Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Departamento de Engenharia Elétrica
Programa de Pós-Graduação em Engenharia Elétrica

Tese de Doutorado

Modelo de Produção da Voz Baseado na Biofísica da Fonação

Raissa Bezerra Rocha

Campina Grande – PB

Fevereiro de 2017

Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Departamento de Engenharia Elétrica
Programa de Pós-Graduação em Engenharia Elétrica

Modelo de Produção da Voz Baseado na Biofísica da Fonação

Raissa Bezerra Rocha

Tese de Doutorado apresentada à Coordenação do Programa de Pós-Graduação em Engenharia Elétrica do Departamento de Engenharia Elétrica da Universidade Federal de Campina Grande como requisito necessário para a obtenção do grau de Doutor em Ciências no Domínio da Engenharia Elétrica.

Área de Concentração: Eletrônica e Telecomunicações

Prof. Dr. Marcelo Sampaio de Alencar
Orientador

Prof. Dr. Wamberto José Lira de Queiroz
Orientador

Campina Grande – PB, Paraíba, Brasil

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

R672c Rocha, Raissa Bezerra.
Modelo de produção da voz baseado na biofísica da fonação / Raissa Bezerra Rocha. ï Campina Grande, 2017.
77 f. : il. color.

Tese (Doutorado em Engenharia Elétrica) ï Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática.
"Orientação: Prof. Dr. Marcelo Sampaio de Alencar, Prof. Dr. Wamberto José Lira de Queiroz".

1. Modelo de Geração de Voz. 2. Transmição de Informação Cicloestacionária. 3. Pulso Glotal de Liljencrants-Fant. 4. Densidade Espectral de Potência do Sinal de Voz. I. Alencar, Marcelo Sampaio de. II. Queiroz, Wamberto José Lira de. III. Universidade Federal de Campina Grande, Campina Grande (PB). III. Título.

CDU 621.391(043)

"MODELO DE PRODUÇÃO DE VOZ BASEADO NA BIOFÍSICA DA FONAÇÃO"

RAISSA BEZERRA ROCHA

TESE APROVADA EM 20/03/2017

MARCELO SAMPAIO DE ALENCAR, Ph.D., UFCG
Orientador(a)

WAMBERTO JOSÉ LIRA DE QUEIROZ, D.Sc, UFCG
Orientador(a)

FRANCISCO MADEIRO BERNARDINO JÚNIOR, D.Sc, UPE
Examinador(a)

ALDEBARO BARRETO DA ROCHA KLAUTAU JÚNIOR, Dr., UFPA
Examinador(a)

BENEDITO GUIMARÃES AGUIAR NETO, Dr.-Ing., MACKENZIE
Examinador(a)

WASLON TERLLIZZIE ARAÚJO LOPES, D.Sc., UFPB
Examinador(a)

GLAUCO FONTGALLAND, Dr., UFCG
Examinador(a)

CAMPINA GRANDE - PB

A Deus, meu bem maior.
A Jesus Cristo, minha fonte inesgotável.
A Nossa Senhora, minha mãe intercessora.
Aos meus pais, Wilson e Gláucia, minha vida.

Agradecimentos

A Deus, cujas palavras me faltam para externar tão grande amor e gratidão. Por todas as bênçãos que me são concedidas, especialmente por eu poder sentir seu amor, sua presença e companhia em todos os momentos da minha vida. Pelo amor incondicional, pelo perdão, pelos sonhos realizados e por tudo o que me negou. Por ser minha fortaleza, minha felicidade e paz diária.

Aos meus pais, Wilson Taveira Rocha e Gláucia Bezerra Rocha, razão maior da minha vida, pelo amor incondicional, renúncia, apoio e fidelidade. Pelo exemplo de caridade, bondade e generosidade. Por estarem ao meu lado em todos os momentos da minha vida. Sem dúvida, são as maiores bençãos que a Sabedoria Divina colocou em meu caminho.

A Thiago Tavares de Alencar, pelo apoio na realização deste trabalho e incentivo em tudo o que faço. Acima de tudo, pelo amor que me dedica.

Aos meus familiares e amigos que me ajudaram e torceram por mim em todas as fases da minha vida. Em especial, ao meu irmão Gláucio Bezerra Rocha, pela parceria neste trabalho e na vida.

Ao meu orientador, Professor Marcelo Sampaio de Alencar, por ter-me aceitado como orientanda desde a época da graduação como aluna de iniciação científica, me incentivando a procurar a pesquisa, fortalecendo o meu desenvolvimento profissional. Em especial, por todos os conselhos e ensinamentos, pelo carinho, atenção e paciência que sempre teve comigo. Pessoa que terá sempre meu respeito e admiração. Que me serviu e sempre servirá de exemplo e referência em todos os momentos da vida. De coração, muito obrigada!!

Ao meu orientador, Professor Wamberto José Lira de Queiroz, pelo inestimável apoio, ideias, discussões, empenho, contribuições e transmissão de conhecimentos que foram fundamentais para o desenvolvimento deste trabalho. Sobretudo pelo exemplo de pessoa e pela amizade, eu agradeço.

Ao Professor Francisco Madeiro Bernadino Júnior pelo apoio e co-orientação na execução deste trabalho.

Agradeço a todos os meus amigos do Iecom, pelo apoio, incentivo, amizade e momentos divertidos. Também agradeço aos amigos do Iecom, que hoje, distantes fisicamente, mas cujos conselhos, ensinamentos e todos os momentos agradáveis ficarão eternizados.

Ao apoio da Universidade Federal de Sergipe (UFS), Universidade Federal de Campina Grande (UFCG), do Instituto de Estudos Avançados em Comunicações (Iecom) e do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

A todos, aceitem meus sinceros agradecimentos.

"Enquanto viver, falarei da Tua bondade,
e levantarei as minhas mãos em oração."
Salmo 63.

Resumo

A busca por novos modelos que representem a biofísica da fonação da voz é importante em aplicações que incluem o processamento do sinal de voz por representar uma ferramenta no conhecimento de característica dos locutores. Esta tese de doutorado apresenta uma nova abordagem para a teoria fonte-filtro de geração de voz, mais precisamente sons sonoros, que realiza a modelagem da voz por meio de três subsistemas independentes: fonte de excitação, trato vocal e radiação dos lábios e narinas. Trata-se de um modelo em que a geração da voz é feita por meio de filtros lineares e invariantes ao deslocamento no tempo e que leva em consideração a física da fonação, a partir da característica cicloestacionária do sinal de voz, proveniente do comportamento de vibração das cordas vocais. É sugerido que a frequência de oscilação das cordas vocais é dada em função da massa e comprimento delas, e que seu valor é alterado principalmente pela tensão longitudinal aplicada a elas. No modelo proposto para geração da voz, o movimento vibratório das cordas vocais é modelado por meio de um de gerador de trem de impulsos cicloestacionário, controlado por um sinal de tensão obtido a partir da forma de onda do sinal de voz. É realizada toda a análise matemática que abrange o novo modelo para a excitação glotal, apresentando-se uma expressão matemática da densidade espectral de potência do sinal que excita a glote, bem como para o sinal de voz, cujos parâmetros podem ser ajustados para emular patologias na glote. Além disso, apresenta-se a análise no domínio da frequência do pulso glotal usado. Para analisar o desempenho do modelo proposto, testes com locução foram realizados e os resultados indicam que o modelo proposto se ajusta bem a geração da voz.

Palavras-Chave: Modelo de geração de voz; Transmissão de informação cicloestacionária; Pulso glotal de Liljencrants-Fant; Densidade espectral de potência do sinal de voz.

Abstract

The search for new models that represent the biophysics of voice phonation is important for applications that include voice signal processing because it represents a tool for getting to know the characteristics of the speakers. This doctoral thesis presents a new proposal for the source-filter theory of voice production, more precisely related to voiced sounds, that performs the voice modelling using three independent subsystems: the excitation source, the vocal tract, the lip and nostrils radiation system. It is a proposal for a model to generate voice using linear and time-invariant systems, and takes into account the phonation physics and the cyclestationarity characteristics of the voice signal, related to the vibrational behavior of the vocal cords. The model suggests that the frequency oscillation of the vocal folds is a function of the mass and length, but controlled by the longitudinal tension applied to them. In the proposed voice generation model, the vibratory movement of the vocal cords is modeled by a cyclestationary train of impulses, controlled by a tension signal obtained from the voice signal waveform. A mathematical analysis encompassing the new model for glottal excitation is accomplished by presenting a mathematical expression of the signal power spectral density which excites the glottis, as well as the voice signal, whose parameters can be adjusted to emulate pathologies in the glottis. Moreover, the analysis of the utilized glottal pulse in the frequency domain is presented. To analyze the performance of the proposed model, tests with locutions were done and the results indicate that the proposed model adjusts well to voice generation.

Keywords: Voice production model; Transmission of cyclostationary information; Glottal pulse of Liljencrants-Fant; Power spectral density of voice signal.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Siglas

DEP	*Power Spectral Density*	Densidade Espectral de Potência
UFCG	*Federal University of Campina Grande*	Universidade Federal de Campina Grande
CNPq		Conselho Nacional de Desenvolvimento Científico e Tecnológico
LPC	*Linear Predictive Coding*	Codificação por Predição Linear
PHS	*Port-Hamiltonian Systems*	
UFS	*Federal University of Sergipe*	Universidade Federal de Sergipe
PB	*Braziliam Portuguese*	Português do Brasil
RPE-LTP	*Regular Pulse Excited-Long Term Predictor*	Preditor de Longo Prazo com Excitação Regular por Pulso
CELP	*Code Excited Linear Prediction*	Predição Linear Excitada por Código
ACELP	*Algebraic Code Excited Linear Predictive*	Preditivo Linear Algébrico Excitado por Código
Iecom	*Institute of Advanced Studies in Communications*	Instituto de Estudos Avançados em Comunicações
QCELP	*Qualcomm Code Excited Linear Predictive*	Predição Linear Excitada por Código Qualcomm
VSELP	*Vector Sum Excited Linear Predictive*	Preditivo Linear Excitado por Soma Vetorial
AMR-NB	*Adaptive Multi-Rate Narrowband*	Multitaxa Adaptativa de Faixa Estreita
AMR-WB	*Adaptive Multirate Wideband*	Multitaxa Adaptativa de Faixa Larga
LF	*Liljencrants-Fant*	
TTS	*Text-to-Speech*	Conversão texto-fala
FDP	*Probability Distribution Function*	Função distribuição de probabilidade

Lista de Símbolos

F_o	Frequência fundamental (frequência média).
t	Constante que representa o tempo contínuo.
n	Constante que representa o tempo discreto.
$x(t)$	Descolamento de massa no tempo do modelo mecânico.
$X(t)$	Processo estocástico cicloestacionário no sentido amplo.
M	Constante de elasticidade e rigidez do modelo de Flanagan e Landgraf.
K	Constante de elasticidade e rigidez do modelo de Flanagan e Landgraf.
B	Constante de elasticidade e rigidez do modelo de Flanagan e Landgraf.
$F(x,t)$	Força média considerada no modelo mecânico.
S_j	Molas não lineares.
k_C	Mola linear.
x_1	Representa o movimento das massas.
x_2	Representa o movimento das massas.
K_j	Rigidez linear.
η_j	Coeficientes referentes à mão-linearidade das molas.
F_1	Força do modelo macânico.
F_2	Força do modelo macânico.
$y[n]$	n-ésima amostra de saída do codificador LPC.
$\hat{x}[n]$	predição linear de $x[n]$.
a_h	k-ésimo coeficiente do preditor LPC.
G_f	Fator de ganho do preditor LPC.
$x[n]$	Entrada do preditor LPC.
p	Ordem do filtro do preditor LPC .
$R_n(z)$	Função de transferência do ressonador.
a_{1n}, a_{2n}, a_{3n}	Coeficientes da frequência central de ressonância.
f_n	Frequência central de ressonância.
B_n	Largura de banda do formante.
T_s	Período de amostragem.
f_1, f_2, f_3	Frequências de ressonadores.
G	Ganho considerado no novo modelo de produção de voz.
F	Frequência de vibração das cordas vocais.

$T(t)$	Tensão submetida às cordas vocais.
α_1	Constante de sensibilidade do processo.
α_2	Constante de sensibilidade do processo.
β	Relação entre as constantes de sensibilidade do processo.
$E(t)$	Derivado do pulso glotal do modelo de Liljencrants-Fant.
$C(t)$	Trem de impulsos cicloestacionário.
$M(t)$	Sinal que estimula tensão nas cordas vocais.
a_o	Coeficiente da série de Fourier.
a_k	Coeficiente da série de Fourier.
b_l	Coeficiente da série de Fourier.
T_o	Período fundamental (período médio).
ω_o	Frequência angular fundamental (frequência média).
$\delta(t)$	Função impulso.
ϕ_o	Fase inicial.
δ_o	Posição inicial dos impulsos.
$\Delta\omega$	Variação de frequência de oscilação das cordas vocais.
M_{max}	Amplitude máxima do sinal aleatório $M(t)$.
ω_m	Frequência máxima de oscilação das cordas vocais.
$\phi(t)$	Fase aleatória que depende de M(t).
$R_c(\tau)$	Função autocorrelação de $C(t)$.
$\varphi_{\omega(t)}$	Função característica de $\omega(t)$.
$S_c(\omega)$	Densidade espectal de potência de $C(t)$.
$f_{M(t)}$	Função distribuição de probabilidade de $M(t)$.
$f_{\Omega(t)}$	Função distribuição de probabilidade de $\Omega(t)$.
$f_X(x)$	Função distribuição de probabilidade Gamma unilateral.
$p_X(x)$	Função distribuição de probabilidade Rayleigh.
Γ	Função Gamma.
k_x	Parâmetro de formato da função distribuição de probabilidade Gamma.
θ	Parâmetro de escala da função distribuição de probabilidade Gamma.
$u(x)$	Função degrau.
σ	Parâmetro de escala da função distribuição de probabilidade Rayleigh.
E_e	Valor máximo que a derivado no pulso glotal de Liljencrants-Fant assume.
E_o	Eo é um fator de escala do pulso glotal de Liljencrants-Fant.
ω_g	Taxa de aumento de amplitudo do modelo de Liljencrants-Fant.
t_o	Instante em que as cordas vocais estão fechadas.
t_e	Instante em que as cordas vocais voltam a estarem fechadas.
α	Fator que determina ω_g.
t_c	Instante final do movimento do pulso glotla de Liljencrants-Fant.
ϵ	Constante de decaimento para a fase de recuperação do pulso glotal de Liljencrants-Fant.
T_a	Eficiência do retorno de fase do pulso glotal de Liljencrants-Fant.

$U(t)$	Modelo do pulso glotal de Liljencrants-Fant.
α_e	Constante da resposta em frequência da derivada do pulso glotal de Liljencrants-Fant.
α_o	Constante da resposta em frequência da derivada do pulso glotal de Liljencrants-Fant.
β_e	Constante da resposta em frequência da derivada do pulso glotal de Liljencrants-Fant.
β_o	Constante da resposta em frequência da derivada do pulso glotal de Liljencrants-Fant.
$e[n]$	Erro de predição.
x_j	Sinal janelado.
w_h	Janela de Hamming.
E	Valor médio quadrático do erro de predição.
$R_{xx}(i)$	Função autocorrelação de $x[n]$.
$H(z)$	Função de transferência do trato vocal.
$V(t)$	Sinal de voz resultante do novo modelo de produçã de voz.
$H(t)$	Resposta do trato vocal no domínio do tempo.
$L(t)$	Resposta da radiação lábios/narinas no domínio do tempo.
$L(\omega)$	Resposta em frequência da radiação lábios/narinas.
$S_V(\omega)$	Densidade espectral de potência do sinal de voz pelo novo modelo de produção de voz.
$E(\omega)$	Resposta em frequência do pulso glotal de Liljencrants-Fant.

CAPÍTULO 1

Introdução

Processamento de voz é um tema bastante estudado pelos pesquisadores de diversas áreas, como engenharia, medicina e fonoaudiologia. A voz é um sinal complexo que representa uma das formas de comunicação mais utilizadas pelos seres humanos. Além disso, a voz é uma boa identidade por mostrar, além do estado emocional, caracteísticas pessoais do locutor, como sexo, idade, *status* social, entre outras.

O estudo da fisiologia do processo de produção de voz e uma representação esquemática que a defina são necessários para a construção de sistemas que envolvam o processamento do sinal de voz, como reconhecimento, segmentação e síntese da fala, bem como sistemas de codificação e identificação de patologias na voz.

A voz é gerada pelo aparelho fonador, que funciona com diferentes configurações para produzir os vários tipos de sons. De forma geral, para que a voz seja formada, são necessários inicialmente dois sinais, um cicloestacionário e um ruidoso. Esses sinais estão relacionados às cordas vocais e servem de base para a construção do sinal de voz.

O funcionamento das cordas vocais é a principal etapa no processo de produção da voz e está diretamente associado à qualidade da voz, uma vez que seu padrão de vibração fornece a informação da frequência fundamental e de seus harmônicos. Esse padrão, uma vez mudado, altera todo o resultado do processo fonatório [13].

O sinal gerado nas cordas vocais tem seu espectro de frequência modelado no trato vocal por meio de frequências determinadas pela configuração instantânea dos músculos que o formam.

A observação do mecanismo de produção da voz iniciou-se no século XVIII, em que estipulavam que a vibração das cordas vocais era produzida pela vibração do ar. Em 1950, Husson propôs que a vibração das cordas vocais é uma consequência de impulsos nervosos individuais, gerados a uma taxa dada pela frequência fundamental, enviados aos músculos vocálicos, obtendo como resultado um força de ar exalado sobre as cordas vocais.

Atualmente, a teoria mais aceita para a descrição do trem de pulsos glotais foi proposta Helmholtz and Muller, aprimorada por van den Berg [14], em 1958, e Titze [15], em 1980, e é denominada teoria aerodinâmica-mioelástica. De acordo com essa teoria, o movimento de abertura e fechamento das cordas vocais está relacionada às propriedades mecânicas dos músculos

que formam as cordas vocais e pelas forças aerodinâmicas que se distribuem ao longo da laringe durante a fonação.

As cordas vocais se movimentam de forma cicloestacionária. Durante um discurso a taxa de vibração das cordas vocais é continuamente mudada a partir da entonação das sentenças. A pergunta "Você está feliz?"apresenta uma entonação crescente, diferentemente da sentença "Estou feliz"que tem um padrão de entonação decrescente. Essa diferença de entonação é justificada pela variação na frequência de oscilação das cordas vocais.

O padrão de vibração das cordas vocais está relacionado ao seu comprimento, massa e tensão. Esses parâmetros estão associados ao sexo e idade do locutor. Os homens, por exemplo, têm suas cordas vocais com comprimento que varia entre 17 a 24 mm, enquanto o comprimento das cordas vocais nas mulheres está na faixa de 13 a 17 mm. Para crianças, esse comprimento é ainda menor, na faixa de 6 a 8 mm. Geralmente, esses comprimentos podem ser alterados em 3 ou 4 mm [16, 17].

Uma corda vocal em particular, com uma determinada massa e comprimento, tem seu padrão de vibração aumentado pelo alongamento e tensão das cordas, que diminui sua massa e aumenta sua elasticidade. Nesse caso, a massa e tensão são mais importantes do que o comprimento na determinação da taxa de vibração das cordas vocais, uma vez que o alongamento diminui a massa e aumenta a tensão provocando um aumento na frequência de oscilação.

Na literatura existem vários trabalhos que propõem uma modelagem para a geração da voz e, consequentemente, para o funcionamento da fonte glotal. Um dos mais difundidos é a teoria fonte-filtro, proposta por Fant em 1970 [18]. No entanto, os modelos de geração de voz encontrados na literatura não levam em consideração o movimento de taxa de vibração variável das cordas vocais da forma como abordado nesta tese.

1.1 Motivação

Por representar uma importante ferramenta de auxílio ao estudo de aplicações em que o sinal de voz está presente, a análise acústica é uma área que atrai cada vez mais pesquisadores. Ela é empregada em vários ramos de processamento do sinal de voz, pois consiste em uma técnica não invasiva de conhecer as particularidades de locutores.

A análise acústica vocal das características temporais e espectrais dos sinais de fala pode ser utilizada em diversas aplicações que levam em consideração as diferentes características encontradas nos vários tipos de sons da fala. Ela permite, por exemplo, obter características como padrões de repetição, taxa de cruzamento por zero, frequência fundamental e seus harmônicos, distribuição de energia em função da frequência, função distribuição de probabilidade (FDP), entre outros.

A partir da análise acústica, sistemas como segmentação de voz, codificação de fala, identificação e classificação de patologias e emulação de distúrbios da voz, entre outros, podem ser desenvolvidos ou aprimorados.

Particulamente no caso da emulação de patologias, várias pesquisas encontradas na literatura objetivam obter métodos que façam a discriminização entre as vozes patológicas e saudá-

veis. Nesse caso, a análise acústica pode ser aliada a técnicas que realizem a observação direta do aparelho fonador, mais precisamente das cordas vocais, com o intuito de obter indicadores que indentifiquem distúrbios na voz.

Várias patologias da voz podem ser detectadas por meio da observação das cordas vocais, que é um dos principais tecidos que envolve a produção da fala. No entanto, para que seja viável a identificação de distúrbios nas cordas vocais, a técnica que possibilita a análise acústica deve ser mais fiel possível na sua representação, durante o processo da fonação.

Nesse contexto, a modelagem matemática que representa o comportamento das cordas vocais ao longo da fonação, como a descrita nesta tese, é uma poderosa técnica de análise acústica. A partir dela, é possível a geração do sinal de voz no domínio do tempo e a estimação da densidade espectral de potência da voz, por meio expressões matemáticas cujos parâmetros podem ser ajustados à emulação de voz saudável e patológica.

1.2 Objetivos

Afim de realizar uma análise acústica mais fiel à geração da voz, esta tese de doutorado tem o objetivo de apresentar um novo modelo matemático para a geração da voz, que, diferentemente dos demais trabalhos encontrados na literatura, inclui o comportamento cicloestácionário das cordas vocais.

O modelo propõe a geração da voz a partir de um sistema linear e invariante no tempo, e considera que o movimento cicloestácionário das cordas vocais é proveniente de mudanças na frequência de oscilação das cordas vocais controlada por um sinal verificado nas cordas vocais, denominado sinal de controle. Nesse caso, a modelagem está baseada no fato de que a variação na frequência de oscilação é proporcional a tensão longitudinal aplicada às cordas vocais.

Esse modelo de produção da voz tem a finalidade de ser utilizado para construção ou melhoramento de sistemas que usam o processamento da fala e, principalmente, na análise acústica para emular patologias na glote por meio de alterações no seu movimento vibratório.

Como, em geral, os sons surdos não são gerados pela vibração das cordas vocais, eles não são apropriados para análise espectral na detecção e classificação de patologias das cordas. Dessa forma, o modelo proposto nesta tese está focado na produção de sons sonoros.

O modelo considera que o fluxo glotal é resultante de uma exitação produzida por um gerador de um trem de impulsos cicloestácionário. Uma vez que um processo linear não gera novas frequências e não amplia a faixa de frequência, o modelo assume que as variações na frequência fundamental estão presentes na forma de onda da voz e são representadas por um sinal que é obtido por meio dos pontos de cruzamento por zero do sinal de voz.

Para alcançar o objetivo proposto, iniciamente é desenvolvida a representação matemática de um gerador de impulsos cicloestácionário que caracteriza a variação da frequência fundamental ao longo de um discurso, regido por um sinal de controle.

Em seguida, a função distribuição de probabilidade do sinal de controle é estimada por meio de um ajuste de curva com as funções Gamma unilateral e Rayleigh. É proposta uma expressão matemática que descreve a densidade espectral de potência (*Power Spectral Density* –

DEP) do pulso glotal utilizado, assim como expressões que caracterizam o comportamento do sinal de voz no domínio do tempo e frequência.

1.3 Estrutura da Tese

Além deste capítulo introdutório, que apresenta uma descrição geral sobre as etapas e características do mecanismo de produção da voz, a importância de métodos de análise acústica da voz e objetivos da tese, há mais quatro capítulos.

O Capítulo 2 relata o estudo sobre a fisiologia da voz humana, bem como o mecanismo biofísico da geração da voz. Também é apresentada uma revisão bibliográfica acerca dos modelos que representam a geração da voz, mais precisamente do funcionamento das cordas vocais. Por fim, são descritos os principais tipos de síntese de voz encontrados na literatura.

O novo modelo de produção da voz é apresentado no Capítulo 3. Nele, a produção da voz é feita a partir de filtros lineares e invariantes no tempo e é exposto o desenvolvimento matemático realizado na modelagem da produção da voz, especialmente para as cordas vocais, com a dedução de uma expressão que representa a densidade espectral de potência do sinal que excita as cordas vocais.

Além disso, também são apresentados as expressões matemáticas no domínio do tempo e frequência para o sinal de voz, bem como a discussão sobre o uso do modelo na emulação de patologias das cordas vocais.

O Capítulo 4 contém os resultados obtidos em todas as etapas do modelo de produção proposto e são expostos todos os parâmetros utilizados na representação de vozes saudáveis. O capítulo é finalizado com a comparação entre o novo modelo de produção da voz e o modelo fonte-filtro de Fant.

As considerações finais e as propostas para trabalhos futuros são apresentas no Capítulo 5. Nesse capítulo é feita uma discursão sobre o modelo proposto, elencadas as principais contribuições da tese e as perspectivas para a continuação do trabalho.

CAPÍTULO 2

A Produção e Síntese de Voz

O conhecimento do comportamento do aparelho fonador durante a fonação é fundamental no desenvolvimento de sistemas que modelem o processo de produção da voz.

Este capítulo tem o objetivo de descrever a base teórica necessária para o desenvolvimento de um modelo vibratório das cordas vocais que pode ser utilizado para síntese de fala, assim como diagnósticos de desordens vocais.

Inicialmente é descrita a fisiologia humana que envolve a geração da voz, bem como o seu funcionamento. Em seguinda, os sons da fala são classificados de acordo com a sua fonte de excitação. Por fim, são apresentadas as principais estratégias de síntese de voz.

2.1 A Fisiologia da Voz Humana

Do ponto de vista fisiológico, a voz é produzida por meio de três principais subsistemas, que formam o aparelho fonador: respiratório, laringeal e articulatório [1, 19].

O subsistema respiratório é constituído pelos pulmões, traqueia, diafragma e brônquios, como ilustra a Figura 2.1. Basicamente, esse subsistema produz um fluxo de ar que fornece energia aerodinâmica aos subsistemas da laringe e articulatório para a geração dos sons.

A laringe, ilustrada na Figura 2.2 é um órgão tubular situado no pescoço, acima da traqueia e abaixo da faringe. Ela possue três funções básicas, que são proteção, respiração e fonação. Inicialmente, a laringe atua como protetora, impedindo que elementos estranhos cheguem ao pulmão, exceto o fluxo de ar. Na respiração, mais precisamente na fase de expiração, as cordas vocais, situadas na laringe, são abduzidas por um conjunto de órgãos até que elas se encostem umas às outras, contribuindo para regular a troca gasosa com o pulmão e a manutenção do equilíbrio ácido-base. A fonação acontece quando há vibração das cordas vocais a partir das suas mudanças de tensão e longitude, além da ampliação da abertura glótica e da intensidade do esforço respiratório [5, 20].

O sistema articulatório, como ilustra a Figura 2.3, consiste da faringe, da língua, do nariz, dos dentes e dos lábios, ou seja, do trato vocal e do trato nasal.

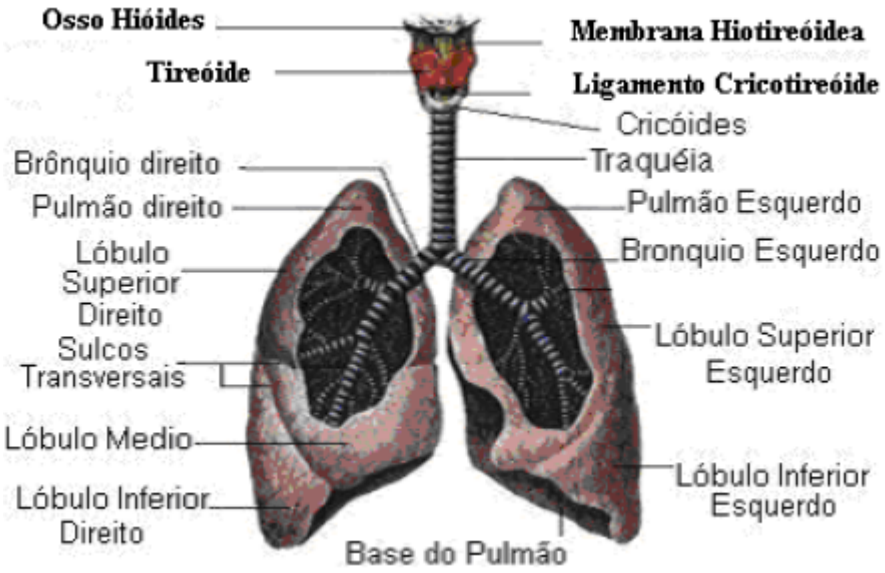

Figura 2.1: Representação esquemático do subsistema respiratório [1].

Figura 2.2: Diagrama esquemático da localização da laringe [2].

2.2 O Mecanismo de Produção da Fala

O processo de produção da fala é composto por três partes fundamentais: fonte de excitação, trato vocal e radiação.

Segundo [21], a voz consiste em uma onda de pressão acústica formada a partir de movimentos voluntários dos órgãos vocais humanos.

A sua geração se inicia a partir do subsistema respiratório, em que um fluxo de ar originado nos pulmões é expelido ultrapassando as cordas vocais. Basicamente, os sons são gerados por dois tipos de excitação. O primeiro é obtido quando há vibração das cordas vocais na passagem do fluxo de ar. No segundo tipo de excitação não há vibração das cordas vocais, consistindo em um turbulência provida pela passagem do ar por uma constrição em alguma região do trato vocal [22].

Como ilustrado na Figura 2.4, as cordas vocais consistem em dois pares de lábios, simetricamente formados por um músculo e um tecido elástico, em que a abertura entre os lábios é denominada glote.

