ao $D6$.

- **Soprano**: A voz soprano é a mais aguda entre as vozes femininas e apresenta um timbre brilhante e cristalino. As sopranos têm a habilidade de alcançar notas altas com clareza e pureza tonal. A extensão vocal de uma soprano pode variar do $D4$ ao $D6$ ou mesmo além.

- **Baixo**: O baixo é a voz mais grave entre as vozes masculinas e é caracterizado por um timbre profundo e ressonante. Os baixos têm uma extensão vocal que normalmente varia do $Mi2$ ao $Mi4$.

- **Barítono**: O barítono possui uma extensão vocal intermediária, situando-se entre o baixo e o tenor. Essa voz é versátil e pode apresentar uma sonoridade rica e encorpada. A extensão vocal típica de um barítono varia do $L2$ ao $F4$.

- **Tenor**: O tenor é a voz mais aguda entre as vozes masculinas e possui um timbre brilhante e penetrante. Os tenores têm a capacidade de alcançar notas altas com potência e expressividade. A extensão vocal de um tenor pode variar do $D3$ ao $Si5$ ou mais.

A classificação vocal é importante tanto para cantores quanto para compositores e regentes, pois permite a seleção adequada de repertório que melhor se adapte às características e habilidades de cada voz. Além disso, a classificação vocal também é útil para a formação de coros e grupos vocais, garantindo uma harmonia equilibrada entre as vozes. A Tabela. 2.1 mostra a classificação vocal dos cantores líricos.

Tabela 2.1: Classificação vocal dos cantores líricos.

Masculino		Feminino	
Categoria Vocal	Extensão	Categoria Vocal	Extensão
Baixo	$D2$ - $F4$	Contralto	$F3$ - $G5$
Barítono	$G2$ - $A4$	Mezzo-Soprano	$G3$ - $A5$
Tenor	$B2$ - $C5$	Soprano	$C4$ - $D6$ / $G6$

2.7 Técnicas de expressão vocal

2.7.1 Ajuste de formantes no canto lírico feminino

Um soprano lírico canta notas musicais cuja f_o é mais alta que a frequência do primeiro formante (F_1) nas vogais faladas. Enquanto o valor médio da f_o na fala feminina adulta está em torno de 220 Hz, as notas emitidas por uma soprano podem estar próximas a 1024 Hz ($D5$). No canto, porém, com o aumento da f_o, aumenta também o espaçamento da série harmônica do som gerado na laringe, diminuindo a concentração de harmônicos em torno dos formantes e perdendo-se os ganhos da ressonância com os formantes [8]. Nesse caso, F_1 não aumenta a amplitude de f_o, e, portanto, o som emitido é mais *fraco*.

Para produzir um ganho de audibilidade, as cantoras sopranos aumentam a abertura da boca e, com isso, aumentam a frequência F_1, fazendo-a coincidir com os valores de f_o,

permitindo que a frequência do primeiro formante reforce a amplitude de f_o [28, 57]. Essa estratégia é conhecida como ajuste de formantes e é utilizada para a produção de altos tons no canto feminino.

Essa técnica de ajuste de formantes é essencial para permitir que cantoras sopranos alcancem tons mais altos de forma mais audível e expressiva. Ao ajustar a abertura da boca e alinhar a frequência F_1 com f_o, elas conseguem otimizar a ressonância vocal e reforçar a amplitude do som fundamental, resultando em uma projeção mais clara e potente das notas musicais.

Esse aumento da frequência do primeiro formante geralmente produz um efeito adverso na inteligibilidade das vogais, particularmente das vogais /e/, /i/, /o/ e /u/, com zonas de tolerância distintas para a perda de inteligibilidade em cada uma delas. A perda de inteligibilidade produzida pelo ajuste de formantes pode ser um problema para a síntese de voz [58, 27].

Cada vogal pode ter uma zona de tolerância diferente para a perda de inteligibilidade, o que pode ser relevante tanto em apresentações ao vivo quanto em síntese de voz, onde é importante garantir a compreensão clara das palavras e das nuances da interpretação.

2.7.2 Formante do cantor

A diferença mais considerável entre as características espectrais dos fonemas vocálicos cantados por cantores e os pronunciados por não cantores está no formante do cantor. É um pico proeminente do envelope do espectro que aparece próximo de 3 kHz (entre $2,2$ kHz e $3,8$ kHz) em todos os espectros de vogais cantadas usando a técnica de cobertura por cantores masculinos e é característico de uma vogal cantada típica. O formante do cantor é gerado a partir do agrupamento dos formantes superiores (terceiro, quarto e quinto) com frequências próximas umas das outras, conforme ilustrado na Figura. 2.13.

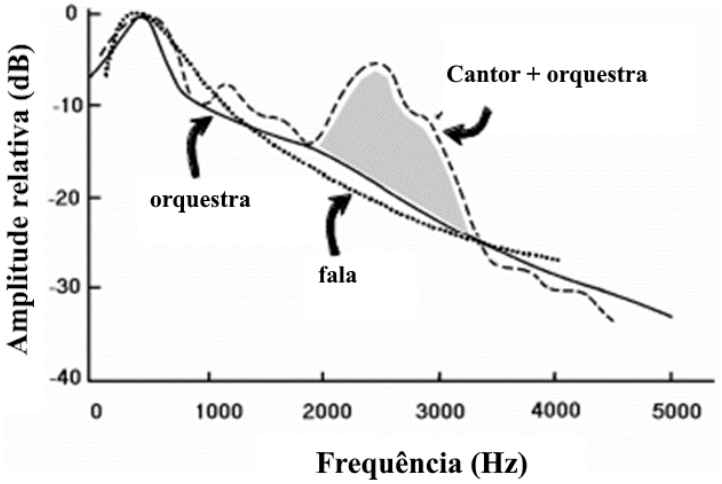


Figura 2.13: Formante do cantor [8].

Na Figura. 2.13, observam-se os espectros médios de longo prazo. Sem o recurso do formante do cantor, os sons da fala são mascarados pela orquestra. A presença do formante do cantor no espectro do som de uma vogal ajuda a voz do cantor a ser ouvida por meio de um acompanhamento orquestral alto [26]. Este fenômeno pode ser explicado considerando que a região de 3 kHz também coincide com a região mais sensível da percepção auditiva humana (onde ocorre uma ressonância com a cavidade do meato acústico do ouvido), sendo amplificada ainda mais pelo processamento auditivo periférico. Além disso, considerando o comportamento passa-altas da irradiação sonora, os harmônicos de alta frequência propagam-se mais eficientemente em linha reta, o que pode auxiliar a plateia a discriminar melhor a voz do cantor no palco [8].

A amplitude da potência espectral do formante do cantor (em dB) depende da classificação vocal. É mais baixo para a categoria de baixo e mais alto para a categoria de tenor. Em relação aos sopranos, a amplitude desse pico é bem menor do que nas demais categorias vocais, podendo até ser considerado um terceiro ou quarto formante normal. Da mesma forma, o nível desse pico também varia em função da intensidade da fonação, efeito derivado da fonte glótica.

Considerando a frequência central, o pico varia dependendo da categoria de voz. Em cantores com categoria vocal de baixo, a frequência central é de cerca de $2,2$ kHz; em barítonos, cerca de $2,7$ kHz; em tenores, cerca de $3,2$ kHz; e em sopranos, em torno de $2,8$ kHz. Essas diferenças de frequência parecem contribuir significativamente para as diferenças de timbres entre essas categorias de voz [26, 59]

2.7.3 Vibrato

O vibrato é uma flutuação periódica na frequência de uma nota comumente utilizada na música. Segundo [60], o vibrato é caracterizado por quatro parâmetros mensuráveis: taxa, extensão, regularidade e forma de onda. A taxa de vibrato (V_{Rate}) determina o número de flutuações por segundo, normalmente variando entre 4 Hz a 14 Hz. A extensão do vibrato (V_{Ext}) está associada à quantidade de variação do tom, com valores médios entre 5 ms a 10 ms. A regularidade do vibrato tende a variar mais durante a fase negativa, ou seja, durante a parte em que a frequência da nota é reduzida. A forma de onda do vibrato é aproximadamente senoidal, apresentando uma curva suave de subida e descida.

A Figura. 2.14 ilustra as características variáveis do vibrato em cantores [9]. Nessa figura, pode-se observar a diversidade no padrão do vibrato entre diferentes cantores e como ele é modulado de acordo com a expressividade vocal e o contexto musical.

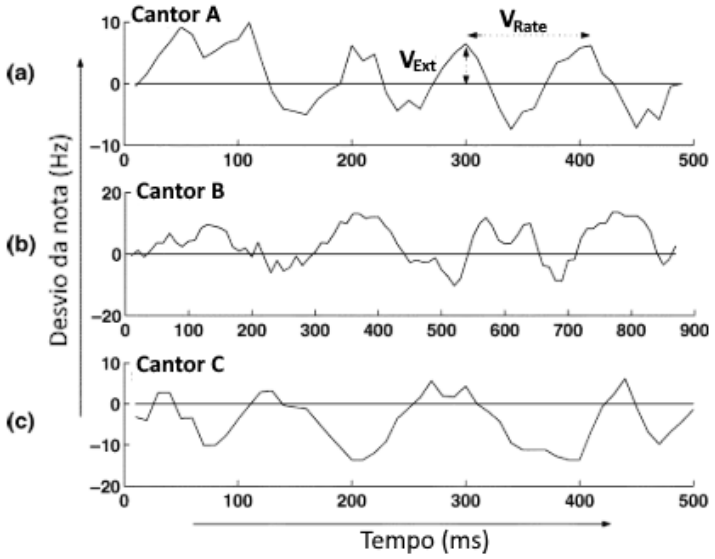


Figura 2.14: Variações nas características do vibrato em três cantores diferentes, observado na nota D6 (1174, 6 Hz), com normalização para 0 no eixo Y. Adaptado de [9].

O vibrato é uma técnica vocal e instrumental amplamente utilizada para adicionar expressividade e emoção à música. Ele pode ser controlado pelo músico para transmitir diferentes nuances de sentimentos e interpretação da música. A taxa do vibrato pode variar de acordo com o estilo musical e a preferência do músico, sendo mais rápido em alguns estilos e mais lento em outros. A extensão do vibrato está diretamente relacionada à variação da frequência da nota, e sua amplitude pode ser ajustada para criar efeitos mais sutis ou mais intensos [61].

Além disso, o vibrato é uma característica importante no desempenho vocal, pois

pode afetar a qualidade do som produzido. Um vibrato bem controlado e adequado pode melhorar a projeção e a qualidade da voz, enquanto um vibrato excessivo ou inadequado pode interferir na clareza e na precisão da afinação.

2.7.4 Trêmulo

Instrumentos que produzem vibrato, como a voz e os instrumentos de corda e sopro, caracterizam-se por uma variação periódica na intensidade do som que é conhecida como trêmulo [62]. O trêmulo ocorre em conjunto com o vibrato, que é a flutuação periódica na frequência da nota, e pode ser descrito por dois parâmetros adicionais: nível de amplificação (A_{Level}) e taxa de trêmulo (T_{Rate}) [9]. Na Figura. 2.15 observa-se o efeito do trêmulo em um sinal de áudio.

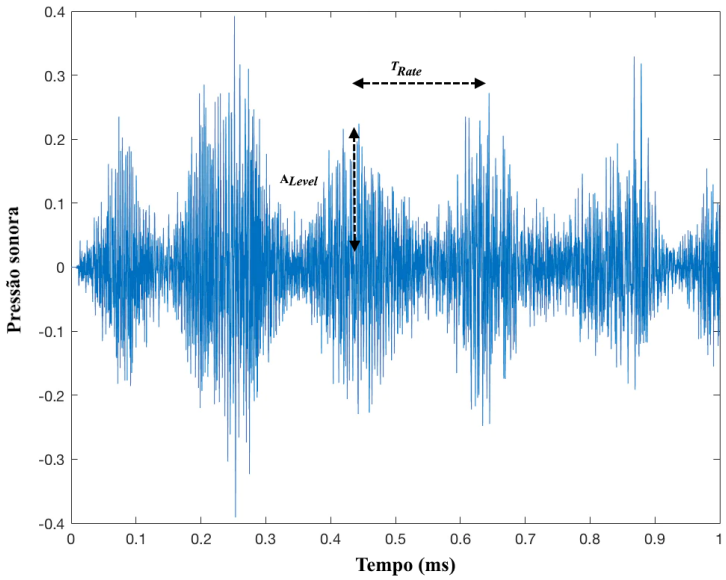


Figura 2.15: O efeito de trêmulo caracterizado pela modulação de amplitude.

O trêmulo é uma variação na amplitude do som produzido pelo instrumento ou pela voz, ocorrendo em sincronia com o vibrato. Enquanto o vibrato afeta a frequência da nota, o trêmulo modula a intensidade do som, criando uma qualidade mais expressiva e emotiva na música. Essa técnica é frequentemente utilizada por cantores e músicos para adicionar emoção e profundidade à execução musical.

Os parâmetros do trêmulo são essenciais para definir o caráter e a intensidade do efeito na música. O nível de amplificação (ou amplitude) refere-se à profundidade da variação na intensidade do som, ou seja, o quanto a intensidade do som é modulada durante o trêmulo. Já a taxa de trêmulo determina a frequência com que essa variação ocorre, ou

seja, quantas flutuações por segundo estão presentes na intensidade do som.

Essa técnica é comum em diferentes gêneros musicais e pode variar de acordo com o estilo de cada interpretação musical. Em alguns casos, o trêmulo é usado de forma mais sutil para adicionar nuances de expressão, enquanto em outras ocasiões, pode ser aplicado de maneira mais acentuada para criar efeitos dramáticos ou emocionais específicos

2.8 Controle de expressão

2.8.1 Frequência fundamental (f_o)

A frequência fundamental (f_o) destaca-se como o parâmetro de controle mais crucial, definindo a identidade da melodia e possibilitando o reconhecimento de uma música, sendo mais influente do que o ritmo ou outros aspectos. Além de sua função melódica, as variações da f_o transmitem estilo musical, expressividade pessoal e características específicas da produção vocal. Portanto, a modelagem precisa da f_o é essencial para uma síntese que soe natural e expressiva, merecendo prioridade máxima. Algumas dessas flutuações surgem de comportamentos não controlados relacionados ao mecanismo vocal e à articulação, independentemente da habilidade do cantor, o que contribui para a naturalidade da voz. Outros tipos de flutuações são deliberados e utilizados como recursos expressivos para interpretar melodias em relação com o estilo de canto e as qualidades estéticas da voz. Isso exige um certo nível de habilidade do cantor para manipulá-las com eficácia.

2.8.2 Jitter

As oscilações das cordas vocais não são exatamente periódicas, e os pulsos glóticos que compõem o sinal glótico apresentam variações na duração de tempo. Essa pequena flutuação aleatória em cada comprimento do ciclo glótico é conhecida como *jitter*, e seu estudo é de grande importância em diferentes áreas relacionadas à geração da voz. O *jitter* é uma medida utilizada para quantificar a irregularidade na periodicidade da vibração das cordas vocais durante a produção da voz.

Mesmo ao se esforçar para cantar uma nota sustentada muito estável, não é possível para um ser humano manter um tom perfeitamente constante durante o canto. O *jitter*, é um dos fatores intrínsecos à voz cantada que contribui para a sua expressividade e singularidade.

O *jitter* é comumente utilizado como uma medida de estabilidade da fonte glótica.

Valores típicos do *jitter* estão entre 0.1 % e 1 % do período fundamental para vozes normais, ou seja, sem a presença de patologias. Quando a flutuação ultrapassa, em média, o valor de 1 %, isso pode indicar a presença de alguma patologia relacionada à voz, como nódulos nas cordas vocais, pólipos ou outras alterações estruturais [50].

A análise do *jitter* é uma ferramenta valiosa em diferentes áreas relacionadas à voz e à fala. Por exemplo, em síntese de voz, modelos de *jitter* podem ser utilizados para melhorar a naturalidade da voz gerada, tornando-a mais próxima do comportamento real das cordas vocais. Em estudos de distúrbios da voz, o *jitter* pode ser utilizado para simular vozes roucas ou para gerar sinais de voz com diferentes características patológicas, o que auxilia na compreensão e no diagnóstico desses distúrbios [43, 49].

Além disso, o *jitter* é relevante para o desenvolvimento e a calibração de algoritmos de processamento de sinal utilizados em análises de voz e fala. Ele também pode ser útil na detecção de ciclos glóticos e na avaliação da qualidade vocal em diferentes situações clínicas e terapêuticas.

2.8.3 Tempo

O Tempo é um elemento crucial na interpretação vocal, especialmente na voz cantada. Refere-se ao controle das dimensões rítmicas do desempenho, incluindo o início e término de cada nota da melodia. No contexto da voz cantada, o tempo abrange a precisão temporal das notas em relação às durações nominais da partitura, a sincronização dos fonemas com as notas e a relação entre ambos, denominada alinhamento temporal. Embora o ritmo seja mais evidente em instrumentos como piano, na voz cantada há complexidades adicionais devido à presença das letras e sílabas. As variações rítmicas naturais e expressivas são comuns na interpretação vocal, resultando em desvios do ritmo da partitura [7].

Capítulo 3

Estimação das Frequências Formantes da Voz Cantada

A análise acústica da voz cantada tem sido um campo de pesquisa fascinante, permitindo uma compreensão mais profunda das características únicas desse modo de expressão vocal. Especificamente, as sopranos líricas, com suas faixas de frequências mais altas, apresentam desafios particulares para a estimativa precisa das frequências formantes. Este capítulo apresenta uma contribuição significativa nesse domínio, revelando dois métodos inovadores desenvolvidos para estimar essas frequências em vogais cantadas por sopranos líricas.

O primeiro método, conhecido como WLP-AME-ADP, emerge como uma solução pioneira para estimar com precisão as frequências formantes da voz cantada, no caso de vogais. Os métodos convencionais de estimativa das frequências dos formantes enfrentam dificuldades em capturar com precisão as características das frequências formantes na voz cantada. Esse desafio é especialmente notável na faixa de frequências mais altas das sopranos líricas, onde os formantes mais baixos podem ser influenciados pelos harmônicos da frequência fundamental. Por outro lado, o método proposto aborda de maneira inovadora a SFI inerente à voz cantada. Ao combinar o algoritmo de WLP com a função de Excitação Principal Atenuada AME, esse método busca contornar as limitações das técnicas existentes. A adaptação da *Predição Linear Ponderada com Excitação Principal Atenuada* (WLP-AME) para a voz cantada é vital, uma vez que a voz cantada apresenta nuances distintas em relação à fala, incluindo maior variação tonal e ressonância. Este método é meticulosamente projetado para considerar esses fatores, permitindo uma estimativa mais precisa das frequências formantes em faixas de frequências mais agudas. Uma das características cruciais do WLP-AME-ADP é sua abordagem inovadora para lidar com a SFI inerente à voz cantada ao estimar o instante de excitação principal do sinal de voz cantado através dos métodos criados especificamente para estimar parâmetros

glotais da voz cantada em altas frequências.

O segundo método, chamado de WLP-HPSV, é especialmente projetado para abordar a complexa natureza dinâmica da resposta do trato vocal, com foco nas vozes de cantores, visando uma estimativa precisa das frequências formantes em vogais cantadas por sopranos líricos. Essa abordagem inovadora expande o WLP para enfrentar os desafios únicos apresentados pelas vozes de sopranos em suas faixas de frequências mais altas. Isso inclui a notável influência da SFI, a degradação da estimativa de frequências formantes devido às harmônicas da frequência fundamental e a ampla variação tonal. O WLP-HPSV adota uma estratégia focada na análise da fase fechada do ciclo glotal, onde a resposta do trato vocal é mais consistente. Isso é realizado através de uma análise detalhada do sinal glotal, desde a estimativa das fases aberta e fechada do ciclo glotal. Dessa forma, obtém-se uma estimativa mais precisa das frequências formantes, especialmente nas faixas de frequências mais altas, onde as mudanças na resposta vocal são mais relevantes. O WLP-HPSV apresenta uma abordagem inovadora para lidar com as particularidades das vozes de sopranos líricos, permitindo uma estimativa mais precisa das frequências formantes em suas faixas de frequências mais agudas.

As adaptações dos métodos WLP-AME-ADP e WLP-HPSV representam contribuições notáveis para o campo da análise acústica da voz cantada. A habilidade de capturar os parâmetros de ressonância do trato vocal das vogais cantadas por sopranos líricos em suas faixas de frequências mais altas é crucial para uma compreensão mais profunda dessa forma de expressão vocal. Esses métodos não somente abrem novas perspectivas na pesquisa da voz cantada, mas também têm o potencial de impactar positivamente o treinamento vocal e o desempenho dos cantores, fornecendo ferramentas precisas para avaliar e aprimorar o desempenho vocal.

3.1 Codificação Preditiva Linear (LPC)

A análise *Codificação Preditiva Linear* (LPC) é um método comumente utilizado para estimar as frequências formantes de um sinal de voz. Ele expressa cada amostra de fala como uma combinação linear de p amostras passadas, de forma a minimizar a soma dos quadrados das diferenças entre as amostras atuais e as amostras passadas. O modelo LPC convencional é representado através de uma simples equação de diferença.

$$s_n = \sum_{k=1}^{p} a_k s_{n-k} + e_n, \tag{3.1}$$

onde s_n representa a n-ésima amostra do sinal de voz, e_n é a n-ésima amostra do residual de predição, a_k é o k-ésimo coeficiente de predição e p é a ordem de predição.

Um dos principais desafios na análise LPC é determinar um conjunto único de coeficientes preditores que caracterizam um modelo *all-pole filter* do sistema do trato vocal no modelo de tempo discreto para a produção de fala. Os métodos de autocorrelação e covariância são as formulações mais amplamente utilizadas na análise LPC para obter coeficientes preditores. Devido à natureza variável no tempo dos sinais de fala, os coeficientes preditores devem ser estimados usando uma abordagem de análise de curto prazo, que identifica os coeficientes que minimizam o erro médio quadrático de previsão em um curto segmento da forma de onda da fala [6].

