Universidade Estadual de Campinas

Faculdade de Engenharia Elétrica e de Computação

Algoritmos OPWI e LDM-GA para Sistemas de Conversão Texto-Fala de Alta Qualidade Empregando a Tecnologia SCAUS

Autor: Edmilson da Silva Morais

Orientador: Prof. Dr. Fábio Violaro

Tese de Doutorado apresentada à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos para obtenção do título de Doutor em Engenharia Elétrica. Área de concentração: **Engenharia de Telecomunicações.**

Banca Examinadora

Fábio Violaro, Dr. ..... FEEC/Unicamp
Fernando Gil Vianna Resende Júnior, PhD. ..... COPPE/UFRJ
Miguel Arjona Ramírez, Dr. ..... Poli/USP
Fernando José Von Zuben, Dr. ..... FEEC/Unicamp
Plínio Almeida Barbosa, Docteur ..... IEL/Unicamp
Jayme Garcia Arnal Barbedo, Dr. ..... FEEC/Unicamp

Campinas, SP, 28 de Julho de 2006

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA E ARQUITETURA - BAE - UNICAMP

M792a

Morais, Edmilson da Silva
Algoritmos OPWI e LDM-GA para sistemas de conversão texto-fala de alta qualidade empregando a tecnologia SCAUS. --Campinas, SP: [s.n.], 2006.

Orientador: Fábio Violaro
Tese (doutorado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Sistemas de telecomunicações. 2. Processamento de sinais – Técnicas digitais. 3. Processamento de sinais. 4. Inteligência artificial. 5. Sistemas de processamento de fala. 6. Lingüística computacional. I. Violaro, Fábio. II. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. III. Título.

Título em Inglês: Algorithm OPWI and LDM-GA for high quality text-to-speech synthesis based on automatic unit selection
Palavras-chave em Inglês: Telecommunication systems, Digital signal processing, Machine Learning techniques, Speech and language science and technology, Computational linguistics
Área de concentração: Telecomunicações e Telemática
Titulação: Doutor em Engenharia Elétrica
Banca examinadora: Fernando Gil Vianna Resende Júnior, Miguel Arjona Ramírez, Fernando José Von Zuben, Plínio Almeida Barbosa, Jayme Garcia Arnal Barbedo
Data da defesa: 28/06/2006

Resumo

Esta Tese apresenta dois novos algoritmos denominados OPWI (*Optimized Prototype Waveform Interpolation*) e LDM-GA (*Linguistic Data Mining Using Genetic Algorithm*). Estes algoritmos são formulados no contexto de sistemas CTF-SCAUS (sistemas de Conversão Texto-Fala empregando a tecnologia de Seleção e Concatenação Automática de Unidades de Síntese). O algoritmo OPWI é apresentado como uma nova alternativa para o módulo de *Back-End* de sistemas CTF-SCAUS, permitindo modificações prosódicas e suavizações espectrais de alta qualidade. O algoritmo LDM-GA foi desenvolvido com o objetivo de minimizar problemas de treinamento, em sistemas CTF-SCAUS, relacionados a distribuições de probabilidade com características LNRE (*Large Number of Rare Events*). Resultados da avaliação dos algoritmos OPWI e LDM-GA são apresentados e discutidos detalhadamente. Além destes dois algoritmos, esta Tese apresenta uma ampla revisão bibliográfica sobre os principais módulos de um sistema CTF-SCAUS, módulos de *Front-End* (Módulo lingüístico), módulo prosódico, módulo de seleção de unidades de síntese e módulo de *Back-End* (Módulo de síntese).

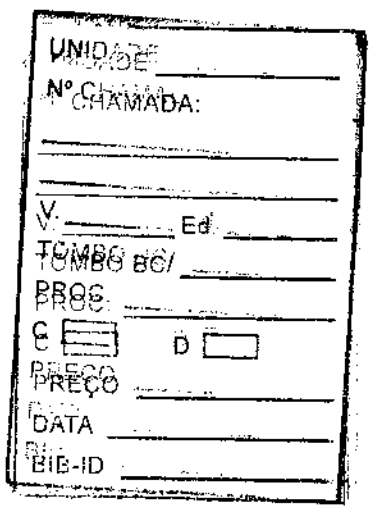
Palavras-chave: Ciência e tecnologia da fala e da linguagem, conversão texto-fala, reconhecimento automático de fala, aprendizado de máquina, lingüística computacional, processamento digital de sinais de fala, algoritmos genéticos, modelos de regressão linear, árvores de classificação e de regressão, modelos ocultos de Markov, algoritmo de Viterbi.

Abstract

This Thesis presents two new algorithms for Unit Selection Based Text-to-Speech systems (USB-TTS). The first algorithm is the OPWI (Optimized Prototype Waveform Interpolation), which was designed to be used as a Back-End module for USB-TTS. The second algorithm is the LDM-GA (Linguistic Data Mining Using Genetic Algorithm), which was designed to minimize training problems related to LNRE (Large Number of Rare Events) distributions. Experimental results and analysis of the OPWI and LDM-GA algorithms are presented in detail. The OPWI algorithm is evaluated under operations of analysis/re-synthesis and prosodic modifications, TSM (Time Scale Modifications) and PSM (Pitch Scale Modifications). The LDM-GA is evaluated in the context of phoneme segmen-

tal duration prediction based on linear regression model. In addition to these two new algorithms (OPWI and LDM-GA), this Thesis presents a large review of the main modules of a USB-TTS system, Front-End Module (Linguistic module), prosodic module, unit-selection module and Back-End module (Synthesis module).