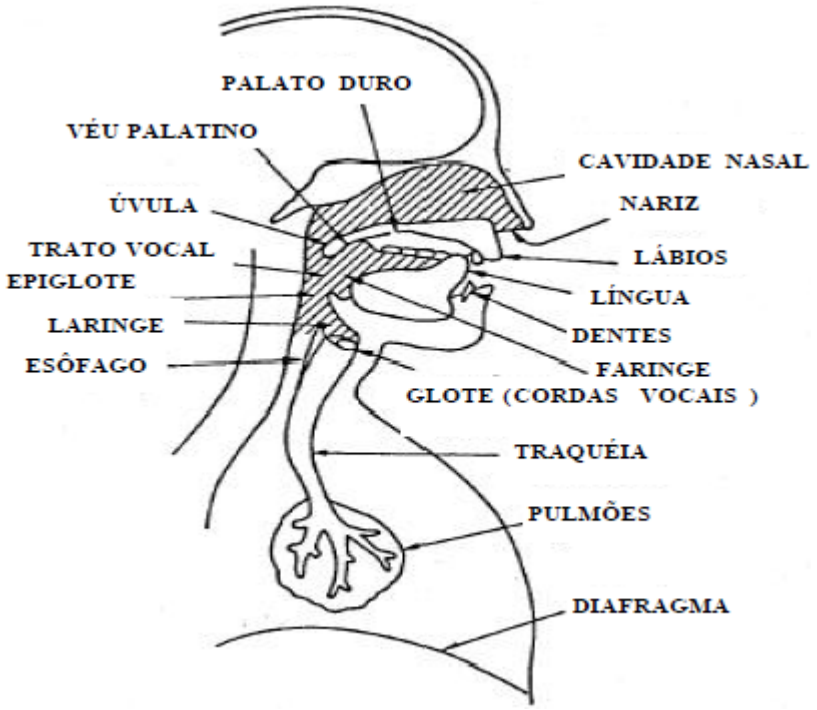

Figura 2.3: Anatomia do aparelho fonador [3, 4]

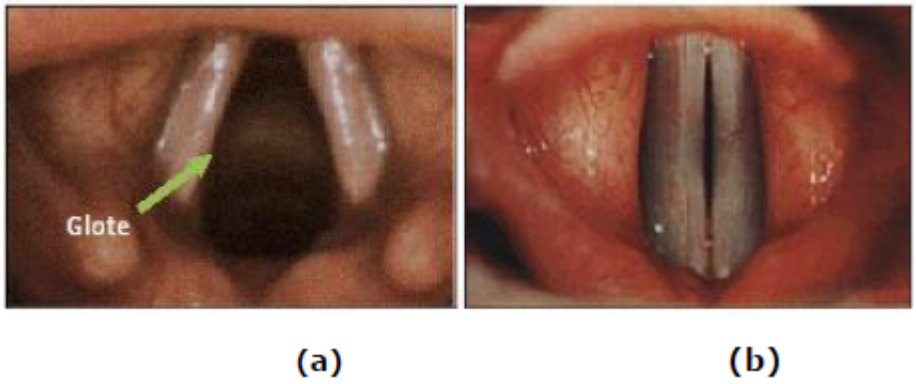

Figura 2.4: Cordas vocais: (a) abertas e (b) fechadas [5].

Ao padrão médio vibratório das cordas vocais dá-se o nome de frequência fundamental (F_o), que corresponde ao número de vibrações das cordas vocais por segundo. A frequência fundamental é determinada com base no comprimento, tensão e massa das cordas vocais, e é dado por [23].

$$F_o = \frac{1}{L_m}\sqrt{\frac{\sigma_c}{\rho}}, \tag{2.1}$$

em que L_m representa o comprimento da membrana em vibração da corda vocal, σ_c a tensão longitudinal e ρ a massa volumétrica do tecido das cordas vocais.

Para a frequência fundamental, é importante considerar os seguintes aspectos. A frequência fundamental, ou primeiro harmônico, aumenta com a pressão subglotal e com a tensão do tecido da prega vocal. Por outro lado, seu valor diminui com o aumento da massa dos tecidos

das cordas vocais. Além disso, seu valor também está relacionado à porção vibrante da corda vocal, obtendo um maior valor para uma curta área de vibração.

Além da fisionomia vibratória, as cordas vocais possuem movimentos verticais e horizontais, que são representados por meio de uma amplitude definida como uma extensão da excursão horizontal das dobras vocais durante a vibração. A amplitude do movimento está relacionada à porção vibrante, rigidez e massa das cordas vocais, além da pressão subglotal. Quanto mais curta for a porção vibrante, quanto maior for a rigidez e massa da prega vocal, menor é a amplitude do movimento. Em contrapartida, quanto maior a pressão subglotal, maior a amplitude do movimento das cordas vocais [20].

O trato vocal, ilustrado na Figura 2.5, consiste em um filtro ressonante que se inicia na abertura entre as cordas vocais e termina nos lábios . O trato nasal começa na úvula e termina nas narinas. O trato vocal e o trato nasal agem como um filtro, em que o sons que por eles se propagam, têm seu espectro de frequência modelado pela seletividade em frequência do filtro.

As frequências de ressonância do trato vocal são denominadas formantes. Essas frequências são determinadas principalmente pela forma e dimensão do trato vocal.

A radiação da voz é a última etapa para sua geração. A voz é irradiada tanto pela cavidade oral quanto pela nasal pelo do acoplamento dos tratos, a partir do abaixamento da úvula.

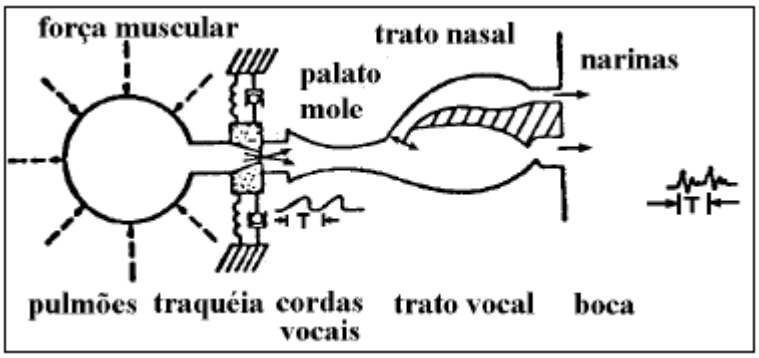

Figura 2.5: Modelo do trato vocal [6].

2.2.1 Classificação dos Sons da Fala

O sons da fala podem ser divididos basicamente em dois grupos: sonoros e surdos.

Na geração do sinais sonoros, a glote, inicialmente fechada, é forçada pela pressão do fluxo de ar vindo dos pulmões, ocasionando na sua abertura. Assim, durante a passagem do ar, as cordas vocais vibram de forma cicloestacionaria produzindo um sequência de pulsos cuja frequência é controlada pela pressão do ar, tensão e comprimento das cordas vocais. Esse tipo de excitação é a base para a geração dos sons vozeados ou sonoros, que no Português do Brasil (PB) são representados pelas vogais.

Diferentemente dos sons vozeados, os sons surdos ou não vozeados são gerados a partir de um excitação formada pela constrição do fluxo de ar na passagem pelo trato vocal, gerando uma turbulência ou ruído de espectro largo. Para o PB, os sons surdos são representados pelas consoantes e classificados com base em quatro critérios: modo de articulação (plosivas, fricativas

e líquidas (róticas e laterais)), quanto ao ponto de articulação (bilabiais, labiodentais, alveolares, palatais e velares), quanto ao papel das cordas vocais (sonoras e surdas) e quanto ao papel das cavidades bucal e nasal (consoantes orais e nasais) [22, 24].

As consoantes fricativas podem ser divididas em dois grupos de acordo com o tipo de excitação utilizada na sua geração: fricativas sonoras e surdas. As fricativas sonoras são caracterizadas por terem uma excitação mista, consistindo na vibração das cordas vocais juntamente a um ponto de constrição em algum local do trato vocal. Por outro lado, na geração das fricativas surdas não há vibração das cordas vocais. No PB, as fricativas sonoras são representados pelos fonemas [v], [z] e [j] e as surdas por [f], [s] e [x] [25].

Os sons plosivos, por sua vez, são decorrentes de um fechamento total do trato vocal. São assim denominados devido ao seu modo de geração, em que há uma explosão correspondente à súbita liberação do ar após o aumento de pressão do fluxo de ar sobre as cordas vocais. Esses sons também podem ser classificados como surdos ou sonoros. As plosivas surdas, caracterizadas pelos fonemas [p], [t] e [k] do PB, não há vibração das cordas vocais, enquanto na criação das plosivas sonoras, como os fonemas [b], [d] e [g] do PB, há uma pequena quantidade de energia nas baixas frequências irradiada através das paredes da garganta no período de constrição total do trato vocal.

As consoantes africadas são formadas a partir da combinação da excitação das fricativas e plosivas. Da mesma forma que as plosivas, as consoantes africadas têm sua excitação gerada por um constrição total em algum ponto do trato vocal, e, após a liberação do ar, tem-se um som caracterizado por um ruído de fricação. No PB há duas consoantes africadas, [T] (surda) e [D] (sonora), que ocorrem a pelo agrupamento das consoantes plosivas [t] ou [d] seguidas pela vogal posterior [i].

O sons nasais são formados por meio de uma excitação glotal, bem como uma constrição em algum ponto do trato vocal. O ar é irradiado pelo trato nasal, porém o trato vocal mantém-se acoplado à faringe e a boca serve como uma cavidade ressonante que emite energia acústica com uma certa frequência. No PB, as três consoantes nasais são: [m], [n] e [N].

As consoantes laterais, representadas no PB pelos fonemas [l] e [L], são assim denominadas pois, após gerar vibração das cordas vocais, o ar vindo dos pulmões percorre pelas laterias da constrição gerada pela língua.

Ao contrário das demais consoantes, as róticas não são geradas por uma constrição no trato vocal. Sua geração dá-se por vibrações na região do estreitamento palatal, bem como das cordas vocais. No PB, as consoantes róticas são representadas pelos fonemas [r] e [R]. Ao contrário das demais consoantes, as róticas não são geradas por uma constrição no trato vocal. Sua geração dá-se por vibrações na região do estreitamento palatal, bem como das cordas vocais. No PB, as consoantes róticas são representadas pelos fonemas [r] e [R].

2.3 Modelagem das Cordas Vocais

O mecanismo mais importante relacionado à produção de voz é o das cordas vocais. O fluxo de ar exerce pressão sobre a glote ocasionando na sua vibração a uma frequência determi-

nada pela massa, comprimento e tensão das cordas vocais. O movimento vibratório acarreta em pulsos de ar que são modificados pelos tratos vocal e nasal para, em seguida, serem irradiados pelas cavidades oral e nasal.

O estudo e modelagem da dinâmica das cordas vocais é objeto de estudos há muitos anos e tem o objetivo de melhor compreender a física básica de voz sonora e fornecer diagnóstico e tratamento para pessoas com distúrbios da voz.

Na literatura, é possível encontrar alguns modelos mecânicos que modulam a passagem do ar, descrevendo tal mecanismo por meio de equações diferenciais.

O primeiro deles, proposto por Flanagan e Landgraf [26] em 1968, descreve o movimento das cordas vocais por modelos mecânicos massa-mola-amortecedor, como ilustrado na Figura 2.6, de massa M, e K e B as constante de elasticidade e rigidez. O deslocamento da massa, $x(t)$, é governada pela Expressão 2.2, em que $F(x,t)$ é a força aplicada ao sistema e é considerada como a média entre as pressões subglotal e supraglotal.

$$M\ddot{x}(t) + B\dot{x}(t) + Kx(t) = F(x,t). \tag{2.2}$$

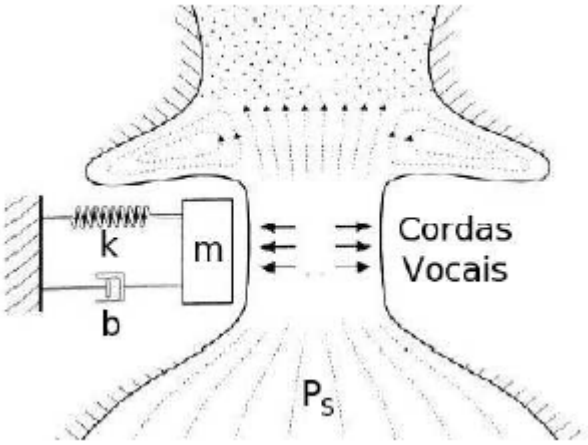

Figura 2.6: Modelo de Flanagan e Landgraf [7, 8].

Com o intuito de aprimorar a modelagem do movimento das cordas vocais, em 1972 Ishizaka e Flanagan [27] propuseram um modelo em que cada uma das cordas vocais é representada por duas massas, como ilustra a Figura 2.7, com as cordas vocais ligadas às paredes da laringe por duas molas não lineares $S1$ e $S2$, e ligadas entre si por uma mola linear k_C.

O modelo considera que as cordas vocais possuem movimento simétrico na direção transversal. Matematicamente, o modelo de duas massas é descrito pela expressão

$$\begin{aligned} M_1\ddot{x}(t) + S_1(x_1) + B_1(x\prime_1) + k_c(x_1 - x_2) &= F_1 \\ M_2\ddot{x}(t) + S_2(x_2) + B_2(\dot{x_2}) + k_c(x_2 - x_1) &= F_2, \end{aligned} \tag{2.3}$$

em que x_1 e x_2 representam o movimento das massas, S_1 e S_2 são as relações das molas não-lineares dadas por

$$S_j(x) = K_j x(1 + \eta_j x^2), \quad \text{para } j = 1,2, \tag{2.4}$$

em que os coeficientes K_j consistem na rigidez linear e η_j são coeficientes positivos que caracterizam a não-linearidade das molas. As forças F_1 e F_2 dependem da pressão subglotal, do fluxo glotal e da área da região entre as cordas vocais.

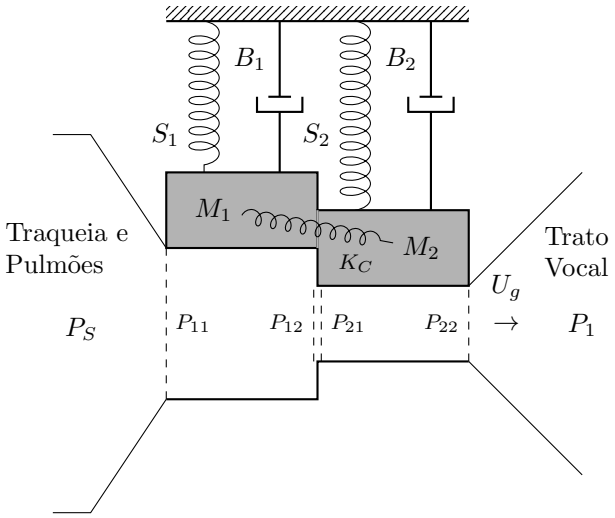

Figura 2.7: Modelo de Ishizaka e Flanagan [8].

Em 1994, Titze [7] considerou a adição de uma terceira massa para modelar o movimento das cordas vocais, apresentado na Figura 2.8. Para as massas, é considerado o movimento perpendicular ao trato vocal. As molas S_1, S_2 e S_3 têm características não lineares e representam as tensões nas cordas vocais [28].

A dinâmica das cordas vocais para o modelo de três massas é descrito pela expressão

$$\begin{aligned} M\ddot{x} + B\dot{x} + S_1(x - x_1) + B_1(\dot{x} - \dot{x_1}) + S_2(x - x_2) + B_2(\dot{x} - \dot{x_2}) &= 0 \\ M_1\ddot{x_1} + S_1(x - x_1) + B_1(\dot{x} - \dot{x_1}) + k_c(x_1 - x_2) &= F_1 \\ M_2\ddot{x_2} + S_2(x - x_2) + B_2(\dot{x} - \dot{x_2}) + k_c(x_2 - x_1) &= F_2. \end{aligned} \tag{2.5}$$

Outras modelagens mais complexas das cordas vocais, visando reproduzir os movimentos vibratórios irregulares, são encontrados na literatura. Em [29] é proposto um modelo de elementos finitos, baseado nas leis da mecânica, para obter características de oscilações das cordas vocais. O modelo leva em consideração uma estrutura mais realista das cordas vocais, com assimetrica do ponto de vista da geometria e tensão nas fonteiras das cordas vocais, além de acomodar não homogeneidades e a característica anisotrópica das cordas vocais. Além disso, o modelo permite simular distúrbios da voz devido à paralisia das cordas vocais.

Os padrões de irregularidades da vibração também são levados em consideração na modelagem das cordas vocais propostas por [30, 31, 32]. Em [30, 31] os modos espaciais domi-

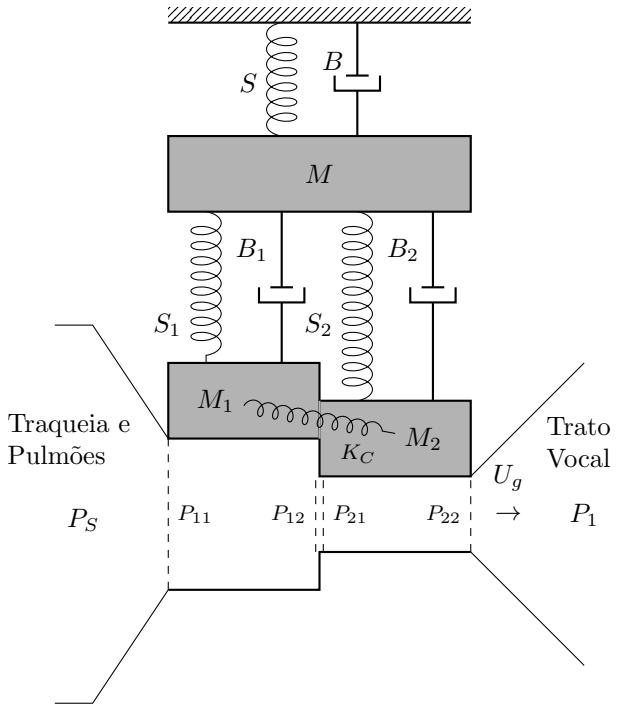

Figura 2.8: Modelo de três massas proposto por Titze [8].

nantes no modelo de elemento finito das cordas vocais são determinados por meio de funções empíricas. É observado que padrões complexos de vibração podem ser explicados por poucos modos de vibração das cordas vocais, bem como modos de alta ordem podem ser estimados pelo modelo em caso de fonação irregular.

Em [33] é relatada a utilização do modelo multi-massa mecânico dependente do tempo para estimar as vibrações pseudoglotais, com objetivo de gerar as vibrações dos tecidos no segmento faringoesofágico para a produção de uma fonte sonora em pacientes cuja laringe foi completamente removida devido ao câncer de laringe. Os resultados indicam que o modelo proposto obtém melhor desempenho na comparação com trabalhos semelhantes encontrados na literatura, com 97,8% a 99,7% de taxa de correlação entre o modelo e o segmento faringoesofágico.

Um modelo sintético baseado na camada epitelial e lâmina própria presentes nas cordas vocais é apresentado em [34]. Os resultados são compatíveis com os obtidos com um modelo de elementos finitos, com o diferencial de apresentar uma menor tensão inicial da glotes, sendo mais próximo da real fonação humana.

O trabalho desenvolvido em [35] apresenta um modelo matemático que representa a laringe durante a fonação. O objetivo consiste em captar fenômenos fisiológicos que ocorrem na laringe durante a fonação. Para isso, os tecidos musculares da laringe são discretizados usando o método de elementos finitos, bem como as equações de Navier-Stokes. Como resultado, tem-se o pulso glotal adquirido de diferentes geometrias de laringes com diferentes propriedades viscoelásticas.

Em [16] é realizado um estudo acerca do comprimento das cordas vocais, a partir do modelo de Fujisaki, além de relacionar tal característica à tensão e frequência fundamental das cordas vocais. Os resultados indicam que o método proposto não é aplicável a todos os locutores,

mas fornecem resultados compatíveis com a literatura [7, 36], atribuindo, para homens, o comprimento das cordas vocais entre 15,2-20,1 mm, e para mulheres 14,3-20,6 mm. Além disso, os autores exaltam que o método pode ser utilizado para estimar o comprimento das cordas vocais em crianças, diferentemente dos métodos clínicos, auxiliando no diagnóstico de doenças.

Na literatura, há também estudos sobre detecção de patologias por meio da modelagem das cordas vocais. Em [37], por exemplo, é descrita uma abordagem não invasiva de detecção de doenças na laringe, além da discriminação da voz sem patologia e voz afetada por edema e outras doenças, como nódulos, cistos e paralisia. Para isso, são utilizados os coeficientes cepstrais, redes neurais e modelos de mistura gaussiana. Os estudos mostram êxito, com uma taxa de 93% de acerto na classificação de voz saudável e taxa de 94% para voz com patologia. Além disso, o método forneceu uma taxa de 76% na discriminação da voz saudável e voz com edema e 85% para outras patologias da voz.

A detecção de patologias na voz por meio de imagens é apresentado em [38]. Para isso, inicialmente são obtidas imagens por meio da videoestroboscopia da laringe e um algoritmo *ad-hoc* de *design* de contorno é utilizado para obter uma segmentação robusta e rápida de imagens, sendo possível a indentificação de patologias e medidas objetivas, como, entre outras, o tamnho do cisto.

O diagnóstico de patologia da voz também é objeto de estudo em [39]. Nesse trabalho, é proposto tal análise por meio da medição do fechamento da glote e ângulos de abertura para diversas patologias. O estudo relata que, na ausência de patologias, o ângulo de abertura está entre 35 e 37 graus, enquanto para cordas vocais com paralisia, o ângulo não excede 30 graus.

Em [40] é investigado o papel do alongamento das cordas vocais na dinâmica dos movimentos glóticos durante a fonação. É proposta a inclusão dessa característica no modelo de duas massas e é observado que um excessivo alongamento das cordas vocais inibe sua vibração, mas que pode haver um alongamento ótimo que maximize as vibrações nas cordas vocais.

A modelagem das cordas vocais por meio da técnica *Port-Hamiltonian Systems* (PHS) é proposta em [41]. Os autores obtiveram êxito ao expressarem o modelo *body-cover* [42] como um PHS e acreditam que a modelagem pode ser empregada em modelos acústicos para a geração da voz.

Os efeitos da massa do pólipo, rigidez, posição e aspectos sobre as frequências naturais e modos de vibração das cordas vocais são estudados em [43] usando um código de elementos finitos. É observado que a massa do pólipo é um fator determinante na frequência natural e que sua posição tem maior influência na frequência do que na rigidez.

Um modelo de síntese de voz que considera a dinâmica do movimento das cordas vocais é a utilizada pelo Codificador por Predição Linear (*Linear Prediction Coding* – LPC). O método de codificação é bastante difundido na literatura e tem a característica de sintetizar a voz baseado em parâmetros extraídos da forma de onda do sinal de voz [44].

O LPC é baseado na modelagem de predição linear, que é fundamentada na aproximação de uma amostra do sinal de fala a partir de uma combinação linear das amostras anteriores. Esse princípio está relacionado com o modelo da produção da voz, ilustrado na Figura 2.9, que

consiste em um sistema linear variante no tempo com dois tipos de excitação, a periódica, usada na geração dos fonemas sonoros, e a ruidosa, para a geração dos fonemas surdos [45, 46].

Periodicamente, para curtos intervalos de tempo em que a fala é considerada estacionária, os coeficientes do preditor usados na combinação linear são computados por minimização da soma das diferenças quadradas entre a amostra de fala atual e a predita linearmente. Esses coeficientes podem ser obtidos por meio da análise LPC, denominados coeficientes LPC, ou por meio de técnicas derivadas dessa análise. Entre os coeficientes utilizados, podem ser citados: coeficientes LPC, cepstrais, cepstrais ponderados, delta cepestrais e delta cepestrais ponderados.

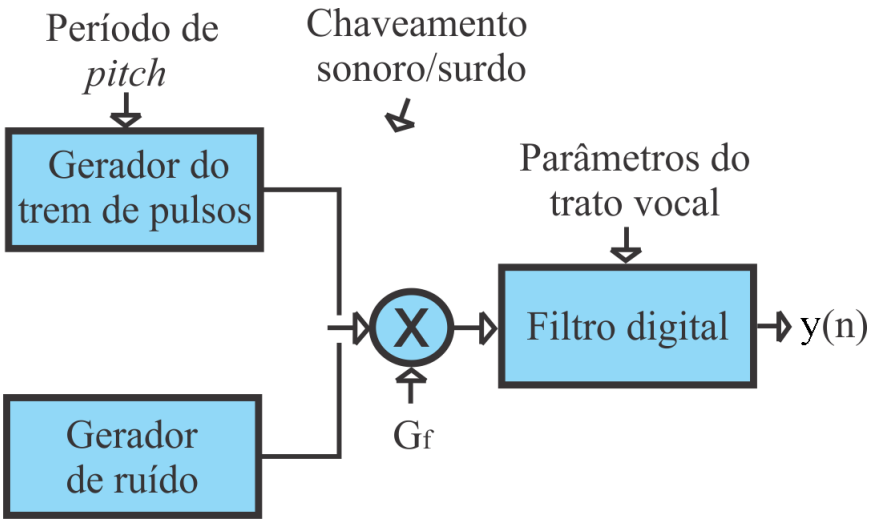

Figura 2.9: Diagrama de blocos do modelo de produção de voz. Adaptado de [9, 6].

Para a geração da fala, o LPC possui dois tipos de excitação. Para gerar o sons sonoros, o LPC considera que as cordas vocais vibram periodicamente. Assim, faz o uso de um trem de impulsos uniformemente espaçados como excitação. Em contrapartida, os sons não vozeados são gerados a partir de uma excitação ruidosa. Além do tipo de excitação, o LPC realiza a síntese da voz com contagem do período para a excitação da voz, fator de ganho e parâmetros do trato vocal que são utilizados como coeficientes do preditor [47, 48, 49].

Na síntese do sinal de voz, o codificador LPC utiliza os parâmetros recebidos periodicamente, que contém informações sobre o modelo e a natureza da excitação, para reconstruir a fala por meio de um modelo matemático dado pela Equação 2.6, construído a partir do ganho e coeficientes do preditor, em que cada amostra é gerada a partir da entrada atual do sistema, somada a uma combinação linear da saída predita do trato vocal [50].

$$y[n] = \sum_{h=1}^{p} a_h y[n-h] + G_f x[n], \tag{2.6}$$

em que $y[n]$ é a n-ésima amostra de saída, a_h é o h-ésimo coeficiente do preditor, G_f é o fator de ganho, $x[n]$ é a entrada amostrada em um tempo n e p é a ordem do modelo.

A codificação por predição linear é a base para diversos algoritmos de codificação relevantes, como o RPE-LTP (*Regular Pulse Excited-Long Term Predictor*), CELP (*Code Excited Linear Prediction*), VSELP (*Vector Sum Excited Linear Predictive*), ACELP (*Algebraic Code Excited Linear*

Predictive), QCELP (*Qualcomm Code Excited Linear Predictive*), AMR-NB (*Adaptative Multi-Rate Narrowband*) e AMR-WB (*Adaptive Multirate Wideband*) [51, 52, 53, 54, 55].

2.4 Técnicas de Síntese de Voz

Sintetizadores de voz são sistemas capazes de reproduzir artificialmente a voz. A voz sintética por ser gerada pela concatenação de unidades acústicas ou por sistemas que incorporem a modelagem do aparelho fonador, como mecanismo de vibração das cordas vocais e característica do filtro representado pelo trato vocal.

A síntese de voz é bastante utilizada em sistemas de conversão texto-fala (*Text-to-Speech*– TTS), com diversas aplicações, tais como: consulta de *e-mails* por telefone, centro de atendimento eletrônico, sistemas de acessibilidade, codificadores de voz, entre outros.

Esta seção apresenta as principais sínteses difundidas na literatura, que são: articulatória, por concatenação e por formantes.

2.4.1 Síntese Articulatória

A produção da voz por meio da síntese articulatória dá-se pela modelagem do aparelho fonador humano. Nesse caso, as características dos articuladores que participam da produção da voz, bem como as do movimento da abertura glotal, como tensão das cordas vocais e pressão nos pulmões são levadas em consideração.

A modelagem da síntese articulatória requer o conhecimento de um conjunto de parâmetros, como area de abertura dos lábios, a constrição formada pela lâmina da língua, a abertura para as cavidades nasais, a área glotal média e a taxa de expansão ou contração do volume na região do trato vocal correspondente à faringe. No entanto, tais parâmetros geralmente são obtidos em 2D, a partir de análises de raios-X, porém o trato vocal real é naturalmente em 3D, dificultando assim a otimização desse tipo de modelo [56, 57].

O trabalho apresentado em [58] propõe uma síntese articulatória utilizando a técnica de regressão supervisionada e dados obtidos a partir de um gerador baseado em MOM, para treinar um modelo que faz a ligação entre a representação acústica da voz e os parâmetros articulatórios do sintetizador. O sintetizador foi avaliado mediante testes de audição, além de análise dos espectrogramas, obtendo bom desempenho ao sintetizar sinais de voz utilizados no treinamento do modelo.

A imitação da voz usando síntese articulatória com a modelagem feita com redes neurais é descrita em [59]. O modelo é construído para sintetizar sílabas e é capaz de fazer o mapeamento articulatório-acústico, e vice-versa, para sequências de consoantes-vogais, incluindo os efeitos co-articulatórios.

2.4.2 Síntese por Concatenação

A síntese concatenativa é caracterizada por produzir o som por meio da junção de segmentos correspondentes a unidades acústicas. Basicamente, esse tipo de síntese é realizada em três etapas, apresentadas na Figura 2.10.