Embora a análise LPC tenha vários benefícios, nem sempre estima com precisão as frequências formantes, especialmente no caso da voz cantada e na fala, com altas frequências [63, 64, 65]. A rápida flutuação das pregas vocais na voz de alta frequência faz com que a modelagem *all-pole filter* se concentre demais na excitação em vez do filtro do trato vocal. Consequentemente, os modelos *all-pole filter* calculados pela análise de LPC são afetados pela fonte glotal, resultando em baixa precisão na estimativa dos formantes. Esse problema pode ser explicado pelas características do domínio de frequência de vozes agudas. A estrutura harmônica dessas vozes é dominada por energia concentrada na f_o e seus múltiplos inteiros mais baixos. Como essas regiões de alta energia se beneficiam do critério de erro de mínimos quadrados, a estimativa dos formantes mais baixos é influenciada pelos harmônicos de frequência. Além disso, a degradação da análise LPC na estimativa dos formantes de vozes de alta frequência pode ser atribuída ao fenômeno de *aliasing*, que ocorre no método de autocorrelação [66, 64, 67]. Alternativamente, podemos usar a Predição Linear Ponderada, discutida a seguir.

3.2 Predição Linear Ponderada (WLP - weighting linear prediction)

A análise WLP é um método baseado em LPC introduzido por Ma et al. [66]. Ele visa calcular modelos *all-pole filter* que são menos afetados por f_o e seus harmônicos. Este método envolve ponderar temporalmente o quadrado do residual na função custo:

$$E = \sum_{n=n_1}^{n_2} e_n^2 \cdot W_n = \sum_{n=n_1}^{n_2} \left(s_n - \sum_{k=1}^{p} a_k s_{n-k} \right)^2 \cdot W_n, \tag{3.2}$$

onde W_n representa a função de ponderação temporal, s_n corresponde ao sinal de fala que está sendo modelado, e $a_k, (1 \leq k \leq p)$ denota os coeficientes preditores. O objetivo é minimizar a função de custo dentro do intervalo especificado definido por n_1 e n_2. No caso do método de autocorrelação, $n_1 = 1$ e $n_2 = N + p$, assumindo que o sinal de fala é considerado zero fora do intervalo $[1, N]$, onde N representa o comprimento do quadro (em amostras) [68]. Os coeficientes do filtro WLP são obtidos minimizando a função de custo ($\partial E/\partial a_i = 0$, $1 \leq i \leq p$) e resolvendo as equações normais resultantes:

$$\sum_{k=1}^{p} a_k \sum_{n=n_1}^{n_2} W_n \cdot s_{n-k} s_{n-i} = \sum_{n=n_1}^{n_2} W_n \cdot s_n s_{n-i}, \quad 1 \leq i \leq p. \tag{3.3}$$

As equações lineares obtidas a partir da Eq. 3.3 são representadas na forma matricial como:

$$\left(\sum_{n=n_1}^{n_2} W_n \cdot \mathbf{s_n} \mathbf{s_n}^T \right) \hat{\mathbf{a}} = \sum_{n=n_1}^{n_2} W_n \cdot s_n \mathbf{s_n}, \tag{3.4}$$

onde T representa a transposição da matriz, o vetor de coeficientes estimados é dado por $\hat{\mathbf{a}} = [\hat{a}1, \hat{a}2, ..., \hat{a}p]^T$, e $\mathbf{sn} = [s_{n-1}, s_{n-2}, ..., s_{n-p}]^T$. É desejável que a diferença entre o vetor de coeficientes estimados $\hat{\mathbf{a}}$ e o vetor de coeficientes verdadeiros $\mathbf{a}$ seja um valor pequeno ou zero. Portanto, a ideia é escolher um W_n apropriado para fazer com que $(\hat{\mathbf{a}} - \mathbf{a})$ diminua.

3.3 Função de Ponderação Temporal

Nas últimas décadas, foram propostas várias tentativas na análise WLP para minimizar os efeitos de tendência da f_o e estimar, com precisão, as frequências dos formantes a partir de sinais de fala aguda. Essas abordagens visam melhorar a estimativa dos parâmetros, selecionando ou ponderando amostras de fala caracterizadas como livres de excitação. Por exemplo, a análise WLP para a fala vozeada, conforme estudada em [66], introduz uma função de ponderação temporal que multiplica o quadrado do resíduo de predição. A função de ponderação é obtida aplicando um limiar na função de *Short-Time Energy* (STE) da sequência de amostras de fala pré-ênfase com atrasos específicos. Este método pondera seletivamente as amostras de fala que se ajustam ao modelo, enquanto diminui o peso daquelas que não se ajustam. No entanto, apesar de representar uma melhoria em relação ao LPC, essa abordagem ainda pode resultar em estimativas imprecisas dos

formantes, o que pode afetar a qualidade geral do modelo WLP.

3.3.1 Função de Excitação Principal Atenuada (AME)

O estudo realizado por Alku et al. [64] desenvolveu uma nova função de ponderação, chamada de função AME, que diminui de forma mais eficaz a contribuição da excitação principal em comparação à função de ponderação STE introduzida por Ma et al. [66]. Isso resulta em modelos mais precisos de WLP da função de transferência do trato vocal. A função AME tem como objetivo reduzir o peso do quadrado do resíduo de predição em várias amostras de tempo localizadas nas proximidades da excitação principal do trato vocal. Essas excitações principais podem ser identificadas usando tanto a *Eletroglotografia* (EGG) [69] quanto técnicas de extração de épocas desenvolvidas para estimar diretamente os GCIs a partir dos sinais de fala acústicos [70]. No estudo deles, Alku et al. calcularam os GCIs usando a EGG diferenciada, quando disponível. No entanto, se apenas o sinal de pressão da fala estiver disponível, a análise WLP-AME deve ser combinada com um método de extração de épocas, como o *Dynamic Programming Projected Phase-Slope Algorithm* (DYPSA), que pode estimar os GCIs a partir do sinal de fala acústico.

A função AME, desenvolvida para um ciclo fundamental, é ilustrada na Figura 3.1. O período fundamental da janela de ponderação AME é denotado por N_0, N_{me} representa o GCI, e o parâmetro de amplitude d determina o nível de atenuação. N_1 e N_2 indicam a duração da seção atenuada da janela e a posição da excitação principal do trato vocal em relação à seção atenuada, respectivamente. O quociente de duração é definido por $DQ = (N_1/N_0) \times 100\%$, e o quociente de posição é definido por $PQ = (N_2/N_1) \times 100\%$.

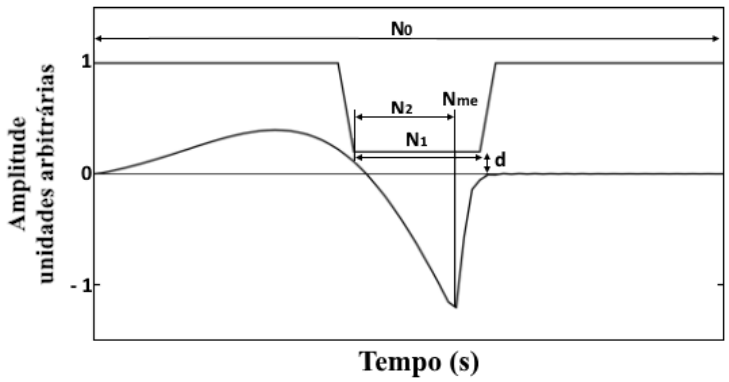


Figura 3.1: Forma de onda da função AME juntamente (na parte inferior) com um fluxo glótico LF diferenciado.

3.3.2 Função de Fase Quase Fechada (QCP)

A função *Fase Quase Fechada* (QCP) é desenvolvida considerando a natureza dinâmica da *Resposta do Trato Vocal* (RTV), que representa a resposta em frequência da cavidade acústica do cantor formada pelas cavidades subglótica e supraglótica. Reconhece-se que a RTV não é invariante no tempo dentro de um único período fundamental, uma vez que seus parâmetros característicos, como frequência de formante e largura de banda, variam ao longo do tempo. Essas variações correspondem à OP e à CP do ciclo glótico, que são controladas pela função de área glótica denotada como $A[n]$ [71]. Durante a CP, a área glótica é zero porque as pregas vocais estão fechadas, resultando em um acoplamento mínimo da excitação ao sinal de canto. Em contraste, durante a OP, as pregas vocais aplicam um efeito de compressão variável no tempo ao sinal de canto, resultando em uma área glótica diferente de zero e um sinal glótico não nulo. Portanto, durante a CP, a RTV permanece invariável, como indicado pelas expressões 3.5 e 3.6.

$$F_c(n) = F_{co}\sqrt{|1 + \alpha A'[n]|}, \tag{3.5}$$

$$BW(n) = BW_o(1 + \beta A[n]), \tag{3.6}$$

onde α e β são constantes, F_{co} e BW_o são os parâmetros constantes da formante durante CP e $A'[n]$ indica a derivada da função da área glótica.

Consequentemente, analisar o envelope espectral do sinal durante a CP é esperado para fornecer um modelo do trato vocal que minimize a contribuição do sinal glótico. Esse conceito forma a base do método QCP utilizado na análise de *Filtragem Inversa da Glote* (GIF), bem como na detecção e estimativa precisa dos formantes da fala de alta frequência [68, 72].

A função QCP, ilustrada na Figura 3.2, é definida por vários parâmetros. N_0 representa o período fundamental, N_{me} indica o GCI, N_{ramp} determina o comprimento da rampa linear, e o parâmetro de amplitude d controla o nível de atenuação. A função QCP incorpora uma transição suave entre seu valor máximo de 1.0 e o valor mínimo de d usando uma rampa linear. A duração dessa rampa pode ser ajustada com base no Quociente de Rampa Proporcional à f_o do tom (RQ), definido como $RQ = (N_{ramp}/N_0) \times 100\%$. Além disso, a duração da seção não atenuada de W_n é representada por N_1, enquanto N_2 indica a posição inicial da seção não atenuada de W_n a partir do GCI anterior. O quociente de

duração é calculado como $DQ = (N_1/N_0) \times 100\%$, e o quociente de posição é definido como $PQ = (N_2/N_0) \times 100\%$.

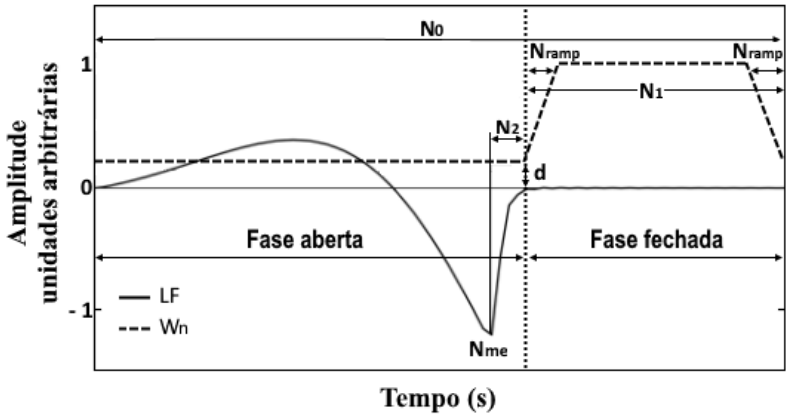


Figura 3.2: Forma de onda da função Quasi Closed Phase (QCP) juntamente com (abaixo) um fluxo glótico diferenciado sintetizado com o modelo de Liljencrants-Fant.

O propósito dos parâmetros DQ e PQ na construção da função QCP difere do seu uso em [64]. Na abordagem WLP-AME, esses parâmetros são usados exclusivamente para reduzir o peso na região em torno do GCI. Nesse contexto, N_1 representa a duração da seção atenuada de W_n, enquanto N_2 indica a posição da excitação principal do trato vocal em relação à seção atenuada.

W_n é igual a 1.0 para todos os instantes de tempo na seção não atenuada (N_1), enquanto W_n é igual a uma pequena constante positiva ($d = 0, d > 0$) na seção atenuada. Conforme indicado em [72], o valor de d utilizado neste estudo é 10^{-5}. O valor não nulo de d é escolhido para evitar problemas relacionados à singularidade da matriz de autocorrelação ponderada [68]. Além disso, é aplicada uma rampa linear com um RQ de 10% para garantir uma transição suave da função QCP.

3.4 Predição Linear Ponderada com Excitação Principal Atenuada Adaptado ao Canto Agudo (WLP-AME-ADP)

Essa seção destaca uma das principais contribuições dessa tese. O método que será descrito aqui é inédito com a adaptação à voz cantada, em particular, em altas frequências, como as sopranos.

O método WLP-AME consiste na combinação do algoritmo WLP com a função AME,

que reduz a ênfase do quadrado do resíduo de predição em pontos próximos à excitação principal do trato vocal [64]. Essa modificação foi motivada pela constatação de que a função STE por si só não conseguiu diminuir adequadamente a contribuição da excitação glótica, resultando em estimativas enviesadas dos formantes calculados pelo WLP [65]. A introdução da função AME como alternativa à função STE melhorou significativamente o desempenho do WLP na estimativa das frequências dos formantes.

Os métodos baseados em WLP, especialmente o WLP-AME, têm demonstrado avanços notáveis na estimativa das frequências dos formantes para fala de alta frequência. No entanto, quando aplicados à voz cantada, especialmente às vozes soprano, sua eficácia é limitada. Isso ocorre porque a voz cantada exibe um maior grau de SFI em comparação com a fala, abrangendo uma ampla gama de tons, variações controladas de tom, durações de frases variáveis, prosódia e uma maior amplitude dinâmica. Essas características singulares da voz cantada apresentam desafios únicos que requerem considerações específicas [30].

Uma das principais contribuições desta tese é a adaptação da análise WLP-AME, originalmente proposta para estimar frequências dos formantes em fala de alta frequência, para a estimação das frequências dos formantes das vogais cantadas por sopranos em suas faixas de frequências mais agudas, mantendo um desempenho robusto e, dando origem ao método WLP-AME-ADP. Essa adaptação inclui a utilização de métodos específicos para a estimativa do GCI, levando em conta as diferenças entre os sinais de fala e canto, com destaque para o aumento do impacto da interação fonte-filtro nos sinais de canto. Os resultados demonstram que o método adaptado proporciona estimativas precisas e consistentes das frequências dos formantes, reforçando a utilidade dessa abordagem para a análise de vozes cantadas. Esse avanço é um passo importante para uma melhor compreensão da acústica complexa do canto, e suas descobertas podem ter implicações significativas para o aprimoramento do treinamento vocal e do desempenho dos cantores.

A adaptação do método WLP-AME foi realizado por meio de um procedimento em etapas, apresentado no diagrama de blocos da Figura 3.3.

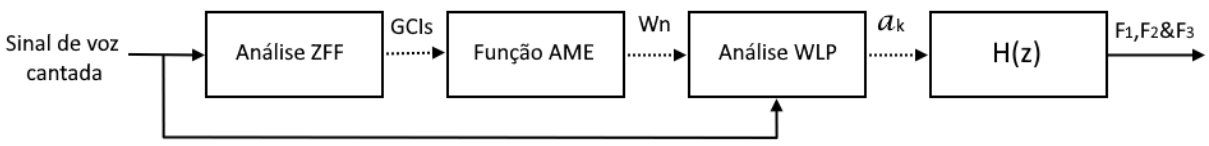


Figura 3.3: Diagrama de blocos do método WLP-AME-ADP.

Primeiramente, o algoritmo WLP-AME-ADP lê uma amostra de sinal de uma vogal

cantada (sustentada) por uma soprano de língua espanhola em seu registro vocal mais agudo a partir de uma base de vozes. Em seguida, devido à ausência de EGG neste estudo, os GCIs são estimados com precisão utilizando a versão modificada do método ZFF proposto por [31], como descrito na subseção 2.2.3. Algoritmos convencionais de extração de épocas comumente utilizados na análise de fala, conforme apresentados em [70] e [73], nem sempre são adequados para capturar os GCIs em voz cantada. Alguns desses algoritmos foram considerados impraticáveis para uso com voz cantada [44]. Portanto, o método ZFF modificado foi empregado para resolver essa questão. É importante destacar que esta é uma das novidades deste estudo, a utilização do método ZFF modificado em conjunto com o algoritmo WLP-AME.

Em uma terceira etapa, o método WLP-AME-ADP realiza a análise para a amostra da vogal cantada, variando os parâmetros da função AME, conforme ilustrado na Figura 3.1. O valor de d é escolhido como uma constante positiva pequena e não nula, no caso, 0.01, para prevenir problemas associados à singularidade da matriz de autocorrelação ponderada [64]. Além disso, os parâmetros DQ e PQ são variados em múltiplos valores para explorar diferentes configurações. Para o parâmetro DQ são utilizados oito valores distintos (10 %, 20 %, 30 %, 40 %, 50 %, 60 %, 70 % e 80 %), permitindo avaliar o efeito de desenfatizar as amostras próximas da excitação principal com diferentes comprimentos da seção atenuada. Já o parâmetro PQ é variado de 0 % a 100 % em incrementos de 10 %, visando examinar a influência da posição da excitação principal em relação à seção atenuada da função AME. Com essa abordagem, são realizadas 88 análises para a amostra da vogal cantada no banco de vozes, abrangendo uma ampla gama de cenários e configurações possíveis da função AME. Essa abordagem minuciosa permite obter *insights* significativos sobre a estimativa das frequências dos formantes em vozes cantadas, bem como a seleção dos melhores parâmetros para um desempenho confiável do método.

Depois, na quarta etapa, o modelo do sistema do trato vocal $H(z)$ baseado em WLP é estimado a partir da amostra da vogal cantada, seguindo a Eq.3.3, utilizando a função AME, W_n. Antes de realizar a estimativa, a amostra da vogal cantada é pré-enfatizada utilizando um filtro *Finite Impulse Response* (FIR) ($p(z) = 1 - 0.98\, z^{-1}$) com o objetivo de realçar as altas frequências do sinal de voz, facilitando sua identificação e estimativa. Por fim, as frequências dos formantes são calculadas resolvendo as raízes de $H(z)$, expresso como:

$$H(z) = \frac{1}{1 - \sum_{k=1}^{p} a_k z^{-k}} \tag{3.7}$$

onde p representa a ordem do modelo WLP e a_k são os coeficientes do modelo obtidos por meio da resolução das equações normais. As frequências dos formantes são, então, extraídas das raízes do polinômio $(1+\sum_{k=1}^{p} a_k z^{-k})$, proporcionando informações essenciais sobre a ressonância do trato vocal durante a emissão da vogal cantada.

3.5 Predição Linear Ponderada adaptada para as Vozes de Canto de Alto Tom (WLP-HPSV)

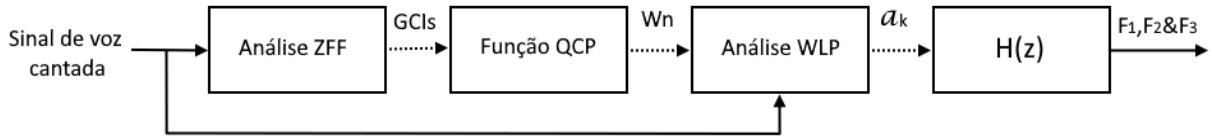


Figura 3.4: Diagrama de blocos do método WLP-HPSV.

Nesta seção, é apresentado o método WLP-HPSV, proposto para estimar as frequências dos formantes na voz cantada, com foco específico nas vogais cantadas por sopranos em seu registro vocal mais agudo, o que constitui uma das principais novidades desta tese. O método WLP-HPSV é fundamentado na análise WLP, originalmente desenvolvida para a estimativa precisa das frequências formantes na fala (mais detalhes na Seção 3.2). No entanto, é amplamente reconhecido que os algoritmos existentes para a estimativa de frequências formantes da fala sofrem de degradação de desempenho quando aplicados à voz cantada, principalmente devido ao aumento da SFI e aos efeitos produzidos pela ampla gama de tons inerente ao canto [31, 29]. Para enfrentar esse desafio, o método WLP-HPSV incorpora um procedimento específico que caracteriza esses dois aspectos-chave presentes na voz cantada. Esse procedimento é ilustrado como um diagrama de bloco na Figura. 3.4 e inclui as seguintes etapas:

Primeiramente, um sinal de voz cantada é extraído de um conjunto de dados contendo amostras de vogais cantadas por uma soprano hispanoparlante em sua tessitura vocal mais alta. Em segundo lugar, é utilizada a versão modificada do método ZFF proposto por Kadiri et al. [31] para estimar com precisão os GCIs a partir da amostra de vogal cantada. O método ZFF modificado envolve a extração de épocas ao passar segmentos curtos do sinal de canto por três ressonadores de frequência zero. Os GCIs são estimados levando em consideração o caso de alta SFI nos sinais. Em terceiro lugar, os GCIs são usados para calcular o período fundamental e construir a função de ponderação, conforme explicado na subseção 3.3.2.

Em quarto lugar, realiza-se a estimativa do modelo do sistema do trato vocal $H(z)$ baseado em WLP a partir da amostra da vogal cantada, seguindo a equação 3.3, utilizando a função QCP, W_n. Assim como no método WLP-AME, antes da estimativa, a amostra da vogal cantada passa por um processo de pré-enfase utilizando um filtro FIR ($p(z) = 1 - 0.98\,z^{-1}$) para realçar as altas frequências do sinal de voz e facilitar sua identificação e estimativa. Em seguida, a análise WLP é desenvolvida. Uma vez estimado o modelo do trato vocal, as frequências dos formantes são calculadas resolvendo as raízes de $H(z)$, fornecendo informações essenciais sobre a ressonância do trato vocal durante a emissão da vogal cantada.

Capítulo 4

Síntese da Voz Cantada Por Sopranos

A *Síntese de Voz Cantada* (SVC) é uma técnica avançada que permite a produção de vozes de canto a partir de partituras musicais e letras. Ao longo dos anos, a literatura tem apresentado diversos sistemas de SVC, cada um com suas próprias abordagens e modelos acústicos. Esses avanços têm contribuído significativamente para a melhoria da qualidade e da naturalidade das vozes cantadas sintetizadas.

Este capítulo representa uma importante contribuição desta tese, pois apresenta o desenvolvimento do um sintetizador de voz cantada, nomeado de *SOPRA-SYNTH*. O foco do *SOPRA-SYNTH* é produzir vogais cantadas no registro agudo específico de sopranos líricos hispanoparlantes.

O *SOPRA-SYNTH* é baseado no método de síntese por formantes, o que permite produzir sinais de voz cantada com características precisas, tais como a frequência, o timbre e a duração das vogais. Através dessa abordagem, buscamos capturar a essência e a autenticidade do canto lírico, proporcionando resultados mais naturais e expressivos.