Key-words: Speech and language science and technology, text-to-speech synthesis, automatic speech recognition, machine learning, computational linguistics, digital signal processing, genetic algorithm, linear regression models, classification and regression trees, hidden Markov models, Viterbi algorithm.



À Jussara

Agradecimentos

Este trabalho de Doutorado teve início em Agosto de 1997, ainda durante o quarto semestre do meu Mestrado em Engenharia Elétrica pela Unicamp. Entretanto, em setembro de 1998 este trabalho foi interrompido, sendo retomado somente em março de 2004. Durante este longo período várias foram as pessoas que contribuiram para a minha formação e conseqüentemente para a realização deste trabalho. Em especial eu gostaria de expressar meus sinceros agradecimentos:

Ao meu orientador Prof. Fábio Violaro pelos ensinamentos, auxílios e conselhos; por viabilizar a realização deste trabalho e pela compreensão, paciência e amizade.

Aos meus pais Expedito e Dirce que sempre se colocaram ao meu lado oferecendo-me apoio e encorajando-me a seguir adiante.

A todas as pessoas que contribuiram para o meu bem estar físico e mental durante a realização desta Tese: os amigos Jaqueline Gonçalves, Pablo Arantes, Maria Luíza, Ana Cristina Matte, Antônio Marcos Araújo e Fabbryccio Cardoso, meu professor de violão Roberto, os instrutores da FEF/Unicamp (Faculdade de Educação Física da Unicamp) dos cursos de extensão em natação, dança de salão e tênis de campo, e o pessoal do Nordestão pelas partidas de futebol nos sábados à tarde.

À INOVA (Agência da Inovação da Unicamp), em especial a Paulo Lemos, Eduardo Grizendi e Prof. Roberto Lotufo, pela oportunidade de participar (desde maio de 2004) do ambiente de pré-incubação de projetos da Unicamp, com um projeto diretamente relacionado ao tema desta Tese.

Ao Prof. Luis Geraldo Meloni, pela oportunidade de trabalhar em um projeto sobre canal de interatividade para o Sistema Brasileiro de TV-Digital (RTP-DSP/2005), durante o período de

Maio de 2005 a Dezembro de 2005.

Aos pesquisadores com quem trabalhei durante o período de afastamento do meu Doutorado, de Setembro de 1998 até Dezembro de 2003: o Dr. Jorge Gonzalez, Jaime Botella e Carlos Nuno da IBM em Sevilha na Espanha; o Dr. Jimmy Krusman da IBM em Heidelberg na Alemanha; o Dr. Paul Taylor, Prof. Steve Isard, Dr. Simon King, Dr. Korin Richmond e Dr. Joe Frankel do CSTR (*Centre for Speech Technology Research*) em Edimburgo, na Escócia; o Prof. Bernd Möbius, Prof. Greg Dogil, Dr. Gregor Müller, Antje Schweitzer, Dra. Michaela Attler e Dr. Norbert Braunschweiler do IMS (Institute for Natural Language Processing) da Universidade de Stuttgart, na Alemanha; a Dra. Kate Knill, Peter Jackson, Dra. Tina Burrows, Dr. Dimitri, Dr. Gabriel Webster, Dr. Marc e Dra. Sabine do STG (*Speech Technology Group*) da Toshiba em Cambrige, na Inglaterra; Saito Morita, Dr. Kagoshima e Dr. Akamine do RDC (*Research and Development Center*) da Toshiba em Kawasaki no Japão; e em especial o Prof. Steve Young da Universidade de Cambridge, na Inglaterra.

À Unicamp, pela minha formação acadêmica, e à CAPES pelo apoio financeiro durante o Doutorado.

À Fapesp, por permitir a continuidade deste trabalho através de um projeto PIPE (Projeto de Inovação Tecnológica em Pequenas Empresas).

Por último, gostaria de agradecer a Deus por sempre ter se colocado ao meu lado, "ajudando-me a combater o bom combate e a guardar a fé".