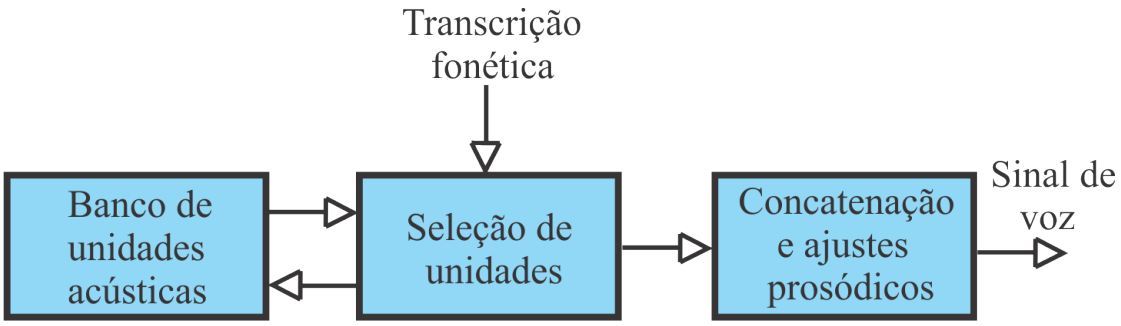

Figura 2.10: Diagrama de blocos da síntese de voz por concatenação. Adaptado de [3].

A primeira etapa consiste na formação do banco de unidades acústicas obtidas por meio da segmentação de locuções previamente gravadas. Em uma síntese por concatenação, a escolha do tamanho das unidades a serem utilizadas no processo de síntese é uma das decisões mais importantes, pois deve representar um compromisso entre inteligibilidade e naturalidade requerida. Várias são as possibilidades de tamanhos e quantidades que podem ser utilizadas.

Um dos segmentos que podem ser usados na síntese por concatenação são os difones, unidades formadas por uma dupla de fones. Ele inicia na metade do primeiro fone e termina na metade do fone seguinte. Sua vantagem consiste em conter inteiramente as transições entre os fones. No entanto, os difones incluem apenas parte dos vários efeitos coarticulatórios da língua falada, o que justifica o uso, mesmo que parcial, de unidades maiores, como os trifones.

Os trifones são segmentos que incluem um fone inteiro e suas transições à esquerda e à direita. Entretanto, devido à grande quantidade de trifones presentes na língua portuguesa, essas unidades são utilizadas como um complemento, para casos de sons especiais, de bancos baseados em unidades menores [60, 61].

Outras unidades que podem ser utilizadas na síntese por concatenação são as metades dos fones, sílabas, demissílabas, palavras e fonemas. A metade dos fones, se estende desde a fronteira entre fones até o ponto médio, ou se estende a partir do ponto médio até o final do fone. Entretanto, essa unidade quando utilizada de forma isolada apresenta dificuldade de representação da coarticulação.

As sílabas podem ser consideradas unidades naturais, uma vez que apresentam a coarticulação entre os fonemas que as formam e são mais importantes que as coarticulações presentes nos segmentos intra-sílabas. No entanto, a ausência dessas coarticulações diminui a qualidade do sinal sintetizado. Outra desvantagem desses segmentos é a sua grande quantidade na língua portuguesa, o que dificulta a construção de um banco utilizando esse tipo de segmento [62].

Com base nos mesmos princípios fonológicos das síbalas, as demissílibas são formadas a partir da divisão das sílabas em duas partes parcialmente sobrepostas, com o pico silábico (núcleo) pertencendo a ambas as partes. Como um exemplo, considere a sílaba tar, que possui

uma demissílaba inicial ta e uma demissílaba final ar. Uma desvantagem do uso desse tipo de segmento é que nem sempre é possível desprezar a interação que ocorre entre os segmentos pertencentes a sílabas diferentes [63, 64].

Além das unidades de concatenação expostas, a síntese de uma sinal de fala também pode ser realizada com palavras ou frases. A desvantagem desse tipo de unidade é o elevado número necessário em um sistema de síntese irrestrita, ou seja, aquela síntese que não é restrita a um conjunto de palavras ou frases.

A síntese por concatenação também pode ser realizada utilizando segmentos fonéticos. Sua vantagem consiste na pequena quantidade de unidades presentes na língua portuguesa, o que permite usar um pequeno banco de voz. Entretanto, a síntese utilizando estas unidades apresenta um comportamento não muito estável, que oscila entre falas sintetizadas com uma alto grau de naturalidade e falas sintetizadas com distorções desagradáveis [65]. Isso ocorre pelo fato de que os pontos de coarticulação passam a ser realizados nas fronteiras dos fones, dificultando a representação precisa do efeito de coarticulação, o que requer várias amostras de uma mesma unidade em diferentes contextos (alofones) [64].

Em uma síntese por concatenação é necessário levar em consideração a variação à qual as unidades de concatenação estão sujeitas de acordo com a posição ocupada dentro de uma frase ou com a entonação aplicada. Por exemplo, no caso da palavra **casa**, a pronúncia do primeiro fonema [a] é diferente do segundo fonema [a]. Assim, para manter uma entonação correta e natural, seria necessário considerar todas as variantes do fonema [a] como unidades de concatenação, que seriam escolhidas em função de regras gramaticais ou semânticas.

Após a escolha do tamanho e tipos de unidades a serem utilizadas na síntese, a próxima etapa consiste na seleção delas a partir da transcrição fonética da locução que se deseja sintetizar. Por fim, é realizada a concatenação das unidades acústicas, com a possibilidade do ajuste da energia, duração e frequência fundamental.

No entanto, algumas desvantagens podem ocorrer na síntese concatenativa: descontinuidades na envoltória espectral e descontinuidades de amplitude, de *pitch* e de fase entre os segmentos. As descontinuidades espectrais ocorrem quando os formantes de segmentos adjacentes não têm os mesmos valores e estão relacionados, principalmente, à coarticulação. Esse problema pode ser atenuado com a suavização das bordas dos segmentos [3, 61, 66].

2.4.3 Síntese por Formantes

A síntese por formantes é baseada na teoria fonte-filtro. Sua realização dá-se por meio de três componentes: fonte de excitação, características de filtragem do trato vocal e característica de radiação para o meio externo.

A fonte de excitação, para sinais sonoros, consiste em um gerador de impulsos uniformemente espaçados por um intervalo de tempo igual ao período de *pitch*. Na produção dos sons surdos, a excitação é representada por um gerador de ruídos.

O trato vocal é um filtro cuja função de transferência é composta por ressonadores cujo objetivo é modelar a frequência e largura de banda de cada formante. A função de transferência de um ressonador é dada por

$$R_n(z) = \frac{a_{1n}}{1 - a_{2n}z^{-1} - a_{3n}z^{-2}}, \tag{2.7}$$

em que a_{1n}, a_{2n} e a_{3n} são coeficientes relacionados à frequência central de ressonância, f_n, e à largura de banda do formante B_n, dada por

$$a_{3n} = -e^{2\pi B_n T_s}, \tag{2.8}$$

em que T_s é o período de amostragem em segundos,

$$a_{2n} = 2e^{-2\pi B_n T_s}\cos\left(2\pi f_n T_s\right), \tag{2.9}$$

e

$$a_{1n} = 1 - a_{3n} - a_{2n}. \tag{2.10}$$

Um sintetizador por formantes pode ser construido em série ou paralelo. A Figura 2.11 ilustra a associação em série dos ressonadores, que tem a vantagem de não necessitar de um ganho específico para cada ressonador, diferentemente da associação em paralelo, como ilustrado na Figura 2.12. A associação em série possui a desvantagem da função de transferência não ser modelada adequadamente para a produção dos sons fricativos e plosivos [4].

Na associação em paralelo, o sinal de excitação é aplicado à todos os ressonadores e suas saídas são somadas. Nesse caso tem-se um controle individual do ganho e da largura de banda de cada formante. Essa configuração produz os sons nasais, fricativos e plosivos, com melhor qualidade do que a estrutura em cascata. Em contrapartida, sua função de transferência não é modelada adequadamente para a produção de vogais.

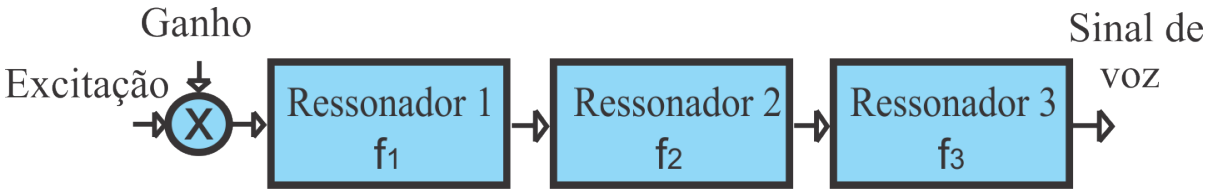

Figura 2.11: Diagrama de blocos de um sintetizador por formantes com configuração em série ou cascata. Adaptado de [3].

Para sintetizar a fala, os sintetizadores por formantes recebem periodicamente informações como amplitude, frequência fundamental do sinal sonoro e frequências e larguras de banda dos formantes. Um exemplo de sintetizador por formante bem difundido na literatura é o de Klatt [67, 68], controlado por 39 parâmetros que são atualizados a cada 5 ms. Nos últimos anos, o sintetizador de Klatt foi utilizado em [69] na geração da voz sintética.

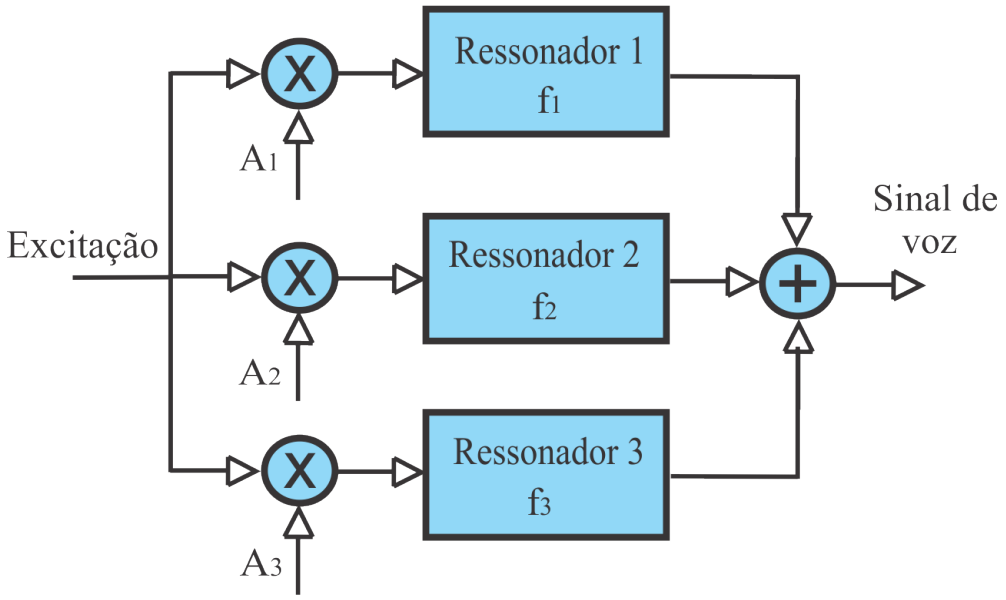

Figura 2.12: Diagrama de blocos de um sintetizador por formantes com configuração em paralelo. Adaptado de [3].

CAPÍTULO 3

Novo Modelo de Produção da Voz

O desenvolvimento de modelos que representem de forma mais fidedigna o processo de produção da voz é fundamental para a teoria de processamento de sinais de voz.

Por meio dessas modelagens é possível implementar ou melhorar algoritmos que abordem segmentação e codificação de voz, bem como aplicações que objetivem a detecção de patologias no aparelho fonador, mais precisamente nas cordas vocais, como é o caso do modelo apresentado nesta tese.

Este capítulo descreve um emulador de voz baseado na biofísica da fonação, que se distingue dos demais trabalhos da literatura por utilizar processos estocásticos e considerar o movimento de oscilação cicloestacionário das cordas vocais na geração dos sons sonoros.

Para isso, inicialmente o modelo fonte-filtro clássico, no qual o novo modelo de produção da voz é baseado, é apresentado. Em seguida, é descrito o desenvolvimento matemático realizado na caracterização da excitação da glote, bem com o modelo do pulso glotal, trato vocal e radiação dos lábios e narinas.

Foi desenvolvido um modelo matemático, que inclui um sistema linear do trato vocal, para sinais cicloestacionários, que modelam o sinal de excitação das cordas vocais, bem como é apresentada uma nova expressão matemática para a formação na voz, no domínio do tempo e da frequência.

A partir das expressões matemáticas obtidas, é possível a modificação dos seus parâmetros para emular a geração de vozes saudáveis e patológicas.

3.1 Modelo de Produção da Voz

Uma das técnicas mais utilizadas para representar a produção da voz é a fonte-filtro. Essa teoria é composta basicamente por três partes, como ilustrado na Figura 3.1: fonte de excitação, trato vocal e radiação.

No modelo do processo de geração da voz, é necessário inicialmente definir o tipo de sinal que se deseja analisar. Os sinais podem ser, basicamente, do tipo sonoro ou surdo e as características da fonte de excitação são definidas a partir das particularidades de cada padrão fonético.

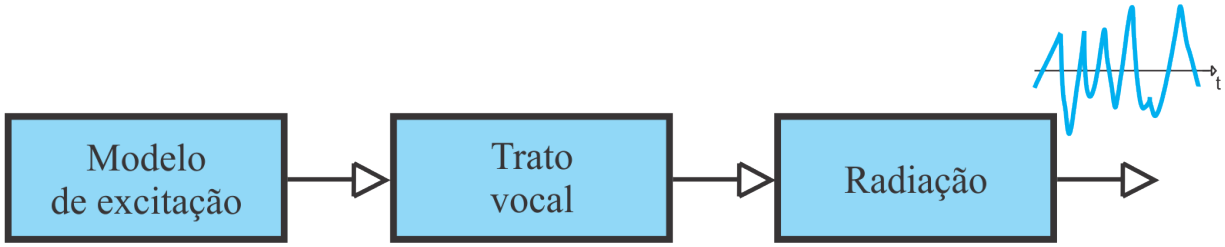

Figura 3.1: Diagrama de blocos básico de um sistema de produção de voz.

A fonte de excitação é a primeira etapa na construção do modelo de geração da voz e está relacionada ao estado da glote, que pode ser vozeado, significando que, pela propagação do fluxo de ar e da aproximação dos músculos que formam a glote, as cordas vocais vibram durante a produção de um determinado som.

Em contrapartida, a glote assume o estado surdo quando não houver vibração das cordas vocais, o que acontece durante a produção de um segmento desvozeado, devido ao fato de que os músculos que a formam estarem completamente separados de forma que o ar passa livremente.

Para os sinais sonoros, o fluxo glotal é formado por um série de harmônicos que são filtrados pelo trato vocal. O trato vocal é um filtro que tem o objetivo de identificar as formantes da estrutura supraglotal, que são modificadas para a geração de cada tipo de som vocálico.

Na etapa de radiação, as baixas frequências sofrem difração nos lábios, enquanto as altas frequências possuem maior diretividade, sendo mais suscetíveis ao efeito de reflexão. O resultado é a amplificação das altas frequências, com ganho médio de 6 dB por oitava [8].

A seguir, são descritos os detalhes sobre o novo modelo de formação da voz, proposto nesta tese, baseado no paradigma fonte-filtro.

3.2 Caracterização do Problema

No processo de produção da voz, uma pressão subglotal causa a separação das cordas vocais e, pelo efeito de Bernoulli, que explica uma queda de pressão supraglotal contra os lados internos de cada dobra vocal, se unem novamente e assim o ar percorre a glote com uma maior velocidade. O ciclo de abertura e fechamento da glote se repete, gerando um trem de pulsos que alimenta o trato vocal. Todo esse processo também só é possível porque as cordas vocais são elásticas [17].

Particularmente, no caso dos sons sonoros, objeto de estudo desta tese, esse procedimento provoca a vibração das cordas vocais. Em média, as cordas vocais vibram a cada período $T_o = 1/F_o\ s$, ou, em outras palavras, as cordas vocais vibram a uma taxa dada pela frequência fundamental, F_o.

No entanto, a frequência de vibração das cordas vocais é constantemente mudada ao se pronunciar diferentes padrões de entonação das sentenças. Dessa forma, uma determinada frequência F produzida por um determinado locutor tem seu valor alterado a todo momento durante um discurso. Nesse caso, no decorrer da pronúncia de uma locução, a frequência fun-

damental é obtida por um breve momento, sendo obtidas frequências maiores ou menores que a frequência média.

O modelo de produção da voz proposto nesta tese é composto por cinco etapas, como apresentadas na Figura 3.2: gerador de pulso, pulso glotal, ganho, trato vocal e radiação.

Diferentemente dos demais trabalhos encontrados na literatura, o novo modelo de formação da voz é baseado na biofísica da fonação e tem a característica de modelar o fluxo glotal levando em consideração o movimento cicloestacionário de vibração das cordas vocais.

Nesse caso, a variação da frequência fundamental no decorrer de um discurso, bem como um parâmetro ganho, relacionado à pressão do ar vindo do diafragma, adicionado ao fluxo glotal, são modelados com o propósito de obter um modelo de geração da voz que permita relacionar os parâmetros com dados biomédicos e que tenha a possibilidade de ajustar os parâmetros para emular patologias da voz.

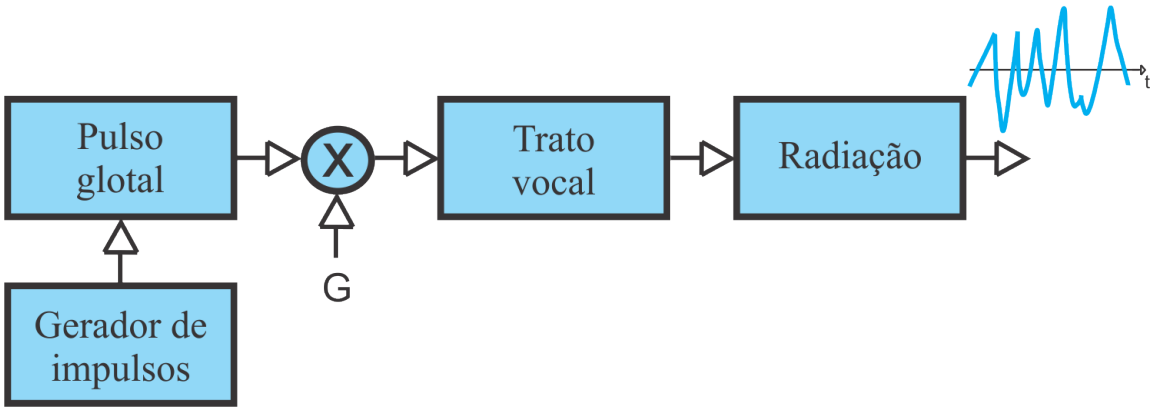

Figura 3.2: Diagrama de blocos do novo modelo de geração da voz.

No modo como a voz é gerada, a frequência de vibração é determinada pela elasticidade, tensão e massa das cordas vocais. Elas vibram naturalmente a uma menor frequência quando são mais compridas e grossas, enquanto as que são mais curtas e finas têm maior frequência de vibração.

Para a fisiologia que regula a frequência de vibração, o alongamento das cordas vocais de um determinado locutor deixa a parte vibratória das cordas vocais mais finas e esticadas, adicionando tensão e provocando aumento na frequência de oscilação das cordas vocais. No entanto, a diminuição da massa e o aumento da tensão são os fatores mais importantes que influenciam o movimento das cordas vocais [17].

A frequência de vibração das cordas vocais é principalmente controlada pela maior ou menor aplicação da tensão longitudinal. De forma secundária, ela é afetada pela tensão vertical obtida pela elevação ou abaixamento da laringe, bem como pela variação da pressão subglotal.

No modelo proposto, a fonte de excitação é controlada por um sinal (sinal de controle), $M(t)$, verificado nas cordas vocais cuja amplitude controla a liberação de pulsos glotais. Esse sinal comanda o mecanismo de vibração das cordas vocais, e matematicamente, a relação entre a variação de frequência de vibração e tensão pode ser expressa por

$$\Delta\omega = \alpha_1 \omega_o T(t), \tag{3.1}$$

em que $\Delta\omega$ é a variação de frequência dos impulsos da sequência de impulsos, $T(t)$ é a tensão mecânica nas cordas vocais e α_1 é uma constante de sensibilidade do processo.

$$M(t) = \alpha_2 T(t), \tag{3.2}$$

em que $M(t)$ é um sinal verificado nas cordas vocais e α_2 é uma constante de sensibilidade do processo.

Dessa forma

$$T(t) = \frac{1}{\alpha_2} M(t). \tag{3.3}$$

$$\Delta\omega = \omega_o \frac{\alpha_1}{\alpha_2} M(t). \tag{3.4}$$

Assim, é possível definir

$$\beta = \frac{\alpha_1}{\alpha_2}, \tag{3.5}$$

em que β é representa a relação entre as constantes de sensibilidade do processo.

A análise da produção dos sinais de voz realizada nesta tese, tendo como protótipo o modelo fonte-filtro, consiste em definir, principalmente, um modelo matemático para a fonte de excitação. A seguir, cada uma das etapas para o novo modelo de formação da voz é descrita.

3.2.1 Gerador de Impulsos Cicloestacionários

No processo de produção dos sons vozeados humanos, um pulso glotal $E(t)$, originado nos pulmões, é forçado através da abertura entre as cordas vocais, denominada glote. Nesse processo, quando as cordas vocais estão sob maior tensão, a glote diminui sua abertura e assim, por estarem sob maior tensão, vibram mais, contribuindo para a geração dos sons agudos da fala. Por outro lado, quando as cordas estão sob menor tensão, sua vibração é menor, contribuindo assim para os sons graves.

Esse processo de geração da voz pode ser modelado matematicamente como a passagem de um trem de impulsos $C(t)$ por um sistema linear e invariante ao deslocamento no tempo de resposta ao impulso igual ao pulso glotal $E(t)$.

O trem de impulsos $C(t)$ pode ser interpretado como a saída de um gerador de impulsos cicloestacionário, uma vez que consiste em uma sequência de impulsos no tempo cujas posições são controladas por um sinal cicloestacionário $M(t)$ que funciona como um sinal modulante.

Um processo estocástico $X(t)$ é cicloestacionário no sentido amplo se sua função média é periódica de período T_o, dada por [70]

$$\eta_X(t) \;=\; \eta_X(t \;+\; nT_o),\ \forall\ t \tag{3.6}$$

em que n é um inteiro arbitrário, e se a função autocorrelação $R_x(t, t+\tau)$ é periódica na variável t com período T_o, ou seja

$$R_X(t, t+\tau) \;=\; R_X(t \;+\; nT_o, t+\tau+nT_o),\ \forall\ t \tag{3.7}$$

Quando o sinal aleatório $M(t)$ não está presente, é possível considerar que as cordas vocais estão sob uma frequência média, F_o, e assim, o sinal $C(t)$ tem impulsos igualmente espaçados por uma duração T_o e pode ser escrito por

$$C(t) = \sum_{k=-\infty}^{\infty} \delta\left(t - kT_o\right) \tag{3.8}$$

Nesse caso, a representação de $C(t)$ em Série de Fourier é dada por

$$C(t) = a_0 + \sum_{k=1}^{\infty} a_k \cos(k\omega_o t) + \sum_{l=1}^{\infty} b_l \text{sen}(l\omega_o t), \tag{3.9}$$

em que

$$\begin{aligned} a_0 &= \frac{1}{T_o}\int_{-\frac{T_o}{2}}^{\frac{T_o}{2}} \delta(t)dt = \frac{1}{T_o}, \\ a_k &= \frac{2}{T_o}\int_{-\frac{T_o}{2}}^{\frac{T_o}{2}} \delta(t)\cos(k\omega_o t)dt = \frac{2}{T_o}, \\ b_l &= \frac{2}{T_o}\int_{-\frac{T_o}{2}}^{\frac{T_o}{2}} \delta(t)\text{sen}(l\omega_o t)dt = 0, \end{aligned} \tag{3.10}$$

de modo que

$$C(t) = \frac{1}{T_o} + \frac{2}{T_o}\sum_{k=1}^{\infty} \cos(k\omega_o t). \tag{3.11}$$

No caso em que o sinal $M(t)$ está presente, pode-se reescrever o sinal $C(t)$ como [71]

$$C(t) = \sum_{k=-\infty}^{\infty} \delta(t + \beta\int_{-\infty}^{t} M(\tau)d\tau - kT_o). \tag{3.12}$$

e sua representação em Série de Fourier é dada por

$$C(t) = \frac{1}{T_o} + \frac{2}{T_o}\sum_{k=1}^{\infty}\cos\left(k\left(\omega_o t + \omega_o\beta\int_{-\infty}^{t} M(\tau)d\tau + \phi_o\right)\right), \tag{3.13}$$

em que

1. A frequência média ω_o correspondente ao período T_o do trem de impulsos não modulado.

2. A fase ϕ_o é diretamente proporcional à posição inicial δ_o dos impulsos, $\phi_o = \omega_o\delta_o$, e é uniformemente distribuída em um intervalo de comprimento 2π.

3. O sinal $M(t)$ é um sinal verificado nas cordas vocais e tem dimensão mV.

4. O parâmetro β pode ser visto como uma constante de sensibilidade do processo de modulação da glote e possui dimensão em V^{-1}.

Para este modelo, a variação de frequência de vibração das cordas vocais também é diretamente proporcional ao sinal $M(t)$, ou seja, $\Delta\omega = \omega_o\beta M(t)$, e essa variação ocorre no intervalo

$$0 \leq \Delta\omega \leq \omega_o\beta M_{max}, \tag{3.14}$$

em que M_{max} é a amplitude máxima do sinal aleatório $M(t)$. Assim sendo, o desvio de frequência é tal que

$$0 \leq \beta\omega_o M(t) \leq \omega_m, \tag{3.15}$$

em que ω_m é a frequência máxima de oscilação das cordas vocais e a amplitude de $M(t)$ varia no intervalo

$$0 \leq M(t) \leq \frac{\omega_m}{\beta\omega_o}. \tag{3.16}$$

Análise Espectral do Sinal C(t)

Antes de iniciar a análise espectral do sinal aleatório $C(t)$ é apropriado considerar

$$\phi(t) = \omega_o\beta\int_{-\infty}^{t} M(\tau)d\tau \tag{3.17}$$

e reescrever $C(t)$ como

$$C(t) = \frac{1}{T_o} + \frac{2}{T_o}\sum_{k=1}^{\infty}\cos\left(k\left(\omega_o t + \phi(t) + \phi_o\right)\right). \tag{3.18}$$

Como ϕ_o é uma variável uniformemente distribuída em um intervalo de comprimento 2π, então o valor esperado de $C(t)$ é uma constante $\frac{1}{T_o}$. Assim, se for possível escrever a autorrelação de $C(t)$ em termos apenas da diferença entre os instantes de observação t e $t+\tau$ então é possível

afirmar que $C(t)$ é um processo aleatório estacionário em sentido amplo. A autocorrelação de $C(t)$ pode ser obtida, como descrito a seguir,

$$\begin{aligned} R_C(\tau) = &\left(\frac{1}{T_o} + \frac{2}{T_o} \sum_{k=1}^{\infty} \cos(k(\omega_o t + \phi(t) + \phi_o)) \right) \\ &\cdot \left(\frac{1}{T_o} + \frac{2}{T_o} \sum_{l=1}^{\infty} \cos(l(\omega_o (t+\tau) + \phi(t+\tau) + \phi_o)) \right). \end{aligned} \tag{3.19}$$

Após realizar o produto dos termos, aplicar o valor esperado e usar o fato que o cosseno composto com uma variável aleatória uniformemente distribuída em um intervalo de comprimento 2π tem média nula, encontra-se que

$$R_C(\tau) = \frac{1}{T_o^2} + \frac{2}{T_o^2} \sum_{k=1}^{\infty} E\left[\cos(k(\omega_o \tau + (\phi(t+\tau) - \phi(t))))\right]. \tag{3.20}$$

De acordo com [72], o processo aleatório $\phi(t+\tau)$ pode ser aproximado por uma estimativa linear de erro médio ao quadrado mínimo, de modo que

$$\phi(t+\tau) \approx \phi(t) + \tau\phi'(t). \tag{3.21}$$

Assim, $R_C(\tau)$ pode ser reescrito como

$$R_C(\tau) = \frac{1}{T_o^2} + \frac{2}{T_o^2} \sum_{k=1}^{\infty} E\left[\cos(k(\omega_o \tau + \tau\phi'(t)))\right]. \tag{3.22}$$

Aplicando a relação de Euler da representação do cosseno, pode-se reescrever $R_C(\tau)$ como

$$\begin{aligned} R_C(\tau) = \frac{1}{T_o^2} &+ \frac{1}{T_o^2} \sum_{k=1}^{\infty} E\left[e^{jk(\omega_o \tau + \tau\phi'(t))}\right] \\ &+ \frac{1}{T_o^2} \sum_{k=1}^{\infty} E\left[e^{-jk(\omega_o \tau + \tau\phi'(t))}\right]. \end{aligned} \tag{3.23}$$

Neste ponto do desenvolvimento é importante lembrar que como $\phi(t)$ é uma fase instantânea definida pela Expressão 3.17, então sua derivada é uma frequência instantânea aleatória, representada por $\omega(t)$. Assim $R_C(t)$ pode ser escrita como

$$\begin{aligned} R_C(\tau) = \frac{1}{T_o^2} &+ \frac{1}{T_o^2} \sum_{k=1}^{\infty} e^{jk\omega_o \tau} E\left[e^{jk\tau\omega(t)}\right] \\ &+ \frac{1}{T_o^2} \sum_{k=1}^{\infty} e^{-jk\omega_o \tau} E\left[e^{-jk\tau\omega(t)}\right]. \end{aligned} \tag{3.24}$$

Os dois valores esperados nesta expressão correspondem, respectivamente, à função característica de $\omega(t)$ e seu conjugado complexo amostradas em $k\tau$. Se essa função for denotada $\varphi_{\omega(t)}(v)$, então $R_C(\tau)$ pode ser escrita como

$$\begin{aligned} R_C(\tau) = \frac{1}{T_o^2} + \frac{1}{T_o^2}\sum_{k=1}^{\infty} e^{jk\omega_o\tau}\varphi_{\omega(t)}(k\tau) \\ + \frac{1}{T_o^2}\sum_{k=1}^{\infty} e^{-jk\omega_o\tau}\varphi^*_{\omega(t)}(k\tau). \end{aligned} \tag{3.25}$$

Por esta expressão, com o auxílio da aproximação $\phi(t+\tau) \approx \phi(t) + \tau\phi'(t)$, é possível escrever $R_C(\tau)$ apenas em função de τ e dizer que $C(t)$ é aproximadamente estacionário em sentido amplo.