Ao longo deste capítulo, se apresenta em detalhe o sistema *SOPRA-SYNTH*, descrevendo suas etapas de desenvolvimento e explicando como utilizamos o modelo Fonte-Filtro (Fant) [37] tanto de forma não interativa quanto interativa. O objetivo é criar sínteses de voz cantada que preservem a naturalidade, expressividade e inteligibilidade, características essenciais para as vogais cantadas por sopranos líricos hispanoparlantes.

Com o *SOPRA-SYNTH*, se espera contribuir para o avanço da síntese de voz cantada por sopranos e para uma maior compreensão das características únicas e desafios enfrentados ao reproduzir a bela e emotiva arte do canto lírico.

4.1 Síntese usando o modelo Fonte-Filtro não interativo

O primeiro modelo de sintetizador desenvolvido neste trabalho é baseado no modelo fonte-filtro, apresentado na seção 2.2. Esse modelo, proposto por Fant [37], é amplamente utilizado para análise da produção de voz, embora seja mais simplificado. Apesar das diferenças nas descrições matemáticas do processo de produção da voz, todas elas caracterizam, de uma forma ou de outra, duas características principais: a fonte glotal, que representa o fluxo de ar através da glote, e a resposta do trato vocal, que corresponde à resposta de frequência da cavidade acústica composta pelas cavidades subglótica e supraglótica. O modelo fonte-filtro não interativo consiste em uma forma de onda representando o sinal glotal e uma função de transferência representando a resposta do trato vocal, o que facilita a obtenção dos parâmetros do modelo a partir de um sinal acústico.

Esse sintetizador é capaz de produzir vogais cantadas por sopranos líricos a partir de parâmetros glotais e de ressonância do trato vocal das cantoras, estimados utilizando algoritmos baseados em modelos matemáticos de estimação de parâmetros glotais, como o método ZFF modificado [31], e os métodos de estimação de frequências formantes, WLP-AME e WLP-HPSV, descritos nas subseções 3.4 e 3.5, respectivamente, que também são novidades desta tese. Esses parâmetros são estimados a partir de uma base de dados que contém vogais cantadas por sopranos líricos em seus registros mais agudos.

A Figura. 4.1 mostra os blocos básicos de construção de sistema de sínteses, ou seja, a criação da base de dados com vozes das cantoras, o módulo de criação da fonte glotal e identificação de parâmetros do modelo pelo método ZFF modificado, o módulo de estimação de parâmetros de ressonância do trato vocal das cantoras e o módulo de síntese que gera o som.

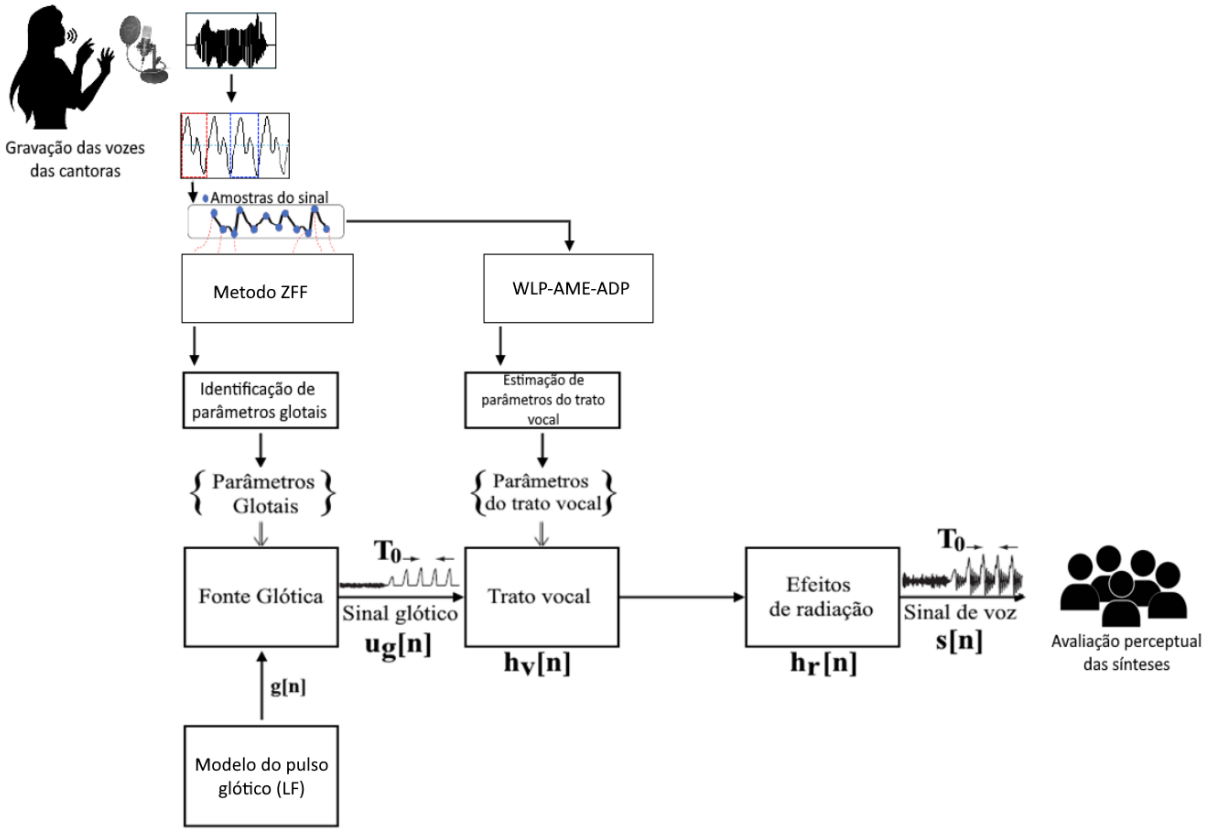


Figura 4.1: Blocos de construção do sistema de síntese de voz cantada por sopranos líricos usando o modelo Fonte-Filtro não interativo.

Os parâmetros glotais estimados (GCI, GOI, f_o) são utilizados para criar o sinal glotal. O sinal glotal é gerado utilizando modelos de pulso glotal como o de Rosenberg e de LF, descritos na subseção 2.2.1. O modelo LF foi especificamente escolhido devido à sua maior aproximação ao sinal glotal real, reproduzindo o comportamento glotal observado no registro agudo das sopranos [15]. Para isso, parâmetros glóticos específicos foram empregados, incluindo o quociente aberto O_q, coeficiente de assimetria, α_m e coeficiente de fase de retorno Q_a, sendo o valor de O_q baseado em medidas obtidas do comportamento glótico no registro agudo das sopranos líricos, conforme investigado em [15]. Na fonte de excitação, são adicionados efeitos acústicos (vibrato e trêmulo) e de perturbação natural (*jitter*) característicos das sopranos líricos.

As frequências formantes e as larguras de faixa estimadas das formantes em estudo são utilizadas para replicar o comportamento de ressonância do trato vocal das sopranos, empregando a equação 2.11. Além disso, os efeitos de radiação nos lábios das sopranos são modelados através da equação 2.14, apresentada na seção 2.3. Com a conclusão dessas etapas, o sintetizador, denominado *SOPRA-SYNTH*, torna-se capaz de produzir sínteses

de voz cantada com características autênticas e expressivas, refletindo a essência do canto lírico das sopranos hispanoparlantes.

4.2 Síntese usando o modelo Fonte-Filtro interativo

O segundo sintetizador desenvolvido nesta pesquisa é baseado no modelo Fonte-Filtro interativo proposto por [71]. Este modelo foi escolhido para sintetizar vogais cantadas por sopranos líricos, uma vez que o modelo Fonte-Filtro não interativo não inclui os efeitos da interação entre a fonte e o filtro que ocorrem na produção da voz cantada.

Uma das principais diferenças entre a fala e a voz cantada é o impacto da interação fonte-filtro. A voz cantada envolve uma interação mais significativa entre a fonte (som produzido principalmente pelas pregas vocais) e o filtro (cavidades ressonantes do trato vocal). Isso contrasta com a maioria das técnicas de processamento de fala, onde essa interação é frequentemente desconsiderada.

A interação dinâmica entre as cavidades ressonantes do trato vocal (subglótica e supraglótica) durante a fonação resulta em variações na RTV a cada ciclo fundamental. Consequentemente, os parâmetros característicos da RTV, como a frequência central da formante ($F_c(n)$) e a largura de faixa ($BW(n)$), também variam ao longo do tempo. Essa abordagem foi explicada detalhadamente na subseção 3.3.2, onde as equações 3.5 e 3.6 foram introduzidas para descrever esse fenômeno. Conforme as equações, o RTV irá variar a cada período fundamental, com sua variação baseada na área glótica.

Na Figura. 4.2, apresenta-se o diagrama de blocos que ilustra o modelo Fonte-Filtro interativo utilizado para sintetizar as vogais cantadas por sopranos líricos. No diagrama, é possível observar que o modelo é composto pelo sinal glotal, a área glótica ($A[n]$), a RTV e a radiação nos lábios da cantora. Cada elemento do modelo Fonte-Filtro interativo é modelado independentemente, utilizando a descrição matemática mais apropriada. Para o sinal glotal, é empregado o modelo de pulso glotal LF, considerado o mais adequado para reproduzir o comportamento glotal observado no registro mais agudo de cada soprano lírico neste estudo. Nesse sentido, os parâmetros do modelo LF são configurados conforme apresentado no modelo Fonte-Filtro não interativo na subseção anterior.

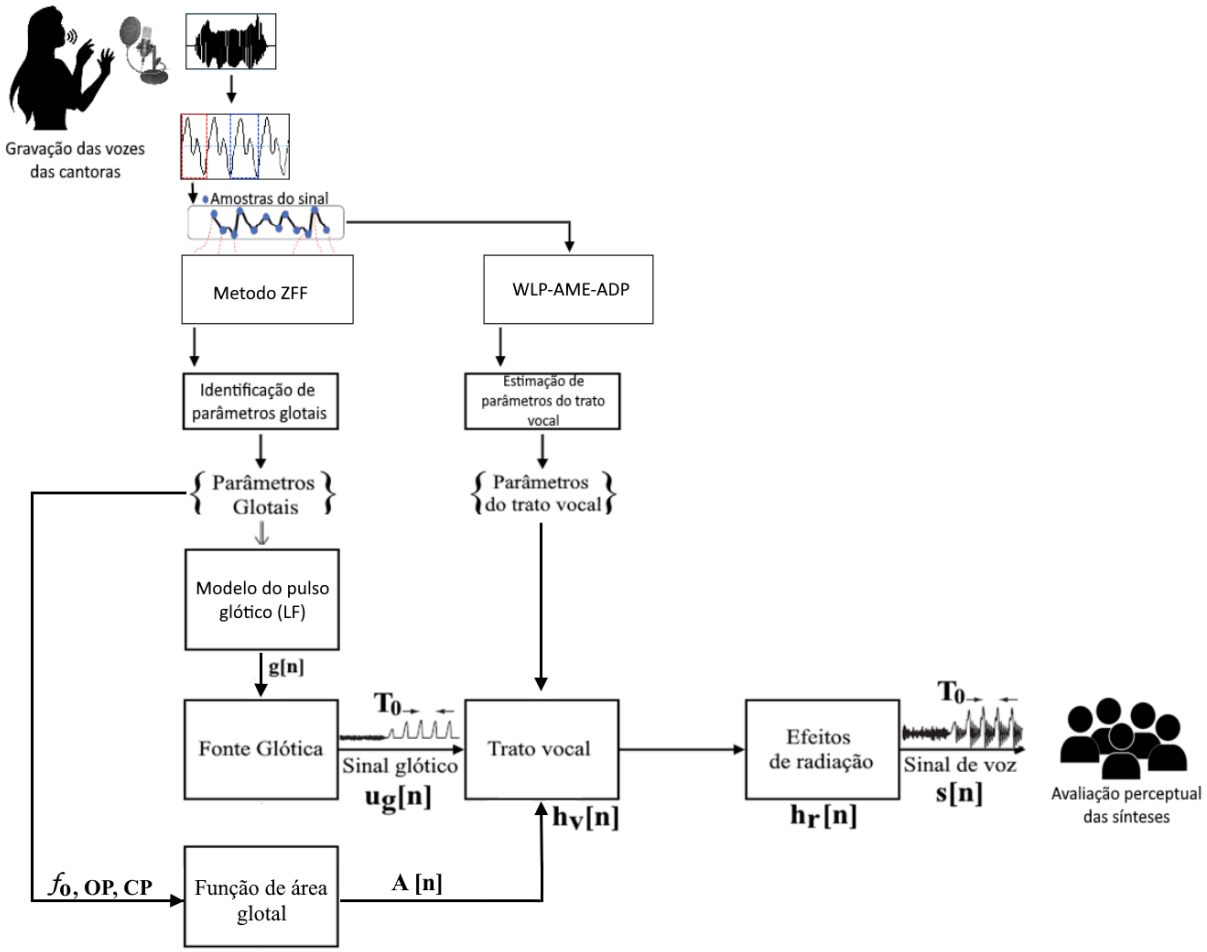


Figura 4.2: Blocos de construção do sistema de síntese de voz cantada por sopranos líricos usando o modelo Fonte-Filtro interativo.

A interação fonte-filtro é incorporada apenas para a primeira frequência formante, uma vez que esse aspecto proporciona uma variação mais perceptível. Essa inclusão é viabilizada por meio da função da área glótica, que regula tanto a frequência central quanto a largura de banda do primeiro formante, F_1. A função da área glótica é matematicamente representada pela seguinte expressão:

$$A[n] = \begin{cases} 0.5[1 - cos(\dfrac{2\pi n}{F_s\, T_0\, OP})] & , \; 0 \leq \dfrac{n}{F_s} \leq T_0\, OP \\ 0 & , \; T_0\, OP < \dfrac{n}{F_s} \leq T_0\,, \end{cases} \tag{4.1}$$

onde T_0 representa o período fundamental do sinal, F_s a frequência de amostragem, e OP é o intervalo de fase aberta relativo. Assim, a interação entre a fonte e o filtro será caracterizada pela duração da fase aberta, OP, e pela variação final da frequência do primeiro formante de pico a pico, ΔF_c. De acordo com a equação 3.5, a máxima variação

da frequência central da formante, $\Delta F_c(n)$máxima, depende de F_{co} e α. Por outro lado, a variação da largura de faixa é determinada por uma variação máxima de largura de faixa, $BW(n)$máxima, que depende de BW_o e β, conforme apresentado na equação 3.6.

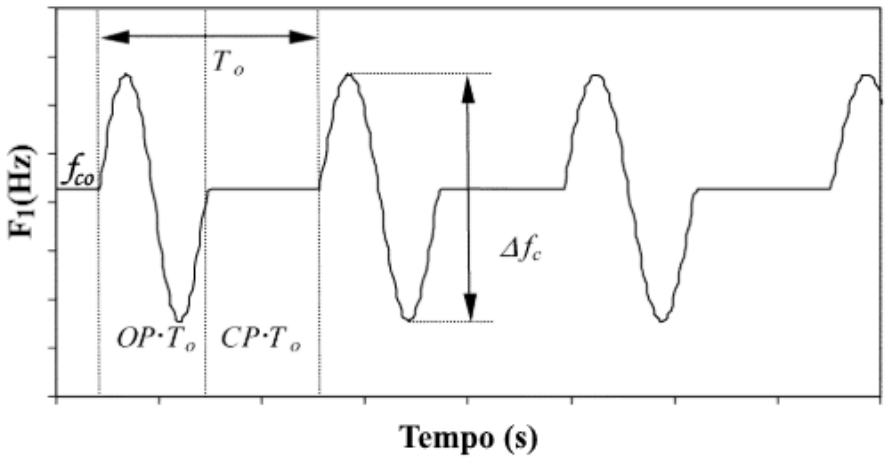


Figura 4.3: Variação da frequência central da primeira formante.

A Figura. 4.3 ilustra o comportamento da variação da frequência central da primeira formante (F_1) ao longo de um período fundamental. Nessa figura, é possível observar o efeito da interação fonte-filtro introduzido durante a fase aberta (OP) do sinal glótico.

De acordo com [71], o efeito da modulação da largura de faixa no sinal da voz cantada é considerado insignificante. Portanto, somente a variação da frequência central da primeira formante (F_1) parece ser relevante para caracterizar a interação entre a fonte e o trato vocal.

Capítulo 5

Metodologia

5.1 Construção da base de vozes

O conjunto de dados utilizado neste estudo é composto por 525 gravações de vozes cantadas, representando as cinco vogais do espanhol (/*a*/, /*e*/, /*i*/, /*o*/ e /*u*/), realizadas por um grupo de oito cantoras líricas hispanoparlantes de nacionalidade colombiana. Esse grupo inclui sete sopranos e uma mezzo-soprano, com idades variando entre 18 e 51 anos. Cada cantora possui um histórico de treinamento lírico específico, com Soprano 1 (S1) tendo 4 anos de treinamento, Soprano 2 (S2), Soprano 3 (S3) e Soprano 4 (S4) com 1 ano, Soprano 5 (S5) com 5 anos, Soprano 6 (S6) com 1 ano, Soprano 7 (S7) com 5 anos e Mezzo-Soprano 1 (M-S1) com 11 anos de treinamento lírico.

Cada arquivo de áudio contém gravações das cinco vogais sustentadas em uma nota musical específica da tessitura de cada cantora. Os arquivos de áudio abrangem desde a série de vogais cantadas na nota musical com o tom mais grave do registro vocal da cantora até a nota musical com o tom mais agudo, seguindo a escala diatônica musical ascendente, resultando em um total de 105 arquivos de áudio, que cobrem toda a tessitura vocal aguda de cada cantora. Dessa forma, as 525 gravações de vozes cantadas são compostas pelas vogais sustentadas em diferentes notas musicais, proporcionando uma ampla variedade de tons vocais para a análise e síntese de voz cantada.

A base de vozes também inclui as vogais do espanhol realizadas em glissandos, que consistem em passagens ascendentes ou descendentes sem interrupções de cada vogal. Além disso, foram gravadas as vogais em escala ascendente, onde as cantoras sustentaram cada vogal da nota mais grave à mais aguda da tessitura de cada uma. A duração de cada vogal sustentada varia entre 5 e 10 segundos. As características específicas do conjunto

de dados são apresentadas na Tabela. 5.1

Tabela 5.1: Gravações por Cantora e Tipo de Vogal

Cantora	Sustentadas	Glissandos	Escala Ascendente
S1	50	–	–
S2	100	–	–
S3	65	–	–
S4	45	–	–
S5	80	5	–
S6	60	5	4
S7	65	5	5
M-S1	60	5	5

Além disso, a base de vozes também contém gravações das vogais faladas /a/, /e/, /i/, /o/ e /u/, realizadas pelas Sopranos 1, 2, 5, 6, 7 e M-S1. Todos os arquivos de áudio foram cuidadosamente gravados em um ambiente de estúdio controlado para garantir a mais alta qualidade e consistência em todo o conjunto de dados.

5.1.1 Especificações técnicas da gravação

As gravações das vozes das cantoras foram realizadas em três estúdios de gravação diferentes, cada um equipado com dispositivos específicos para capturar as vozes com alta qualidade e precisão. A *Estação de Áudio Digital* (DAW) *Pro Tools* foi utilizada em todos os estúdios, garantindo a consistência e confiabilidade dos dados. A duração das sessões de gravação variou em cada estúdio, com uma média de uma hora por sessão. Profissionais especializados em gravação de áudio conduziram todas as gravações, assegurando a qualidade e o rigor em cada etapa do processo.

Tabela 5.2: Especificações técnicas por estúdio de gravação

Cantora	Estúdio de Gravação	Interface de Áudio	Data da Sessão
S1	Estúdio 1	Clarett 8Pre USB - Focusrite	9-03-20
S2	Estúdio 1	Clarett 8Pre USB - Focusrite	14-03-20
S3	Estúdio 2	Clarett 8Pre USB - Focusrite	9-05-22
S4	Estúdio 2	Clarett 8Pre USB - Focusrite	9-05-22
S5	Estúdio 3	Presonus Studio 1824c	6-05-23
S6	Estúdio 3	Presonus Studio 1824c	20-05-23
S7	Estúdio 3	Presonus Studio 1824c	20-05-23
M-S1	Estúdio 3	Presonus Studio 1824c	20-05-23

Além disso, as gravações foram realizadas diretamente em um *Solid-State Drive* (SSD). Para a captação das vozes, utilizou-se o microfone Shure KSM32/SL, um modelo de *Cardioid Condenser*, posicionado frontalmente a uma distância de 15 cm da boca de cada cantora. Adicionalmente, um filtro *anti-pop* foi colocado a meio caminho entre o microfone e a cantora, como é apresentado na Figura 5.1.

Figura 5.1: Processo de gravação da voz da Soprano 4 com os equipamentos utilizados no estúdio de gravação 2: (1) Microfone *Shure KSM32/SL* (Condenser, Digital), (2) Fones de ouvido para monitoramento, (3) *Anti-pop*, (4) Cabine de gravação, (5) Estação de Áudio Digital (DAW).

5.2 Características dos Sinais de Áudio

As características de áudio dos sinais variam para cada soprano. Na Tabela. 5.3 são apresentadas as especificações técnicas dos sinais de áudio.

Tabela 5.3: Características de Áudio dos Sinais das Cantoras Sopranos

Cantora	Taxa de Amostragem	Canais	Profundidade de Bits
S1	48 kHz	Mono	24
S2	44.1 kHz	Mono	16
S3	44.1 kHz	Stereo	16
S4	44.1 kHz	Stereo	16
S5	44.1 kHz	Stereo	16
S6	44.1 kHz	Stereo	24
S7	44.1 kHz	Stereo	24
M-S1	44.1 kHz	Stereo	24

Todos os sinais de áudio foram gravados em formato *Waveform Audio File* (WAV), sem compressão de áudio, garantindo alta qualidade e fidelidade nos arquivos de áudio utilizados. Para obter informações mais detalhadas sobre o processo de gravação das vogais cantadas, incluindo as condições de gravação e as especificações técnicas, consulte Barrientos et al.[74] e acesse o conjunto de dados no seguinte link: `https://www.dropbox.com/sh/esk5ckp5ij7fuln/AABLrKG4dGAPPSU-O7__F2nTa?dl=0`.

5.3 Determinação do sinal de análise

Para a estimativa dos parâmetros glotais e das frequências dos formantes, um sinal de áudio foi selecionado a partir das gravações das sopranos presentes na base de vozes descrita anteriormente. A escolha do sinal se baseou na necessidade de uma vogal cantada representativa, que oferecesse estabilidade suficiente para a análise precisa dos parâmetros desejados. Após uma revisão minuciosa das gravações, um trecho do sinal foi selecionado, centrando-se na região em que o som apresentou maior estabilidade. Esse trecho foi extraído usando a plataforma *Matlab* e, em seguida, analisado para verificar a presença de ruídos e artefatos que pudessem afetar negativamente os resultados. A Figura 5.2 mostra o sinal de voz cantada da vogal /*a*/ realizada pela soprano 1, destacando as partes mais relevantes para a análise e indicando o trecho escolhido para a estimativa dos parâmetros

glotais e das frequências dos formantes.

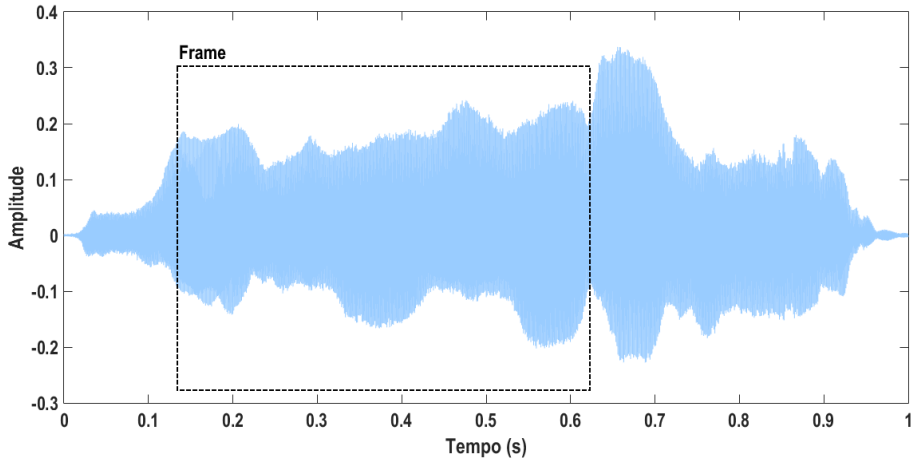


Figura 5.2: Exemplo de sinal de voz cantada da vogal /*a*/ realizada pela soprano 1. O trecho de sinal destacado representa o *frame* selecionado para análise e estimação dos parâmetros glotais e de ressonância. As transições do sinal, como ataques e liberações, foram omitidas nesta etapa para garantir a precisão das estimativas.

A determinação do trecho adequado do sinal de voz cantada, conhecido como *frame*, para a análise e estimação dos parâmetros glotais e de ressonância, é um aspecto crucial deste estudo. A escolha criteriosa do *frame* é essencial, uma vez que trechos instáveis ou transições abruptas podem comprometer a precisão das estimativas dos parâmetros. Por outro lado, é necessário limitar a duração do *frame* para capturar apenas as informações relevantes do sinal de voz cantada. O *frame* selecionado deve ser suficientemente longo para conter as informações relevantes do sinal, porém não tão extenso a ponto de abranger transições ou mudanças significativas. É importante ressaltar que as transições do sinal, como ataques e liberações (*releases*), não são consideradas na determinação do *frame*, assim como variações abruptas do sinal. Os *frames* das vogais cantadas, usados para estimar os parâmetros glotais e de ressonância do trato vocal, estão disponíveis no seguinte link: `https://www.dropbox.com/sh/4icnu5dptg8b2c8/AACA0GDUVXU4r0QPSUvcIinBa?dl=0`.

5.4 Síntese de voz cantada usando modelo Fonte-Filtro não interativo

Uma abordagem empregada para a síntese das vogais cantadas por sopranos líricos hispanoparlantes baseia-se no modelo Fonte-Filtro não interativo, conforme exposto na subseção 4.1. Nesta abordagem, são contemplados dois enfoques para o desenvolvimento da

síntese das vogais cantadas, incluindo variações de tons sustentados com efeitos de vibrato e trêmulo. Para modelar a produção das vogais, optou-se pela teoria Fonte-Filtro Fant [36], apresentada na seção 2.2. Nesse contexto, a fonte é representada pelos modelos de pulso glotal de Rosenberg e de LF [39, 40]. Os efeitos do trato vocal foram desenvolvidos utilizando um modelo de filtro *all-pole filter*, conforme descrito na subseção 2.2.4, considerando as ressonâncias do trato vocal (formantes) correspondentes aos polos da função de transferência do sistema discreto do trato vocal, $V_k(z)$, mostrada na Equação 2.11.

No primeiro enfoque de síntese, as frequências dos formantes e as larguras de faixa foram estimadas das vogais cantadas, utilizando a plataforma de análise de voz, o *software Praat*, a partir dos sinais de áudio de duas cantoras sopranos na base de vozes. No segundo enfoque de síntese, as frequências dos formantes e suas larguras de faixa foram estimadas com precisão utilizando o método WLP-HPSV, a partir das vogais cantadas por diferentes sopranos na base de vozes. Finalmente, todos os sons gerados pelo primeiro enfoque de síntese foram avaliados por um grupo de ouvintes, ressaltando a qualidade em termos de inteligibilidade e naturalidade dos sons produzidos. É importante destacar que todos os sons gerados pelos dois enfoques de síntese estão disponíveis para análise e avaliação.

5.4.1 Síntese não interativa usando parâmetros estimados de forma geral

Neste tipo de síntese, consideram-se as cinco vogais cantadas pelas sopranos líricas $S1$ e $S2$. Os valores médios de f_{o} para as notas musicais cantadas por cada soprano foram estimados utilizando a plataforma *Praat versão* 5.3. Considerando esse método de estimação de f_{o}, a $S1$ produziu as vogais cantadas desde a nota musical $F4$ até $E5$ (349 Hz - 659 Hz), enquanto a $S2$ gerou vogais cantadas desde a nota musical $E3$ até $F5$ (167 Hz - 698 Hz), conforme ilustrado na Figura. 5.3.

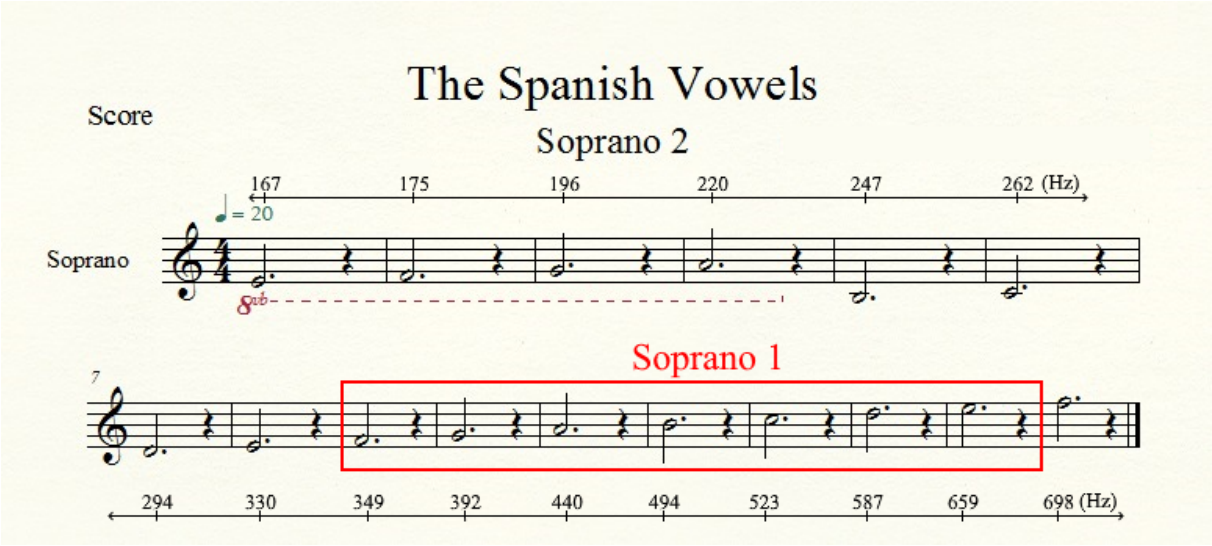


Figura 5.3: Partitura musical para as vogais do espanhol cantadas pela soprano 1 (emoldurada em tinta vermelha) e pela soprano 2 (em tinta preta).

O sinal glótico foi gerado utilizando o modelo de pulso glótico de Rosenberg, com parâmetros configurados para concordar com o comportamento glótico observado nas sopranos em registros agudos [15]. A frequência fundamental f_o empregada na geração do sinal correspondeu a um dos tons da tessitura de cada cantora soprano. A convolução do pulso glótico completo com um trem de impulsos resultou no sinal glótico utilizado nas sínteses vocais.

Para adicionar expressividade às sínteses vocais, foram aplicados efeitos de vibrato e trêmulo. Esses efeitos foram configurados com base em características acústicas comuns à identificação de cantores [9]. O vibrato foi ajustado para criar variações naturais na frequência, enquanto o trêmulo foi configurado para proporcionar modulações sutis. Ambos os efeitos foram aplicados de maneira apropriada para enriquecer a expressividade das sínteses.

A partir dos segmentos de análise (*frames*) extraídos dos sinais de áudio de cada vogal cantada, foram estimados os valores das primeiras cinco frequências dos formantes, juntamente com suas respectivas larguras de faixa. Para realizar essa estimação de parâmetros, foi utilizado o programa de análise de fala *Praat*. Além disso, os valores de f_o de cada tom também foram estimados por meio do programa *Praat*, com o intuito de serem posteriormente empregados pelo sintetizador de voz cantada *SOPRA-SYNTH*.

As sínteses foram separadas por notas musicais e correspondem às que foram comuns entre as duas sopranos durante as gravações iniciais e submetidas a um grupo de pessoas para que pudessem fazer uma avaliação. O teste de escuta foi realizado por dez brasileiros

e dez colombianos, com idades entre 20 e 53 anos, sendo oito mulheres e doze homens. O mesmo grupo classificou as vozes sintetizadas em relação à naturalidade e, também, à inteligibilidade, atribuindo notas, variando de 0, no pior caso, a 5, no caso em que a síntese apresentava-se excelente.

A informação foi obtida através de um questionário desenvolvido para classificar a naturalidade e a inteligibilidade das sínteses das vogais cantadas, enviado a cada um dos ouvintes do grupo. Os dados obtidos são de natureza quantitativa, organizados em tabelas e apresentados em gráficos de barras. A média aritmética foi utilizada como medida de tendência central e permitiu identificar o modo como os dados dos testes perceptuais foram distribuídos. O desvio padrão foi utilizado como medida de dispersão.

5.4.2 Síntese não interativa usando parâmetros estimados de forma precisa

Nesse segundo enfoque, aprimora-se a síntese de voz cantada ao empregar o modelo Fonte-Filtro não interativo. Esta abordagem inclui a incorporação de parâmetros glotais e de ressonância do trato vocal, os quais são obtidos com precisão a partir de um conjunto diversificado de vogais cantadas por diferentes sopranos líricas, selecionadas a partir da base de vozes desta tese. Por meio da aplicação de técnicas específicas desenvolvidas para o tratamento da voz cantada em tons de alta frequência, ocorre um refinamento significativo do modelo Fonte-Filtro não interativo, o que, por sua vez, culmina em uma notável melhoria da qualidade e das sutilezas das sínteses resultantes.

Para obter uma estimativa precisa dos parâmetros de ressonância do trato vocal e glotal, as vogais cantadas na faixa de frequências mais altas de cada uma das sopranos líricas hispanoparlantes ($S1$, $S2$, $S3$, $S4$, $S5$, $S6$ e $S7$) foram consideradas. Os valores médios de f_o para os tons das notas musicais entoadas por cada soprano foram estimados utilizando o método ZFF modificado.

O sinal glótico foi gerado utilizando o modelo de pulso glótico de LF, conforme detalhado na Seção 2.2.1.2. Foram adotados parâmetros específicos desse modelo, considerando o comportamento observado da glote em frequências agudas das sopranos [15]. A frequência fundamental f_o utilizada para criar o sinal glótico correspondeu a uma das notas na tessitura vocal das sopranos, conforme registrado na base de dados. O sinal glótico final foi obtido por meio da convolução do pulso glótico completo com um trem de impulsos.

A abordagem para adicionar expressividade às sínteses das vogais por meio da inclusão do vibrato e do trêmulo foi semelhante à estratégia anterior de Síntese não interativa, que empregou parâmetros estimados de forma geral [9]. Essas técnicas foram aplicadas com base em valores específicos que seguem características acústicas reconhecidas, contribuindo para uma maior autenticidade das sínteses vocais.

As frequências dos primeiros cinco formantes, juntamente com suas larguras de faixa correspondentes, foram estimadas a partir das vogais $/a/, /i/e/u/$ entoadas pelas sopranos líricas nos tons mais agudos que puderam produzir utilizando suas técnicas vocais. Essas vogais foram selecionadas como o grupo de estudo central nessa abordagem de síntese. Essas estimativas foram realizadas de forma precisa, empregando o método WLP-HPSV, que é apresentado como uma contribuição inovadora desta tese. Tais valores, bem como as frequências fundamentais (f_{o}) associadas a cada nota musical, foram registrados em uma planilha que será empregada em fases subsequentes da síntese vocal.

5.5 Síntese de voz cantada usando modelo Fonte-Filtro interativo

Nesse terceiro enfoque, realiza-se a exploração da síntese de voz cantada por sopranos líricos usando o modelo Fonte-Filtro interativo, como uma abordagem de síntese com grande novidade nesta tese. Usando este modelo de síntese se procura produzir as vogais cantadas com maior naturalidade, considerando a variação da RTV associada à SFI implícita na voz de canto.

Para desenvolver essa abordagem se usam os mesmos parâmetros de ressonância do trato vocal, estimados com precisão a partir do conjunto diversificado de vogais cantadas pelas diferentes sopranos líricas também exploradas na síntese não interativa.

O sinal glótico foi gerado usando o modelo de pulso glótico de LF, conforme detalhado na Seção 2.2.1.2. A configuração dos parâmetros do pulso glótico foi ajustada para replicar o comportamento glótico observado nas sopranos na faixa de tons mais agudos [15]. Com o objetivo de alcançar essa fidelidade, a f_{o} e o valor do parâmetro OQ em cada ciclo glótico foram estimados por meio do método ZFF modificado, em conjunto com um algoritmo de estimação de GOIs a partir do sinal de áudio das vogais cantadas, extraídas da base de vozes. Os coeficientes de assimetria α_m, fixado em 0.5, e o coeficiente de fase de retorno Q_a, ajustado para 0.2, foram determinados para obter uma forma de onda simétrica, levando em consideração o mecanismo laríngeo $M3$ utilizado pelas sopranos em seus registros mais

agudos [75]. Desta maneira, o sinal glótico foi gerado visando se aproximar com maior precisão do sinal glótico produzido pela cantora durante a produção vocal.

Para esta abordagem mais recente de síntese, a incorporação de elementos expressivos como o vibrato e o trêmulo seguiu uma abordagem semelhante à estratégia anterior de Síntese não interativa. A aplicação dessas técnicas foi baseada em valores específicos alinhados com características acústicas reconhecidas [9], o que contribuiu para conferir maior autenticidade às sínteses vocais, enriquecendo suas nuances e expressividade.

No contexto desta nova abordagem, a obtenção das frequências dos primeiros cinco formantes e suas correspondentes larguras de banda foi conduzida a partir das vogais cantadas pelas sopranos líricas escolhidas como foco central deste estudo. Essas estimativas foram realizadas com extrema precisão, por meio da aplicação do método WLP-HPSV, uma contribuição inovadora apresentada nesta tese. As informações obtidas, incluindo a f_{o} associada a cada nota musical, foram meticulosamente registradas em uma planilha dedicada para uso futuro nas etapas posteriores da síntese vocal.

Capítulo 6

Resultados

6.1 Síntese não interativa usando parâmetros estimados de forma geral

Neste estudo, sintetizamos vogais cantadas por sopranos líricas usando uma abordagem não interativa com parâmetros gerais. Através da plataforma *Praat*, estimamos f_o para notas musicais específicas de cada cantora. O modelo de pulso glótico de Rosenberg e os efeitos de vibração e tremor foram aplicados para aprimorar a expressividade. Usando o programa *Praat*, estimamos as frequências dos cinco primeiros formantes e suas larguras de banda. Esses resultados confiáveis e expressivos são documentados no artigo [74].

6.1.1 Geração do sinal glótico com efeito vibrato e trêmulo

O sinal glótico foi gerado por meio do modelo de pulso glótico de Rosenberg, conforme detalhado na Subseção 2.2.1.1. Para obter os parâmetros deste modelo, foi considerado um valor de $OQ = 0,78$, que se alinha com o comportamento glótico observado no registro agudo das sopranos [15]. Com essas premissas, os instantes relativos de abertura e fechamento glótico foram calculados como $\alpha_1 = 58\,\%$ e $\alpha_2 = 20\,\%$. A frequência fundamental, f_o, usada para recriar o sinal glótico correspondeu a uma das notas na tessitura das cantoras sopranos, como registrada na base de vozes. Utilizando as informações do pulso glótico completo, o sinal glótico foi gerado através da convolução do pulso com um trem de impulsos.

Na Figura. 6.1 apresenta-se o sinal glótico, $u_g[n]$, formado por uma sequência de 7 pulsos produzidos com o modelo do pulso glótico de Rosenberg.

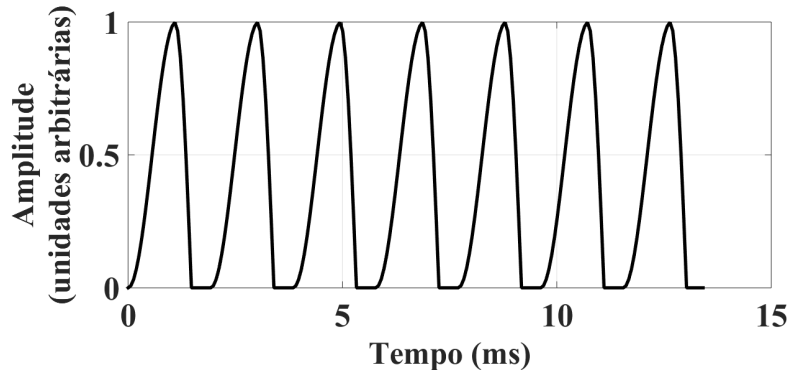


Figura 6.1: Sinal glótico gerado através do modelo de Rosenberg e considerando os parâmetros: $f_o = 520$ Hz, $\alpha_1 = 58\,\%$ e $\alpha_2 = 20\,\%$.

Os valores dos parâmetros utilizados para gerar o efeito de vibrato na voz cantada foram: $V_{Ext} = 0,4$ ms e $V_{Rate} = 4,5$ Hz, seguindo as características acústicas correspondentes ao vibrato para a identificação de cantores [9]. Com respeito à geração do efeito trêmulo, a taxa do trêmulo segue a mesma faixa da taxa do vibrato: $T_{Rate} = 4,5$ Hz, assim como o nível de amplificação segue a faixa da extensão do vibrato: $A_{Level} = 0,4$ ms.

6.1.2 Estimação de parâmetros de ressonância do trato vocal

Os valores das primeiras cinco frequências de ressonância do trato vocal, juntamente com suas respectivas larguras de faixa, foram obtidos por meio do programa de análise e síntese de fala *Praat*, na versão 5.3. O *Praat* utiliza o método LPC para estimar as frequências dos formantes. O LPC é uma técnica amplamente empregada na análise de sinais de fala e áudio para modelar os espectros de frequência. Além disso, tanto as frequências fundamentais de cada nota musical entoada quanto as frequências dos formantes foram registradas em uma planilha, visando seu uso futuro na síntese da voz cantada. Essas informações podem ser obtidas em https://cutt.ly/ayFS4qr.

6.1.3 Avaliação perceptual de naturalidade das sínteses

Os dados recolhidos na avaliação perceptual de naturalidade foram organizados numa tabela. Essa informação pode ser obtida em: https://cutt.ly/gh0cDAH. As linhas da tabela referem-se às avaliações dos ouvintes e as colunas referem-se ao tom sustentado por cada soprano. O gráfico de barras da Figura. 6.2 apresenta os resultados da avaliação perceptual de naturalidade na síntese da voz cantada com os parâmetros de ressonância

das sopranos 1 e 2, no intervalo $F4 - E5$.

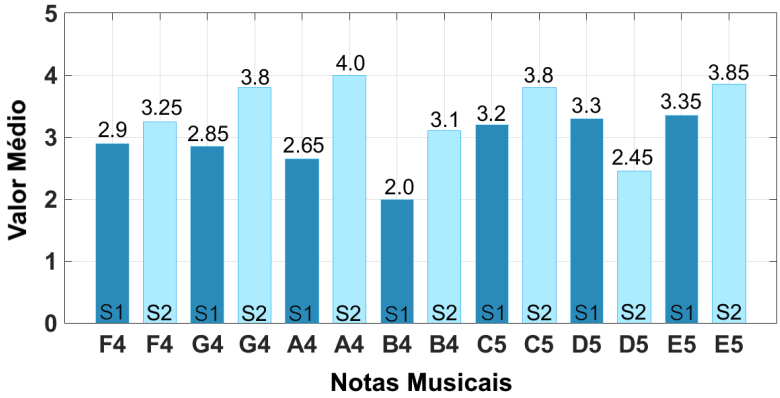


Figura 6.2: Gráfico de barras do teste perceptual das sínteses de vogais cantadas por sopranos considerando a média de naturalidade.

A maioria das sínteses com os dados da soprano 2 obtiveram uma média de naturalidade superior em comparação às sínteses da voz cantada com os dados da soprano 1. As notas musicais $G4$, $A4$, $C5$ e $E5$ produzidas com as frequências de ressonância da soprano 2 obtiveram uma média de naturalidade de $3,8$, $4,0$, $3,8$ e $3,85$ respectivamente. As sínteses das vogais cantadas apresentaram características robóticas devido às qualidades do sinal glótico: a forma de onda do pulso e ausência de *jitter*. Por outro lado, a adição das caraterísticas acústicas de vibrato e trêmulo contribuíram satisfatoriamente à naturalidade das sínteses.