A densidade espectral de potência de $C(t)$ pode ser obtida calculando a Transformada de Fourier de $R_C(\tau)$. Usando a definição de função característica e a definição de transformada de Fourier de sinais definidos em tempo contínuo, pode-se escrever a DEP de $C(t)$ como

$$\begin{aligned} S_C(v) = \frac{1}{T_o^2}\sum_{k=1}^{\infty}\int_{-\infty}^{\infty} f_\Omega(\omega)\int_{-\infty}^{\infty} e^{-j(v-k\omega_o-k\omega)\tau}d\tau d\omega \\ + \frac{1}{T_o^2}\sum_{k=1}^{\infty}\int_{-\infty}^{\infty} f_\Omega(\omega)\int_{-\infty}^{\infty} e^{-j(v+k\omega_o+k\omega)\tau}d\tau d\omega \\ + \frac{2\pi}{T_o^2}\delta(v). \end{aligned} \tag{3.26}$$

Usando o fato, da teoria do impulso de área unitária, que

$$\delta(t-t_o) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-j\omega(t-t_o)}d\omega, \tag{3.27}$$

pode-se reescrever a Expressão da DEP $S_C(v)$ como

$$\begin{aligned} S_C(v) = \frac{2\pi}{T_o^2}\sum_{k=1}^{\infty}\int_{-\infty}^{\infty} f_\Omega(\omega)\delta(v-k\omega_o-k\omega)d\omega \\ + \frac{2\pi}{T_o^2}\sum_{k=1}^{\infty}\int_{-\infty}^{\infty} f_\Omega(\omega)\delta(v+k\omega_o+k\omega)d\omega \\ + \frac{2\pi}{T_o^2}\delta(v). \end{aligned} \tag{3.28}$$

Usando então as propriedades da filtragem e do escalonamento no tempo do impulso de área unitária pode-se reescrever $S_C(\nu)$ como

$$\begin{aligned} S_C(\nu) = &\frac{2\pi}{T_o^2}\sum_{k=1}^{\infty}\frac{1}{k}f_\Omega\left(\frac{\nu}{k}-\omega_o\right) \\ &+\frac{2\pi}{T_o^2}\sum_{k=1}^{\infty}\frac{1}{k}f_\Omega\left(\frac{-\nu}{k}-\omega_o\right) \\ &+\frac{2\pi}{T_o^2}\delta(\nu), \end{aligned} \tag{3.29}$$

que pode ainda ser reescrita como

$$S_C(\omega) = \frac{2\pi}{T_o^2}\delta(\omega)+\frac{2\pi}{T_o^2}\sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty}\frac{1}{|k|}f_\Omega\left(\frac{\omega}{k}-\omega_o\right). \tag{3.30}$$

Usando o fato que

$$\phi(t) = \omega_o\beta\int_{-\infty}^{t}M(\tau)d\tau \tag{3.31}$$

e que

$$\Omega(t) = \omega_o\beta M(t) \tag{3.32}$$

então

$$f_{\Omega(t)}(\omega) = \frac{1}{\omega_o\beta}f_{M(t)}\left(\frac{\omega}{\omega_o\beta}\right). \tag{3.33}$$

Substituindo esse resultado em 3.34, pode-se reescrever $S_C(\omega)$ como

$$S_C(\omega) = \frac{2\pi}{T_o^2}\delta(\omega)+\frac{2\pi}{\beta\omega_o T_o^2}\sum_{\substack{k=-\infty \\ k\neq 0}}^{\infty}\frac{1}{|k|}f_{M(t)}\left(\frac{\omega}{k\beta\omega_o}-\frac{1}{\beta}\right). \tag{3.34}$$

Análise da Distribuição de $M(t)$

Após a análise espectral do sinal de excitação $C(t)$, o próximo passo é analisar a distribuição de probabilidade do sinal $M(t)$.

Para $M(t)$, é proposto um processo aleatório estacionário contínuo com distribuição de probabilidade igual à distribuição dos pontos de cruzamento por zero do sinal de voz.

Essa proposta tem como base o modelo de produção da voz, que é gerada por um sistema linear e invariante no tempo, no qual não há geração de novas frequências quando o fluxo glotal atravessa o trato vocal. Nesse caso, a variação da tensão e, consequentemente, da frequência de oscilação das cordas vocais está diretamente ligada aos pontos de cruzamento por zero obtido na forma de onda da voz.

A frequência fundamental consiste em uma frequência média alcançada por cada orador. No entanto, em um discurso, a taxa de variação das cordas vocais pode ser maior ou menor que

a frequência fundamental. Diante do exposto, duas distribuições de probabilidade são usadas para modelar a variação de amplitude do sinal de controle M(t): Gamma unilateral e Rayleigh.

1. Distribuição Gamma Unilateral

   A distribuição Gamma unilateral pode ser caracterizada pela FDP

$$f_X(x) = \frac{1}{\Gamma(k_x)\theta^{k_x}} x^{k_x-1} e^{-\frac{x}{\theta}} u(x), \tag{3.35}$$

   em que k_x e θ representam, respectivamente, o parâmetro de formato e escala e $u(x)$ a função degrau unitário.

   A Figura 3.3 ilustra o comportamento da FDP da distribuição Gamma unilateral para três valores dos parâmteros k e θ.

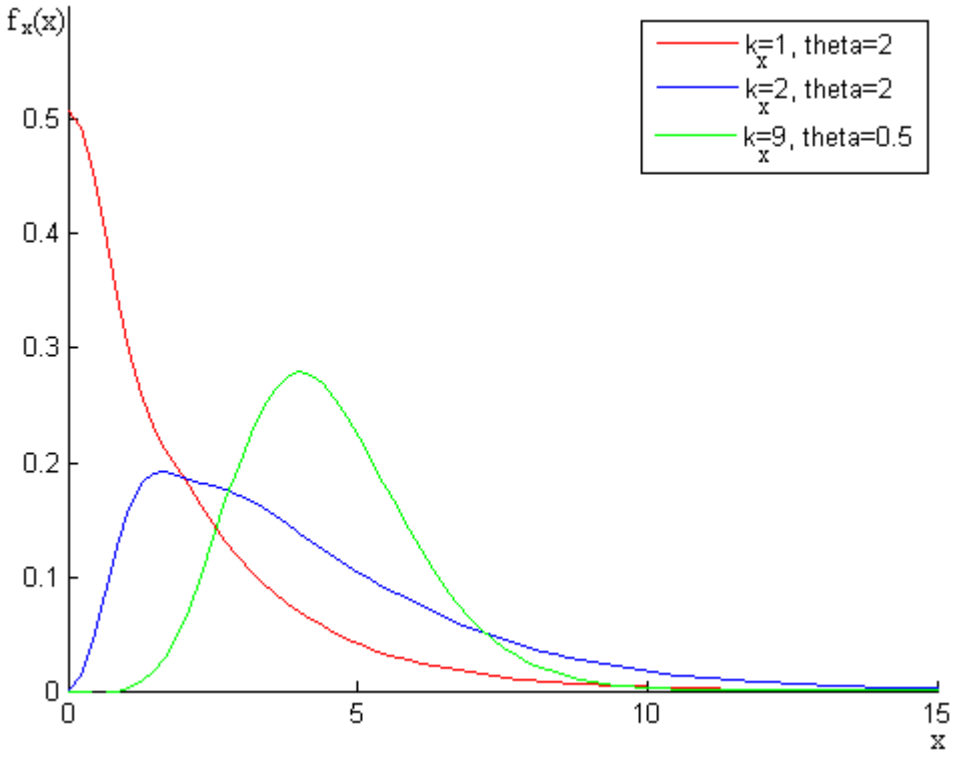

Figura 3.3: Comportamento da FDP da distribuição Gamma.

2. Distribuição Rayleigh

   A distribuição Rayleigh pode ser caracterizada pela FDP

$$p_X(x) = \frac{x}{\sigma^2} e^{-\frac{x^2}{(2\sigma^2)}} \text{para } x \geq 0, \tag{3.36}$$

   em que σ é o parâmetro de escala da FDP Rayleigh.

   A Figura 3.4 apresenta a FDP Rayleigh unilateral para três valores do parâmetro σ

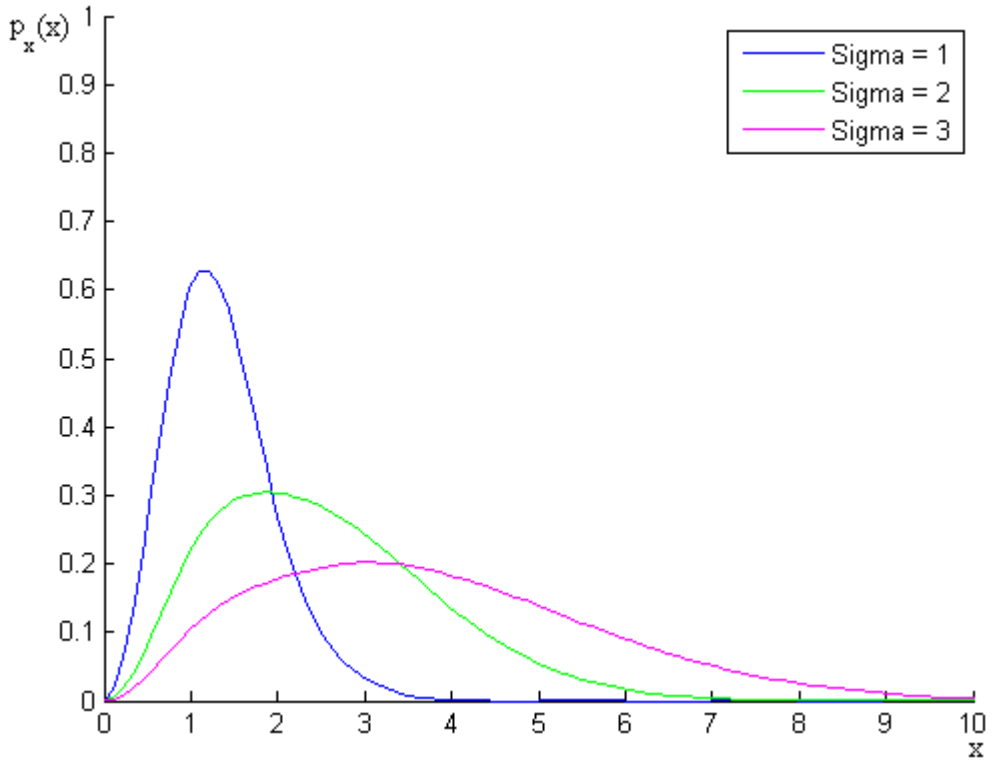

Figura 3.4: Comportamento da FDP da distribuição Rayleigh.

3.2.2 Modelo do Pulso Glotal

Na literatura é possível encontrar alguns modelos matemáticos para representar analiticamente o pulso glotal. Apesar das diferentes quantidades de parâmetros entre os modelos, todos possuem semelhanças em suas características, como representar o pulso glotal sempre positivo ou nulo, considerar o fluxo glotal quase periódico e como uma função contínua no tempo.

Além disso, as funções que representam o fluxo glotal são diferenciáveis no tempo, com exceção dos instantes de abertura e fechamento da glote, e a fase de abertura é maior que a de fechamento da glote. O modelo glotal de Rosenberg [73], o de Fant [74], o de Liljencrants-Fant (LF) [75], e o de Klatt [76], são alguns dos modelos de fluxo glotal encontrados na literatura.

Por ser utilizado em diversas modelagens da voz, o modelo da derivada do pulso glotal de Liljencrants-Fant foi escolhido para ser usado nesta tese.

O modelo LF representa a derivada do fluxo glotal e é dividido em dois segmentos, cujas fases, modelos e parâmetros estão ilustradas nas Figuras 3.5 e 3.6, respectivamente.

O primeiro compreende o processo de abertura das cordas vocais. Esse segmento inicializa no instante t_o, que é quando as cordas vocais estão fechadas, até o instante t_e, quando a glote volta ao seu estado incial, após sua abertura, cuja derivada assume seu valor máximo negativo, $-E_e$ [10]. Matematicamente, esse segmento é dado por

$$E(t) = E_o e^{\alpha t} sen(\omega_g t),\ t_o \leq t \leq t_e, \tag{3.37}$$

em que ω_g é a taxa de aumento da amplitude, determinada por α, e E_o é um fator de escala para alcançar uma área de balanço.

O segundo segmento do pulso glotal consiste em uma função exponencial que modela a fase de retorno da excitação principal até a fase de fechamento total [12]. Esse segmento começa no instante t_e e termina no instante t_c, cuja duração é T_b. Matematicamente, esse segmento pode ser descrito por

$$E(t) = \frac{-E_e}{\epsilon T_a}\left(e^{-\epsilon(t - t_e)} - e^{-\epsilon T_b}\right),\ t_e \leq t \leq t_c, \tag{3.38}$$

em que ϵ é uma constante de decaimento para a fase de recuperação da exponencial.

O principal parâmetro para o segundo segmento é T_a, que representa a eficiência para o retorno de fase. Para o modelo de Liljencrants-Fant, o fluxo glotal, $U(t)$, é dado por

$$U(t) = \begin{cases} \frac{E_o e^{\alpha t} sen\left(\omega_g t - \arctan\frac{\omega_g}{\alpha}\right)}{\sqrt{\alpha^2 + \omega_g}} + \frac{E_o \omega_g}{\alpha^2 + \omega_g}, & t_o \leq t \leq t_e, \\ \frac{E_e}{\epsilon^2 T_a}\left[e^{-\epsilon(t-t_e)} + \epsilon e^{-\epsilon T_b}\left(t - \left(t_c + \frac{1}{\epsilon}\right)\right)\right], & t_e \leq t \leq t_c. \end{cases} \tag{3.39}$$

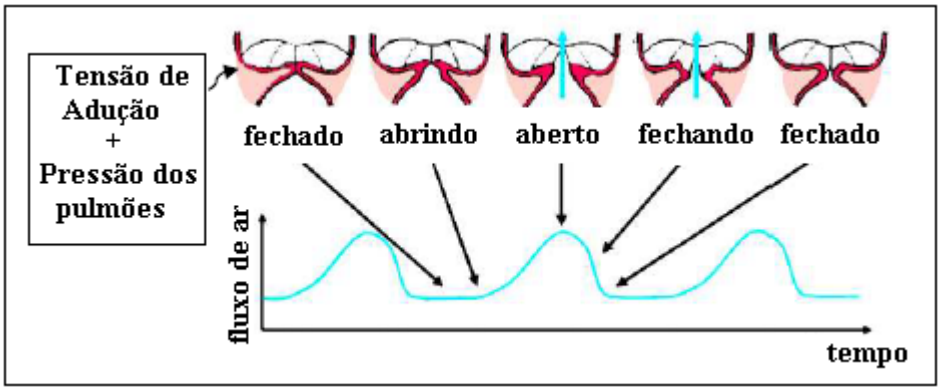

Figura 3.5: Ilustração das fases das cordas vocais no momento da fonação. Adaptado de [10, 11].

Análise Espectral do Pulso Glotal

Segundo o modelo LF, o pulso glotal no domínio do tempo, é dado por

$$E(t) = \begin{cases} E_o e^{\alpha t} sen(\omega_g t), & t_o \leq t \leq t_e, \\ \frac{-E_e}{\epsilon T_a}\left(e^{-\epsilon(t - t_e)} - e^{-\epsilon T_b}\right), & t_e < t \leq t_c. \end{cases} \tag{3.40}$$

Sua representação no domínio da frequência é dada pela Expressão D.27, apresentada no Anexo D.

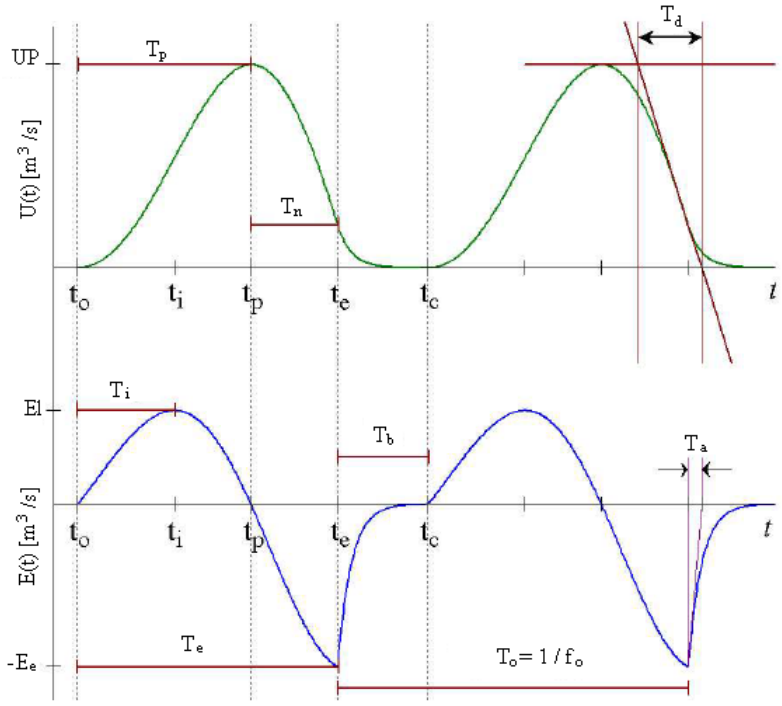

Figura 3.6: Ilustração do pulso (E(t)) do modelo LF e seu pulso de fluxo glotal (U(t)). Adaptado de [12].

$$
\begin{aligned}
E(\omega) = & \frac{\sqrt{\alpha_e^2 + \beta_e^2\omega^2}\, e^{-j\left(\omega t_e - tg^{-1}\left(\frac{\beta_e}{\alpha_e}\omega\right)\right)}}{(\omega + j\alpha)^2 - \omega_g^2} \\
& - \frac{\sqrt{\alpha_o^2 + \beta_o^2\omega^2}\, e^{-j\left(\omega t_o - tg^{-1}\left(\frac{\beta_o}{\alpha_o}\omega\right)\right)}}{(\omega + j\alpha)^2 - \omega_g^2} \\
& + \frac{E_e}{\epsilon T_a}(t_c - t_e) e^{\frac{-j\omega}{2}(t_c + t_e)} \\
& . \left[e^{-\epsilon T_b} \mathrm{Sa}\left(\frac{(t_c - t_e)}{2}\omega\right) - \frac{1}{2} e^{-\frac{\epsilon}{2}(t_c - t_e)} \mathrm{Sa}\left(j\frac{(t_c - t_e)}{2}(\epsilon + j\omega)\right)\right].
\end{aligned}
\tag{3.41}
$$

3.2.3 O Trato Vocal

Trato vocal é o espaço compreendido entre as cordas vocais e os lábios. No processo de geração da voz, o fluxo glotal é a entrada para o trato vocal, cujos músculos causam a movimentação dos articuladores que por sua vez, mudam a forma do trato vocal causando a produção de diferentes sons.

Comparado ao movimento das cordas vocais, o trato vocal muda de forma relativamente devagar. O tempo mínimo necessário para que os nervos e músculos consigam variar as articulações que participam da formação da fala corresponde à duração de um fone. Essa duração é da ordem de 50 ms, o que representa a emissão de 20 fones por segundo [25, 77].

Durante a produção de voz, o trato vocal é excitado por um gerador de pulsos produzidos pelas cordas vocais para a formação dos sons sonoros, e, no caso dos sons não sonoros, por ar turbulento passando através das constrições do trato. O trato vocal atua como um filtro ressonante, cujas diferentes configurações de articuladores definem as distintas frequências formantes, que têm o objetivo de moldar o espectro de frequência do som que se propaga através das suas cavidades. No geral, para a geração dos fones, são necessárias de três a cinco formantes.

As frequências formantes estão relacionadas às características biofísicas do trato vocal de cada ser humano. O movimento dos articuladores durante a fonação foi objeto de estudo em 1971, por Lindblom e Sundberg [78]. Por imagens de raio-x, eles verificaram que a primeira formante está relacionada ao deslocamento, no sentido vertical, da língua, ou seja, da abertura da mandíbula. Por outro lado, o segundo formante está relacionado ao movimento da língua no sentido horizontal, enquanto o terceiro fomante ao grau de obstrução formado entre a faringe e língua. O quatro formante está associado à posição vertical da laringe. Além disso, o estudo relata que a posição dos lábios pode aumentar o comprimento do trato vocal, ocasionando em uma diminuição das frequências formantes.

Nesta tese, as características de frequência do trato vocal são estimadas pelo método LPC, que representa amostras futuras pela combinação linear de amostras precedentes, além de determinar a frequência fundamental, espectro, formantes, entre outros parâmetros [47, 48, 49].

No LPC, a predição linear é representada pelo modelo

$$\hat{x}[n] = \sum_{h=1}^{p} a_h x[n-h], \tag{3.42}$$

em que $\hat{x}[n]$ é a predição linear de $x[n]$ e p é a ordem do preditor, que varia entre entre 8 a 14, sendo atribuído a ordem 10 na maioria das aplicações.

A diferença entre a amostra atual, $x[n]$, e a predita, $\hat{x}[n]$, é denominada erro residual ou erro de predição. Esse erro é dado por

$$e[n] = x[n] - \hat{x}[n]. \tag{3.43}$$

Dessa forma,

$$e[n] = x[n] - \hat{x}[n] = x[n] - \sum_{h=1}^{p} a_h x[n-h]. \tag{3.44}$$

O problema da estimação do trato vocal pelo método LPC consiste em determinar os coeficientes mais adequados de forma que o erro de predição seja o menor possível. No LPC, com a excessão da periodicidade, as propriedades espectrais da fala são representadas nos coeficientes LPC [44]. A literatura dispõe de vários métodos para a obtenção dos parâmetros LPC, como o método da covariância, o método da autocorrelação, a formulação do filtro inverso, a formulação da estimação espectral, a formulação da máxima verossimilhança e a formulação do produto interno [9].

No entanto, um dos métodos mais utilizados na literatuda para o cálculo dos coeficientes do preditor, também utilizado nesta tese, é o da autocorrelação. Inicialmente, o sinal é janelado

usando a janela de Hamming em intervalos que assegure a sua estacionariedade, ou seja, 20 ms, sendo matematicamente representado por

$$x_j[n] = x[n]w_h[n], \tag{3.45}$$

em que $x_j[n]$ representa o sinal janelado e $w_h[n]$ a janela de Hamming.

O método da autocorrelação é baseado na minimização do valor médio quadrático do erro de predição. Assim,

$$E = \sum_{n=0}^{\infty}[e[n]]^2 = \sum_{n=0}^{\infty}[x_j[n] - \sum_{h=1}^{p} a_h x_j[n-h]]^2. \tag{3.46}$$

A Expressão 3.46 é minimizada fazendo

$$\frac{\partial E}{\partial a_h} = 0, \text{ para } h = 1,2,...,p. \tag{3.47}$$

e produzindo p equações lineares,

$$\sum_{n=-\infty}^{\infty} x_j[n-i]x_j[n] = \sum_{h=1}^{p} a_h \sum_{n=-\infty}^{\infty} x_j[n-i]x_j[n-h], \text{ para } i = 1,2,...,p. \tag{3.48}$$

Uma vez que

$$R_{xx}(i) = \sum_{n=1}^{N-1} x[n]x[n-i], \text{ para } i = 1,2,...,p, \tag{3.49}$$

em que N representa o número de amostras no quadro em análise, então

$$\sum_{h=1}^{p} a_h R_{xx}(i-h) = R_{xx}(i), \text{ para } i = 1,2,...,p. \tag{3.50}$$

Os coeficientes do preditor são obtidos a partir da solução das Expressões 3.50, utilizando o método de Levinson-Durbin [6].

Após a estimação dos coefientes LPC, a função de transferência do trato vocal é obtida por

$$H(z) = \frac{1}{1-\sum_{h=1}^{p} a_h z^{-h}}, \tag{3.51}$$

que consiste em um filtro apenas de pólos, que pode ser escrita em um círculo de raio unitário, tal que $z = e^{-j\omega}$, como

$$H(z) = \frac{1}{1-\sum_{h=1}^{p} a_h e^{-jh\omega}}. \tag{3.52}$$

3.2.4 Densidade Espectral de Potência para o Sinal de Voz

O protótipo para a geração de voz proposto nesta tese é baseado no modelo fonte-filtro, cuja principal diferença é a modelagem da vibração cicloestacionária das cordas vocais. O modelo apresenta a voz como um sinal produzido a partir de um sistema linear e invariante no tempo, no intervalo no qual a voz é considerada cicloestacionária, tipicamente de 16 a 32 ms, sendo possível estimar o comportamento do sinal de voz no domínio da frequência.

O sistema fonte-filtro proposto considera a geração da voz a partir de etapas independentes, que são: modelo de excitação, trato vocal e radiação.

Para o novo modelo de produção da voz, o modelo de excitação leva em consideração o movimento cicloestacionário das cordas vocais, dado em função dos seus parâmetros físicos como tensão, massa e comprimento. Uma vez que, para um determinado orador, a massa das cordas vocais é fixa e o comprimento varia de forma moderada, é considerado que a vibração das cordas vocais está fortemente relacionadas à tensão longitudinal aplicada a elas.

Nesse contexto, é considerada que a frequência de oscilação das cordas vocais é diretamente proporcional à tensão em que elas são submetidas, controladas por um sinal cujas características está presente na forma de onda do sinal de voz.

O sinal de controle é dado no domínio do tempo, cujo período é inversamente proporcional à tensão longitudinal aplicada às cordas vocais. O novo modelo de produção de voz considera, então, que o sinal de voz é resultado da convolução entre o sinal decorrente do gerador de impulsos cicloestacionário controlado pela tensão, pulso glotal, resposta ao impulso do trato vocal e radiação nos lábios, como ilustrado na Figura 3.2 e matematicamente apresentado por

$$V(t) = GC(t) * E(t) * H(t) * L(t), \tag{3.53}$$

em que $V(t)$ representa o sinal de saída, $C(t)$ o trem de impulsos controlado pelo sinal de tensão, $E(t)$ o pulso glotal, $H(t)$ a resposta ao impulso do trato vocal, G um ganho positivo relacionado à potência do ar proveniente do diafragma e $L(t)$ o efeito da radiação.

O efeito da radiação nos lábios e narinas é conjuntamente representado por um filtro passa-altas aproximado por uma derivada de primeira ordem no domínio do tempo, significando que a derivada do fluxo glotal é a excitação para o trato vocal. A etapa de radiação amplifica as altas frequências com ganho médio de 6 dB por oitava e matematicamente é dado por

$$L(z) = 1 - \alpha z^{-1}, \tag{3.54}$$

em que α é o coeficiente de radiação dos lábios/narina que, usualmente, assume valores entre 0,95 e 0,99 para que o zero fique localizado dentro do círculo unitário no plano z.

O modelo proposto assume que cada um dos subsistemas para a geração da voz é um filtro linear e invariante no tempo. Nesse cenário, a cada etapa do processo de geração, a densidade

espectral de potência resultante é dada pela multiplicação do sinal de entrada pelo módulo ao quadrado da resposta em frequência do filtro [79, 80].

Dessa forma, a densidade espectral de potência da voz dada pelo novo modelo de produção é dada pela Expressão 3.55, em que $S_c(\omega)$ representa a DEP do trem de impulsos que excita as cordas vocais, $|E(\omega)|^2$, resposta em frequência do modelo do pulso glotal, $|G|^2$ ganho associado à potência do ar e $|H(\omega)|^2$ e $|L(\omega)|^2$ a seletividade em frequência do trato vocal e efeito da radiação, respectivamente.

$$S_V(\omega) = G^2 S_c(\omega)|E(\omega)|^2|H(\omega)|^2|L(\omega)|^2. \tag{3.55}$$

O próposito do modelo de produção de voz desenvolvido é a possibilidade de, a partir das suas expressões matemáticas, realizar ajustes de parâmetros para emular patologias nas cordas vocais.

As Expressões 3.53 e 3.55 produzidas pelo novo modelo, no domínio do tempo e frequência respectivamente, representam a descrição em tempo e frequência do sinal de saída do emulador de voz.

Uma vez que, diante de patologias nas cordas vocais, sua massa e seu comportamento vibratório são alterados, os parâmetros das expressões obtidas, como função distribuição de probabilidade, constante de sensibilidades e representação do trem de impulsos cicloestacionário no tempo, podem ser ajustados de forma a se adaptarem a uma voz saudável, bem como às características de cada patologia.

CAPÍTULO 4

Resultados

A fim de analisar o desempenho do novo modelo de geração de voz, três locuções de um orador masculino e três locuções de um orador feminino foram aleatoriamente selecionadas. As locuções são provenientes de bancos de dados de voz, gravadas por locutores paulistas, do interior do Estado de São Paulo. As sentenças foram gravadas a uma taxa de $22,05 \times 10^3$amostras/s e quantizadas com 16 *bits* por amostra, para o orador masculino, e 32 *bits* por amostra para o orador feminino. As locuções têm, em média, três segundos e foram gravadas com o mínimo de ruído possível.

São apresentados resultados para cada etapa de geração da voz com o novo modelo de geração baseado na teoria fonte-filtro, com o diferencial de modelar o movimento cicloestacionário das cordas vocais. Todo o processamento é realizado no intervalo em que o sinal de voz é considerado estacionário, ou seja, a cada 20 ms, com particionamento utilizando a janela de Hamming.

Embora o modelo proposto nesta tese seja relacionado à geração de sons sonoros, os testes foram realizados com locuções, que incluem sons surdos e sonoros. Essa escolha foi feita pelo fato da locução conter maior quantidade de amostras ao se comparar com um único fonema, otimizando os resultados expostos.

4.1 Sinal de Controle Cicloestacionário

A forma de onda do sinal de voz é o resultado do processo de geração da fala. Basicamente, a fala é formada a partir de uma excitação cicloestacionária ou um ruído de espectro largo, semelhante ao ruído branco.

As excitações são modeladas pelo trato vocal, composto pela faringe, laringe, boca, língua, dentes, entre outros, que formam a função de transferência do trato vocal, para produzir a forma de onda da voz. Por meio dela, características da fala podem ser observadas, como a presença de sinais sonoros e surdos.

Os sinais sonoros têm na sua forma de onda uma cicloestacionariedade provida pelo tipo de excitação no seu processo de geração. Além disso, esse sinais apresentam energia elevada

e baixa taxa de cruzamento por zero. Por sua vez, os sinais surdos, devido ao seu processo de composição, possuem características ruidosas, baixa energia e alta taxa de cruzamento por zero.

Além dessas características, é possível observar na forma da onda da fala segmentos delimitados por pontos onde o sinal de voz cruza o zero.

O novo modelo de geração da voz assume que a frequência de vibração das cordas vocais é diretamente proporcional à tensão aplicada à elas, estimulada por um sinal de controle que representa o sinal de tensão medido nas cordas vocias.

Para o modelo, o sinal que controla o movimento cicloestacionário das cordas vocais está presente na forma de onda do sinal de voz e é representado pelos pontos de cruzamento por zero. Nesse contexto, o sinal que controla a atividade de abertura e fechamento das cordas vocais é obtido por meio da fragmentação da forma de onda da voz em cada ponto de cruzamento por zero, sendo realizado pela passagem do sinal por um quantizador de dois níveis, de acordo com 4.1, em que a cada amostra do sinal de voz é associada a um nível específico, a depender se ela assume valor maior ou menor que o limiar.

$$\text{sgn}(s(n)) = \begin{cases} 1, & s(n) \geq 0, \\ -1, & s(n) < 0. \end{cases} \tag{4.1}$$

O processo de quantização resulta na representação do sinal de voz por meio de um vetor formada por regiões constituídas por sequências de 1s e -1s, cuja transições consistem nos pontos de cruzamento por zero do sinal de voz.

O sinal de fala passa a ser então descrito por 4.3, em que o vetor M_N exprime a quantidade de amostras contidas em cada segmentos de curta duração delimitados pelos instantes de cruzamento por zero, estimada pela contagem do número de 1s ou -1s contida em cada sequência.

$$M_N = [m_1,\ m_2,\ m_3,\ \ldots,\ m_i,\ \ldots, m_N],\ \text{para}\ i = 1,2,...,N, \tag{4.2}$$

em que m_i representa o parâmetro de duração que descreve a quantidade de amostras na seção i e N consiste na quantidade de cruzamento por zero.

A partir do vetor M_N é possível encontrar o vetor T_N que representa o período ou intervalo de tempo a cada ponto de cruzamento por zero. Dessa forma, o vetor T_N é obtido pela multiplicação do vetor M_N pelo período de amostragem, T_s, e dada por

$$T_N = [T_1,\ T_2,\ T_3,\ \ldots, T_N],\ \text{para}\ i = 1,2,...,N \tag{4.3}$$

em que T_i representa o período da sessão i.

4.1.1 Análise da Função de Distribuição de Probabilidade do Sinal de Controle

Com o propósito de estabelecer a representação espectral do sinal de voz, faz-se necessário caracterizar o sinal de controle a partir da estimativa da sua função densidade de probabilidade a fim de descrever o comportamento no domínio da frequência do gerador de impulsos cicloestacionário usando a Expressão 3.34.

Durante uma elocução, as cordas vocais têm a maior probabilidade de oscilar a uma taxa dada pela frequência fundamental, que consiste em uma frequência específica de cada locutor, determinada pelo comprimento, massa e principalmente pela tensão aplicada às cordas vocais. No entanto, no decorrer de um discurso, as cordas vocais podem atingir uma taxa de vibração maior ou menor que a taxa estabelecida pela frequência fundamental, com maior probabilidade para valores acima dela.

O sinal de controle, que representa o movimento de oscilação das cordas vocais, tem a característica de ser cicloestacionário e possui sua distribuição de probabilidade especificada por um pico que representa a probabilidade de variação na frequência fundamental. Além disso, a distribuição do sinal de controle apresenta valores de probabilidades maiores para valores maiores que a frequência fundamental, sendo também possível a detecção de valores de probabilidades para taxa menores que a frequência fundamental.

Nesse caso, o sinal de controle possui um comportamento tal que sua função de distribuição de probabilidade pode ter estimada por meio das distribuições de probabilidade Gamma unilateral e Rayleigh, que foram escolhidas por apresentarem comportamento semelhante, com curvas que estimam um sinal que assume um determinado valor com maior probabilidade e valores maiores ou menores que aquele, características típicas do sinal de controle.

Dessa forma, para cada uma das seis locuções de testes, apresentadas no Apêndice B, realiza-se o levantamento da quantidade de amostras presentes nos pontos de cruzamento por zero, representada pelo vetor M_N, em que o modelo de geração proposto assume com uma representação da taxa de vibração das cordas vocais e, em seguida, é realizada a estimativa dos parâmetros das distribuições Gamma unilateral e Rayleigh.

Nas Figuras 4.1 a 4.12 são apresentados os ajustes do vetor M_N às distribuições de probabilidade Rayleigh e Gamma unilateral.

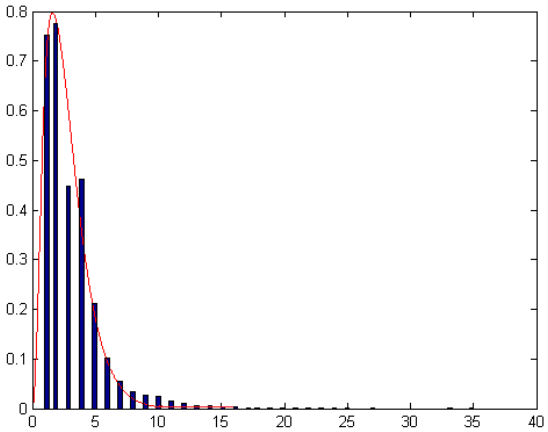

Figura 4.1: Ajuste de curva do histograma do vetor M_N, com a função de distribuição de probabilidade Gamma unilateral, utilizando os parâmetros $k_x = 2,7$ e $\theta = 1$, para a locução 1 (voz masculina).

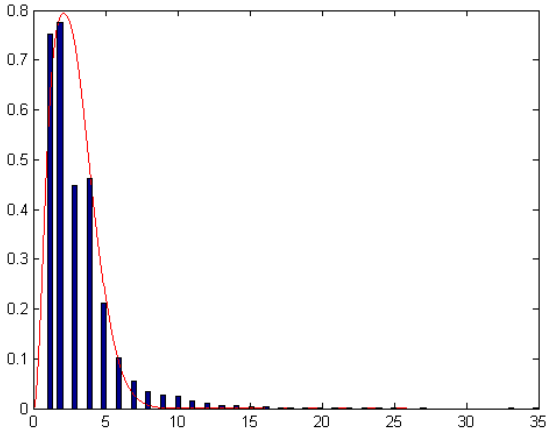

Figura 4.2: Ajuste de curva do histograma do vetor M_N, com a função de distribuição cumulativa Rayleigh, utilizando o parâmetro $\sigma = 2,2$, para a locução 1 (voz masculina).

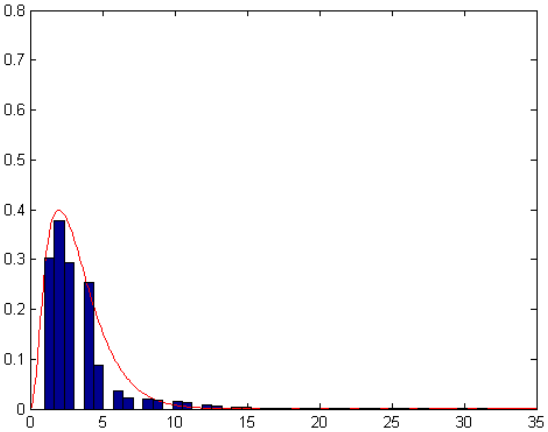

Figura 4.3: Ajuste de curva do histograma do vetor M_N, com a função distribuição de probabilidade Gamma unilateral, utilizando os parâmetros $k_x = 1,7$ e $\theta = 1$, para a locução 2 (voz masculina).

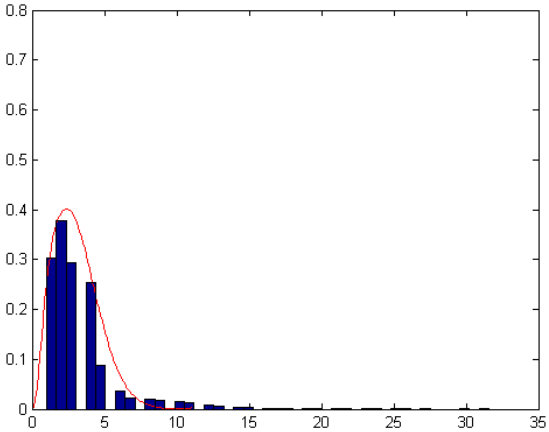

Figura 4.4: Ajuste de curva do histograma do vetor M_N, com a função distribuição de probabilidade Rayleigh, utilizando o parâmetro $\sigma = 2$, para a locução 2 (voz masculina).

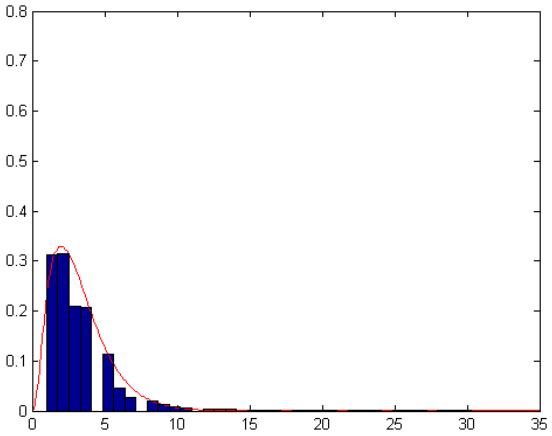

Figura 4.5: Ajuste de curva do histograma do vetor M_N, com a função distribuição de probabilidade Gamma unilateral, utilizando os parâmetros $k_x = 2$ e $\theta = 1$, para a locução 3 (voz masculina).

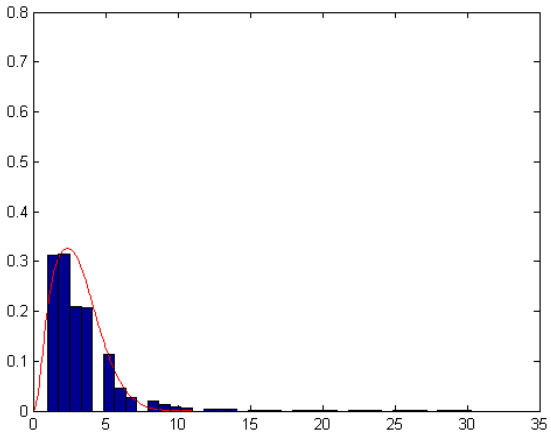

Figura 4.6: Ajuste de curva do histograma do vetor M_N, com a função distribuição de probabilidade Rayleigh, utilizando o parâmetro $\sigma = 2,7$, para a locução 3 (voz masculina).

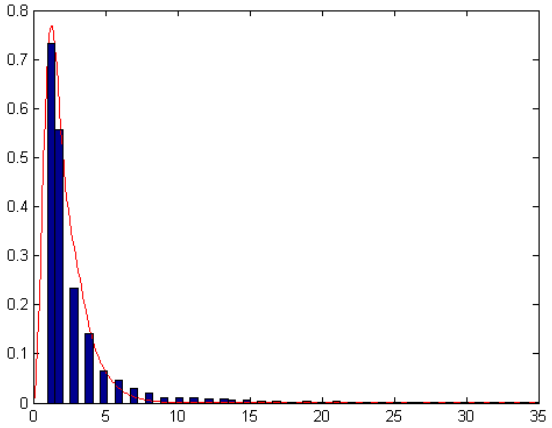

Figura 4.7: Ajuste de curva do histograma do vetor M_N, com a função distribuição de probabilidade Gamma unilateral, utilizando os parâmetros $k_x = 2$ e $\theta = 1$, para a locução 4 (voz feminina).

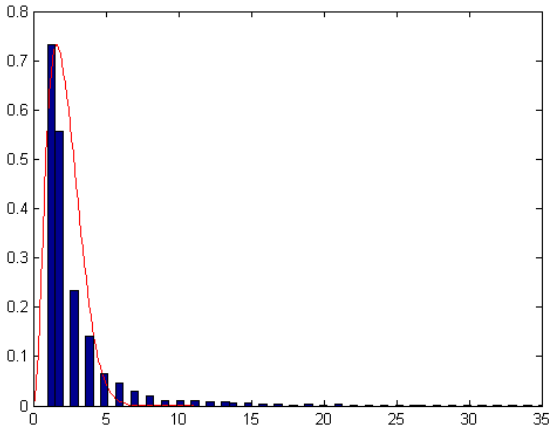

Figura 4.8: Ajuste de curva do histograma do vetor M_N, com a função distribuição de probabilidade Rayleigh, utilizando o parâmetro $\sigma = 2,2$, para a locução 4 (voz feminina).

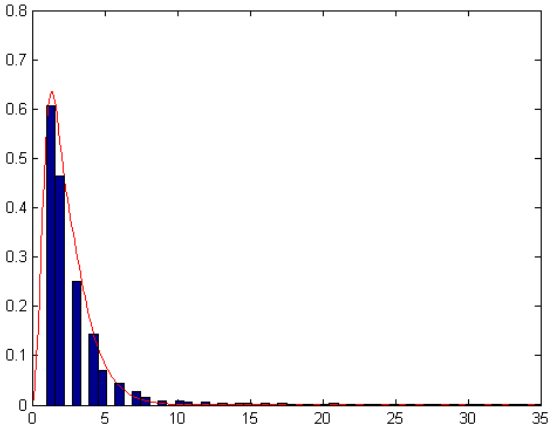

Figura 4.9: Ajuste de curva do histograma do vetor M_N, com a função distribuição de probabilidade Gamma unilateral, utilizando os parâmetros $k_x = 2,3$ e $\theta = 1$, para a locução 5 (voz feminina).

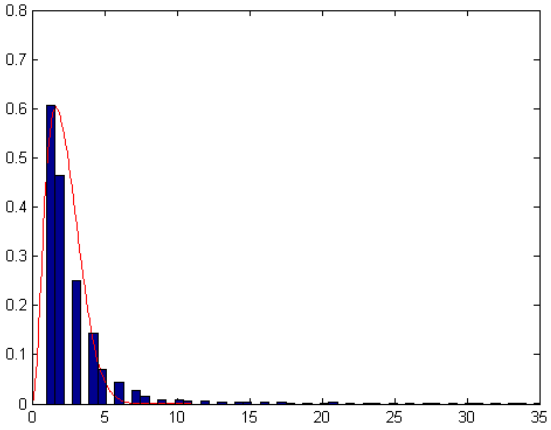

Figura 4.10: Ajuste de curva do histograma do vetor M_N, com a função distribuição de probabilidade Rayleigh, utilizando o parâmetro $\sigma = 2,6$, para a locução 5 (voz feminina).

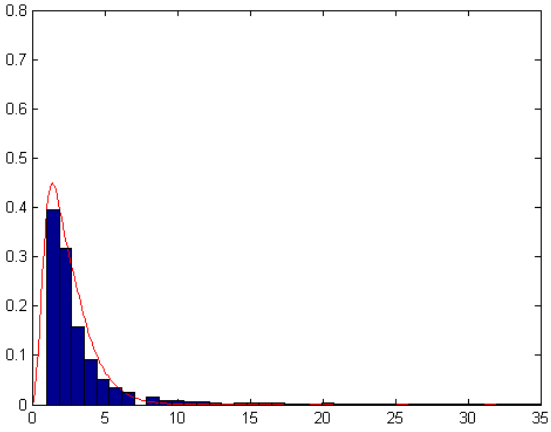

Figura 4.11: Ajuste de curva do histograma do vetor M_N, com a função distribuição de probabilidade Gamma unilateral, utilizando os parâmetros $k_x = 2$ e $\theta = 1$, para a locução 6 (voz feminina).

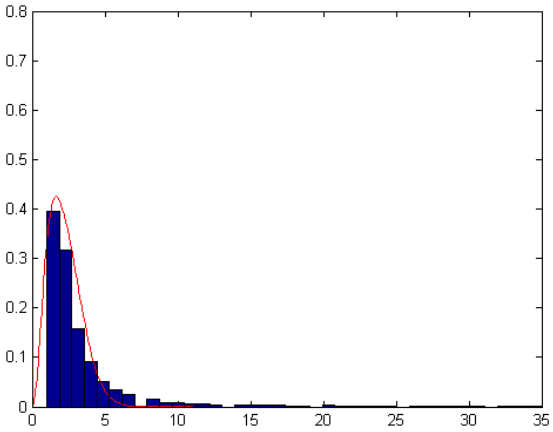

Figura 4.12: Ajuste de curva do histograma do vetor M_N, com a função distribuição de probabilidade Rayleigh, utilizando o parâmetro $\sigma = 2,6$, para a locução 6 (voz feminina).

4.2 Gerador de Impulsos Cicloestacionário

O modelo de geração da voz proposto nesta tese de doutorado baseia-se na física da fonação, considerando a cicloestacionariedade do sinal de voz, ocasionado pela oscilação das cordas vocais.

Nesse cenário, o protótipo das cordas vocais para a geração da voz dá-se pela geração de um trem de impulsos cicloestacionário, que excita as cordas vocais, proporcionando um fluxo glotal cicloestacionário.

O gerador de impulsos tem como resultado uma sequência de impulsos cujo espaçamento é cicloestacionário e estabelecido pelo sinal de controle que estimula a tensão nas cordas vocais ocasionando na sua oscilação. A Figura 4.13 ilustra o sinal de saída do gerador, $C(t)$, formado a partir de cada elemento do vetor T_N, que representa uma medida de espaçamento entre os impulsos.

Figura 4.13: Exemplo de saída do gerador de trem de impulsos, $C(t)$, com espaçamento cicloestacionário obtidos a partir do vetor T_N.

O sinal $C(t)$ pode ser visto como um esquema de modulação por posição de pulsos usado para transmitir à glote a informação de tensão longitudinal, em que o espaçamento ou período entre os impulsos no tempo é inversamente proporcional à tensão e, consequentemente, à frequência de oscilação das cordas vocais. Matematicamente, essa relação é dada por

$$\Delta\omega = \omega_o \beta M(t), \tag{4.4}$$

em que β é uma constante de proporcionalidade entre o sinal de tensão e o sinal que representa a frequência de vibração das cordas vocais. Nesta tese, por melhor se ajustar aos resultados obtidos, o valor utilizado para a constante de sensibilidade foi $\beta = 0,1\ V^{-1}$.

Afim de estimar a DEP final do sinal de voz, a seguir, as Figuras 4.14 a 4.25 apresentam as densidades espectrais de potência, $S_c(\omega)$, do trem de impulsos cicloestacionário para cada uma das seis locuções de teste, obtidas com a Expressão 3.34.

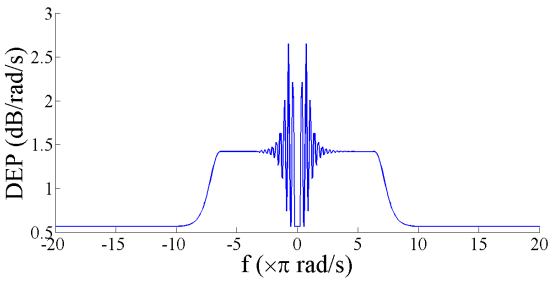

Figura 4.14: Densidade espectral de potência para o trem de impulsos cicloestacionário obtido com a função distribuição de probabilidade Gamma unilateral para a locução 1 (voz masculina).

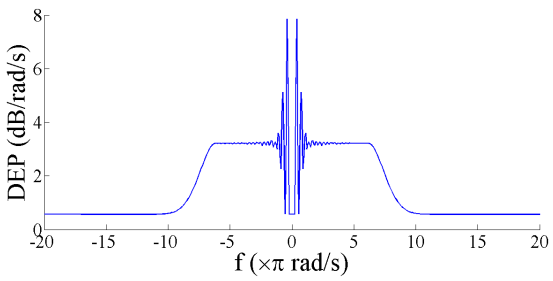

Figura 4.15: Densidade espectral de potência para o trem de impulsos cicloestacionário obtido com a função distribuição de probabilidade Rayleigh para a locução 1 (voz masculina).

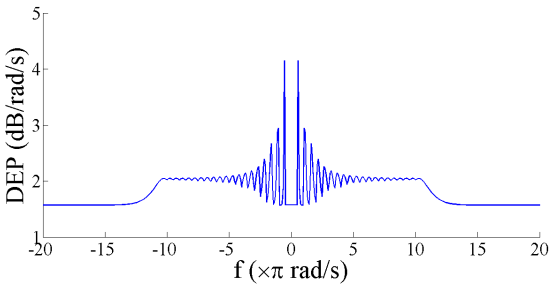

Figura 4.16: Densidade espectral de potência para o trem de impulsos cicloestacionário obtido com a função distribuição de probabilidade Gamma unilateral para a locução 2 (voz masculina).

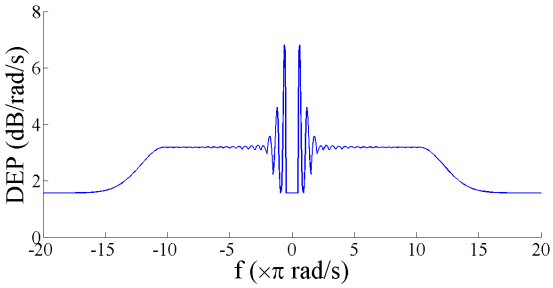

Figura 4.17: Densidade espectral de potência para o trem de impulsos cicloestacionário obtido com a função distribuição de probabilidade Rayleigh para a locução 2 (voz masculina).

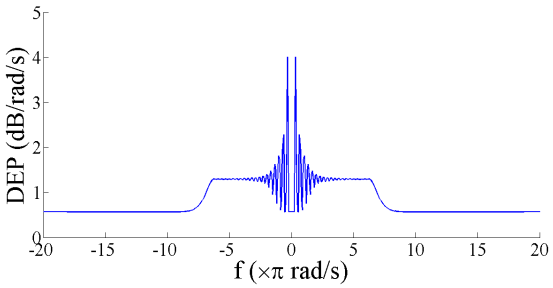

Figura 4.18: Densidade espectral de potência para o trem de impulsos cicloestacionário obtido com a função distribuição de probabilidade Gamma unilateral para a locução 3 (voz masculina).

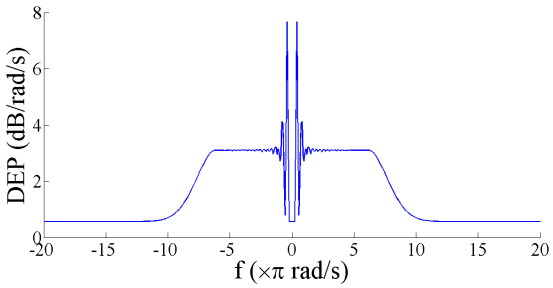

Figura 4.19: Densidade espectral de potência para o trem de impulsos cicloestacionário obtido com a função distribuição de probabilidade Rayleigh para a locução 3 (voz masculina).

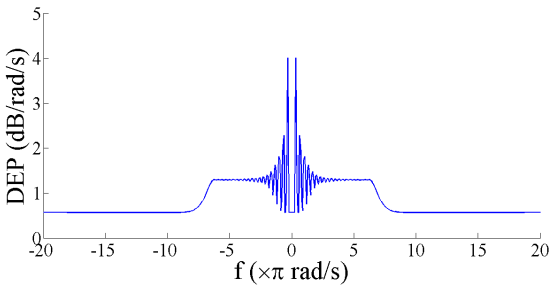

Figura 4.20: Densidade espectral de potência para o trem de impulsos cicloestacionário obtido com a função distribuição de probabilidade Gamma unilateral para a locução 4 (voz feminina).

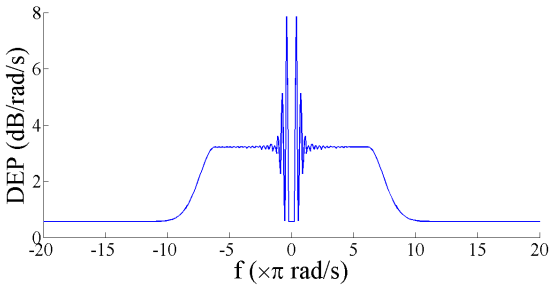

Figura 4.21: Densidade espectral de potência para o trem de impulsos cicloestacionário obtido com a função distribuição de probabilidade Rayleigh para a locução 4 (voz feminina).

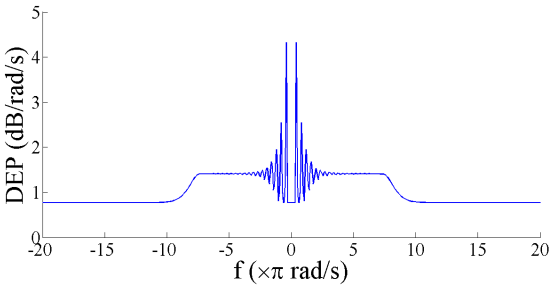

Figura 4.22: Densidade espectral de potência para o trem de impulsos cicloestacionário obtido com a função distribuição de probabilidade Gamma unilateral para a locução 5 (voz feminina).

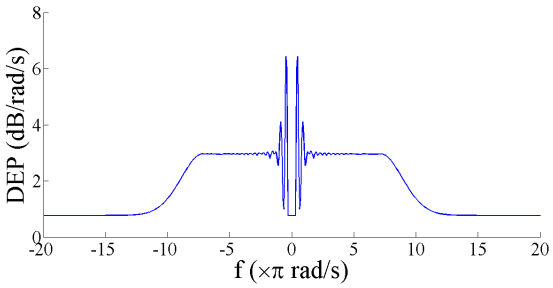

Figura 4.23: Densidade espectral de potência para o trem de impulsos cicloestacionário obtido com a função distribuição de probabilidade Rayleigh para a locução 5 (voz feminina).

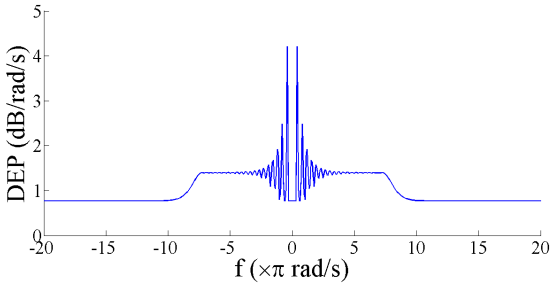

Figura 4.24: Densidade espectral de potência para o trem de impulsos cicloestacionário obtido com a função distribuição de probabilidade Gamma unilateral para a locução 6 (voz feminina).

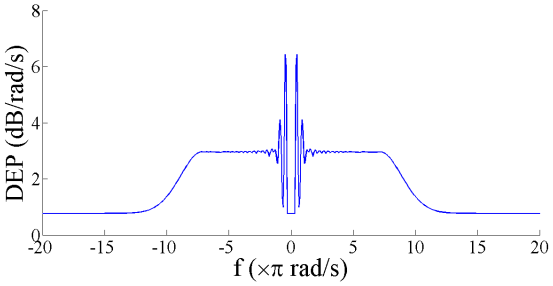

Figura 4.25: Densidade espectral de potência para o trem de impulsos cicloestacionário obtido com a função distribuição de probabilidade Rayleigh para a locução 6 (voz feminina).

4.3 Análise Temporal e Espectral Após a Glote

Para o novo modelo de geração da voz, as cordas vocais são excitadas por um trem de impulsos cicloestacionários, caracterizando de forma mais adequada a geração do sinal de voz. As cordas vocais, nesta tese, são modeladas por meio da derivada do pulso glotal de Liljencrants-Fant, cuja resposta ao impulso é dada pelas Expressões 3.37 e 3.38.

No domínio do tempo, o fluxo glotal é representado por uma sequência de pulsos glotais com espaçamento cicloestacionário entre pulsos adjacentes, resultante da convolução entre a resposta ao impulso das cordas vocais e do trem de impulsos cicloestacionário. Matematicamente, o fluxo glotal $S_Y(t)$, ilustrado na Figura 4.26, pode ser escrito como

$$S_Y(t) = C(t) * E(t). \tag{4.5}$$

Para a análise no domínio da frequência, inicialmente é necessário calcular transformada de Fourier da derivada do pulso glotal. Nesta tese, é proposta a dedução matemática para essa DEP, discriminado no Apêndice D, a partir da transformada de Fourier apresentada na Expressão D.27.

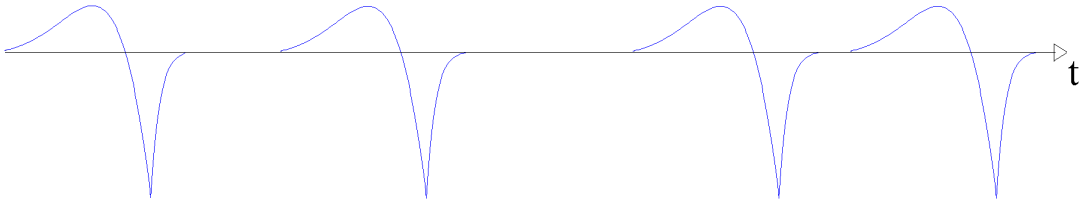

Figura 4.26: Exemplo do fluxo glotal obtido após a convolução do trem de impulso cicloestacionário com a derivada da resposta ao impulso do modelo da glote de Liljencrants-Fant.

A Figura 4.27 ilustra a comparacação entre a DEP simulada da derivada do pulso glotal de Liljencrants-Fant e a DEP obtida com a Expressão D.27, proposta nesta tese, em que é possível observer a concordância com outros trabalhos da literatura [81, 82, 83, 84].