A partir dos resultados do teste de escuta observou-se que a maioria das notas musicais que apresentaram a menor média de naturalidade correspondem às sínteses geradas com as frequências de ressonância da soprano 1, ainda que as qualidades do sinal glótico sejam as mesmas que as utilizadas nas sínteses das vogais produzidas com as frequências de ressonância da soprano 2. Na Tabela. 6.1 apresenta-se o desvio padrão, σ, dos dados correspondentes às avaliações perceptuais de naturalidade.

Tabela 6.1: Desvio padrão dos dados das avaliações perceptuais de naturalidade.

Soprano 1							
Tom	$F4$	$G4$	$A4$	$B4$	$C5$	$D5$	$E5$
σ	1,21	1,27	1,14	1,45	1,32	1,22	1,23

Soprano 2							
Tom	$F4$	$G4$	$A4$	$B4$	$C5$	$D5$	$E5$
σ	1,16	0,83	1,03	1,33	1,01	1,15	0,99

Os desvios calculados para as notas musicais $G4$ e $E5$ sustentadas pela soprano 2 permitiram identificar uma menor variabilidade dos dados e, portanto, uma maior confiabilidade nos resultados obtidos.

6.1.4 Avaliação perceptual de Inteligibilidade das sínteses

Os dados recolhidos na avaliação perceptual de inteligibilidade foram organizados em tabelas e classificados de acordo com a nota musical sustentada ($F4$ - $E5$). As colunas da tabela referem-se às cinco vogais espanholas e as linhas referem-se às avaliações dos ouvintes para as sínteses com parâmetros de ressonância das sopranos 1 e 2. Os dados podem ser obtidos acessando o seguinte link: https://cutt.ly/sh2v7Hp.

O gráfico de barras da Figura. 6.3 apresenta as cinco vogais da língua espanhola com maior média de inteligibilidade entre todas as sínteses submetidas no teste perceptual, e os áudios correspondentes a essas sínteses podem ser ouvidos em https://cutt.ly/4yVGUcr.

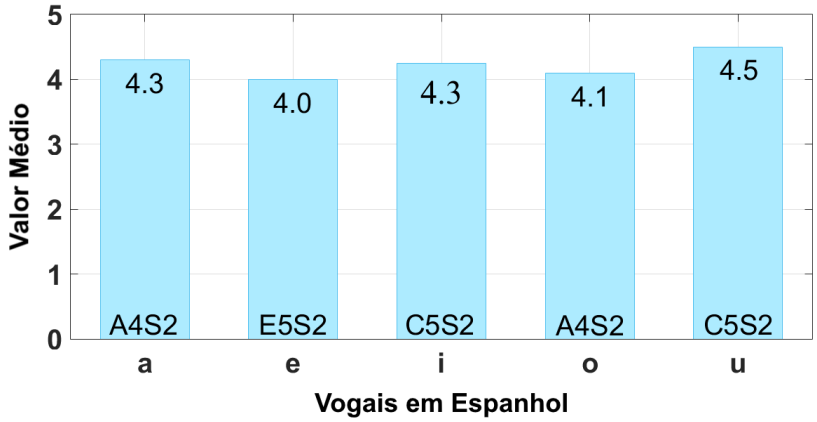


Figura 6.3: Gráfico de barras do teste perceptual das síntese de vogais cantadas por sopranos considerando a média de inteligibilidade.

As sínteses das vogais espanholas apresentadas na Figura. 6.3 são de grande interesse para uma subsequente análise espectral, pois não só alcançaram a maior média de inteligibilidade más também foram sustentadas nas notas musicais ($A4$, $C5$ e $E5$) com a maior média de naturalidade entre todas as sínteses submetidas aos testes perceptuais.

Na Tabela. 6.2 apresenta-se o desvio padrão, σ, dos dados correspondentes às avaliações perceptuais de inteligibilidade das cinco vogais espanholas.

Tabela 6.2: Desvio padrão dos dados das avaliações perceptuais de inteligibilidade.

Vogal	Tom	Cantora	Média	σ
$/a/$	$A4$	$S2$	4,3	0,86
$/e/$	$E5$	$S2$	4,0	0,92
	$F4$	$S1$	4,0	1,72
$/i/$	$C5$	$S2$	4,3	1,02
$/o/$	$A4$	$S2$	4,1	0,85
$/u/$	$C5$	$S2$	4,5	0,83

Os desvios para as avaliações perceptuais da inteligibilidade das vogais espanholas apresentadas na Figura. 6.3 permitiram identificar uma menor variabilidade dos dados em torno da média ($\sigma \leq 1,02$).

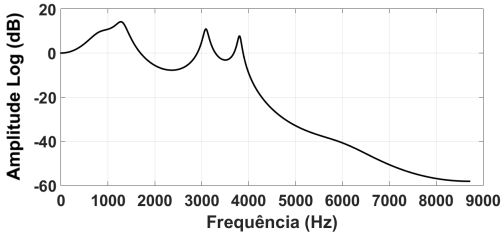
Observa-se que a vogal $/e/$ produzida com os parâmetros de ressonância da soprano 2 apresentou um desvio padrão de $0,92$ sendo inferior ao calculado para a mesma vogal

produzida com os parâmetros de ressonância da soprano 1. Devido a esse resultado, considerou-se uma maior confiabilidade na inteligibilidade da vogal /e/ produzida com os parâmetros de ressonância da soprano 2.

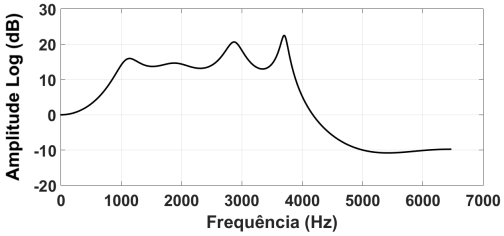
6.1.5 Formante do cantor em sopranos

Na Figura. 6.4 apresentam-se as respostas em frequência das cinco vogais espanholas que alcançaram a maior média de inteligibilidade no teste perceptual.

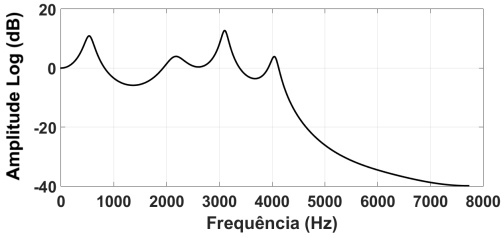
A resposta em frequência foi obtida em 1024 amostras abrangendo o círculo unitário completo e usando uma janela retangular sem sobreposição entre as amostras.



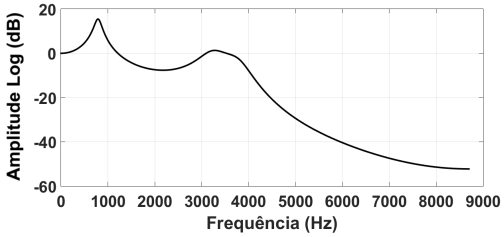
(a) Vogal /a/.



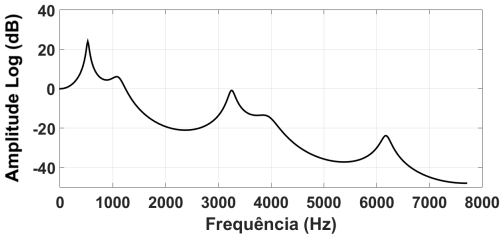
(b) Vogal /e/.



(c) Vogal /i/.



(d) Vogal /o/.



(e) Vogal /u/.

Figura 6.4: Resposta em frequência para as cinco vogais espanholas cantadas que apresentaram a maior média de inteligibilidade.

Na resposta em frequência da vogal /e/ observa-se que, em torno de 3 kHz, a energia acústica eleva-se para acima de 20 dB, porém a média de inteligibilidade dessa vogal é a

mais baixa de todas as vogais, evidenciando-se que ocorre um efeito adverso na inteligibilidade da vogal quando a energia acústica do som se eleva.

É importante notar um pico do envelope espectral em torno de 3 kHz, causado pelo agrupamento das frequências formantes superiores (F_3, F_4, F_5). Em relação a esse agrupamento, F_5 está longe em frequência de F_4, portanto, não participa adequadamente da construção do formante do cantor. Esta é uma característica comum nas cinco respostas em frequência desenvolvidas.

As cinco primeiras frequências formantes (em Hz) correspondem às vogais espanholas que alcançaram a maior média de inteligibilidade e aparecem na Tabela. 6.3. Assim, pode-se observar que os valores de F_1 das vogais /e/, /i/, /o/ e /u/ são superiores aos valores de F_1 da voz falada que foram apresentados no triangulo articulatório das vogais espanholas da Figura. 2.6. Esse resultado deve-se ao fato que a cantora ajustou a F_1 das vogais aos valores de f_o das notas musicais sustentadas.

No caso da vogal /a/, a F_1 da voz falada foi superior à f_o da nota musical sustentada e, portanto, significa que a cantora não precisou ajustar a F_1. [76].

Tabela 6.3: Frequências formantes (em Hz) utilizadas para produzir as vogais espanholas cantadas.

Tom	f_o	Vogal	F_1	F_2	F_3	F_4	F_5
A4	435	a	808	1304	3088	3808	6025
E5	646	e	1091	1891	2873	3706	6869
C5	515	i	536	2159	3104	4051	6276
A4	435	o	790	1551	3226	3795	5928
C5	513	u	526	1108	3251	3947	6179

As larguras de faixa das cinco primeiras frequências formantes que foram utilizadas para produzir as vogais espanholas e que alcançaram a maior média de inteligibilidade nesta pesquisa são mostradas na Tabela. 6.4 e representam os efeitos da perda do trato vocal [6].

Em particular, as larguras de faixa das frequências formantes utilizadas para produzir a vogal /u/ são relativamente baixas (menos de $200\,Hz$), refletindo baixas perdas no trato vocal da soprano 2 e, portanto, a propagação do som através do trato vocal é mais eficiente que nas outras vogais (/a/, /e/, /i/, /o/). Essa característica influi nos resultados positivos de naturalidade e inteligibilidade da vogal /u/ sustentada pela soprano 2 na nota

musical $C5$.

Tabela 6.4: Largura de faixa (em Hz) das frequências dos formantes utilizadas para produzir as vogais espanholas cantadas.

Tom	Vogal	Bw_1	Bw_2	Bw_3	Bw_4	Bw_5
A4	a	282	126	53	40	808
E5	e	195	357	146	59	998
C5	i	90	202	68	76	1411
A4	o	77	1676	291	286	2153
C5	u	23	116	62	198	87

6.2 Síntese não interativa de vogais usando parâmetros estimados de forma precisa

Nesse refinamento da síntese vocal, o modelo Fonte-Filtro não interativo é aplicado, incorporando parâmetros glotais e de ressonância do trato vocal, obtidos com precisão por meio do método WLP-HPSV. Esse método inovador, desenvolvido como parte desta pesquisa, utiliza várias vogais cantadas por diversas sopranos líricas da base de vozes. Essa abordagem representa um avanço significativo na criação de sínteses vocais autênticas e expressivas.

6.2.1 Geração do sinal glótico com efeito vibrato e trêmulo

O sinal glótico foi gerado por meio do modelo de pulso glótico de LF, conforme detalhado na Seção 2.2.1.2. Os parâmetros desse modelo foram selecionados para estar em concordância com o comportamento observado da glote das sopranos em frequências agudas [15]. Para alcançar esse objetivo, foram utilizados parâmetros específicos do sinal glótico, incluindo o quociente de abertura $OQ = 0,78$, o coeficiente de assimetria $\alpha_m = 0,5$ e o coeficiente de fase de retorno $Q_a = 0,2$. A frequência fundamental, f_o, utilizada para gerar o sinal glótico correspondeu a uma das notas na tessitura vocal das sopranos, conforme registrado na base de dados. Os valores médios de f_o para as notas musicais cantadas por cada soprano foram calculados por meio do método ZFF modificado, como apresentados no seguinte link: `https://www.dropbox.com/scl/fi/gy1o58gwtfkhmrv5w893j/Resultados-de-fo-estimadas.xlsx?rlkey=y32npd60a7jdblfeirzpmp24p&dl=0`. O sinal glótico final foi obtido por meio da convolução do pulso glótico completo com um trem

de impulsos.

Os valores dos parâmetros utilizados para incorporar o efeito de vibrato à voz cantada foram escolhidos de acordo com o enfoque anterior (Síntese não interativa usando parâmetros estimados de forma geral). Foram adotados os seguintes valores: $V_{Ext} = 0,4$ ms e $V_{Rate} = 4,5$ Hz, seguindo as características acústicas correspondentes ao vibrato para a identificação de cantores [9]. No que diz respeito à geração do efeito trêmulo, a taxa do trêmulo foi mantida na mesma faixa da taxa do vibrato, ou seja, $T_{Rate} = 4,5$ Hz, enquanto o nível de amplificação do trêmulo seguiu a extensão do vibrato: $A_{Level} = 0,4$ ms

A Figura 6.5 ilustra o sinal glótico modulado com vibrato e trêmulo, denotado como $u_g[n]$. Esse sinal é composto por uma sequência de pulsos gerados utilizando o modelo de pulso glótico de LF.

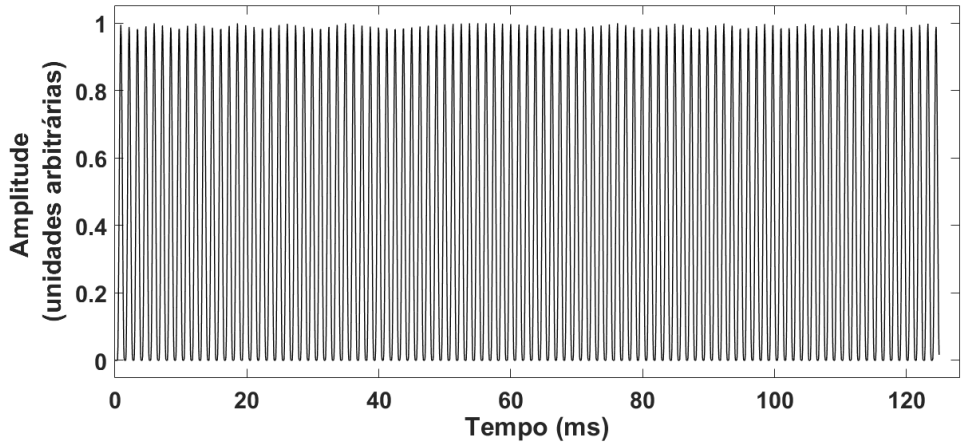


Figura 6.5: Sinal glótico com vibrato e trêmulo gerado pelo modelo de LF utilizando os parâmetros $OQ = 0,78$, $\alpha_m = 0,5$, $Q_a = 0,2$ e $f_o = 791,37$ Hz, correspondente à nota G5 cantada pela Soprano 1.

6.2.2 Estimação de parâmetros de ressonância do trato vocal com precisão

A aplicação do método WLP-HPSV na estimativa das frequências dos primeiros cinco formantes e suas larguras de banda, a partir das vogais $/a/, /i/e/u/$ cantadas pelas sopranos líricas, resultou em um conjunto de dados confiáveis. Esses valores são essenciais para a criação de sínteses vocais mais autênticas e expressivas, pois essas características formantes desempenham um papel fundamental na definição da qualidade e singularidade da voz. Além disso, as frequências fundamentais associadas a cada

nota musical foram registradas, o que permitirá a incorporação precisa dessas informações no processo subsequente de síntese de voz. A utilização desse método inovador fortalece a base metodológica desta pesquisa e contribui para a qualidade geral das sínteses vocais produzidas. Os valores resultantes podem ser acessados em: `https://www.dropbox.com/sh/kj9mplj2dqf7xgh/AAAUcZ4hbemiEBnm66AtZtsSa?dl=0`.

Finalmente, as sínteses das vogais cantadas por sopranos líricas foram obtidas através de um processo refinado, como apresentado nesta abordagem de síntese não interativa. Utilizando o modelo do filtro vocal conhecido como *all-pole filter*, foram produzidas as sínteses das vogais cantadas. Essas sínteses são particularmente confiáveis devido à precisão alcançada na estimativa dos parâmetros de ressonância do trato vocal, os quais desempenham um papel crucial na qualidade e singularidade das vozes sintetizadas.

6.2.3 Sínteses das vogais cantadas por sopranos líricos

As sínteses resultantes dessa abordagem estão disponíveis para acesso através do seguinte link: `https://www.dropbox.com/sh/sgiclm9ito6vedm/AABf3yQfwGJv9Urqr5tdFZ6Va?dl=0`. A confiabilidade dessas sínteses é reforçada pela comparação com o enfoque anterior, que utilizava o método LPC para estimar frequências formantes em altas frequências. O método atual supera as limitações do LPC e proporciona resultados mais precisos e confiáveis, especialmente para as vogais cantadas em frequências altas, contribuindo significativamente para a qualidade geral das sínteses vocais produzidas.

Neste contexto, é evidente que as sínteses alcançam um nível mais elevado de naturalidade e inteligibilidade em comparação com a abordagem da síntese não interativa, onde os parâmetros são estimados de maneira mais generalizada. A obtenção precisa desses parâmetros desempenha um papel fundamental na melhoria dos resultados obtidos

Adicionalmente, com o objetivo de exemplificar a aplicação prática do enfoque de síntese não interativa utilizando parâmetros cuidadosamente estimados, foram realizadas sínteses das vogais $/a/, /i/e/u/$ concatenadas. As sínteses resultantes demonstraram qualidade e fidelidade, validando a precisão das estimativas dos parâmetros utilizados. Um aspecto relevante abordado neste processo é a tentativa de replicar a variação da duração dos fonemas em diferentes tons musicais e vogais. Esta abordagem procura reproduzir as variações temporais características das vogais ao longo das notas musicais e fonemas. Para ouvir os exemplos das sínteses das vogais e entender melhor esse processo, os leitores são convidados a acessarem o seguinte link `https://www.dropbox.com/sh/hoa8quh6ftnumhb/AADTbjLK2fwliGqNWnaPFFdOa?dl=0`.

6.3 Síntese de vogais cantadas usando modelo Fonte-Filtro interativo

Neste terceiro enfoque, explora-se a síntese da voz cantada de sopranos líricas usando o modelo Fonte-Filtro interativo, uma abordagem inovadora introduzida neste estudo. No entanto, os resultados das sínteses obtidas não atingiram as expectativas desejadas. As sínteses apresentaram certo nível de ruído indesejável que dá uma característica robotizada e com perda de inteligibilidade da vogal, provavelmente devido à inadequação do algoritmo utilizado para modelar o filtro do trato vocal.

No processo, usou-se o algoritmo apresentado no anexo C. Para a geração do sinal glótico, aplicou-se o modelo de pulso glótico de LF,estimando adequadamente a f_o e o parâmetro OQ de cada ciclo glótico, e ajustando cuidadosamente seus parâmetros para corresponder ao comportamento glótico nas notas mais agudas, contribuindo para uma síntese mais autêntica do sinal glótico. Elementos expressivos como vibrato e trêmulo foram adicionados. As frequências dos primeiros cinco formantes foram obtidas com alta precisão usando o método WLP-HPSV, porém as sínteses não atingiram a naturalidade desejada, apontando para a necessidade de aprimorar a modelagem do filtro do trato vocal. Mais detalhes e amostras das sínteses estão disponíveis em `https://www.dropbox.com/sh/8a477hp6wa9bci4/AAAKmImUR-E-rxFjULTZPstga?dl=0`.

A tese poderia terminar com a síntese não interativa usando parâmetros estimados de forma precisa, o que já forneceu resultados muito bons, porém, o autor decidiu dar um passo adiante e discutir, também, o caso para modelo Fonte-Filtro interativo.

Essa técnica deveria dar uma síntese mais próxima ainda das vozes reais, mas a falta de um filtro adequado deixou os resultados aquém do esperado. Decidiu-se, neste trabalho, deixar esse método registrado e o algoritmo documentado para que futuros pesquisadores possam se basear nele e tentar melhorar os resultados.

Capítulo 7

Conclusões

Neste estudo, foram realizadas sínteses de vogais cantadas por sopranos líricas utilizando abordagens não interativas com parâmetros estimados tanto de forma geral quanto com precisão. Os resultados obtidos permitiram obter várias conclusões relevantes:

A aplicação do modelo de pulso glótico de Rosenberg, juntamente com os efeitos de vibração e tremor, contribuiu para uma maior expressividade nas sínteses das vogais. A incorporação desses elementos resultou em vozes mais naturais e melódicas, uma vez que as sínteses de voz cantada com esse modelo glotal normalmente apresenta um som robotizado.

A utilização do método LPC na estimação das frequências dos primeiros formantes foi aceitável, porém, o método inovador e inédito WLP-HPSV trouxe resultados mais precisos e confiáveis, especialmente nas frequências altas. Esses parâmetros desempenham um papel crucial na qualidade e singularidade das vozes sintetizadas.

A abordagem de síntese não interativa com parâmetros estimados de forma geral já resultou em vozes cantadas aceitáveis, porém, a abordagem refinada com parâmetros estimados de forma precisa demonstrou um aumento significativo na naturalidade e inteligibilidade das sínteses. Isso é particularmente notável nas frequências mais altas.

A adição dos efeitos de vibrato e trêmulo contribuiu positivamente para a expressividade das sínteses vocais, melhorando a naturalidade das vozes produzidas. A configuração desses parâmetros de acordo com as características acústicas dos cantores mostrou-se eficaz.

A análise da resposta em frequência das vogais cantadas revelou que as características espectrais estão em concordância com a literatura, demonstrando que as frequências formantes estão alinhadas com as características das vogais do espanhol. Isso reforça a

autenticidade das sínteses.

A aplicação do método WLP-HPSV e WLP-AME-ADP para estimar parâmetros de ressonância do trato vocal representa um avanço significativo na criação de sínteses vocais autênticas e expressivas. Esses métodos inovadores proporcionaram resultados mais confiáveis e precisos, ampliando as possibilidades de síntese vocal realista com sintetizadores por formantes com o *SOPRA-SYNTH*.

Os resultados das avaliações perceptuais indicaram que as sínteses baseadas nos parâmetros da soprano 2 obtiveram, em geral, melhores avaliações de naturalidade e inteligibilidade em comparação com a soprano 1. Isso sugere que a seleção de parâmetros baseados nas características individuais das cantoras influencia positivamente os resultados.