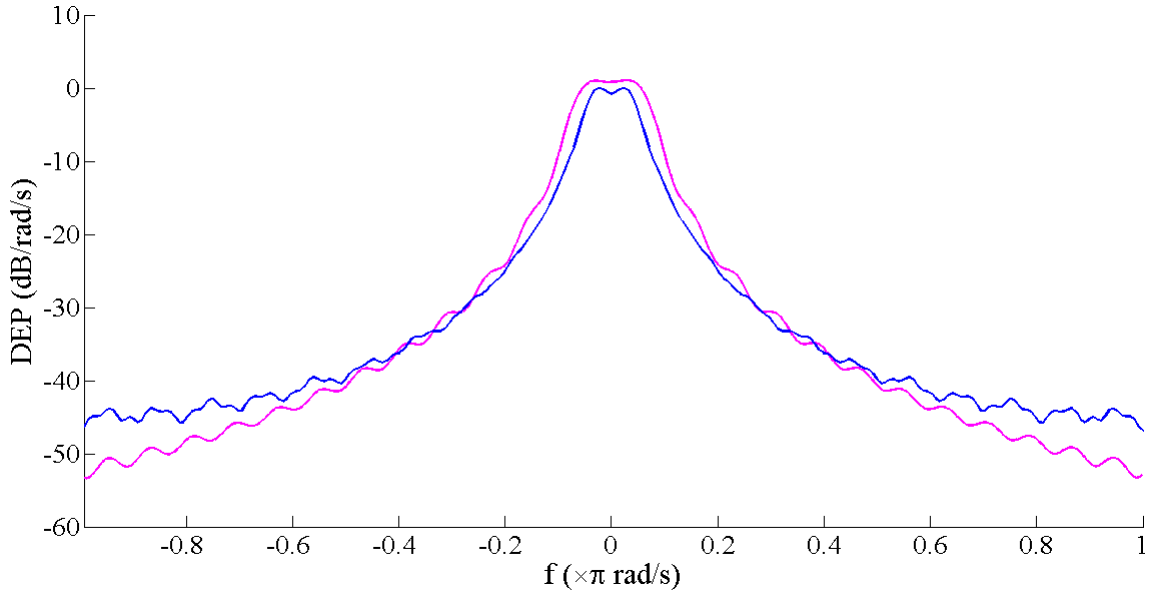

Figura 4.27: Comparação entre a densidade espectral de potência simulada (azul) e a obtida com a expressão proposta (lilás) nesta tese para a derivada da resposta ao impulso do modelo da glote de Liljencrants-Fant.

Uma vez que o modelo assume que a voz é gerada por meio de um sistema linear e invariante no tempo, a DEP, após a passagem pelas cordas vocais pode ser escrita como

$$S_Y(\omega) = |E(\omega)|^2 S_C(\omega), \tag{4.6}$$

em que $E(\omega)$ representa a transformada de Fourier do pulso glotal $E(t)$. Como $E(t)$ foi considerado um sinal resposta ao impulso de um sistema linear e invariante ao deslocamento no tempo, então $E(\omega)$ representa a resposta em frequência desse sistema. Esse desenvolvimento permite afirmar que, na saída da glote, o espectro dos sinais observados em uma janela de tempo que justifique a estacionaridade em sentido amplo pode ser ajustado ao espectro $S_Y(\omega)$.

As Figuras 4.28 a 4.39 ilustram as densidades espectrais de potência, $S_Y(\omega)$, obtidas para cada uma das locuções de testes.

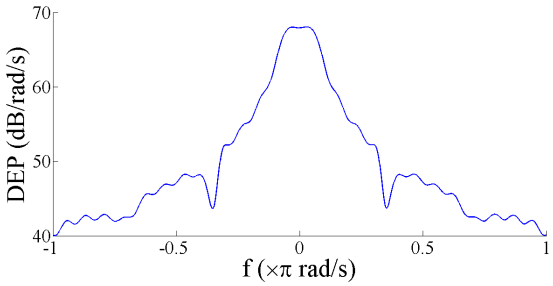

Figura 4.28: Densidade espectral de potência, $S_Y(\omega)$, para o sinal após a glote, obtida com o trem de impulsos sendo modelado com a distribuição de probabilidade Gamma unilateral, para a locução 1 (voz masculina).

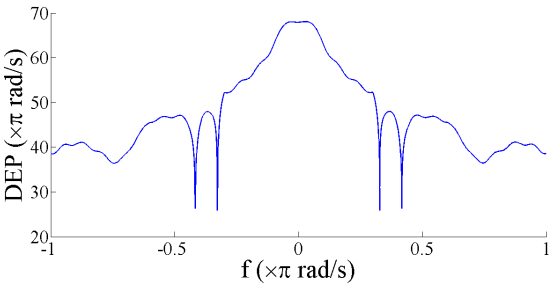

Figura 4.29: Densidade espectral de potência, $S_Y(\omega)$, para o sinal após a glote, obtida com o trem de impulsos sendo modelado com a distribuição de probabilidade Rayleigh, para a locução 1 (voz masculina).

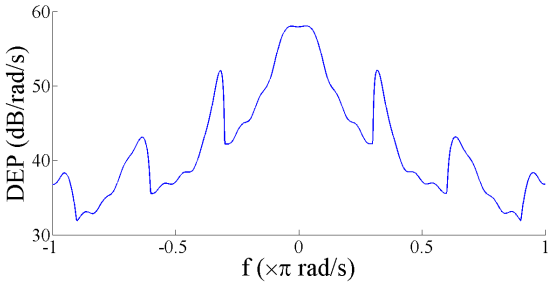

Figura 4.30: Densidade espectral de potência, $S_Y(\omega)$, para o sinal após a glote, obtida com o trem de impulsos sendo modelado com a distribuição de probabilidade Gamma unilateral, para a locução 2 (voz masculina).

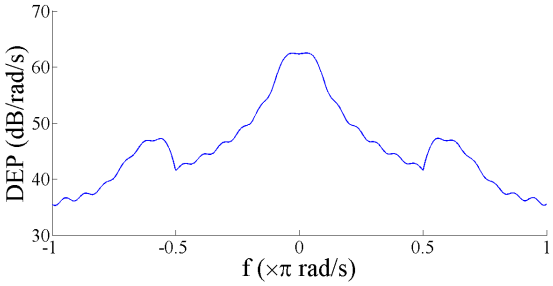

Figura 4.31: Densidade espectral de potência, $S_Y(\omega)$, para o sinal após a glote, obtida com o trem de impulsos sendo modelado com a distribuição de probabilidade Rayleigh, para a locução 2 (voz masculina).

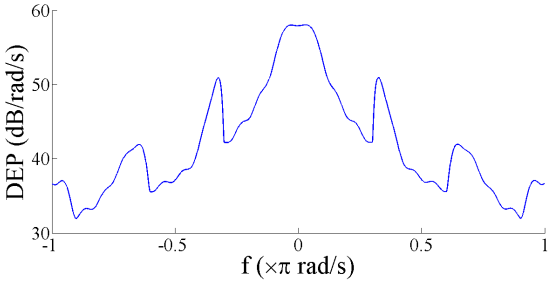

Figura 4.32: Densidade espectral de potência, $S_Y(\omega)$, para o sinal após a glote, obtida com o trem de impulsos sendo modelado com a distribuição de probabilidade Gamma unilateral, para a locução 3 (voz masculina).

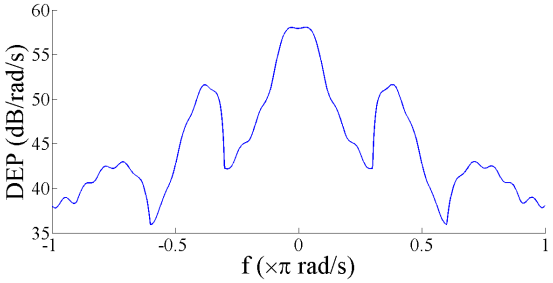

Figura 4.33: Densidade espectral de potência, $S_Y(\omega)$, para o sinal após a glote, obtida com o trem de impulsos sendo modelado com a distribuição de probabilidade Rayleigh, para a locução 3 (voz masculina).

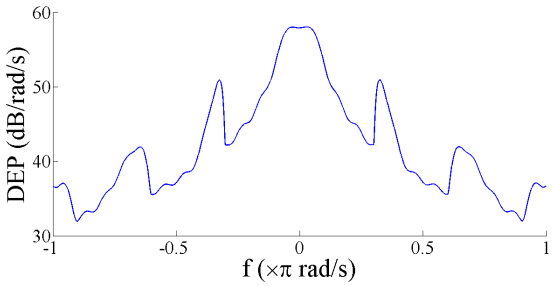

Figura 4.34: Densidade espectral de potência, $S_Y(\omega)$, para o sinal após a glote, obtida com o trem de impulsos sendo modelado com a distribuição de probabilidade Gamma unilateral, para a locução 4 (voz feminina).

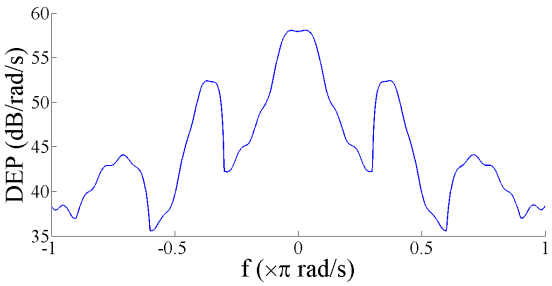

Figura 4.35: Densidade espectral de potência, $S_Y(\omega)$, para o sinal após a glote, obtida com o trem de impulsos sendo modelado com a distribuição de probabilidade Rayleigh, para a locução 4 (voz feminina).

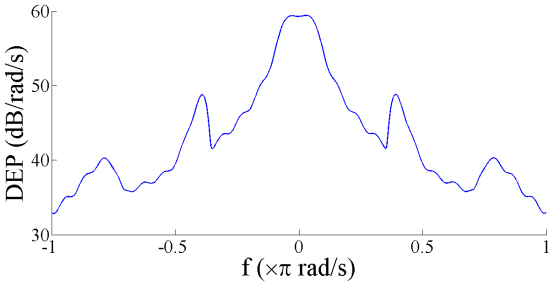

Figura 4.36: Densidade espectral de potência, $S_Y(\omega)$, para o sinal após a glote, obtida com o trem de impulsos sendo modelado com a distribuição de probabilidade Gamma unilateral, para a locução 5 (voz feminina).

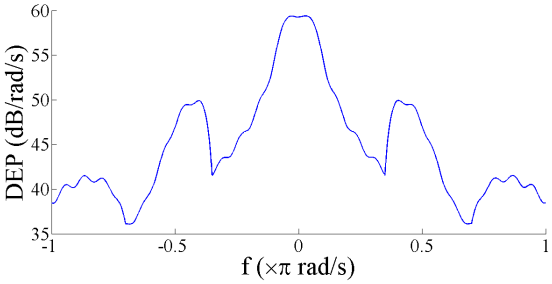

Figura 4.37: Densidade espectral de potência, $S_Y(\omega)$, para o sinal após a glote, obtida com o trem de impulsos sendo modelado com a distribuição de probabilidade Rayleigh, para a locução 5 (voz feminina).

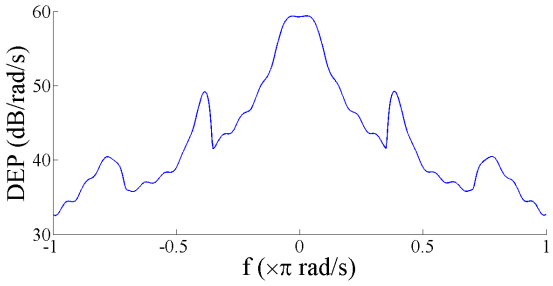

Figura 4.38: Densidade espectral de potência, $S_Y(\omega)$, para o sinal após a glote, obtida com o trem de impulsos sendo modelado com a distribuição de probabilidade Gamma unilateral, para a locução 6 (voz feminina).

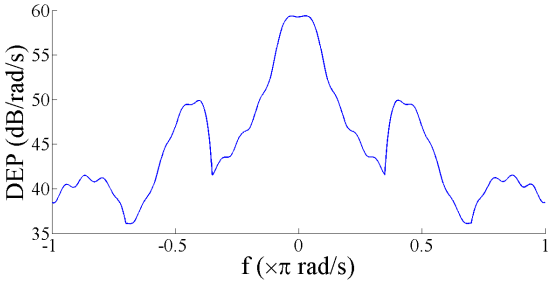

Figura 4.39: Densidade espectral de potência, $S_Y(\omega)$, para o sinal após a glote, obtida com o trem de impulsos sendo modelado com a distribuição de probabilidade Rayleigh, para a locução 6 (voz feminina).

4.4 Análise Espectral do Trato Vocal

Após a passagem pela glote, o fluxo glotal representa a entrada do trato vocal, que tem a função de filtrar a partir de uma função de transferência determinada pela posição dos articuladores no momento da fonação de cada fonema.

Como mencionado no Capítulo 3, a estimação do espectro de magnitude da seletividade em frequência do trato vocal é obtida por meio da análise LPC. Uma vez que o modelo trata a geração da voz como um modelo linear e invariante no tempo, a resposta em frequência para o trato vocal é dada por $|H(\omega)|^2$, em que $H(\omega)$ consiste na resposta em frequência obtida pela representação LPC.

As Figuras 4.40 a 4.45 apresentam a magnitude do espectro de frequência, $|H(\omega)|^2$, para o modelo de geração da voz proposto para cada uma das seis locuções de teste.

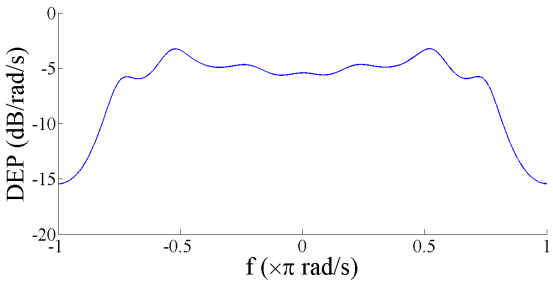

Figura 4.40: Estimação do trato vocal para a locução 1 (voz masculina).

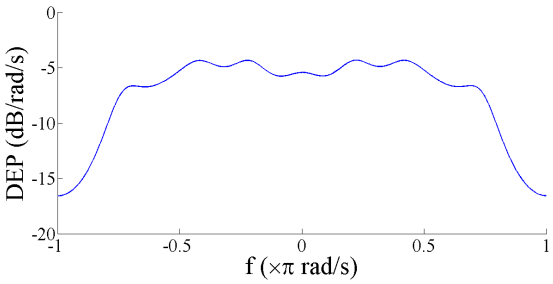

Figura 4.41: Estimação do trato vocal para a locução 2 (voz masculina).

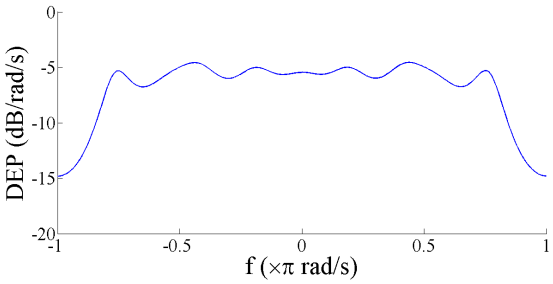

Figura 4.42: Estimação do trato vocal para a locução 3 (voz masculina).

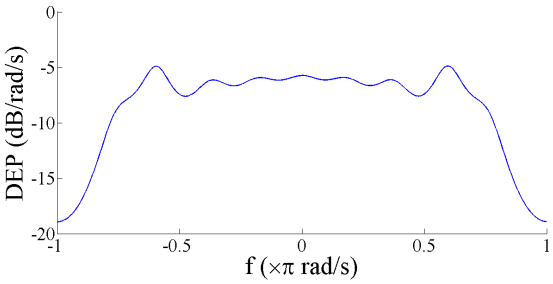

Figura 4.43: Estimação do trato vocal para a locução 4 (voz feminina).

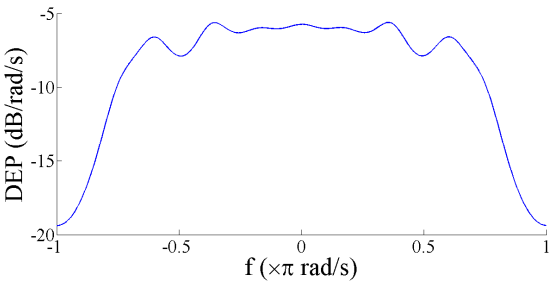

Figura 4.44: Estimação do trato vocal para a locução 5 (voz feminina).

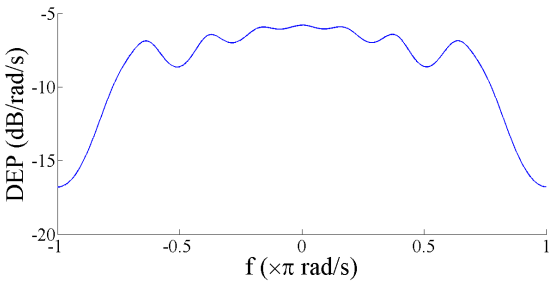

Figura 4.45: Estimação do trato vocal para a locução 6 (voz feminina).

4.5 Análise Temporal e Espectral Final

A representação temporal e espectral do sinal de voz é uma análise acústica primordial para as aplicações quem incluem o processamento do sinal de voz. O modelo desta tese propõe que a voz é produzida por um sistema linear e invariante no tempo, para intervalos que ela é considerada estacionária no sentido amplo.

No domínio do tempo, o sinal gerado é representado pela Expressão 4.7, repetida abaixo por conveniência.

$$V(t) = GC(t) * E(t) * H(t) * L(t). \tag{4.7}$$

A Figura 4.46 ilustra um exemplo do sinal resultante do modelo, que apresenta características do sinal de voz, em que se observa os pulsos glotais com espaçamentos cicloestacionário, determinado pelo sinal de tensão que controla o movimento das cordas vocais, modificados pelos trato vocal e radiação dos lábios e narinas.

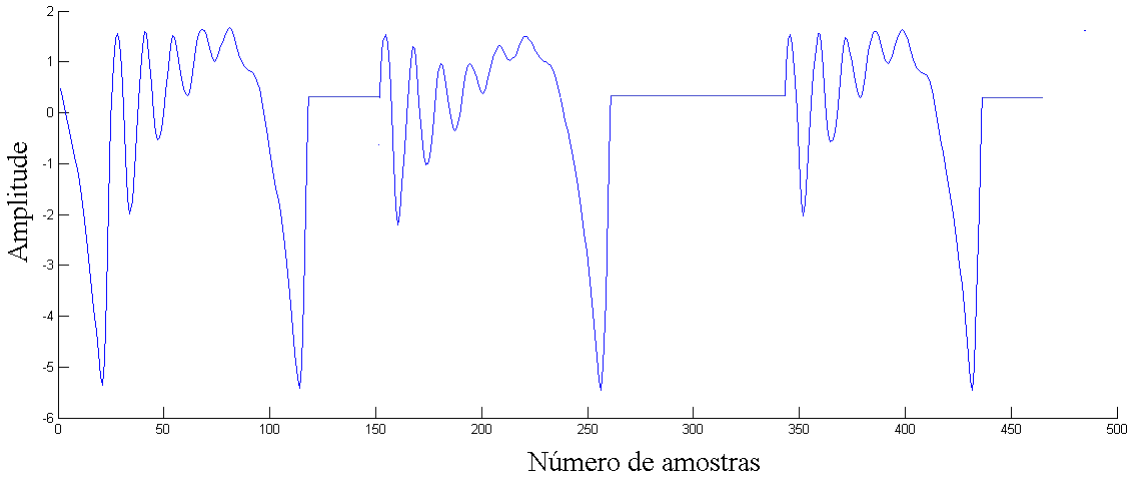

Figura 4.46: Exemplo do sinal no domínio do tempo obtido pelo modelo de geração de voz proposto.

A análise espectral do sistema linear e invariante no tempo fornece a densidade espectral de potência do sinal de saída como a multiplicação da DEP do sinal de entrada pelo módulo ao quadrado da resposta em frequência do filtro.

Desse modo, para a formação da expressão geral da DEP do sinal de voz, inicialmente tem-se a DEP do trem de impulsos cicloestacionário, que faz a excitação das cordas vocais. Após as cordas vocais, a DEP é formada pela multiplicação da densidade espectral de potência do trem de impulsos pelo módulo ao quadrado da resposta em frequência da derivada do pulso glotal. Em seguida, um ganho, que representa a potência do diafragma é adicionado e, na DEP resultante, é adicionado o módulo do ganho ao quadrado. Por fim, há o modelamento da DEP do trato vocal e radiação dos lábios e narinas, resultando na DEP final do sinal de voz, apresentada na Expressão 4.8 e repetida abaixo por conveniência.

$$S_V(\omega) = G^2 S_C(\omega)|E(\omega)|^2|H(\omega)|^2|L(\omega)|^2, \tag{4.8}$$

em que $H(\omega)$ e $L(\omega)$ representam, respectivamente, a transformada de Fourier do trato vocal e da radiação dos lábios e narinas. Uma vez que $H(t)$ e $L(t)$ consistem em um sinal resposta ao impulso de um sistema linear e invariante ao deslocamento no tempo, as suas transformadas de Fourier representam as respostas em frequência de cada sistema.

As Figuras 4.47 até a 4.58 ilustram a comparação entre as densidades espectrais de potência obtidas pelo modelo de geração proposto, para as funções densidades de probabilidade Gamma unilateral e Rayleigh, e também as densidades simuladas. A Tabela 4.1 apresenta o ganho para as seis locuções de teste.

Tabela 4.1: Valores do ganho $|G|^2$ para cada uma das locuções de teste.

Locução	**Ganho** ($\times 10^{-9}$)
Locução 1	3
Locução 2	5
Locução 3	6
Locução 4	7
Locução 5	6
Locução 6	5

A partir da observação das figuras, é possível perceber que a DEP proporcionada pelo novo modelo de geração da voz se ajusta bem ao comportamento de frequência das locuções de teste.

No processo de filtragem, a DEP do trem de impulsos cicloestacionário é filtrada e passa a ter largura de banda do sinal de voz. Sua influência causa oscilações na DEP do fluxo glotal, ou seja, após a passagem do trem de impulsos pela glote, que não existe no modelo fonte-filtro clássico. As oscilações são provenientes do movimento cicloestacionário das cordas vocais, proporcionando melhor ajuste ao se comparar com a DEP do sinal de voz.

Os resultados permitem afirmar que o espectro dos sinais testados em uma janela de tempo estacionária no sentido amplo pode ser ajustado ao espectro $S_V(\omega)$, com o uso das duas funções de distribuições de probabilidade. No entanto, a distribuição Gamma unilateral, de forma geral, se mostrou melhor no ajuste à DEP das locuções testadas.

Com o modelo proposto, é possível realizar uma estimativa espectral para análises patológicas. Perante uma patologia, as cordas vocais se comportam de forma irregular, e a presente proposta de geração da voz é um método capaz de emular patologias, pois modela o comportamente delas.

Outra vantagem é que a estimação das oscilações das cordas vocais, por meio de gerador de impulsos cicloestacionário, também pode ser usada na diferenciação entre sons sonoros e surdos, uma vez que esses dois tipos de sons se distinguem na taxa de cruzamento por zero.

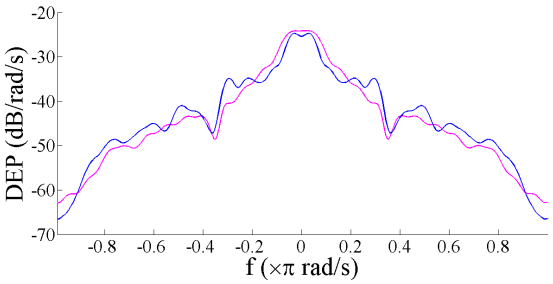

Figura 4.47: Comparação entre a densidade espectral de potência da locução 1 (azul) e sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Gamma unilateral.

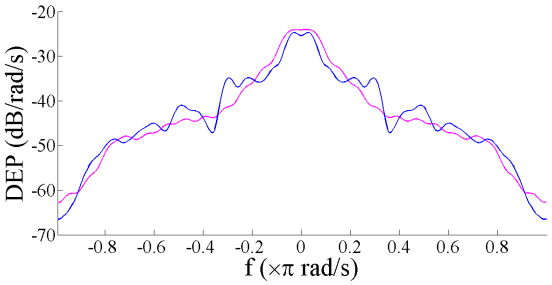

Figura 4.48: Comparação entre a densidade espectral de potência da locução 1 (azul) e sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Rayleigh.

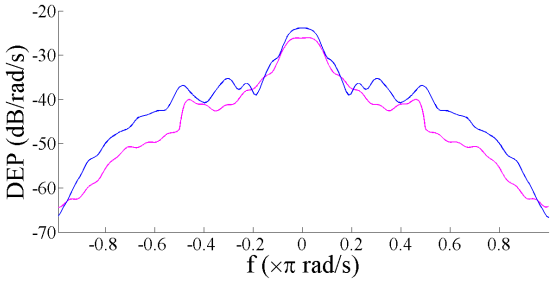

Figura 4.49: Comparação entre a densidade espectral de potência da locução 2 (azul) e sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Gamma unilateral.

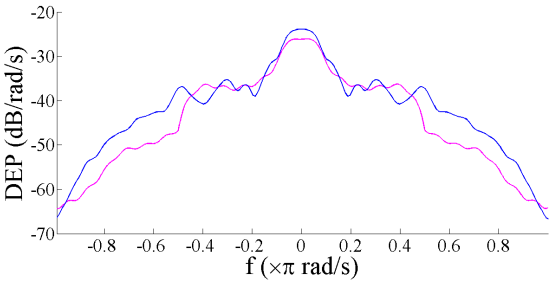

Figura 4.50: Comparação entre a densidade espectral de potência da locução 2 (azul) e sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Rayleigh.

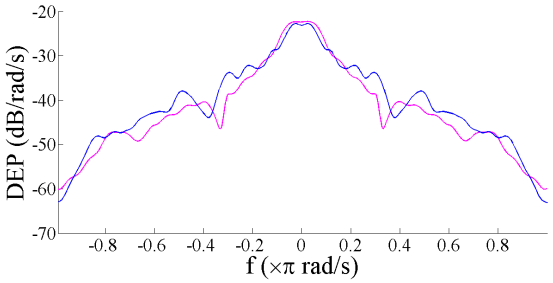

Figura 4.51: Comparação entre a densidade espectral de potência da locução 3 (azul) e sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Gamma unilateral.

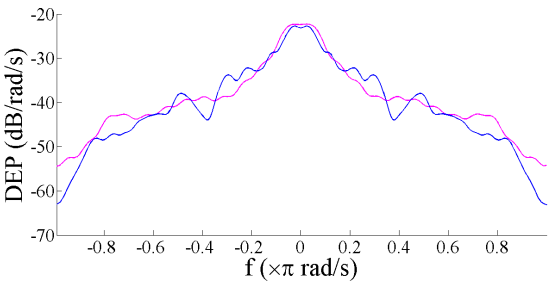

Figura 4.52: Comparação entre a densidade espectral de potência da locução 3 (azul) e sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Rayleigh.

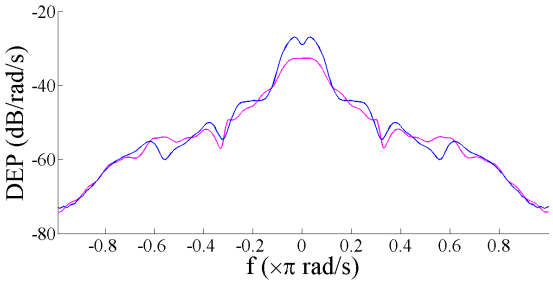

Figura 4.53: Comparação entre a densidade espectral de potência da locução 4 (azul) e sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Gamma unilateral.

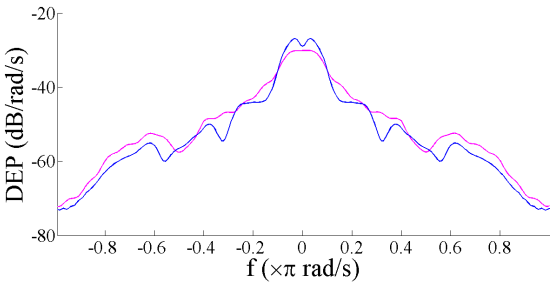

Figura 4.54: Comparação entre a densidade espectral de potência da locução 4 (azul) e sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Rayleigh.

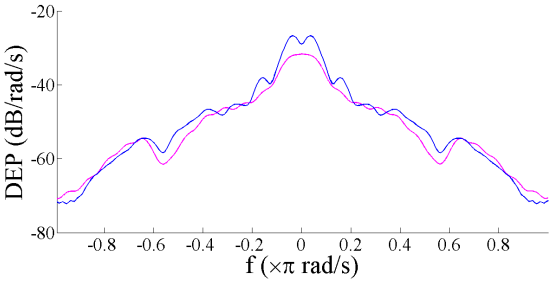

Figura 4.55: Comparação entre a densidade espectral de potência da locução 5 (azul) e sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Gamma unilateral.

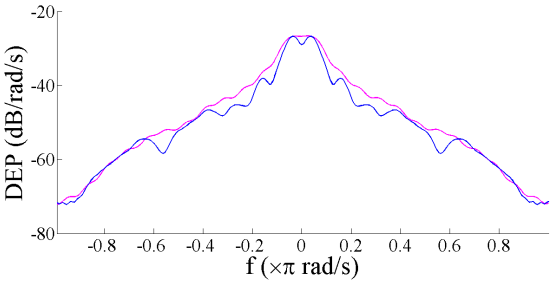

Figura 4.56: Comparação entre a densidade espectral de potência da locução 5 (azul) e sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Rayleigh.

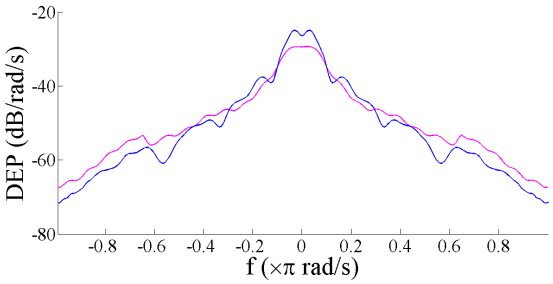

Figura 4.57: Comparação entre a densidade espectral de potência da locução 6 (azul) e sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Gamma unilateral.

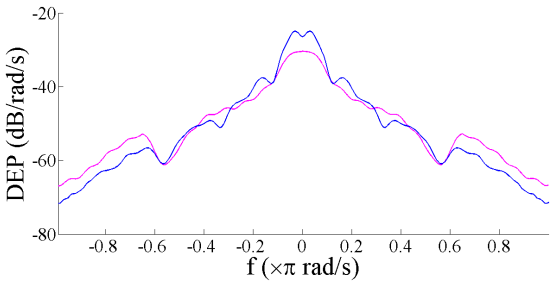

Figura 4.58: Comparação entre a densidade espectral de potência da locução 6 (azul) e sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Rayleigh.

4.6 Comparação com o Modelo Clássico de Geração da Voz

O modelo clássico de geração da voz pela teoria fonte-filtro, na qual muitos trabalhos na literatura são baseados, foi proposto por Fant em 1970 [18]. Segundo Fant, a representação da geração da voz é realizada por meio da convolução entre um modelo de pulso glotal, da resposta ao impulso do trato vocal e do efeito da radiação dos lábios e narinas.

Devido ao fato da presente tese ser focada na geração de sons vocálicos, a modelagem da fonte de excitação é dada pela derivada do pulso glotal de Liljencrants-Fant, não sendo utilizados, portanto, testes com excitação ruidosa.

Para manter compatilidade com os resultados apresentados para o modelo proposto nesta tese, o trato vocal para o modelo fonte-filtro clássico foi estimado pela análise LPC, bem como os efeitos da radiação dos lábios e narinas pela Expressão 3.54.

No domínio do tempo, Fant propõe a representação da voz por

$$V_c(t) = E(t) * H(t) * L(t). \tag{4.9}$$

e no domínio da frequência como

$$V_c(\omega) = E(\omega)H(\omega)L(\omega). \tag{4.10}$$

A sequência de Figuras 4.59 a 4.70 apresentam a comparação entre as densidades espectrais de potência obtidas pela simulação de elocuções, pelo novo modelo de geração da voz e pelo modelo fonte-filtro clássico

Observando as figuras, é possível notar que o novo modelo de geração da voz estima a DEP da voz de forma mais adequada em comparação com o modelo de Fant. Nota-se que a geração do sinal de voz é melhor representado quando se considera o movimento cicloestacionário das cordas vocais na geração de sons sonoros, proporcionando uma representação mais fiel às oscilações do espectro da voz.

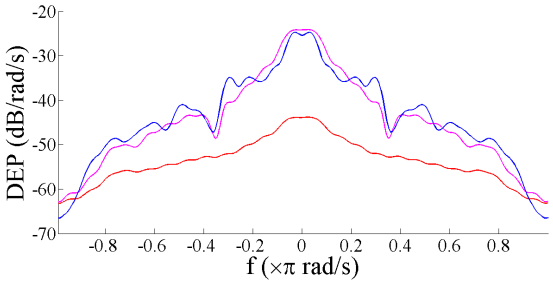

Figura 4.59: Comparação entre a densidade espectral de potência da locução 1 (azul), sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Gamma unilateral e a DEP obtida com o modelo fonte-filtro clássico (vermelho).

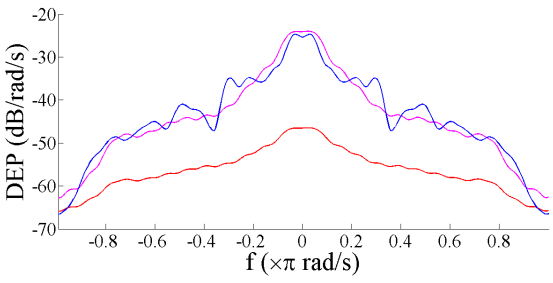

Figura 4.60: Comparação entre a densidade espectral de potência da locução 1 (azul), sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Rayleigh e a DEP obtida com o modelo fonte-filtro clássico (vermelho).

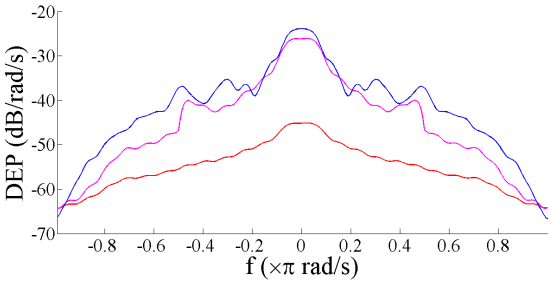

Figura 4.61: Comparação entre a densidade espectral de potência da locução 2 (azul), sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Gamma unilateral e a DEP obtida com o modelo fonte-filtro clássico (vermelho).

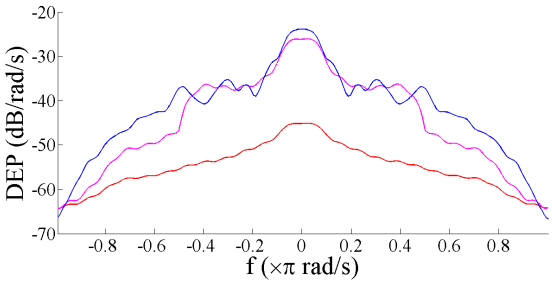

Figura 4.62: Comparação entre a densidade espectral de potência da locução 2 (azul), sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Rayleigh e a DEP obtida com o modelo fonte-filtro clássico (vermelho).

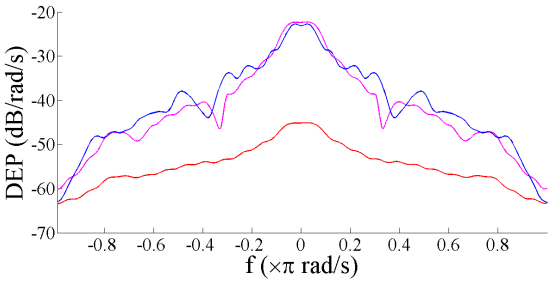

Figura 4.63: Comparação entre a densidade espectral de potência da locução 3 (azul), sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Gamma unilateral e a DEP obtida com o modelo fonte-filtro clássico (vermelho).

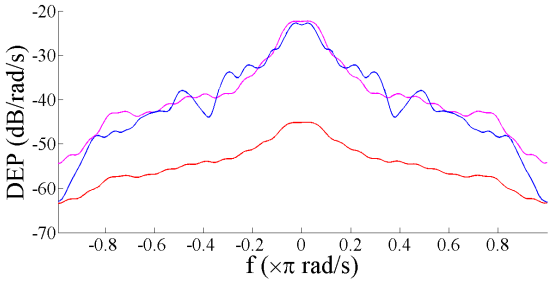

Figura 4.64: Comparação entre a densidade espectral de potência da locução 2 (azul), sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Rayleigh e a DEP obtida com o modelo fonte-filtro clássico (vermelho).

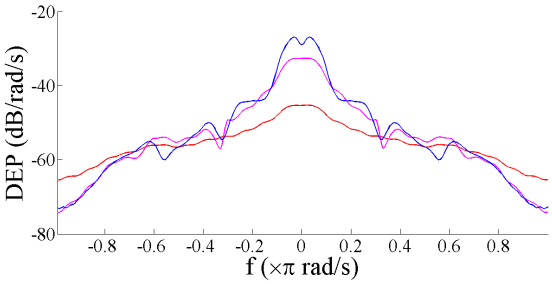

Figura 4.65: Comparação entre a densidade espectral de potência da locução 4 (azul), sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Gamma unilateral e a DEP obtida com o modelo fonte-filtro clássico (vermelho).

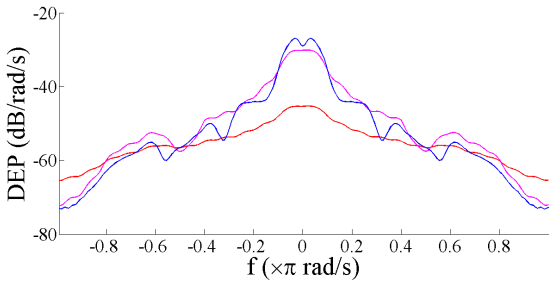

Figura 4.66: Comparação entre a densidade espectral de potência da locução 4 (azul), sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Rayleigh e a DEP obtida com o modelo fonte-filtro clássico (vermelho).

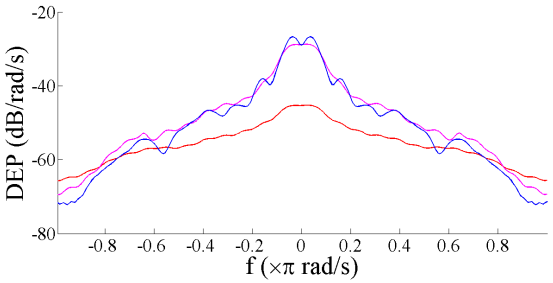

Figura 4.67: Comparação entre a densidade espectral de potência da locução 5 (azul), sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Gamma unilateral e a DEP obtida com o modelo fonte-filtro clássico (vermelho).

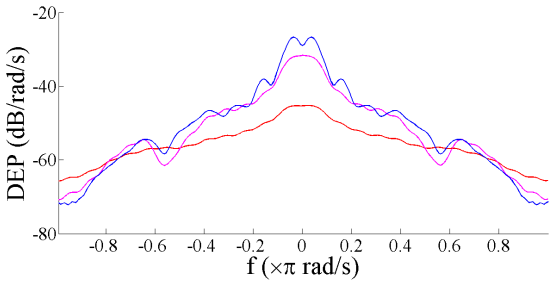

Figura 4.68: Comparação entre a densidade espectral de potência da locução 5 (azul), sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Rayleigh e a DEP obtida com o modelo fonte-filtro clássico (vermelho).

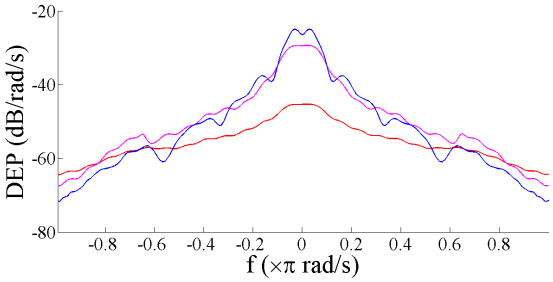

Figura 4.69: Comparação entre a densidade espectral de potência da locução 6 (azul), sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Gamma unilateral e a DEP obtida com o modelo fonte-filtro clássico (vermelho).

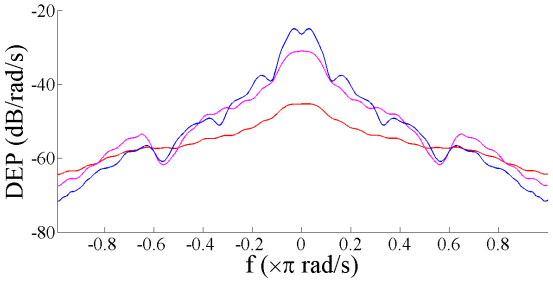

Figura 4.70: Comparação entre a densidade espectral de potência da locução 6 (azul), sua estimação (lilás) pelo novo modelo de geração da voz, utilizando a função distribuição de probabilidade Rayleigh e a DEP obtida com o modelo fonte-filtro clássico (vermelho).

CAPÍTULO 5

Considerações Finais e Propostas para Trabalhos Futuros

Devido à sua grande importância, a modelagem do processo de geração do sinal de voz ainda continua sendo tema de interesse dos pesquisadores.

A literatura fornece basicamente três tipos de síntese de voz, em que um dos mais difundidos é a síntese por formantes, baseada na teoria fonte-filtro proposta em 1970 por Fant. Essa síntese considera que o sinal de voz é formado a partir de três subsistemas independentes, que são fonte de excitação, trato vocal e radiação dos lábios e narinas.

A fonte de excitação é a primeira etapa do sistema de síntese de voz e pode fundamentalmente ser um trem de impulsos, para sinais sonoros, ou um ruído de espectro largo, para sons ruidosos. Por ser uma das principais ferramentas na geração da voz, muitos pesquisadores propuseram modelos para a representação da fonte de excitação, mais precisamente sobre as cordas vocais.

A princípio, as cordas vocais foram modeladas por equações diferenciais que descrevem modelos mecânicos massa-mola-amortecedor, com uma, duas ou três massas. Posteriormente, outras técnicas como as que incluem o modelo da camada epitelial das cordas vocais, representação matemática da larige, entre outras, foram propostas, inclusive métodos que têm objetivo de detecção de patologias nas cordas vocais, como os que utilizam, por exemplo, os coeficientes Cepstral, redes neurais, modelos de mistura gaussianas e ângulo de abertura.

Seguidamente à fonte de excitação, tem-se o trato vocal que é formado por articuladores que, a cada fonação, se configuram de forma distinta, realizando uma filtragem no fluxo glotal, a partir da seletividade em frequência gerada pela disposição dos ressonadores. A modelagem da radiação dos lábios e narinas consiste na última fase da geração da voz. De acordo com a literatura, seu efeito resulta no acréscimo médio de 6 dB por oitava nas altas frequências.

Esta tese apresenta um novo método de síntese de um sinal com características da voz baseado na teoria fonte-filtro. O objetivo é o desenvolvimento de uma teoria mais fiel à biofísica da fonação, com o intuito de possibilitar a detecção e classficação de patologias nas cordas vocais a partir de suas características específicas.

Por serem mais apropriados à emulação de patologias, o estudo desta tese está voltado para a geração dos sons sonoros, uma vez que no seu processo de geração está incluido o comportamento das cordas vocais.

O novo modelo de geração propõe que os sons vocálicos são formados a partir de uma oscilação cicloestacionária das cordas vocais, representada por uma frequência média de oscilação, que é a frequência fundamental, além de frequências acima e aquém dela. Esse movimento é proporcional à massa, comprimento das cordas vocais e, principalmente, a um sinal de tensão longitudinal.

Para um dado locutor, o novo modelo considera fixos os parâmetros de massa e comprimento das cordas vocais e propõe que o sinal de tensão que rege o movimento cicloestacionário das corvas vocais é diretamente proporcional à frequência de oscilação e está presente na forma de onda do sinal de voz, a cada ponto de cruzamento por zero.

Nesse cenário, é realizada a análise matemática do processo de excitação das cordas vocais a partir de sua representação por um trem de impulsos cicloestacionário resultante de um gerador.

O modelo realiza a síntese de um sinal com característica da fala considerando que o sistema fonte-filtro é constituído a partir de subsistemas lineares e invariantes no tempo, para intervalos de tempo em que o sinal de voz é considerado estacionário. Nesse caso, no domínio do tempo, o fluxo após a glote é descrito pela convolução entre o modelo do pulso glotal e o trem de impulsos cicloestacionário.

Nesta tese, a resposta da glote foi modelada pela derivada do modelo do pulso glotal de Liljencrants-Fant. Para a geração do sinal de voz, o fluxo glotal é convoluído com a resposta do impulso do trato vocal e radiação dos lábios e narinas. Além disso, considera-se um ganho para a representação da potência do ar proveniente do diafragma.

A tese apresenta uma nova formulação matemática para a densidade espectral de potência da voz, em função da DEP do gerador de impulsos cicloestacionário, que é dada em função da distribuição de probabilidade do sinal de tensão. É proposta a representação da função distribuição de probabilidade do sinal de tensão e, consequentemente da frequência de oscilação das cordas vocais, por meio das função Gamma unilateral e Rayleigh.

A fim de avaliar o desempenho do modelo proposto de síntese do sinal com característica da voz, seis elocuções foram selecionadas de forma aleatória, sendo três de um locutor masculino e três de um locutor feminino. São apresentados os resultados de saída para cada subsistema da geração da voz.

A partir da observação dos resultados, é possível perceber que a geração da voz apresenta melhores resultados com a inclusão do movimento cicloestacionário das cordas vocais. O modelo de geração de voz desenvolvido é promissor para a emulação de vozes patológicas pois realiza a modelagem da voz por meio de parâmetros relacionados à biofísica da fonação.

Pretende-se, assim, a partir da estimação de parâmetros como a função distribuição de probabilidade do sinal que controla o movimento das cordas vocais, massa, comprimento e frequência média de oscilação, a emulação de distúrbios na laringe, uma vez que é sabido que a presença de patologias, como edemas, nódulos, pólipos e cistos, causam modificações nas cordas

vocais, como o aumento de massa, provocando uma vibração irregular. Além disso, propicia o mau fechamento da glote, proporcionando o surgimento de componentes ruidosas, ocasionando modificações significativas no sons sonoros.

5.1 Principais Contribuições

As principais contribuições desta tese estão resumidas a seguir.

1. Nesta tese é discutida a síntese de um sinal com características da voz baseada na modelo fonte-filtro. De acordo com a literatura, a teoria fonte-filtro data do ano de 1970 e foi proposta por Fant. Ela tem a característica de modelar a geração da voz por três subsistemas independentes, que representam modelos para o pulso glotal, trato vocal e radiação dos lábios e narinas.

   Neste trabalho, uma nova abordagem para o modelo fonte-filtro é apresentada. É proposta a geração da voz por meio de subsistemas lineares e invariantes no tempo, em intervalo de tempo em que o sinal de voz é considerado estacionário no sentido amplo, levando em consideração o movimento cicloestacionário das cordas vocais, obtido no decorrer de um discurso. Dessa forma, cada subsistema é modelado por um filtro cuja densidade espectral de potência de saída é dada pela multiplicação da DEP do seu sinal de entrada pelo módulo ao quadrado da resposta em frequência de cada filtro.

   Nesse cenário, é apresentada uma expressão matemática que rege o comportamento da energia no domínio da frequência para o sinal de voz. A densidade espectral de potência do sinal de voz é então descrita em termos das densidades espectrais de potência do sinal de excitação da glote, do módulo ao quadrado da resposta em frequência do modelo de pulso glotal considerado, de um ganho que representa a potência do ar vindo do diafragma durante a fonação, e do módulo ao quadrado da resposta em frequência do trato vocal e radiação dos lábios e narinas.

2. Outra contribuição desta tese é a análise matemática de um gerador de impulsos cicloestacionário.

   É sabido que, durante uma elocução, as cordas vocais vibram a uma frequência média, denominada frequência fundamental. No entanto, no decorrer de um discurso, a frequência de vibração pode atingir uma taxa maior ou menor que a frequência média. Para o sinal de voz, esse movimento é cicloestacionário e modelado por um gerador que fornece um trem de impulsos cujo espaçamento tem característica cicloestacionária.

   No âmbito desta tese, o gerador é utilizado na modelagem de um sinal que controla a tensão a qual as cordas vocais são submetidas quando estimuladas por um sinal de controle. A frequência de oscilação das cordas vocais é diretamente proporcional à tensão. Dessa forma, a saída do gerador representa o sinal de excitação das cordas vocais.

   É apresentado um desenvolvimento matemático para a densidade espectral de potência do trem de impulsos cicloestacionário. A expressão obtida é em função da frequência

fundamental, da constante de sensibilidade do processo de modulação e dos parâmetros da função distribuição de probabilidade do sinal de controle.

3. A nova descrição da relação entre a forma de onda do sinal e a frequência de oscilação das cordas vocais é uma outra contribuição desta tese.

   A nova descrição da síntese de um sinal com característica de fala apresentada nesta tese considera que a voz é formada por um sistema linear invariante no tempo. Nesse caso, por se tratar de um sistema linear, durante o processo de geração da voz, não há ampliação da faixa de frequência.

   Nesse cenário, o trem de impulsos cicloestacionário herda da forma de onda do sinal de voz o movimento oscilatório das cordas vocais. Dessa forma, é considerado que as cordas vocais realizam o movimento de abertura e fechamento a cada ponto de cruzamento por zero obtido da forma de onda do sinal de voz.

4. Proposta do uso das distribuições Gamma unilateral e Rayleigh no ajuste do sinal que coordena o fechamento e abertura das cordas vocais

   Como mencionado, o sinal de controle é obtido da forma de onda do sinal de voz, nos pontos de cruzamento por zero. Esse sinal estimula a tensão na glote, proporcionando o movimento cicloestacionário das cordas vocais.

   Pela característica da cicloestacionariedade da frequência de vibração das cordas vocais, que tem maior probabilidade de vibrarem a uma taxa média, mais que também podem atingir taxas maiores ou menores de vibração, é proposto o ajuste da FDP do sinal $M(t)$ com as funções Gamma unilateral e Rayleigh, por apresentarem comportamento similar a do sinal de controle.

   A estimação dos parâmetros da FDP é importante na diferenciação de vozes saudáveis e patológicas.

5. É apresentada a transformada de Fourier da derivada do pulso glotal utilizado nesta tese, que é o modelo de Liljencrants-Fant.

5.2 Propostas para Trabalhos Futuros

Com o intuito de dar continuidade à pesquisa, são sugeridos novos testes para análise do desempenho do modelo descrito, estudo detalhado das características dos distúrbios das cordas vocais e análise da emulação de patologias por meio do modelo fonte-filtro apresentado.

1. Como trabalho futuro, pretende-se realizar novos testes de desempenho do modelo de geração da voz proposto, com elocuções de diferentes locutores, de regiões e nacionalidades distintas, bem como locuções com diferentes taxas de amostragens.

   Além disso, pretende-se observar o desempenho do modelo mediante novas modelagens do pulso glotal, como os modelos glotais de Rosenberg, de Fant e o de Klatt.

2. O estudo detalhado das particularidades de patologias como edemas, nódulos, pólipos e cistos é tema para trabalho futuro. O objetivo é observar características relacionadas a cada doença para obtenção de parâmetros que sejam adequados a identificação e classsificação de patologias pelo modelo de síntese da voz desenvolvido.

   Devido ao fato do modelo ter como base a biofísica da fonação no seu desenvolvimento, o seu uso na análise da diferenciação entre a voz saudável e a voz com alguma patologia é promissor.

   Para que a emulação de distúrbios sejam feitas pelo modelo proposto, é necessário a estimação de parâmetros para a função distribuição de probabilidade do sinal que controla a tensão, a partir da forma de voz gerada para cada doença, bem como a estimação da constante de sensibilidade da sequência de pulsos e a frequência fundamental.

   Além disso, a partir da análise dos parâmetros, deve-se definir faixas de valores e modelos para as particulariedade de cada distúrbio e verificar a influência no modelo proposto ao se analisar vozes patológicas masculinas, femininas e infantis. Por fim, pretende-se também realizar a comparação dos resultados obtidos na emulação de patologias com outras técnicas apresentadas na literatura.

APÊNDICE A

Segmentos Fonéticos do Português Brasileiro

Símbolo	Classificação	Exemplos
p	Oclusiva bilabial desvozeada	**P**ato
b	Oclusiva bilabial vozeada	**B**ala, **B**arco
t	Oclusiva alveolar desvozeada	**T**apa, **T**elha
d	Oclusiva alveolar vozeada	**D**ata, **D**ado
k	Oclusiva velar desvozeada	**C**apa, **C**arro
g	Oclusiva velar vozeada	**G**ato
t∫	Africada alveopalatal desvozeada	**T**ia
dZ	Africada alveopalatal desvozeada	**D**ia
f	Ficativa labiodental desvozeada	**F**aca, **F**ala, **F**arelo
v	Ficativa labiodental vozeada	**V**ento, **V**aca
s	Ficativa alveolar vozeada	**S**ala, Ca**ç**a, **C**ebola
z	Ficativa alveolar vozeada	Ca**s**a, **Z**ero
∫	Ficativa alveopalatal desvozeada	**Ch**á, A**ch**a
Z	Ficativa alveopalatal vozeada	**J**á
x	Fricativa velar desvozeada	**R**ata
R	Fricativa velar vozeada	Ca**r**ga
m	Nasal bilabial vozeada	**M**ala, **M**arca
n	Nasal alveolar vozeada	**N**ada, **N**ervo
$\tilde{y}$	Nasal palatal vozeada	Ba**nh**a, Arra**nh**ado
r	Tepe alveolar vozeado	Ca**r**a, P**r**ata
l	Lateral alveolar vozeada	**L**ata, P**l**ana, **L**uz
w	Lateral alveolar vozeada velarizada	Sa**l**ta, Ma**u**
λ	Lateral palatal vozeada	Ma**lh**a, Caval**lh**eiro

Símbolo	Classificação	Exemplos
i	Vogal alta anterior não-arredondada	Vi
$\tilde{i}$	Vogal alta anterior não-arredondada nasal	Vim
e	Vogal média-alta anterior não-arredondada	Ipê
$\tilde{e}$	Vogal média-alta anterior não-arredondada	Tempo
é	Vogal média-baixa anterior não-arredondada	Pé
a	Vogal baixa central não-arredondada	Pá
$\tilde{a}$	Vogal baixa central não-arredondada nasal	Lã
ó	Vogal média-baixa posterior arredondada	Avó
o	Vogal média-alta posterior arredondada	Avô
$\tilde{o}$	Vogal média posterior arredondada nasal	Tom
u	Vogal alta posterior arredondada	Jacu
$\tilde{u}$	Vogal alta posterior arredondada nasal	Jejum
I	Vogal alta anterior não-arredondada	Vi
i~	Vogal alta posterior arredondada	Vim
ê	Vogal média-baixa central	Ipê

APÊNDICE B

Locuções Utilizadas no Teste de Desempenho do Modelo de Produção de Voz

1. Cada aluno fez a sua avaliação.
2. A essa altura todos estavam emocionados.
3. A atriz recebeu o prêmio com entusiasmo.
4. Com o recuo do dólar, segmentos que tiveram um aumento exagerado estão voltando atrás.
5. Estou desencantado com o lixo na atividade política, afirmou.
6. Para a partida de domingo, contra o Palmeiras, pela taça Libertadores.

APÊNDICE C

Publicações

C.1 Artigos completos publicados em periódicos

1. R. B. Rocha, V. V. Freire, F. Madeiro, M. S. Alencar. Sistema de Segmentação de Fala Baseado na Observação do Pitch. Revista de Tecnologia da Informação e Comunicação, v. 4, p. 36-41, 2014.

C.2 Capítulos de livros publicados

1. R. B. Rocha, M. S. Alencar, F. M. B. Junior. Introdução à segmentação de voz.. In: Isabela do Rêgo Barros; Karl Heinz Efken; Moab Acioli; Nadia Azevedo; Ranata da Fonte; Roberta Caiado; Wanilda Cavalcanti.. (Org.). Ensino, texto e discurso.. 1ed.Curitiba: Editora CRV, 2014, v. , p. 93-102.

C.3 Trabalhos completos publicados em anais de congressos

1. R. B. Rocha, V. V. Freire, M. S. Alencar. Voice Segmentation System Based On Energy Estimation. In: 22 European Signal Processing Conference, 2014, Lisboa. Anais do 22 European Signal Processing Conference, 2014.

2. R. B. Rocha, G. B. Rocha, M. S. Alencar. Using Segmentation in the Development of a Speech Recognition System for Brazilian Portuguese. In: International Workshop on Telecommunications, 2013, Santa Rita do Sapucaí. V International Workshop on Telecommunications, 2013.

C.4 Resumos publicados em anais de congressos

1. R. B. Rocha, V. V. Freire, F. M. B. Junior, M. S. Alencar. Sistema de Segmentação Automático Aplicado ao Sinal de Voz. In: Encontro Anual do Iecom em Comunicações, Redes e Criptografia, 2013, Recife. Encontro Anual do Iecom em Comunicações, Redes e Criptografia, 2013.

2. R. B. Rocha, G. B. Rocha, M. S. Alencar. Acerca da Segmentação de Sinais de Voz. In: Encontro Anual do Iecom em Comunicações, Redes e Criptografia, 2012, Campina Grande. Encontro Anual do Iecom em Comunicações, Redes e Criptografia, 2012.

C.5 Artigos submetidos em periódicos

1. R. B. Rocha, W. J. L. Queiroz, M. S. Alencar. Multidimensional Speech Segmentation Technique. Journal of Communications and Information Systems, 2016.

APÊNDICE D

Análise Espectral do Pulso Glotal

Neste anexo é apresentado o desenvolvimento matemático realizado para obtenção da transformada de Fourier da derivada do pulso glotal de Liljencrants-Fant, utilizado neste trabalho.