Este estudo fornece *insights* importantes para a pesquisa em síntese vocal, especialmente em contextos de cantores líricos. As abordagens apresentadas podem ser adaptadas e expandidas para outros tipos de vozes e gêneros musicais.

As sínteses obtidas pela abordagem refinada demonstraram uma qualidade excepcional e autenticidade na reprodução das vogais cantadas por sopranos líricas. Isso pode abrir portas para aplicações em música, teatro e produção audiovisual.

Embora a abordagem inovadora de síntese da voz cantada utilizando o modelo fonte-filtro interativo tenha mostrado potencial, os resultados obtidos não atenderam às expectativas devido a ruídos indesejáveis e falta de naturalidade nas sínteses. Apesar da cuidadosa estimação de parâmetros e da incorporação de elementos expressivos, as limitações na modelagem do filtro do trato vocal foram evidentes.

Este estudo estabelece uma base sólida para futuras explorações na síntese vocal não interativa. Pode-se considerar a expansão para diferentes tipos de vozes, a otimização dos parâmetros de efeitos adicionais e a investigação de técnicas de síntese mais avançadas.

Em resumo, este estudo demonstrou que a precisão na estimativa de parâmetros de ressonância do trato vocal e a incorporação de efeitos expressivos têm um impacto significativo na qualidade, naturalidade e inteligibilidade das sínteses vocais. A pesquisa contribuiu para o avanço da síntese vocal realista e autêntica, abrindo possibilidades para aplicações variadas nas áreas de música e produção audiovisual.

Capítulo 8

Trabalhos Futuros

8.1 Avaliação perceptual comparativa entre abordagens de síntese

Embora este estudo tenha gerado vozes sintéticas por meio do modelo Fonte-Filtro não interativo, usando estimação precisa de parâmetros, uma importante avaliação perceptual comparativa entre essa abordagem e o enfoque anterior de estimativas de parâmetros de forma geral ainda não foi realizada. A ausência dessa avaliação perceptual restringe a possibilidade de efetuar uma comparação objetiva entre os dois métodos e determinar suas eficácias relativas em termos de naturalidade e inteligibilidade vocal [74].

Como um trabalho futuro essencial, recomenda-se a condução de avaliações perceptuais abrangentes. Essas avaliações poderiam incluir análises de naturalidade vocal, inteligibilidade das sínteses e preferências dos ouvintes. Com a realização desses estudos comparativos, será possível estabelecer uma base sólida para a avaliação objetiva das vantagens e limitações de cada abordagem. Isso, por sua vez, contribuirá para uma compreensão mais profunda das técnicas de síntese de voz cantada e fornecerá diretrizes valiosas para aprimoramentos futuros.

8.2 Explorar desafios na síntese de voz cantada usando o modelo Fonte-Filtro Interativo

No escopo deste trabalho, abordamos a desafiadora tarefa da síntese de vogais cantadas por sopranos líricos utilizando o modelo Fonte-Filtro interativo. Esta abordagem inovadora prometia produzir resultados mais naturalísticos e autênticos, levando em con-

sideração as complexas variações da RTV e a Interação Fonte-Filtro SFI presente nas vozes de canto [71].

No entanto, na prática, encontramos um obstáculo inesperado. Ao aplicar o modelo Fonte-Filtro interativo para a síntese de vogais cantadas, nos deparamos com a presença de um ruído indesejado que comprometeu a qualidade das sínteses. Esse ruído, até então não solucionado, afetou a autenticidade e a fidelidade das vozes geradas, tornando-se um desafio significativo a ser enfrentado.

Esta questão emerge como uma oportunidade para pesquisas futuras. Investigar a origem e a natureza desse ruído, além de desenvolver estratégias adequadas para sua mitigação ou eliminação, torna-se uma meta crucial para aprimorar a síntese de voz cantada usando o modelo Fonte-Filtro interativo. Compreender as complexas interações entre os elementos do modelo e as características das vozes de sopranos líricos permitirá avanços substanciais na qualidade e realismo das sínteses, levando a uma representação mais fiel das nuances expressivas e da riqueza timbrística presentes nas vozes de canto. Nesse sentido, futuras investigações podem explorar métodos de modelagem mais refinados, técnicas de processamento avançadas e otimização dos parâmetros do modelo para superar esse desafio e alcançar resultados de alta qualidade na síntese de vozes cantadas.

8.3 Explorar técnicas de modelagem para efeitos de vibrato na síntese de voz cantada

Outra área de interesse para futuras pesquisas diz respeito à exploração de técnicas avançadas para a modelagem de efeitos de vibrato. O vibrato é uma característica expressiva essencial na interpretação vocal, adicionando riqueza e emotividade à voz cantada. No entanto, a incorporação precisa e controlada de diferentes tipos de vibrato, tanto em termos de velocidade quanto de amplitude, é uma tarefa complexa.

O estudo da modelagem do vibrato pode ser dividido em várias dimensões, incluindo a análise e extração de padrões de vibrato de vozes de sopranos líricos, a investigação de técnicas de síntese que reproduzam de maneira convincente as variações sutis e complexas do vibrato, e a exploração de estratégias para aplicar o vibrato de forma seletiva em diferentes segmentos do sinal de voz [61].

A exploração de algoritmos de aprendizado de máquina pode desempenhar um papel fundamental na captura das variações individuais de vibrato entre diferentes sopranos líricos e em diferentes estilos musicais. Isso pode permitir a criação de modelos persona-

lizados de vibrato que se ajustem às características únicas de cada voz e interpretação.

Dessa forma, investigar abordagens inovadoras para a modelagem do vibrato na síntese de voz cantada não apenas aprimoraria a expressividade das sínteses geradas, mas também abriria caminho para a criação de vozes virtuais altamente autênticas e emocionalmente envolventes. A pesquisa nessa área tem o potencial de impactar significativamente a qualidade e a diversidade das sínteses vocais, enriquecendo a experiência musical e artística como um todo [9].

8.4 Modelos estocásticos para a voz cantada

Um outro passo interessante e importante, é o de incluir modelos estocásticos de *jitter* também nas vozes cantadas, como feito para vogais sustentadas em [49, 43, 50]. Incluindo a possível identificação de patologias e, particularmente, o envelhecimento da voz, fenômeno importante para quem sua a voz profissionalmente, com destaque para os cantores líricos.

Referências

[1] DIAS, S. de O. Estimation of the glottal pulse from speech or singing voice. 2012.

[2] GONZÁLEZ, R. D. Producción de la voz y el habla: la fonación. 2014.

[3] BERNARDONI, N. H. *Etude de la source glottique en voix parlée et chantée: modélisation et estimation, mesures acoustiques et électroglottographiques, perception*. Tese (Doutorado) — Université Pierre et Marie Curie-Paris VI, 2001.

[4] MARSOLA, M.; BAE, T. *Canto Uma Expressão*. [S.l.]: Irmãos Vitale, 2000.

[5] SALAS, M. Aplicaciones del análisis acústico en los estudios de la voz humana. *Seminario Internacional de Acústica. Santiago*, 2003.

[6] RABINER, L. R.; SCHAFER, R. W. *Theory and applications of digital speech processing*. [S.l.]: Pearson, 2011. 1042 p. ISBN 0136034284.

[7] ARDAILLON, L. *Synthesis and expressive transformation of singing voice*. Tese (Doutorado) — Paris 6, 2017.

[8] VIEIRA, M. N. Uma introdução à acústica da voz cantada. *Seminário Música Ciência Tecnologia*, v. 1, n. 1, 2004.

[9] NWE, T. L.; LI, H. Exploring vibrato-motivated acoustic features for singer identification. *IEEE Transactions on Audio, Speech and Language Processing*, v. 15, p. 519–530, 2 2007. ISSN 15587916.

[10] D'ALESSANDRO, N.; WOODRUFF, P.; FABRE, Y.; DUTOIT, T.; BEUX, S. L.; DOVAL, B.; D'ALESSANDRO, C. Realtime and accurate musical control of expression in singing synthesis. *Journal on Multimodal User Interfaces 2007 1:1*, Springer, v. 1, p. 31–39, 3 2007. ISSN 1783-8738.

[11] NOSE, T.; KANEMOTO, M.; KORIYAMA, T.; KOBAYASHI, T. Hmm-based expressive singing voice synthesis with singing style control and robust pitch modeling. *Computer Speech & Language*, Academic Press, v. 34, p. 308–322, 11 2015. ISSN 0885-2308.

[12] NISHIMURA, M.; HASHIMOTO, K.; OURA, K.; NANKAKU, Y.; TOKUDA, K. Singing voice synthesis based on deep neural networks. *Interspeech*, p. 2478–2482, 9 2016.

[13] CHANDNA, P.; BLAAUW, M.; BONADA, J.; GóMEZ, E. Wgansing: A multi-voice singing voice synthesizer based on the wasserstein-gan. *European Signal Processing Conference*, European Signal Processing Conference, EUSIPCO, v. 2019-September, 9 2019. ISSN 22195491.

[14] MELLODY, M.; HERSETH, F.; WAKEFIELD, G. H. Modal distribution analysis, synthesis, and perception of a soprano's sung vowels. *Journal of Voice*, Mosby, v. 15, p. 469–482, 12 2001. ISSN 0892-1997.

[15] GARNIER, M.; HENRICH, N.; CREVIER-BUCHMAN, L.; VINCENT, C.; SMITH, J.; WOLFE, J. Glottal behavior in the high soprano range and the transition to the whistle register. *The Journal of the Acoustical Society of America*, Acoustical Society of AmericaASA, v. 131, p. 951, 1 2012. ISSN 0001-4966.

[16] BONADA, J.; SERRA, X. Synthesis of the singing voice by performance sampling and spectral models. *IEEE Signal Processing Magazine*, Institute of Electrical and Electronics Engineers Inc., v. 24, p. 67–79, 2007. ISSN 10535888.

[17] JOLIVEAU, E.; SMITH, J.; WOLFE, J. Tuning of vocal tract resonance by sopranos. *Nature 2004 427:6970*, Nature Publishing Group, v. 427, p. 116–116, 1 2004. ISSN 1476-4687.

[18] SUNDBERG, J.; LINDBLOM, B.; HEFELE, A. M. Voice source, formant frequencies and vocal tract shape in overtone singing. a case study. *Logopedics Phoniatrics Vocology*, Taylor & Francis, p. 1–13, 2021. ISSN 16512022.

[19] BEUX, S. L.; DOVAL, B. Real-time calm synthesizer new approaches in hands-controlled voice synthesis. *Proceedings of the conference on New interfaces for musical expression*, p. 266–271, 2006.

[20] JOLIVEAU, E.; SMITH, J.; WOLFE, J. Vocal tract resonances in singing: The soprano voice vocal tract resonances in singing: The soprano voice. *J Acoust Soc Am*, v. 116, p. 2434–2439, 2004.

[21] LONI, D. Y.; SUBBARAMAN, S. Formant estimation of speech and singing voice by combining wavelet with lpc and cepstrum techniques. *9th International Conference on Industrial and Information Systems, ICIIS 2014*, Institute of Electrical and Electronics Engineers Inc., 2 2015.

[22] ALKU, P.; KADIRI, S. R.; GOWDA, D. Refining a deep learning-based formant tracker using linear prediction methods. *Computer Speech & Language*, Academic Press, v. 81, p. 101515, 6 2023. ISSN 0885-2308.

[23] CHOWDHURY, H. A.; RAHMAN, M. S. Formant estimation from speech signal using the magnitude spectrum modified with group delay spectrum. *Acoustical Science and Technology*, ACOUSTICAL SOCIETY OF JAPAN, v. 42, n. 2, p. 93–102, 2021.

[24] SUNDBERG, J. Perceptual aspects of singing. *Journal of Voice*, Mosby, v. 8, p. 106–122, 6 1994. ISSN 0892-1997.

[25] SUNDBERG, J. The acoustics of the singing voice. *Scientific American*, v. 236, p. 82–91, 1977.

[26] SUNDBERG, J. Vocal tract resonance in singing. *The NATS Journal*, v. 44, p. 11–20, 1988.

[27] JULIáN, P. P. de. Modificación o aggiustamento de las vocales españolas en el canto lírico. *Estudios de fonética experimental*, XXV, p. 263–293, 2016. ISSN 2385-3573.

[28] SUNDBERG, J. Research on the singing voice in retrospect. *TMH-QPSR*, v. 45, p. 11–22, 2003.

[29] PALAPARTHI, A.; MAXFIELD, L.; TITZE, I. R. Estimation of source-filter interaction regions based on electroglottography. *Journal of Voice*, Mosby, v. 33, p. 269–276, 5 2019. ISSN 0892-1997.

[30] ECHTERNACH, M.; HERBST, C. T.; KöBERLEIN, M.; STORY, B.; DöLLINGER, M.; GELLRICH, D. Are source-filter interactions detectable in classical singing during vowel glides? *The Journal of the Acoustical Society of America*, AIP Publishing, v. 149, p. 4565–4578, 6 2021. ISSN 0001-4966.

[31] KADIRI, S. R.; YEGNANARAYANA, B. Analysis of singing voice for epoch extraction using zero frequency filtering method. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, Institute of Electrical and Electronics Engineers Inc., v. 2015-August, p. 4260–4264, 8 2015. ISSN 15206149.

[32] DÍAZ, C. La producción de la voz: estructuras anatómicas y biomecánica laríngea. In: *X Congreso Argentino y V Latinoamericano de Educación Física y Ciencias (La Plata, 2013)*. [S.l.: s.n.], 2013.

[33] TITZE, I. R. *Principles of voice production*. 1. ed. [S.l.]: Prentice Hall, 1994.

[34] SUNDBERG, J.; ROSSING, T. D. *The science of singing voice*. [S.l.]: Acoustical Society of America, 1990.

[35] VENTURA, J. A. P. d. S. Biofeedback da voz cantada. 2011.

[36] FANT, G. Acoustic theory of speech production.'sgravenhage: Mouton. *The Netherlands*, 1960.

[37] FANT, G. The source filter concept in voice production. *STL-QPSR*, Citeseer, v. 1, n. 1981, p. 21–37, 1981.

[38] DEGOTTEX, G. *Glottal source and vocal-tract separation*. Tese (Doutorado) — Université Pierre et Marie Curie-Paris VI, 2010.

[39] ROSENBERG, A. E. Effect of glottal pulse shape on the quality of natural vowels. *The Journal of the Acoustical Society of America*, Acoustical Society of America, v. 49, n. 2B, p. 583–590, 1971.

[40] FANT, G.; LILJENCRANTS, J.; LIN, Q.-g. A four-parameter model of glottal flow. *STL-QPSR*, Citeseer, v. 4, n. 1985, p. 1–13, 1985.

[41] DOVAL, B.; D'ALESSANDRO, C.; HENRICH, N. The spectrum of glottal flow models. *Acta acustica united with acustica*, S. Hirzel Verlag, v. 92, n. 6, p. 1026–1046, 2006.

[42] HENRICH, N.; D'ALESSANDRO, C.; DOVAL, B. Glottal flow models: waveforms, spectra and physical measurements. In: *Forum Acusticum*. [S.l.: s.n.], 2002. p. 1.

[43] CATALDO, E.; BAHIANO, D. Stochastic models of glottal pulses from the rosenberg and liljencrants-fant models with unified parameters. *Computer Speech & Language*, Elsevier, p. 1–20, 2021.

[44] BABACAN, O.; DRUGMAN, T.; BERNARDONI, N. H.; DUTOIT, T.; HENRICH, N.; DUTOIT, T. A quantitative comparison of glottal closure instant estimation algorithmson a large variety of singing sounds. p. 1–5, 8 2013.

[45] KRISHNAN, K. S. G.; GOVIND, D. Comparison of glottal closure instant estimation algorithms for singing voices in indian context. *2017 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2017*, Institute of Electrical and Electronics Engineers Inc., v. 2017-January, p. 1447–1452, 11 2017.

[46] KLATT, D. H. Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America*, AIP Publishing, v. 67, p. 971–995, 3 1980. ISSN 0001-4966.

[47] FEUGèRE, L. Synthèse par règles de la voix chantée contrôlée par le geste et applications musicales. Université Pierre et Marie Curie - Paris VI, 9 2013.

[48] CATALDO, E.; SOIZE, C. Stochastic mechanical model of vocal folds for producing jitter and for identifying pathologies through real voices. *Journal of biomechanics*, Elsevier, v. 74, p. 126–133, 2018.

[49] CATALDO, E.; SOIZE, C. Voice signals produced with jitter through a stochastic one-mass mechanical model. *Journal of voice*, Elsevier, v. 31, n. 1, p. 111–e9, 2017.

[50] CATALDO, E.; MONTEIRO, L.; SOIZE, C. A novel source-filter stochastic model for voice production. *Journal of Voice*, Elsevier, 2021.

[51] KENMOCHI, H.; OHSHITA, H. Vocaloid-commercial singing synthesizer based on sample concatenation. *Interspeech*, p. 4009–4010, 2007.

[52] KOB, M.; HENRICH, N.; HERZEL, H.; HOWARD, D.; TOKUDA, I.; WOLFE, J. Analysing and understanding the singing voice: recent progress and open questions. *Current bioinformatics*, Bentham Science Publishers, v. 6, n. 3, p. 362–374, 2011.

[53] LÓPEZ, P. R. F. *Características vocales de la voz cantada de un grupo de estudiantes de canto de música popular contemporánea de la ciudad de Lima.* Tese (Doutorado) — Pontificia Universidad Catolica del Peru-CENTRUM Catolica (Peru), 2019.

[54] APFELBACH, C. S. Tessa: A novel matlab program for automated tessitura analysis. *Journal of Voice*, Elsevier, 2020.

[55] HERTEGARD, S.; GAUFFIN, J.; SUNDBERG, J. Open and covered fiberoptics, inverse singing as studied by means of filtering, and spectral analysis. *J Voice*, v. 4, p. 220–230, 1990.

[56] FACAL, M. L. *La voz del cantante: estudio comparativo del análisis objetivo y subjetivo de la voz hablada y cantada.* [S.l.]: Librería Akadia Editorial, 2005.

[57] FRIČ, M.; PAVLECHOVÁ, A. Listening evaluation and classification of female singing voice categories. *Logopedics Phoniatrics Vocology*, Taylor & Francis, v. 45, n. 3, p. 97–109, 2020.

[58] SOUZA, G. V. S. de; DUARTE, J. M. T.; VIEGAS, F.; SIMÕES-ZENARI, M.; NEMR, K. An acoustic examination of pitch variation in soprano singing. *Journal of Voice*, Elsevier, v. 34, n. 4, p. 648–e41, 2020.

[59] MILLHOUSE, T. J.; CLERMONT, F. Perceptual characterisation of the singer's formant region: a preliminary study. In: *Proceedings of the Eleventh Australian International Conference on Speech Science and Technology*. [S.l.: s.n.], 2006. p. 253–258.

[60] HIRANO, M.; HIBI, S.; HAGINO, S. Physiological aspects of vibrato. *Vibrato*, San Diego, CA: Singular, p. 9–33, 1995.

[61] NESTOROVA, T.; BRANDNER, M.; GINGRAS, B.; HERBST, C. T. Vocal vibrato characteristics in historical and contemporary opera, operetta, and schlager. *Journal of Voice*, Elsevier Inc., v. 0, 2023. ISSN 18734588.

[62] REGNIER, L.; PEETERS, G. Singing voice detection in music tracks using direct voice vibrato detection. In: IEEE. *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.], 2009. p. 1685–1688.

[63] MAGI, C.; POHJALAINEN, J.; BäCKSTRöM, T.; ALKU, P. Stabilised weighted linear prediction. *Speech Communication*, North-Holland, v. 51, p. 401–411, 5 2009. ISSN 0167-6393.

[64] ALKU, P.; POHJALAINEN, J.; VAINIO, M.; LAUKKANEN, A.-M.; STORY, B. H. Formant frequency estimation of high-pitched vowels using weighted linear prediction. *The Journal of the Acoustical Society of America*, Acoustical Society of AmericaASA, v. 134, p. 1295, 8 2013. ISSN 0001-4966.

[65] LIU, L.; SHIMAMURA, T. Pitch-synchronous linear prediction analysis of high-pitched speech using weighted short-time energy function. *Journal of Signal Processing*, Research Institute of Signal Processing, Japan, v. 19, p. 55–66, 3 2015. ISSN 1342-6230.

[66] MA, C.; KAMP, Y.; WILLEMS, L. F. Robust signal selection for linear prediction analysis of voiced speech. *Speech Communication*, v. 12, p. 69–81, 1993.

[67] FREIN, R. D. Power-weighted lpc formant estimation. *IEEE Transactions on Circuits and Systems II: Express Briefs*, Institute of Electrical and Electronics Engineers Inc., v. 68, p. 2207–2211, 6 2021.

[68] AIRAKSINEN, M.; RAITIO, T.; STORY, B.; ALKU, P. Quasi closed phase glottal inverse filtering analysis with weighted linear prediction. *IEEE Transactions on Audio, Speech and Language Processing*, v. 22, p. 596–607, 3 2014. ISSN 15587916.

[69] THOMAS, M. R.; NAYLOR, P. A.; GUDNASON, J. Estimation of glottal closing and opening instants in voiced speech using the yaga algorithm. *IEEE Transactions on Audio, Speech and Language Processing*, v. 20, p. 82–91, 2012.

[70] NAYLOR, P. A.; KOUNOUDES, A.; GUDNASON, J.; BROOKES, M. Estimation of glottal closure instants in voiced speech using the dypsa algorithm. *IEEE Transactions on Audio, Speech and Language Processing*, v. 15, p. 34–43, 1 2007.

[71] ARROABARREN, I.; CARLOSENA, A. Inverse filtering in singing voice: A critical analysis. *IEEE Transactions on Audio, Speech and Language Processing*, v. 14, p. 1422–1431, 7 2006. ISSN 15587916.

[72] GOWDA, D.; AIRAKSINEN, M.; ALKU, P. Quasi-closed phase forward-backward linear prediction analysis of speech for accurate formant detection and estimation. *The Journal of the Acoustical Society of America*, AIP Publishing, v. 142, p. 1542–1553, 9 2017. ISSN 0001-4966.

[73] MURTY, K. S. R.; YEGNANARAYANA, B. Epoch extraction from speech signals. *IEEE Transactions on Audio, Speech and Language Processing*, v. 16, p. 1602–1613, 11 2008. ISSN 15587916.

[74] BARRIENTOS, E.; CATALDO, E. Synthesis of sung spanish vowels in lyrical singing by sopranos. *IEEE Latin America Transactions*, v. 19, 2021. ISSN 15480992.

[75] ROUBEAU, B.; HENRICH, N.; CASTELLENGO, M. Laryngeal vibratory mechanisms: The notion of vocal register revisited. *Journal of Voice*, Mosby, v. 23, p. 425–438, 7 2009. ISSN 0892-1997.

[76] WEISS, R.; JR, W. B.; MORIS, J. Singer's formant in sopranos: fact or fiction? *Journal of Voice*, Elsevier, v. 15, n. 4, p. 457–468, 2001.

APÊNDICE A – Sintetizador não interativo usando o modelo do pulso glotal de Rosenberg

```
clear;
clc;
%-------------------------------------------------------------
Av=1; %Maximum amplitude of the Glottal Flow GFM
Tfs=7;     %Sustained Phonation Time
alpha1=58; %Porcentagen do tempo de apertura relativa 58
alpha2=20; %Porcentagen do tempo de fechamento relativa 20
nfft=1024;
fs=16000;  %sample frequency
alphaT=101; %Total alpha initialized to an exceeded value
%-------------------------------------------------------------
%Vocal Tract
%-------------------------------------------------------------
filename='E5_S1.xlsx';
name=strtok(filename, '.');
formants = xlsread(filename,'C3:G7');
bdw= xlsread(filename,'J3:N7');
F0_s= xlsread(filename,'B3:B7');
H= input('Enter the vowel 1(/a/),2(/e/),3(/i/),4(/o/),5(/u/):');
fmts=formants(H,:);
bandwidths=bdw(H,:);
F0=F0_s(H); %Fundamental Frequency
%-------------------------------------------------------------
T0=1/F0; %Fundamental Period
Tp=(alpha1/100)*T0; %Glottal opening time
TN=(alpha2/100)*T0; %Glottal closing time
```

```
%-------------------------------------------------------------
%Rosenberg's Glottal Pulse Approximation
%-------------------------------------------------------------
Ts=1/fs;
t1=0:Ts:Tp;
t2=(Tp+Ts):Ts:(Tp+TN);
t3=(Tp+TN+Ts):Ts:T0;
t=[t1 t2 t3];
%-------------------------------------------------------------
x1=cos(pi*t1/Tp);
gn1=(Av/2)*(1-x1);
%-------------------------------------------------------------
x2=pi*(t2-Tp);
gn2=Av*cos(x2/(2*TN));
%-------------------------------------------------------------
gn3=zeros(1,length(t3));
gn=[gn1 gn2 gn3];
%-------------------------------------------------------------
%Rosenberg's Glottal Pulse Derivate.
%-------------------------------------------------------------
udg=diff(gn)/Ts;
%-------------------------------------------------------------
%Comb Dirac
%-------------------------------------------------------------
bin=fix((Tfs-T0)/(T0));
LD=(length(gn)*bin)+2;
Dirac=zeros(LD,1);
Dirac(1:length(gn):end)=1;
%-------------------------------------------------------------
%Convolution of Glottal Flow and Dirac
%-------------------------------------------------------------
yk=conv(gn,Dirac);
%-------------------------------------------------------------
%Voice signal modulation method
%-------------------------------------------------------------
```

```
%Tsn=(length(yk)-1)*Ts;
%m1=0:Ts:Tsn;
%m2=sin(pi*(1/Tsn)*m1);
%m2=m2(:);
%yk=yk.*m2;
%---------------------------------------------------------------
%The vibrato effect is created
%---------------------------------------------------------------
rate=4.5; %Vibrato rate in(Hz, use the range 2-5 Hz
Depth=0.0004; %Delay-line width, use the range 0.5-2 ms
vib= vibrato(yk, fs, rate, Depth);
%---------------------------------------------------------------
% Generate vowel impulse response
%---------------------------------------------------------------
if F0 > fmts(1)
   msg = msgbox('F0 > F1','Warning','warn');
end
xin=[1 zeros(1,499)];
    for resonance=1:5
        f=fmts(resonance);
        bw=bandwidths(resonance);
        num=1 -2*exp(-bw*2*pi/fs)*cos(2*pi*f/fs)+ exp(-4*bw*pi/fs);
        den=[1 -2*exp(-bw*2*pi/fs)*cos(2*pi*f/fs) exp(-4*bw*pi/fs)];
        yout=filter(num, den, xin);
        xin=yout;
    end
%---------------------------------------------------------------
% Approximation of radiation effects on lips
%---------------------------------------------------------------
r=0.95;
B1=[1 -r];
A1=[1];
yout=filter(B1, A1, yout);
%---------------------------------------------------------------
zout=conv(yout,vib);
```

```
%-------------------------------------------------------------
%The tremolo effect is created
%-------------------------------------------------------------
Fc =4.5;
alpha_t =0.002;
trem= tremolo(zout, Fc, fs, alpha_t);
ys=trem;
%-------------------------------------------------------------
%Normalization of the voice signal
%-------------------------------------------------------------
maximo=max(abs(ys));
ys=ys/maximo;
sound(ys,fs)
%-------------------------------------------------------------
%Plot of Glottal Pulse
%-------------------------------------------------------------
figure(1)
subplot(2,1,1);
plot(t*1000,gn,'g');
title(['Forma de onda do pulso glotal de Rosenberg']);
xlabel({'Time (ms)'})
ylabel('Glottal flow (cm^3/s)')
hold on;
%-------------------------------------------------------------
%plot Rosenberg's Glottal Pulse Derivate
%-------------------------------------------------------------
subplot(2,1,2);
plot(t(2:length(t))*1000,udg,'k');
title(['Forma de onda da derivada do pulso glotal de Rosenberg']);
xlabel({'Time (ms)'})
ylabel('Derivative glottal flow (cm^3/s^2)')
hold on;
%-------------------------------------------------------------
%Recording a sung vogal
%-------------------------------------------------------------
```

```
fname=strcat('i_Rsbg_',num2str(name),'.wav');
wavwrite(ys,fs,fname);
```

APÊNDICE B – Sintetizador não interativo usando o pulso glotal com o modelo de LF

```
clear;
clc;
%-------------------------------------------------------------------
Av=1;   %Maximum amplitude of the Glotal Flow GFM(Tp)
Tfs=7;  %Sustained phonation time
nfft=1024; %The n-point
Oq=0.78; %Open quotient(ranging 0 < Oq < 1)
m=0.5;  %Asymmetry coefficient (ranging between 0.5 (symmetrical)
        %and 1 (So asymmetrical)
Qa=0.2; %Return phase coefficient (ranging 0.1 < Qa < 0.7)
%-------------------------------------------------------------------
%Vocal Tract Parameters
%-------------------------------------------------------------------
filename='C6_S5.xlsx';
name=strtok(filename, '.');
formants = xlsread(filename,'C3:G5');
bdw= xlsread(filename,'J3:N5');
F0_s= xlsread(filename,'B3:B5');
H= input('Enter the vowel 1(/a/),2(/i/),3(/u/):');
fmts=formants(H,:);
bandwidths=bdw(H,:);
F0=F0_s(H); %Fundamental Frequency
T0=1/F0; %Fundamental Period
%-------------------------------------------------------------------
%The parameter Epsilon (e) is defined by an implicit equation
%-------------------------------------------------------------------
```

```
syms e
pe= solve ( e*Qa*(1-Oq)*T0 == 1-exp(-e*(1-Oq)*T0),e);
E=double(pe);
%-----------------------------------------------------------------------
%The parameter (a) is defined by an implicit equation
%-----------------------------------------------------------------------
syms a
x0=pi/(m*Oq*T0);
x1=((a^2)+(x0^2));
x2=exp(-a*Oq*T0);
x3=cos(pi/m);
x4=sin(pi/m);
x5=T0*(1-Oq);
x6=exp(E*T0*(1-Oq));
x7=1/E;
pa=solve((1/x1)*(a+x0*((x2-x3)/x4))==(x5/(x6-1))-x7,a);
A=double(pa);
%-----------------------------------------------------------------------
%The parameter (Ee) is obtained by to solve the next equation:
%-----------------------------------------------------------------------
x8= exp(-A*Oq*T0);
x9=sin(pi/m);
x10=pi/(m*Oq*T0);
x11=(A^2)+(x10^2);
x12=x10*exp(A*m*Oq*T0);
Ee=(-Av*x9*x11)/(x8*(x10+x12));
%-----------------------------------------------------------------------
%Glottal Flow LF Model--First part
%-----------------------------------------------------------------------
fs=16000; % Sample rate
Ts=1/fs;
t1=0:Ts:Oq*T0;
y1= Ee*exp(-A*Oq*T0);
y2=sin(pi/m)*(A^2+(pi/(m*Oq*T0))^2);
y3=pi/(m*Oq*T0);
```

```
y4=A.*exp(A.*t1).*sin(y3.*t1);
y5=y3.*exp(A.*t1).*cos(y3.*t1);
Ug1=-(y1/y2).*(y3+y4-y5);
%------------------------------------------------------------------------
%Glottal Flow LF Model--Second part
%------------------------------------------------------------------------
t2=(Oq*T0)+Ts:Ts:T0;
y6=1/(E*Qa*(1-Oq)*T0);
y7=(1-exp(E*(T0-t2)))./E;
Ug2=-Ee*(y6-1)*(T0-t2+y7);
%------------------------------------------------------------------------
%Derivate by Nathalie--First part
%------------------------------------------------------------------------
z1=exp(A.*(t1-Oq*T0));
z2=sin(y3.*t1);
z3=sin(pi/m);
UGD1=-Ee.*z1.*(z2./z3);
%------------------------------------------------------------------------
%Derivate by Nathalie--Second part
%------------------------------------------------------------------------
z4=1/(E*Qa*(1-Oq)*T0);
z5=exp(E*Oq*T0);
z6=(exp(-E.*t2)-exp(-E*T0));
UGD2=-(Ee*z4)*z5.*z6;
%------------------------------------------------------------------------
%Glottal Flow LF Model and Derivate
%------------------------------------------------------------------------
ug=[Ug1 Ug2];
UGD=[UGD1 UGD2];
ugd=diff(ug)/Ts;
t=[t1 t2];
%------------------------------------------------------------------------
%Comb Dirac
%------------------------------------------------------------------------
bin=fix((Tfs-T0)/(T0));
```

```
LD=(length(ug)*bin)+2;
Dirac=zeros(LD,1);
Dirac(1:length(ug):end)=1;
%-------------------------------------------------------------------------
%Convolution of Glottal Flow and Dirac
%-------------------------------------------------------------------------
pulse_train=conv(ug,Dirac);
%-------------------------------------------------------------------------
%Variations of the Subglottal Pressure is created
%-------------------------------------------------------------------------
% Tsn=(length(pulse_train)-1)*Ts;
% m1=0:Ts:Tsn;
% m2=sin(pi*(1/Tsn)*m1);
% m2=m2(:);
% pulse_train=pulse_train.*m2;
%-------------------------------------------------------------------
%The vibrato effect is created
%-------------------------------------------------------------------------
rate=4.5; % Vibrato rate in(Hz, use the range 2-5 Hz
Depth=0.0004; %Delay-line width, use the range 0.5-2 ms
vib= vibrato(pulse_train, fs, rate, Depth);
%-------------------------------------------------------------------------
% Generate impulse response
%-------------------------------------------------------------------------
if F0 > fmts(1)
    msg = msgbox('F0 > F1','Warning','warn');
end
xin=[1 zeros(1,499)];
    for resonance=1:5
        f=fmts(resonance);
        bw=bandwidths(resonance);
        num=1 -2*exp(-bw*2*pi/fs)*cos(2*pi*f/fs)+ exp(-4*bw*pi/fs);
        den=[1 -2*exp(-bw*2*pi/fs)*cos(2*pi*f/fs) exp(-4*bw*pi/fs)];
        yout=filter(num, den, xin);
        xin=yout;
```

```
    end
%-------------------------------------------------------------------------
%Radiation Model
%-------------------------------------------------------------------------
%Aproximação aos efeitos de radiação nos lábios
%-------------------------------------------------------------------------
r=0.95;
B1=[1 -r];
A1=[1];
yout=filter(B1, A1, yout);
%-------------------------------------------------------------------------
% convolve vowel impulse response with excitation
%-------------------------------------------------------------------------
zout=conv(yout,vib);
%-------------------------------------------------------------------------
%The tremolo effect is created
%-------------------------------------------------------------------------
Fc =4.5;
alpha_t =0.002;
trem= tremolo(zout, Fc, fs, alpha_t);
ys=trem;
%-------------------------------------------------------------------------
%normalizing the output signal
%-------------------------------------------------------------------------
maximo=max(abs(ys));
ys=ys/maximo;
sound(ys,fs)
%-------------------------------------------------------------------------
% convert from frequency and bandwidth to all-pole digital transfer
% function coefficients
%-------------------------------------------------------------------------
zk=exp(-2*pi*bandwidths/fs);
fkn=cos(2*pi*fmts/fs);
%-------------------------------------------------------------------------
% evaluate vocal tract log magnitude response; save in HT
```

```
    HR=[];
    HT=zeros(1,nfft);
    HP=(complex(ones(1,nfft),zeros(1,nfft)))';
    M=5;
    for k=1:M
        B1=1-2*zk(k)*fkn(k)+zk(k)*zk(k);
        A1=[1 -2*zk(k)*fkn(k) zk(k)*zk(k)];
        [Tr,W]=freqz(B1,A1,nfft,'whole',fs);
        HP=HP.*Tr;
        HL=20*log10(abs(Tr));
        HT=HT+HL';
    end
%-------------------------------------------------------------------
%Figure of Glottal Pulse
%-------------------------------------------------------------------
figure (1)
plot(t*1000,ug);
title(['Forma de onda do pulso glotal de LF']);
xlabel({'Time (ms)'})
ylabel('Amplitude (arbitrary units)')
%-----------------------------------------------------------------
%Figure of Glottal Signal with vibrato
%-----------------------------------------------------------------
Tx=[0:Ts:(Ts*(100*length(ug)))-Ts];
xk=vib(1:100*length(ug));
plot(Tx*1000,xk,'k');
title(['Forma de onda do pulso glotal de LF']);
xlabel({'Time (ms)'})
ylabel('Amplitude (arbitrary units)')
%-----------------------------------------------------------------
%Figure Nathalie's Glottal Pulse Derivate
%-------------------------------------------------------------------
plot(t*1000,UGD);
title(['Forma de onda da derivada do pulso glotal de LF']);
xlabel({'Time (ms)'})
```

```
ylabel('Amplitude (arbitrary units)')
%-------------------------------------------------------------------------
%Figure vocal tract log magnitude response, HT.
%-------------------------------------------------------------------------
plot(W(1:nfft/2+1),HT(1:nfft/2+1),'r--','LineWidth',2);
xlim([0 6e3])
xlabel('Frequência (Hz)'),ylabel('Magnitude Log (dB)');
grid on;
hold on;
%-----------------------------------------------------------------
%Gravando a vogal cantada
%-----------------------------------------------------------------
fname=strcat(num2str(name),'_u','.wav');
wavwrite(ys,fs,fname);
```

APÊNDICE C – Sintetizador interativo usando o pulso glotal com o modelo de LF

```
clear;
clc;
%-------------------------------------------------------------------------
filename='06_SR2_a.mat';%filename is .mat file
name=strtok(filename, '.');
% Precise Resonance Parameter Estimation using ZFF and WLP-HPSV
[fmts,bandwidths,Note,gci,goi,fs,f0_mean]=glottal(filename);
% Estimate Open Phase (OP) and Close Phase (CP)
%of the glottal signal from the soprano's sung vowel
CP=[];
OP=[];
for i=1:length(goi)
    CP(i)=goi(i)-gci(i);
    if gci(i)==gci(end)
       break;
    else
       OP(i)=gci(i+1)-goi(i);
    end
end
CP=CP(2:end);
NO=[];
for j=1:length(goi)
    if goi(j)==goi(end)
        break;
    end
    NO(j)=goi(j+1)-goi(j);
```

```
end
T0=N0/fs; %Fundamental Period
Oq=[];%Open quotient(ranging 0 < Oq < 1)
TP=[];
for i=1:length(CP)
    TP(i)=CP(i)+OP(i);
    Oq(i)=OP(i)./TP(i);
end
%-------------------------------------------------------------------
Av=1;   %Maximum amplitude of the Glotal Flow GFM(Tp)
Tfs=7;  %Sustained phonation time
nfft=1024; %The n-point
m=0.5;  %Asymmetry coefficient
Qa=0.2; %Return phase coefficient
F0=1./T0;%Fundamental Frequency
%Glottal signal
Gs=[];
for j=1:length(Oq)
    [ug]=LF_pulse(fs,Av,m,Qa,T0(j),Oq(j));
    Gs=[Gs ug];
end
%-------------------------------------------------------------------
%Comb Dirac - used to adjust the glottal signal duration to match
%the phoneme's duration Tfs
%-------------------------------------------------------------------
%T0_mean=1/f0_mean;
Tmed=length(Gs)/fs;
bin=fix((Tfs-Tmed)/(Tmed));
LD=(length(Gs)*bin)+2;
Dirac=zeros(LD,1);
Dirac(1:length(Gs):end)=1;
%-------------------------------------------------------------------
%Convolution of Glottal Flow and Dirac
%-------------------------------------------------------------------
% pulse_train=conv(Gs,Dirac);
```

```
vib=conv(Gs,Dirac);
%-------------------------------------------------------------------------
%The vibrato effect is created
%-------------------------------------------------------------------------
rate=4.5; %Vibrato rate in(Hz, use the range 2-5 Hz
Depth=0.0004; %Delay-line width (milliseconds) value
vib= vibrato(pulse_train, fs, rate, Depth);
%-------------------------------------------------------------------------
%Area glottal
%-------------------------------------------------------------------------
Ag=[];
for j=1:length(OP)
    [A]=glot_area(fs,T0(j),OP(j));
    Ag=[Ag A];
end
Ts=1/fs;
An=diff(Ag)/Ts;
%-------------------------------------------------------------------------
% Generate impulse response
%-------------------------------------------------------------------------
if F0 > fmts(1)
    msg = msgbox('F0 > F1','Warning','warn');
end
alph=0.1;
Anp=(An*alph)+1;
x=abs(Anp);
Fc1=fmts(1)*sqrt(x);

xin=[1 zeros(1,499)];
    for resonance=1:5
        if resonance==1
           f=Fc1;
        else
           f=fmts(resonance);
        end
```

```
        bw=bandwidths(resonance);
        num=1 -2*exp(-bw*2*pi/fs)*cos(2*pi*f/fs)+ exp(-4*bw*pi/fs);
        den=[1 -2*exp(-bw*2*pi/fs)*cos(2*pi*f/fs) exp(-4*bw*pi/fs)];
        yout=filter(num, den, xin);
        xin=yout;

    end
%-----------------------------------------------------------------------
%Metodo2% Aproximação aos efeitos de radiação nos lábios
%-----------------------------------------------------------------------
r=0.95;
B1=[1 -r];
A1=[1];
yout=filter(B1, A1, yout);
%-----------------------------------------------------------------------
% convolve vowel impulse response with excitation
%-----------------------------------------------------------------------
zout=conv(yout,vib);
ys=conv(yout,vib);
%-----------------------------------------------------------------------
%The tremolo effect is created
%-----------------------------------------------------------------------
Fc =4.5;
alpha_t =0.002;
trem= tremolo(zout, Fc, fs, alpha_t);
ys=trem;
%-----------------------------------------------------------------------
%normalizing the output signal
%-----------------------------------------------------------------------
maximo=max(abs(ys));
ys=ys/maximo;
sound(ys,fs)
%-------------------------------------------------------------
%Recording the sung vowel
%-------------------------------------------------------------
```

```
fname=strcat(num2str(name),'.wav');
wavwrite(ys,fs,fname);
```

Onde glottal(), corresponde a uma função que estima os parametros glotais e de ressonância do trato vocal pelo ZFF modificado e o método WLP-HPSV dado pelo seguinte código:

```
function[fmts,bws,Note,gci,goi,Fs,f0_mean]=glottal(filename)

%%% script for using the Weighted Linear Prediction adapted to the
%%% High-Pitched Singing Voices (WLP-HPSV) %%%%%%

load(filename) % Load up frame, with Fs = 12000 Hz

wl=0.0750*Fs; % Length of frame of 75 ms(in samples)
p = 10; % LP order

s=frame(1:wl);%Singing voice signal to analysis (frame 75ms)
[gci,N0] = f_ZFF(s,Fs); %GCI estimation

dy_cpfrac=0.2;%presumed closed phase fraction of larynx cycle
goi=simplegci2goi(gci,dy_cpfrac);

[Note, f0_mean]=Music_notes(N0,Fs); %Musical note sung by the soprano

%% Use this to manually select PQ and DQ values
DQ=0.3; %DQ= (T1/T0)*100%
PQ=0.0; %PQ= (T2/T0)*100%
wn = makeW(s,p,DQ,PQ,gci,f0_mean,Fs);

%% The pre-emphasis filter is a high pass filter

x2 = filter([1 -0.98],1,s); % Pre-emphasis H(z)=1-azˆ-1
h = hamming(wl);
%% Section 6: LP analisys by the autocorrelation criterion. Using v_lpcauto_2.m
```

```
a=wlp(h.*x2,wn,p);
a=fixzeroes(a,0,Fs); % Check and fix to guarantee stabile all-pole filter
[fmts,bws]=formants(a,Fs);
end
```