Seja o modelo de Liljencrants-Fant dado por

$$E(t) = \begin{cases} E_o e^{\alpha t}\sin(\omega_g t), & t_o \leq t \leq t_e, \\ \frac{-E_e}{\epsilon T_a}\left(e^{-\epsilon(t - t_e)} - e^{-\epsilon T_b}\right), & t_e < t \leq t_c. \end{cases} \tag{D.1}$$

Se $t_o \leq t \leq t_e$, $E(\omega)$ pode ser escrita como

$$\begin{aligned} E(\omega) = &\left(\omega_g e^{-(-\alpha + j\omega)t_o}\cos(\omega_g t_o) - \alpha e^{-(-\alpha + j\omega)t_o}\sin(w_g t_o) + \right. \\ &+ j\omega e^{-(-\alpha + j\omega)t_o}\sin(w_g t_o) - \omega_g e^{-(-\alpha + j\omega)t_e}\cos(\omega_g t_e) + \\ &\left. + \alpha e^{-(-\alpha + j\omega)t_e}\sin(w_g t_e) - j\omega e^{-(-\alpha + j\omega)t_e}\sin(w_g t_e)\right) \\ &/\left(\omega^2 + 2j\alpha\omega - \alpha^2 - \omega_g^2\right), \end{aligned} \tag{D.2}$$

que pode ser reescrita como

$$\begin{aligned} E(\omega) = &\left(-e^{(\alpha - j\omega)t_o}\left(\omega_g\cos(\omega_g t_o) - \alpha\sin(\omega_g t_o)\right)\right. \\ &+ j\omega e^{(\alpha - j\omega)t_o}\sin(\omega_g t_o) - e^{(\alpha - j\omega)t_e}\left(\omega_g\cos(\omega_g t_e) - \alpha\sin(\omega_g t_e)\right) \\ &\left. - j\omega e^{(\alpha - j\omega)t_e}\sin(\omega_g t_e)\right) \\ &/\left(\omega^2 + 2j\alpha\omega - \alpha^2 - \omega_g^2\right). \end{aligned} \tag{D.3}$$

$$
\begin{aligned}
E(\omega) = & \left(-e^{(\alpha - j\omega)t_o} \sqrt{\alpha^2 + \omega_g^2} \left(\frac{\omega g}{\sqrt{\alpha^2 + \omega_g^2}} \cos(\omega_g t_o) - \frac{\alpha}{\sqrt{\alpha^2 + \omega_g^2}} \sin(\omega_g t_o) \right) \right. \\
& - e^{(\alpha - j\omega)t_e} \sqrt{\alpha^2 + \omega_g^2} \left(\frac{\omega g}{\sqrt{\alpha^2 + \omega_g^2}} \cos(\omega_g t_e) - \frac{\alpha}{\sqrt{\alpha^2 + \omega_g^2}} \sin(\omega_g t_e) \right) \\
& \left. + j\omega e^{(\alpha - j\omega)t_o} \sin(\omega_g t_o) - j\omega e^{(\alpha - j\omega)t_e} \sin(\omega_g t_e) \right) \\
& / \left(\omega^2 + 2j\alpha\omega - \alpha^2 - \omega_g^2 \right).
\end{aligned}
\tag{D.4}
$$

$$
\begin{aligned}
E(\omega) = & \left(-e^{(\alpha - j\omega)t_o} \sqrt{\alpha^2 + \omega_g^2} \cos(\omega_g t_o + \theta) - e^{(\alpha - j\omega)t_e} \sqrt{\alpha^2 + \omega_g^2} \cos(\omega_g t_e + \theta) \right. \\
& \left. + j\omega \left(e^{(\alpha - j\omega)t_o} \sin(\omega_g t_o) - e^{(\alpha - j\omega)t_e} \sin(\omega_g t_e) \right) \right) \\
& / \left(\omega^2 + 2j\alpha\omega - \alpha^2 - \omega_g^2 \right).
\end{aligned}
\tag{D.5}
$$

$$
\begin{aligned}
E(\omega) = & -\left(e^{-j\omega t_o} \alpha_o - e^{-j\omega t_e} \alpha_e + j\omega e^{-j\omega t_o} \beta_o - j\omega e^{-j\omega t_e} \beta_e \right) \\
& / \left(\omega^2 + 2j\alpha\omega - \alpha^2 - \omega_g^2 \right),
\end{aligned}
\tag{D.6}
$$

em que

$$
\alpha_o = e^{\alpha t_o} \sqrt{\alpha^2 + \omega_g^2} \cos(\omega_g t_o + \theta), \quad \theta = tg^{-1} \left(\frac{\alpha}{\omega_g} \right),
\tag{D.7}
$$

$$
\alpha_e = e^{\alpha t_e} \sqrt{\alpha^2 + \omega_g^2} \cos(\omega_g t_e + \theta),
\tag{D.8}
$$

$$
\beta_o = e^{\alpha t_o} \sin(\omega_g t_o),
\tag{D.9}
$$

$$
\beta_e = e^{\alpha t_e} \sin(\omega_g t_e).
\tag{D.10}
$$

Assim

$$
E(\omega) = -\frac{\left[e^{-j\omega t_o} \left(\alpha_o + j\omega\beta_o \right) - e^{-j\omega t_e} \left(\alpha_e + j\omega\beta_e \right) \right]}{\left(\omega + j\alpha \right)^2 - \omega_g^2}.
\tag{D.11}
$$

Ou ainda

$$E(\omega) = \frac{\left[e^{-j\omega t_e}\left(\alpha_e + j\omega\beta_e\right) - e^{-j\omega t_o}\left(\alpha_o + j\omega\beta_o\right)\right]}{\left(\omega + j\alpha\right)^2 - \omega_g^2}. \tag{D.12}$$

Ou

$$E(\omega) = \frac{\left(\cos(\omega t_e) - j\sin(\omega t_e)\right)\left(\alpha_e + j\omega\beta_e\right) - \left(\cos(\omega t_o) - j\sin(\omega t_o)\right)\left(\alpha_o + j\omega\beta_o\right)}{\omega^2 + 2j\alpha\omega - \alpha^2 - \omega_g^2}. \tag{D.13}$$

Ou

$$\begin{aligned} E(\omega) = &\left(\alpha_e\cos(\omega t_e) + j\beta_e\omega\cos(\omega t_e) - j\alpha_e\sin(\omega t_e) + \right. \\ &+ \omega\beta_e\sin(\omega t_e) - \alpha_o\cos(\omega t_o) - j\beta_o\omega\cos(\omega t_o) \\ &\left. + j\alpha_o\sin(\omega t_o) - \omega\beta_o\sin(\omega t_o)\right) \\ &/\left(\omega^2 + 2j\alpha\omega - \alpha^2 - \omega_g^2\right). \end{aligned} \tag{D.14}$$

Assim,

$$\begin{aligned} E(\omega) = &\left[\sqrt{\alpha_e^2 + \beta_e^2\omega^2}\cos\left(\omega t_e - tg^{-1}\left(\frac{\beta_e\omega}{\alpha_e}\right)\right)\right. \\ &- \sqrt{\alpha_o^2 + \beta_o^2\omega^2}\cos\left(\omega t_o - tg^{-1}\left(\frac{\beta_o\omega}{\alpha_o}\right)\right) \\ &- j\sqrt{\alpha_e^2 + \beta_e^2\omega^2}\sin\left(\omega t_e - tg^{-1}\left(\frac{\beta_e\omega}{\alpha_e}\right)\right) \\ &\left. + j\sqrt{\alpha_o^2 + \beta_o^2\omega^2}\sin\left(\omega t_o - tg^{-1}\left(\frac{\beta_o\omega}{\alpha_o}\right)\right)\right]\Big/ \\ &\left(\omega^2 + 2j\alpha\omega - \alpha^2 - \omega_g^2\right). \end{aligned} \tag{D.15}$$

Se $t_e < t \leq t_c$,
então

$$E(\omega) = \frac{-E_e}{\epsilon T_a}e^{\epsilon t_e}\int_{t_e}^{t_c} e^{-\epsilon t}e^{-j\omega t}dt + \frac{E_e}{\epsilon T_a}e^{-\epsilon T_b}\int_{t_e}^{t_c} e^{-j\omega t}dt. \tag{D.16}$$

$$E(\omega) = \frac{-E_e}{\epsilon T_a}e^{\epsilon t_e}\frac{e^{-(\epsilon + j\omega)t}}{-(\epsilon + j\omega)} + \frac{E_e}{\epsilon T_a}e^{-\epsilon T_b}\frac{e^{-j\omega t}}{-j\omega}. \tag{D.17}$$

$$E(\omega) = \frac{E_e}{\epsilon T_a} e^{\epsilon t_e} \frac{e^{-(\epsilon + j\omega)t_c} - e^{-(\epsilon + j\omega)t_e}}{(\epsilon + j\omega)} - \frac{E_e}{\epsilon T_a} e^{-\epsilon T_b} \frac{e^{-j\omega t_c} - e^{-j\omega t_e}}{-j\omega}. \tag{D.18}$$

$$E(\omega) = \frac{E_e}{\epsilon T_a} e^{\epsilon t_e} \frac{e^{-(\epsilon + j\omega)\frac{t_c}{2}}}{(\epsilon + j\omega)} \left[e^{-(\epsilon + j\omega)\frac{t_c}{2}} - e^{-(\epsilon+j\omega)t_e+(\epsilon + j\omega)\frac{t_c}{2}} \right] - \frac{E_e}{\epsilon T_a} e^{-\epsilon T_b} \frac{e^{-j\omega\frac{t_c}{2}}}{j\omega} \left[e^{-j\omega\frac{t_c}{2}} - e^{-j\omega)t_e + j\omega\frac{t_c}{2}} \right]. \tag{D.19}$$

$$\begin{aligned} E(\omega) = {} & \frac{E_e}{\epsilon T_a} e^{\epsilon t_e} e^{-(\epsilon + j\omega)\frac{t_c}{2}} \frac{e^{-(\epsilon + j\omega)\frac{t_e}{2}}}{(\epsilon + j\omega)} \\ & . \left[e^{-(\epsilon + j\omega)\frac{t_c}{2}} e^{(\epsilon + j\omega)\frac{t_e}{2}} - e^{-(\epsilon + j\omega)\frac{t_e}{2}} e^{(\epsilon + j\omega)\frac{t_c}{2}} \right] \\ & - \frac{E_e}{\epsilon T_a} e^{-\epsilon T_b} \frac{e^{-j\omega\frac{t_c}{2}} e^{-j\omega\frac{t_e}{2}}}{j\omega} \\ & . \left[e^{-j\omega\frac{t_c}{2}} e^{j\omega\frac{t_e}{2}} - e^{-j\omega\frac{t_e}{2}} e^{j\omega\frac{t_c}{2}} \right]. \end{aligned} \tag{D.20}$$

$$\begin{aligned} E(\omega) = {} & \frac{E_e}{\epsilon T_a} e^{\epsilon t_e} \frac{e^{-(\epsilon + j\omega)\frac{t_e+t_c}{2}}}{(\epsilon + j\omega)} \\ & . \left[e^{-(\epsilon + j\omega)\frac{t_c-t_e}{2}} - e^{(\epsilon + j\omega)\frac{t_c-t_e}{2}} \right] \\ & - \frac{E_e}{\epsilon T_a} e^{-\epsilon T_b} \frac{e^{-j\omega\frac{t_c+t_e}{2}}}{j\omega} \\ & . \left[e^{-j\omega\frac{t_c-t_e}{2}} - e^{j\omega\frac{t_c-t_e}{2}} \right]. \end{aligned} \tag{D.21}$$

$$\begin{aligned} E(\omega) = {} & -\frac{E_e}{\epsilon T_a} e^{\epsilon t_e} \frac{e^{-\epsilon(\frac{t_e+t_c}{2})}}{(\epsilon + j\omega)} e^{-j\omega(\frac{t_e+t_c}{2})} \\ & . \left[e^{(\epsilon + j\omega)\frac{t_c-t_e}{2}} - e^{-(\epsilon + j\omega)\frac{t_c-t_e}{2}} \right] \\ & \frac{E_e}{\epsilon T_a} e^{-\epsilon T_b} \frac{e^{-j\omega\frac{t_c+t_e}{2}}}{j\omega} \\ & . \left[e^{-j\omega\frac{t_c-t_e}{2}} - e^{j\omega\frac{t_c-t_e}{2}} \right]. \end{aligned} \tag{D.22}$$

$$\begin{aligned} E(\omega) \;=\; & -\frac{E_e}{\epsilon T_a} e^{\epsilon \frac{(t_c - t_e)}{2}} \frac{e^{-j\omega(\frac{t_e+t_c}{2})}}{(\epsilon \;+\; j\omega)} \\ & .\mathrm{senh}\left(\frac{t_c - t_e}{2}(\epsilon + j\omega)\right) \\ & + \frac{2E_e}{\epsilon T_a} e^{-\epsilon T_b} \frac{e^{-j\omega \frac{t_c+t_e}{2}}}{\omega} \\ & .\mathrm{senh}\left(\frac{t_c - t_e}{2}\omega\right). \end{aligned} \tag{D.23}$$

$$\begin{aligned} E(\omega) \;=\; & -\frac{E_e}{2\epsilon T_a}(t_c - t_e) e^{\frac{\epsilon}{2}(t_c - t_e)} e^{\frac{-j\omega}{2}(t_c + t_e)} \frac{\sin(j\frac{(t_c - t_e)}{2}(\epsilon + j\omega))}{j\frac{t_c - t_e}{2}(\epsilon + j\omega)} \\ & + \frac{2E_e}{\epsilon T_a} e^{-\epsilon T_b}(t_c - t_e) \frac{\sin(\frac{(t_c - t_e)\omega}{2})}{\frac{(t_c - t_e)\omega}{2}}. \end{aligned} \tag{D.24}$$

$$\begin{aligned} E(\omega) \;=\; & -\frac{E_e}{2\epsilon T_a}(t_c - t_e) e^{\frac{\epsilon}{2}(t_c - t_e)} e^{\frac{-j\omega}{2}(t_c + t_e)} \mathrm{Sa}\left(j\frac{(t_c - t_e)}{2}(\epsilon + j\omega)\right) \\ & + \frac{2E_e}{\epsilon T_a} e^{-\epsilon T_b}(t_c - t_e)\mathrm{Sa}\left(\frac{(t_c - t_e)}{2}\omega\right), \end{aligned} \tag{D.25}$$

em que,

$$\mathrm{Sa} \;=\; \frac{\sin(\mathrm{x})}{x}. \tag{D.26}$$

Dessa forma, a transformada de Fourier da derivada do pulso glotal é dada por

$$\begin{aligned} E(\omega) \;=\; & \frac{\sqrt{\alpha_e^2 \;+\; \beta_e^2\omega^2}\, e^{-j\left(\omega t_e \;-\; tg^{-1}\left(\frac{\beta_e}{\alpha_e}\omega\right)\right)}}{(\omega \;+\; j\alpha)^2 - \omega_g^2} \\ & - \frac{\sqrt{\alpha_o^2 \;+\; \beta_o^2\omega^2}\, e^{-j\left(\omega t_o \;-\; tg^{-1}\left(\frac{\beta_o}{\alpha_o}\omega\right)\right)}}{(\omega \;+\; j\alpha)^2 - \omega_g^2} \\ & + \frac{E_e}{\epsilon T_a}(t_c - t_e) e^{\frac{-j\omega}{2}(t_c + t_e)} \\ & .\left[e^{-\epsilon T_b}\mathrm{Sa}\left(\frac{(t_c - t_e)}{2}\omega\right) - \frac{1}{2}e^{-\frac{\epsilon}{2}(t_c - t_e)}\mathrm{Sa}\left(j\frac{(t_c - t_e)}{2}(\epsilon + j\omega)\right)\right]. \end{aligned} \tag{D.27}$$

Referências Bibliográficas

[1] C. D. P. Crovato. Classificação de Sinais de Voz Utilizando a Transformada Wavelet Packet e Redes Neurais Artificiais. Dissertação de mestrado, Faculdade de Engenharia da Universidade do Porto, 2004.

[2] M. E. Dajer. Padrões Visuais de Sinais de Voz através de Técnica de Análise de Não-Linear. Dissertação de mestrado, Escola de Engenharia de São Carlos, 2006.

[3] M. L. da C. Neto. Um Modelo para Geração de Prosódia de Palavras em Conversores Texto-Fala para a Língua Portuguesa Falada no Brasil. Tese de doutorado, Universidade Federal de Campina Grande, Abril 2004.

[4] F. S. Pacheco. Técnicas de Processamento de Sinais para Alteração de Parâmetros Prosódicos Aplicadas a um Sistema de Conversão Texto-Fala para a Língua Portuguesa Falada no Brasil. Dissertação de mestrado, Universidade Federal de Santa Catarina, 2001.

[5] S. L. do N. C. Costa. Análise Acústica, Baseada no Modelo Linear de Produção da Fala, para Discriminação de Vozes Patológicas. Tese de doutorado, Universidade Federal de Campina Grande, 2008.

[6] L. R. Rabiner and R. W. Schafer. *Digital Processing of Speech Signals*. Prentice Hall, New Jersey, 1978.

[7] I. R. Titze. *Principles of Voice Production*. Prentice Hall, New Jersey, 1994.

[8] A. S. Brandão. Modelagem Acústica da Produção da Voz Utilizando Técnicas de Visualização de Imagens Médicas Associadas a Métodos Numéricos. Tese de doutorado, Universidade Federal Fluminense, Niterói, 2011.

[9] J. M. Fechine. Reconhecimento Automático de Identidade Vocal Utilizando Modelagem Híbrida: Paramétrica e Estatística. Tese de doutorado, Universidade Federal da Paraíba, Campina Grande, Brasil, 2000.

[10] S. de O. Dias. Estimation of the Glottal Pulse from Speech or Singing Voice. Master's thesis, School of Engineering of the University of Porto, 2012.

[11] R. J. F. dos Santos. Avaliação de Pacientes com Paralisia Unilateral das Pregas Vocais. Dissertação de mestrado, Universidade de Aveiro, 2009.

[12] C. Gobl. The Voice Source in Speech Communication. Doctoral thesis, School of Engineering of the University of Porto, 2003.

[13] F. R. H. Andrade. Análise do Fluxo Glotal em Modelo da Laringe Baseado em Tomografia Computadorizada. Dissertação de mestrado, Escola de Engenharia de São Carlos, 2013.

[14] I. v. d. Berg. Myoelastic-Aerodynamic Theory of Voice Production. *J. Speech Hear*, pages 227–244, 1958.

[15] I. R. Titze. Comments on the Myoelastic-Aerodynamic Theory of Phonation. *The Journal of the Acoustical Society of America.*, 23:495–510, 1980.

[16] T. B. Patel and H. A. Patil. Novel Approach for Estimating Length of the Vocal Folds using Fujisaki Model. *Chinese Spoken Language Processing.*, pages 308–312, 2014.

[17] L. J. Raphael, G. J. Borden, and K. S. Harris. *Speech Science Primer.* Lippincott Williams Wilkins, 2011.

[18] G. Fant. *Acoustic Theory of Speech Production.* The Hague, 1970.

[19] R. D. Kent and C. Read. The Acoustics Analysis of Speech. *Singular Publishing Group Inc.*, 1992.

[20] S. M. Zitta. *Análise Perceptivo-Auditiva e Acústica em Mulheres com Nódulos Vocais.* Centro Federal de Educação Tecnológica - CEFET-PR, 2005.

[21] J. G. Proakis J. R. Deller and J. H. L. Hansen. *Discrete-Time Processing of Speech Signals.* Prentice-Hall, 1993.

[22] A. M. Selmini. Sistema Baseado em Regras para o Refinamento da Segmentação Automática de Fala. Tese de doutorado, Universidade Estadual de Campinas, Campinas, Brasil, Agosto de 2008.

[23] P. A. F. P. Fernandes. Modelo do Sistema de Produção de Voz Aplicável à Detecção de Anomalias nas Cordas Vocais. Dissertação de mestrado, Faculdade de Engenharia da Universidade do Porto, 2004.

[24] L. F. M. P. Coelho. Etiquetagem Automática de Sinais de Fala Segmentação e Classificação Fonética. Dissertação de mestrado, Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, Fevereiro de 2005.

[25] E. D. S. Paranaguá. Segmentação Automática do Sinal de Voz Para Sistemas de Conversão Texto-Fala. Tese de doutorado, Universidade Federal do Rio de Janeiro, Março 2012.

[26] J. Flanagan and L. Landgraf. Self-Oscillating Source for Vocal-Tract Synthesizers. *IEEE Transactions on Audio and Eletroacoustics*, 16(1):57–64, 1968.

[27] K. Ishizaka and J. Flanagan. Synthesis of Voiced Sounds from Two-Mass Model of the Vocal Cords. *Bell System Technical Journal*, 51:1233–1268, 1972.

[28] E. Cataldo, R. Sampaio and L. Nicolato. Uma Discussão sobre Modelos Mecânicos de Laringe para Síntese de Vogais. *Engevista*, 6(1):47–57, 2004.

[29] F. Alipour, D. A. Berry, and I. R. Titze. A Finite-Element Model of Vocal-Fold Vibration. *The Journal of the Acoustical Society of America*, December 2000.

[30] D. A. Berry, H. Herzel, I. R. Titze, and K. Krischer. Interpretation of Biomechanical Simulations of Normal and Chaotic Vocal Fold Oscillations with Empirical Eigenfunctions. *Journal of the Acoustical Society of America*, 95(6):3595–3604, 1994.

[31] D. A. Berry and I. R. Titze. Normal Modes in a Continuum Model of Vocal Fold Tissues. *Journal of the Acoustical Society of America*, 100(5):3345–3354, 1996.

[32] J. C. Lucero. Oscillation Hysteresis in a Two-Mass Model of the Vocal Folds. *Journal of Sound and Vibration*, 282(3-5):1247–1254, 2005.

[33] B. Hüttner and M. Döllinger and G. Luegmair and U. Eysholdt and A. Ziethe and E. Gurlek. Parameter Optimization for a Time-Dependent Multi-Mass Model for the Pharyngo-Esophageal Segment. *Proceedings of Models and Analysis of Vocal Emissions for Biomedical Applications*, 2011.

[34] S. L. Thomson and P. R. Murray. Self-Oscillating, Multi-Layer Numerical and Aartificial Vocal Fold Models with Thin Epithelial and Loose Cover Layers. *Proceedings of Models and Analysis of Vocal Emissions for Biomedical Applications*, 2011.

[35] M. O. Rosa. Laringe Digital. Tese de doutorado, Universidade de São Paulo, 2002.

[36] A. E. Aronson and D. M. Bless. Clinical voice disorders. *Thieme Medical Publishers*, 2009.

[37] J. V. de M. L. Marinus, J. M. F. R. de Araújo, H. M. Gomes, and S. C. Costa. On the Use of Cepstral Coefficients, Multilayer Perceptron Networks and Gaussian Mixture Models for Vocal Fold Edema diagnosis. *Biosignals and Biorobotics Conference.*, pages 1–6, 2013.

[38] A. M. Zorrilla, N. El-Zehiry, B. G. Zapirain, and A. Elmaghraby. Pathological Vocal Folds Diagnosis Using Modified Active Contour Models. *Information Sciences Signal Processing and their Applications.*, pages 504–507, 2010.

[39] A. Mendez, B. Garcia, J. Vicente, I. Ruiz, and K. Sanchez. Objective Model of Vocal Folds, Based on Glottal Closure, Opening Angles and Morphologic Criteria. *Signal Processing and Its Applications.*, pages 1–4, 2007.

[40] Y. Zhang anda M. F. Regner and J. J. Jiang. Theoretical Modeling and Experimental High-Speed Imaging of Elongated Vocal Folds. *IEEE Transactions on Biomedical Engineering.*, 58 (10):2725–2731, October 2011.

[41] M. Encina anda J. Yuz, M. Zañartu, and G. Galindo. Voice Fold Modeling Through the port-Hamiltonian Systems Approach. *IEEE Conference on Control Applications*, pages 1558–1563, 2015.

[42] B. H. Story and I. R. Titze. Voice Simulation with a Body-Cover Model of the Vocal Folds. *Journal of the Acoustical Society of America*, 97(2):1249–1260, 1995.

[43] R. Greiss, J. Rocha, and E. Matida. Modal Analysis of a Parameterized Model of Pathological Vocal Fold Vibration. *IEEE EMBS International Student Conference*, pages 1–4, 2016.

[44] A. Meireles. *Análise Acústica da Fala.* Editora Cortez, 2015.

[45] W. C. A. Costa. Reconhecimento de Fala Utilizando Modelos de Markov Escondidos (HMM's) de Densidades Contínuas. Dissertação de mestrado, Universidade Federal da Paraíba, Junho de 1994.

[46] J. P. R. Teixeira. Modelização Paramétrica de Sinais para Aplicação em Sistemas de Conversão texto-Fala. Dissertação de mestrado, Universidade do Porto, Outubro 1995.

[47] R. F. B. Sotero. Novas Abordagens para Codificação de Voz e Reconhecimento Automático de Locutor Projetadas Via Mascaramento Pleno em Frequência por Oitava. Dissertação de mestrado, Universidade Federal de Pernambuco, 2009.

[48] E. L. F. da Silva. Estimativas de Comportamento Vocálico de Locutores e um Novo Sistema de Separação Silábica. Dissertação de mestrado, Universidade Federal de Pernambuco, 2012.

[49] L. R. Rabiner and B. Juang. *Fundamentals on Speech Recognition.* Prentice Hall, 1996.

[50] S. Furui. *Digital Speech Processing, Synthesis, and Recognition*. Tokai University Press, 1985.

[51] L. M. Silva. Contribuições para a Melhoria da Codificação CELP a Baixas Taxas de Bits. Tese de doutorado, Pontifícea Universidade Católica do Rio de Janeiro, 1996.

[52] ITU-T Rec. G.729. General Aspects of Digital Transmission Systems Terminal Equipments – Coding of Speech at 8 kbits/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Prediction (CS-ACELP). 1996.

[53] B. Bessette, R. Salami, R. Lefebvre, M. Jelínek, J. Rotola-Pukkila, J. Vainio, H. Mikkola, and K. Järvinen. The Adaptive Multirate Wideband Speech Codec (AMR-WB). *IEEE Transactions on Speech and Audio Processing*, 10(8):620 – 636, November 2002.

[54] R. Salami, R. Lefebvre, A. Lakaniemi, K. Kontola, S. Bruhn, and A. Taleb. Extended AMR-WB for High-Quality Audio on Mobile Devices. *IEEE Communications Magazine*, pages 90 – 97, May 2006.

[55] R. da S. Maia. Codificação CELP e Análise Espectral da Voz. Dissertação de mestrado, Universidade Federal do Rio de Janeiro, Março de 2000.

[56] F. P. O. Araújo. Imitação da Voz Humana através do Processo de Análise-por-Síntese utilizando Algoritmo Genético e Sintetizador de Voz por Formantes. Tese de doutorado, Universidade Federal de Santa Catarina, 2015.

[57] J. Brito. Genetic Learning of Vocal Tract Area Functions for Articulatory Synthesis of Spanish Vowels. *Applied Soft Computing*, pages 1035–1043, 2007.

[58] I. S. Howard and M. A. Huckvale. Training a Vocal Tract Synthesizer to Imitate Speech Using Distal Learning. *Proceedings of InterSpeech*, 2005.

[59] A. Philippsen, F. R. Reinhart, and B. Wrede. Learning how to Speak: Imitation-based Refinement of Syllable Production in an Articulatory-Acoustic Model. *IEEE Int. Conf. on Development and Learning and on Epigenetic Robotics (ICDL)*, pages 187–192, 2014.

[60] P. Taylor. *Text-to-Speech Synthesis*. Cambridge University Press, 2009.

[61] T. Dutoit. *A Introduction to Text-to-Speech Synthesis*. Academic Publishers, 2011.

[62] K. W. A. X. da Silva. Sistema de Conversão Texto-Fala com Busca Otimizada de Unidades Acústicas em Banco de Voz. Dissertação de mestrado, Universidade Federal do Rio de Janeiro, Dezembro 2011.

[63] F. O. Simões. Implementação de um Sistema de Conversão Texto-Fala para o Português do Brasil. Tese de mestrado, Unicamp, 1999.

[64] V. L. Latsch. Construção de um Banco de Unidades para Síntese da Fala por Concatenação no Domínio Temporal. Dissertação de mestrado, Universidade Federal do Rio de Janeiro, Abril de 2005.

[65] E. da S. Morais. Algoritmo OPWI e LDM-GA para Sistemas de Conversão Texto-Fala de Alta Qualidade Empregando a Tecnologia SCAUS. Tese de doutorado, Unicamp, 2006.

[66] E. A. M. Klabbers. Segmental and Prosodic Improvements to Speech Generation. Tese de doutorado, Technische Universiteit Eindhoven, Netherlands, 2000.

[67] D. Klatt. Software for a Cascade / Parallel Formant Synthesizer. *Journal of the Acoustical Society of America*, 1980.

[68] D. Klatt and L. Klatt. Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Speakers. *Journal of the Acoustical Society of America*, 1990.

[69] G. K. Anumanchipalli, Y. C. Cheng, J. Fernandez, X. Huang, Q. Mao, and A. W. Black. KlaTTStat: Knowledge-based Statistical Parametric Speech Synthesis. *7th ISCA Workshop on Speech Synthesis*, 2010.

[70] C. J. L. Pimentel. *Comunicação Digital*. Brasport, 2007.

[71] V. Turcu and A. Pop. On the Use of Pulse Position Modulation Theory in the Study of Period Variability Phenomena. *Interplay of Periodic, Cyclic and Stochastic Variability in Selected Areas of the H-R Diagram.*, pages 373–376, 2003.

[72] M. S. de Alencar. *Communications Systems*. Editora Springer, 2005.

[73] G. Degottex. Glottal Source and Vocal-Tract Separation. Estimation of Glottal Parameters, Voice Transformation and Synthesis using a Glottal Model. Tese de doutorado, Université Paris, 2010.

[74] G. Fant. Vocal-Source Analysis – A Progress Report. *TL-QPSR*, pages 31–53, 1979.

[75] J. Liljencrants G. Fant and Qi guaq Lin. A Four Parameter Model of Glottal Flow. *TL-QPSR*, pages 1–13, 1985.

[76] D. Klatt and L. Klatt. Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers. *Jurnal of the Acoustical Society of America*, pages 820–857, 1990.

[77] D. O'Shauqhnessy. Modern Methods of Speech Synthesis. *IEEE Circuits and Systems Magazine*, 7(3):6–23, 2007.

[78] B. Lindblom and J. Sundberg. Acoustical Consequences of Lip, Tongue, Jaw and Laryny Movement. *Journal of the Acoustical Society of America*, pages 1166–799, 1971.

[79] M. S. de Alencar. *Probabilidade e Processos Estocásticos*. Editora Érica, 2009.

[80] B. P. Lathi. *Modern Digital and Analog Communication Systems.* Oxford University Press, 2009.

[81] G. Fant. The LF-model Revisited. Transformations and Frequency Domain Analysis. *Quarterly Progress and Status Report (STL-QPSR)*, 36(2-3):119–156, 1995.

[82] G. Fant and K. Gustafson. LF-Frequency Domain Analysis. *Quarterly Progress and Status Report (STL-QPSR)*, 37(2):135–138, 1996.

[83] B. Doval, C. d'Alessandro, and N. Henrich. The Spectrum of Glottal Flow Models. *Acta Acustica united with Acustica*, 92(1):1–21, 2006.

[84] J. Kane, M. Kane, and C. Gob. A Spectral LF Model Based Approach to Voice Source Parameterisation. *Interspeech Conference*, 2010.