APÊNDICE D – Algoritmo para a síntese de uma sequência de vogais em diferentes notas musicais

```
clc;
clear;
close all;
%-------------------------------------------------------------------
Av=1;   %Maximum amplitude of the Glotal Flow GFM(Tp)
Oq=0.78; %Open quotient
m=0.5;  %Asymmetry coefficient
Qa=0.2; %Return phase coefficient
q=1;     %q=1 generates a deterministic sound
fs=16000;% sampling frequency
ys=[]; % Sung voice signal
vow='';
mus='';
%-------------------------------------------------------------------
No=['G5,A5 ou B5';'G5,A5 ou B5';'D5,E5 ou F5';
    'E5,F5 ou G5';'A5,B5 ou C6';'E5,F5 ou G5';
    'F5,G5 ou A5'];
S=input(['Digite a cantora','1(S1),2(S2),3(S3),4(S4),5(S5),6(S6),7(S7):']);
A=No(S,:);
tons=input('Digite o numero de tons (entre 1 e 3 tons): ');
%-------------------------------------------------------------------
for i=1:tons

    F=input(['Digite o tom ',A,' entre aspas:']);
```

```
    filename=strcat(F,'_','S',num2str(S),'.xlsx');
    formants = xlsread(filename,'C3:G5');
    bdw= xlsread(filename,'J3:N5');
    F0_s= xlsread(filename,'B3:B5');
    H= input('Enter the vowel 1(/a/),2(/i/),3(/u/):');
    fmts=formants(H,:);
    bandwidths=bdw(H,:);
    F0=F0_s(H); %Fundamental Frequency
    T0=1/F0; %Fundamental Period
    Tfs=input('Digite a duração do fonema:');
    [vib]= LF_main(Tfs,Av,Oq,m,Qa,fs,F0,q);
%-------------------------------------------------------------------------
    % Generate impulse response
%-------------------------------------------------------------------------
    xin=[1 zeros(1,499)];

        for resonance=1:5
            f=fmts(resonance);
            bw=bandwidths(resonance);
            num=1 -2*exp(-bw*2*pi/fs)*cos(2*pi*f/fs)+ exp(-4*bw*pi/fs);
            den=[1 -2*exp(-bw*2*pi/fs)*cos(2*pi*f/fs) exp(-4*bw*pi/fs)];
            yout=filter(num, den, xin);
            xin=yout;
        end
%-------------------------------------------------------------------------
    %Aproximação aos efeitos de radiação nos lábios
%-------------------------------------------------------------------------
    r=0.95;
    B1=[1 -r];
    A1=[1];
    yout=filter(B1, A1, yout);
%-------------------------------------------------------------------------
    % convolve vowel impulse response with excitation
%-------------------------------------------------------------------------
    zout=conv(yout,vib);
```

```
%-------------------------------------------------------------------------
    %The tremolo effect is created
%-------------------------------------------------------------------------
    Fc =4.5;
    alpha_t =0.002;
    trem= tremolo(zout, Fc, fs, alpha_t);
%-------------------------------------------------------------------------
    %normalizing the output signal
%-------------------------------------------------------------------------
    maximo=max(abs(trem));
    ys1=trem/maximo;
%-------------------------------------------------------------------------
    %Variations of the Subglottal Pressure is created
%-------------------------------------------------------------------------
    Ts=1/fs;
%-------------------------------------------------------------------------
    cx=length(ys1);
    mx=ones(cx,1);
    Tsn=(cx*0.001)*Ts;
    mx1=0:Ts:Tsn;
    mx2=sin(pi*(1/Tsn)*mx1);
    mx2=mx2(:);
    cx1=length(mx2);
    mx(1:cx1)=mx2;
    mx((cx-cx1+1):cx)=mx2;
%-------------------------------------------------------------------------
    ys2=ys1.*mx;
%-------------------------------------------------------------------------
    ysc=zeros(40,1);
    ys=[ys; ysc; ys2];
%-------------------------------------------------------------------------
    if H==1
       vow1='a';
    elseif H==2
       vow1='i';
```

```
    else
       vow1='u';
    end
    vow=[vow vow1];
    mus=[mus F '_'];
end

sound(ys,fs)
%-------------------------------------------------------------------------
%Recording the sung vowel
%-------------------------------------------------------------------------
fname=strcat('S',num2str(S),'_',mus,vow,'.wav');
wavwrite(ys,fs,fname);
%-------------------------------------------------------------------------
```

Onde LF_main(),corresponde a uma função que gera o sinal glotal pelo modelo de LF e com vibrato dado o seguinte código:

```
function [vib]= LF_main(Tfs,Av,Oq,m,Qa,fs,F0,q)

clc;
T0=(1/F0)*q;  %Período Fundamental with some jitter
%-------------------------------------------------------------------------
%The parameter Epsilon (e) is defined by an implicit equation
%-------------------------------------------------------------------------
syms e
pe= solve ( e*Qa*(1-Oq)*T0 == 1-exp(-e*(1-Oq)*T0),e);
E=double(pe);
%-------------------------------------------------------------------------
%The parameter (a) is defined by an implicit equation
%-------------------------------------------------------------------------
syms a
x0=pi/(m*Oq*T0);
x1=((a.^2)+(x0.^2));
x2=exp(-a*Oq*T0);
```

```
x3=cos(pi/m);
x4=sin(pi/m);
x5=T0*(1-Oq);
x6=exp(E*T0*(1-Oq));
x7=1/E;

pa=solve((1/x1)*(a+x0*((x2-x3)/x4))==(x5/(x6-1))-x7,a);
A=double(pa);
%-------------------------------------------------------------------------
%The parameter (Ee) is obtained by to solve the next equation:
%-------------------------------------------------------------------------
x8= exp(-A*Oq*T0);
x9=sin(pi/m);
x10=pi/(m*Oq*T0);
x11=(A^2)+(x10^2);
x12=x10*exp(A*m*Oq*T0);
Ee=(-Av*x9*x11)/(x8*(x10+x12));
%-------------------------------------------------------------------------
%Glottal Flow LF Model--First part
%-------------------------------------------------------------------------
Ts=1/fs; % fs=Sample rate
t1=0:Ts:Oq*T0;
y1= Ee*exp(-A*Oq*T0);
y2=sin(pi/m)*(A^2+(pi/(m*Oq*T0))^2);
y3=pi/(m*Oq*T0);
y4=A.*exp(A.*t1).*sin(y3.*t1);
y5=y3.*exp(A.*t1).*cos(y3.*t1);
Ug1=-(y1/y2).*(y3+y4-y5);
%-------------------------------------------------------------------------
%Glottal Flow LF Model--Second part
%-------------------------------------------------------------------------
t2=(Oq*T0)+Ts:Ts:T0;

y6=1/(E*Qa*(1-Oq)*T0);
y7=(1-exp(E*(T0-t2)))./E;
```

```
Ug2=-Ee*(y6-1)*(T0-t2+y7);
%-------------------------------------------------------------------------
%Derivate by Nathalie--First part
%-------------------------------------------------------------------------
z1=exp(A.*(t1-Oq*T0));
z2=sin(y3.*t1);
z3=sin(pi/m);
UGD1=-Ee.*z1.*(z2./z3);
%-------------------------------------------------------------------------
%Derivate by Nathalie--Second part
%-------------------------------------------------------------------------
z4=1/(E*Qa*(1-Oq)*T0);
z5=exp(E*Oq*T0);
z6=(exp(-E.*t2)-exp(-E*T0));
UGD2=-(Ee*z4)*z5.*z6;
%-------------------------------------------------------------------------
%Glottal Flow LF Model and Derivate
%-------------------------------------------------------------------------
ug=[Ug1 Ug2];
UGD=[UGD1 UGD2];
ugd=diff(ug)/Ts;
t=[t1 t2];
%-------------------------------------------------------------------------
%Comb Dirac
%-------------------------------------------------------------------------
bin=fix((Tfs-T0)/(T0));
LD=(length(ug)*bin)+2;
Dirac=zeros(LD,1);
Dirac(1:length(ug):end)=1;
%-------------------------------------------------------------------------
%Convolution of Glottal Flow and Dirac
%-------------------------------------------------------------------------
pulse_train=conv(ug,Dirac);
%-------------------------------------------------------------------------
%The vibrato effect is created
```

```
%-----------------------------------------------------------------------
rate=4.5; %Vibrato rate in(Hz, use the range 2-5 Hz
Depth=0.0001; %Delay-line width (milliseconds) value
vib= vibrato(pulse_train, fs, rate, Depth);
end
```

APÊNDICE E – Algoritmo do vibrato

```
close all; clear; clc;
%Basic delay of input sample in sec
Delay=Width;

%Basic delay in samples
DELAY=round(Delay*SAMPLERATE);

%Modulation width in samples
WIDTH=round(Width*SAMPLERATE);
if WIDTH>DELAY
   error('de1ay greater than basic delay');
   return;
end

%Modulation frequency in samples
MODFREQ=Modfreq/SAMPLERATE;

% # of samples in WAV-file
LEN=length(x) ;

% Length of the entire delay
L=2+DELAY+WIDTH*2;

% Memory allocation for delay
Delayline=zeros(L,1);

% Memory allocation for output vector
```

```
y=zeros(size(x)) ;

for n= 1:(LEN-1)
   M=MODFREQ;
   MOD=sin(M*2*pi*n);
   ZEIGER= 1+DELAY+WIDTH*MOD;
   i= floor(ZEIGER);
   frac=ZEIGER-i;
   Delayline= [x(n);Delayline(1:L-1)];

%Linear Interpolation
y(n,1)=Delayline(i+1)*frac+Delayline(i)*(1-frac);
```

APÊNDICE F – Algoritmo para estimar frequência fundamental pelo método de autocorrelação

```
function [F0_Ac_mean]=period (data,Fs)
%% Fundamental frequency estimation
%% Section 1: Read and window the analyzed sample:
%1.1. Read the audio file and sampling rate
%[data, Fs] = audioread('u_C5_S2.WAV');%
%info=audioinfo('u_C5_S2.WAV');%'B3_S2_a.WAV'

%1.2. Resample the audio signal to the desired frequency (16kHz).
% Fs_new=16000;
% data_new=resample(data,Fs_new,Fs);

%1.3.Split the data sequence into windows. Use windowing.m
frame_length = round(0.03*Fs); %In samples for 30ms
hop_size = round(0.0125*Fs); % In samples for 25% overlap
window_types = {'rect','hann','cosine','hamming','Gaussian'};
frame_matrix = windowing(data, frame_length, hop_size, window_types{4});
%%Section 2:Fundamental frequency estimation with the autocorrelation method

%2.1. Define minimum and maximum values for the F0 search range, and the
%threshold value for Voiced/Unvoiced decision.
f0_max = 1048; % In Hz (C6=1048 Hz)
f0_min = 131; % In hz (C3=131 Hz)
vuv_threshold_ac = 0.0;
```

```
%2.2.Write a loop through frame_matrix that calls the function autocorr
%to obtain the F0 estimates for each frame

f0vec_ac = zeros(1,size(frame_matrix,2));    % Allocate f0 vector for
                                             % autocorr method
ac_peak_vec = zeros(1,size(frame_matrix,2)); % Allocate ac peak vector

for iFrame = 1:size(frame_matrix,2)
    [f0vec_ac(iFrame), ac_peak_vec(iFrame)]= autocorr(frame_matrix(:,iFrame),
    Fs_new, f0_min, f0_max, vuv_threshold_ac);
end

F0_Ac_mean=mean(f0vec_ac);
```

APÊNDICE G – Algoritmo para determinar a nota musical do tom cantado pela Soprano

```
close all; clear; clc;

T0=N0*(1/Fs);
F0=1./T0;

%Block used to discard any infinite value of fo
ind=isinf(F0);
for k=1:length(F0)
    if ind(k)==0
       ind(k)=1;
    else
        ind(k)=0;
    end
end

% Returns the standard deviation of the elements of fo
f0=F0.*ind;
s=std(f0,'omitnan');
f0=mean(f0, 'omitnan');

oct_0=[0,0,0,0,0,0,0,0,0,28,29,31];
oct_1=[33,35,37,39,41,44,46,49,52,55,58,62]; %Primeira oitava
oct_2=[65,69,73,78,82,87,92,98,104,110,117,123]; %Segunda oitava
oct_3=[131,139,147,156,167,175,185,196,208,220,233,247]; %Terça oitava
```

```
oct_4=[262,277,294,311,330,349,370,392,415,440,466,494]; %Quarta oitava
oct_5=[523,554,587,622,659,698,740,784,831,880,932,988]; %Quinta oitava
oct_6=[1047,1109,1175,1245,1319,1480,1568,1661,1760,1865,1976]; %Sexta oitava
oct_7=[2093,2218,2349,2489,2637,2794,2960,3136,3322,3520,3729,3951]; %Setima oitava
oct_8=[4186,0,0,0,0,0,0,0,0,0,0,0];

octs=[oct_0,oct_1,oct_2,oct_3,oct_4,oct_5,oct_6,oct_7,oct_8];
x=find(octs);
u_notes=octs(x);
notes_1=u_notes(1:(length(u_notes)-1));
notes_2=u_notes(2:length(u_notes));

notes=[notes_1;notes_2];

val_means=mean(notes);

freqs=[u_notes(1),val_means,u_notes(end)];
mus_0={'A0','A#0','B0'};
mus_1={'C1','C#1','D1','D#1','E1','F1','F#1','G1','G#1','A1','A#1','B1'};
mus_2={'C2','C#2','D2','D#2','E2','F2','F#2','G2','G#2','A2','A#2','B2'};
mus_3={'C3','C#3','D3','D#3','E3','F3','F#3','G3','G#3','A3','A#3','B3'};
mus_4={'C4','C#4','D4','D#4','E4','F4','F#4','G4','G#4','A4','A#4','B4'};
mus_5={'C5','C#5','D5','D#5','E5','F5','F#5','G5','G#5','A5','A#5','B5'};
mus_6={'C6','C#6','D6','D#6','E6','F6','F#6','G6','G#6','A6','A#6','B6'};
mus_7={'C7','C#7','D7','D#7','E7','F7','F#7','G7','G#7','A7','A#7','B7'};
mus_8='C8';

mus=[mus_0,mus_1,mus_2,mus_3,mus_4,mus_5,mus_6,mus_7,mus_8];

mn=0;
j=2;
for i=1:length(mus)-1
    if f0>=freqs(1) && f0<freqs(j)
       mn=mus(i);
       break;
```

```
    else
       j=j+1;
    end
end
```

APÊNDICE H – Algoritmo do método ZFF modificado

```
%1 Read the audio file and sampling rate
info = audioinfo('02_SR1_i.wav');
[ssn, fs]=audioread('02_SR1_i.wav');

%2 Resample the audio signal to the desired frequency (16kHz).
Fs=10000;
Ts=1/Fs;
sn=resample(ssn,Fs,fs);

%1.The speech signal s[n] is differentiated
xn=sn-delayseq(sn,1);

%2.The differential signal is passed through a cascade of three zero
%frequency resonators.

b=1;
a=[1 -6 15 -20 15 -6 1];

xin=[1];% zeros(1,200)
y0 = filter(b,a,xin);
y0=conv(xn,y0,'same');

%3.Calculation of the period of the average tone for segments
%of 30ms of xn.
%The period the tone is estimated by the autocorrelation method
```

```
T0=period(xn,Fs);

%4. The trend in y0 is removed
yn=trend_x(y0,Fs,T0);

%5.Checking the positive polarity of the signal.
yn_p=polarity(yn);

%6. Zero crossing instants are extracted.

[gcix,s]=v_zerocros(yn_p,'b');

%7. The NPZCs are selected from the gcix vector considering that
%a PNZC cannot be too close to an NPZC.

cros=zeros(length(gcix),2);

for j=1:length(s)
    if s(j)<0
       s(j)=0;
    else
       s(j)=1;
    end
end
cros(:,1)=gcix;
cros(:,2)=s;

% The gcis function is used to discriminate false zero crossings.

gci=gcis(cros, T0, Fs);

%8. Code block used to extract the fundamental period
gci2=gci;
N0=zeros(1,length(gci));
for i=2:length(gci)
```

```
    N0(i-1)=gci2(i)-gci(i-1);
end
gci=gci.';
end
```

APÊNDICE I – Algoritmo do método WLP-HPSV para análise exaustivo de parâmetros QCP

```
%%% script for using the Weighted Linear Prediction adapted to the
%%% High-Pitched Singing Voices (WLP-HPSV) %%%%%%
%%% Script used to SYNTHETIC VOWELS

close all; clear; clc;

filename='10_SR6_u.mat';
name=strtok(filename, '.');
load(filename) % Load up frame, with Fs = 12000 Hz

% 2. Resample the audio signal to the desired frequency (10kHz).
% Comment the block to Fs=12kHz
% fs=10000;
% frame=resample(frame,fs,Fs);
% Fs=fs;

wl=0.0750*Fs; % Length of frame of 75 ms(in samples)
p = 10; % LP order
s=frame(1:wl);%Singing voice signal to analysis (frame 75ms)
[gci,N0] = f_ZFF(s,Fs); %GCI estimation

[Note, f0_mean]=Music_notes(N0,Fs); %Musical note sung by the soprano

%The pre-emphasis filter is a high pass filter
```

```
x2 = filter([1 -0.98],1,s); % Pre-emphasis H(z)=1-azˆ-1
h = hamming(wl);

%%% Analysis
in=input('Type of analyse 1(LPC), 2(WLP-AME):');
dx=1;
filename='Estimation_of_Formants_WLP-HPSV_Tese_S7.xlsx';
sheet=strcat(num2str(name));

if Fs==12000
   F0cell=strcat('R',num2str(3));
   xlswrite(filename,f0_mean,sheet,F0cell);
   M_notecell=strcat('S',num2str(3));
   xlswrite(filename,Note,sheet,M_notecell);
elseif Fs==10000
   F0cell=strcat('AL',num2str(3));
   xlswrite(filename,f0_mean,sheet,F0cell);
   M_notecell=strcat('AM',num2str(3));
   xlswrite(filename,Note,sheet,M_notecell);
end

if in==1
   a2 = lpc(h.*x2,p); % Conventional LP analysis for reference
   [fmts,bws]=formants(a2,Fs);

   if Fs==12000
      xlRange=strcat('S',num2str(13));
      xlswrite(filename,fmts,sheet,xlRange);
   elseif Fs==10000
      xlRange=strcat('S',num2str(22));
      xlswrite(filename,fmts,sheet,xlRange);
   end

else
   ps=0.1;
```

```
    [DQ,PQ]= pmts(ps);
    for i=1:1:length(DQ)
        wn = makeW(s,p,DQ(i),PQ(i),gci,f0_mean,Fs);
        a=wlp(h.*x2,wn,p);
        a=fixzeroes(a,0,Fs); %
        [fmts,bws]=formants(a,Fs);
        indx=11+dx;

        if Fs==12000
           xlRange=strcat('E',num2str(indx));
           xlswrite(filename,fmts.',sheet,xlRange);
           dx=dx+1;
        elseif Fs==10000
           xlRange=strcat('Y',num2str(indx));
           xlswrite(filename,fmts.',sheet,xlRange);
           dx=dx+1;
        end
    end
end
```

APÊNDICE J – Algoritmo do método WLP-AME-ADP para análise exaustivo de parâmetros AME

```
%% Formant frequency estimation of high-pitched vowels using weighted
%linear prediction

close all; clear; clc;
%% Reading, Resampling and remove start and end of the audio signal

% 1. Read the audio file and sampling rate
info = audioinfo('06_SR1_i.wav');
[data, fs]=audioread('06_SR1_i.wav');

% 2. Resample the audio signal to the desired frequency
Fs=12000;
Ts=1/Fs;
data_new=resample(data,Fs,fs);

% 3. Remove start and end transients samples
% N=length(data_new);
% sg=data_new((0.3*N):(0.7*N)); % Signal for analysis
sg=data_new;

% 4. Pre-emphasis of the signal
%Pre-emphasis is used to lift the spectrum at the higher frequencies so
%that the modeling of formants of different magnitudes would be more even.
%A typical way to achieve this is to use a first-order FIR filter
```

```
%The pre-emphasis filter is a high pass filter
%b=[1 -0.98];
%x_n=filter(b,1,sg); %H(z)=1-az^-1
x_n=filter([1 -exp(-2*pi*50/Fs)],1,sg);

% 5. Determine LP order
% p=ceil((0.001*Fs)+2); %rule of thumb for formant estimation
p=10;

% 6. Weighting function

[gci,N0] = f_ZFF(sg,Fs);
gci=gci.';
ki=gci(1);
ke=gci(end);
T0=N0*(1/Fs);
F0=1./T0;

% 7. Block used to discard any infinite value of fo
ind=isinf(F0);
for k=1:length(F0)
    if ind(k)==0
       ind(k)=1;
    else
        ind(k)=0;
    end
end

f0=F0.*ind;
s=std(f0,'omitnan');
f0=mean(f0, 'omitnan');
Note=Music_notes(f0);

% 8. Split the data sequence into frames.
frame_length = round(0.075*Fs); %In samples for 30ms_F40.075*Fs
```

```
hop_size = round(0.01875*Fs); % In samples for 25% overlap 0.0075_F4
lpcskip=0;

% 9. Input data
in=input('Type of analyse 1(LPC), 2(WLP-AME):');
d=0.01; %Amplitude of the attenuated section
tme=gci/Fs;

dx=1;
filename='Formant estimation_Natural_Vowels.xlsx';
sheet='Vowel i_F5';

F0cell=strcat('R',num2str(3));
xlswrite(filename,f0,sheet,F0cell);
M_notecell=strcat('S',num2str(3));
xlswrite(filename,Note,sheet,M_notecell);

if in==1
   [Ar,e,k]=v_lpcauto_2(x_n(ki:ke), p,[hop_size frame_length lpcskip],'m','');
   [fmts,bws]=resonances(Ar,p,Fs);

   if Fs==12000
      xlRange=strcat('S',num2str(13));
      xlswrite(filename,fmts,sheet,xlRange);
   else
      xlRange=strcat('S',num2str(22));
      xlswrite(filename,fmts,sheet,xlRange);
   end

else
   for DQ=0:0.1:1
       for PQ=0:0.1:1
         [wn,sn]=alku(DQ,PQ,d,tme,x_n,T0,f0,Fs);
         [Ar,e,k]=v_lpcauto_2(sn, p,[hop_size frame_length lpcskip],'m','',wn);
         [fmts,bws]=resonances(Ar,p,Fs);
```

```
            indx=11+dx;

            if Fs==12000
               xlRange=strcat('E',num2str(indx));
               xlswrite(filename,fmts.',sheet,xlRange);
               dx=dx+1;
            else
               xlRange=strcat('Y',num2str(indx));
               xlswrite(filename,fmts.',sheet,xlRange);
               dx=dx+1;
            end
         end
      end
   end

   % 10. Print the formant frequencies and their respective bandwidths.
   for jj=1:length(fmts)
       fprintf('F%d %.1f Hz, Bw%d %.3f Hz\n',jj,fmts(jj), jj, bws(jj));
   end
```