Sumário

Lista de Figuras xvii

Lista de Tabelas xxv

Lista de Símbolos xxvii

Lista de Acrônimos xxix

Trabalhos Publicados Pelo Autor xxxi

1 Introdução 1

1.1 Considerações Iniciais 1

1.2 Tecnologia SCAUS 2

1.2.1 Arquitetura de um Sistema CTF-SCAUS 3

1.2.2 Síntese em Domínio Irrestrito × Síntese em Domínios Restritos 5

1.3 Treinamento de Sistemas CTF-SCAUS 5

1.4 Problemas com Sistemas CTF-SCAUS 6

1.4.1 Distribuições LNRE 6

1.4.2 Nível de Qualidade Vocal × Nível de Estabilidade Vocal 7

1.5 Motivações e Objetivos 8

1.6 Estrutura da Tese 9

2 *Front-End*: Módulo Lingüístico 11

2.1 Introdução 11

2.2 Análise do Texto 11

2.2.1 Determinação da Estrutura do Texto 12

2.2.2 Normalização do Texto 13
2.2.3 Análise Lingüística 14
2.3 Análise Fonética 18
2.3.1 Desambiguação Homográfica 18
2.3.2 Análise Morfológica 18
2.3.3 Conversão Grafema-Fonema 18
2.4 Léxico 19
2.4.1 Compressão do Léxico 19
2.5 Considerações Finais 20

3 Módulo Prosódico **21**
3.1 Introdução 21
3.2 Módulo Prosódico 22
3.2.1 Aspectos Paralingüísticos: Estilo de Elocução 23
3.2.2 Aspectos Fonológicos: Prosódia Simbólica/Abstrata 23
3.2.3 Aspectos Prosódicos: Realização Acústica 25
3.3 Modelagem da Entoação: Contorno de F_0 26
3.3.1 Modelos Baseados em Tons 26
3.3.2 Modelos Perceptivos 27
3.3.3 Modelos Superposicionais 28
3.3.4 Modelos de Estilização Acústica 30
3.4 O Modelo Entoacional de Paul Taylor: Modelo Tilt 30
3.4.1 Processo de Estilização 30
3.4.2 Análise 32
3.4.3 Síntese 34
3.5 O Modelo Entoacional de Kagoshima 36
3.5.1 Processo de Estilização 36
3.5.2 Análise 36
3.5.3 Síntese 38
3.6 Considerações Finais 39

4 Módulo de Seleção Automática de Unidades de Síntese **41**

4.1 Introdução 41

4.2 Unidades de Síntese 42

4.3 Seleção de Unidades de Síntese 43

4.3.1 Funções de Custo Fonético-Prosódico 45

4.3.2 Funções de Custo Concatenativo 45

4.3.3 Custo Total 47

4.4 Clusterização de Unidades de Síntese 47

4.4.1 Métricas para Estimar a Distância Entre Unidades de Síntese 49

4.4.2 Técnicas de Poda 52

4.4.3 *Lookup Tables* 53

4.5 Corpus e Inventário de Unidades de Síntese 53

4.5.1 Definição da Aplicação 53

4.5.2 Projeto 53

4.5.3 Seleção de Locutores 54

4.5.4 Gravações 54

4.5.5 Segmentação e Etiquetagem 55

4.5.6 Inventário de Unidades de Síntese 55

4.6 Considerações Finais 56

5 Algoritmo LDM-GA: Formulação Teórica **57**

5.1 Introdução 57

5.2 Formulação do Problema de Predição da Duração Segmental da Fala 59

5.2.1 Conceitos Fundamentais e Definições 59

5.2.2 Problemas na Modelagem da Duração Segmental da Fala 60

5.3 Descrição e Análise do Corpus Utilizado 61

5.3.1 Fatores e Níveis 61

5.3.2 Conjunto de Fones Utilizados 62

5.3.3 Freqüência de Ocorrência dos Fones na Base de Dados 62

5.3.4 Dimensionalidade do Espaço de *Fatores* Lingüísticos 65

5.3.5 Histogramas dos Fones da Base de Dados 65

5.4 Regressão Linear Multivariável Empregando QMTI 66

5.4.1 Quantificação e Modelagem 66
5.4.2 Efeito dos Fatores Lingüísticos 67
5.5 Algoritmo LDM-GA 67
5.5.1 Princípios de Operação do Algoritmo LDM-GA 67
5.5.2 Estimação das Topologias Ótimas dos Modelos de Regressão 68
5.5.3 Estimação das Topologias Intermediárias dos Modelos de Regressão 69
5.5.4 Operações Envolvidas na Estimação das Topologias Intermediárias e Ótimas . . 71
5.5.5 Construção da Árvore de *Clusterização* 73
5.5.6 Principais Operações na Construção da Árvore de Clusterização 74
5.5.7 Seleção das Classes de Fones Ótimas 76
5.6 Considerações Finais 76

6 Algoritmo LDM-GA: Resultados Experimentais e Análises **79**
6.1 Introdução 79
6.2 Esparsidade do Espaço de *Fatores* Lingüísticos 80
6.3 Modelagem por Fones 81
6.3.1 Topologias Ótimas Segundo o Método LDM-GA 81
6.3.2 Topologias Ótimas Segundo o Método MANOVA 82
6.3.3 Análise do Operador Regra Majoritária 84
6.3.4 Resultados Comparativos: QMTI/Ph + LDM-GA × QMTI/Ph Cheio 86
6.3.5 Resultados Comparativos: QMTI/Ph + LDM-GA × QMTI/Ph + MANOVA . 88
6.3.6 Resultados Comparativos: QMTI/Ph + LDM-GA × RT/Ph 89
6.4 Modelagem por Classes de Fones 91
6.4.1 Árvore de Clusterização 91
6.4.2 Seleção das Classes de Fones Ótimas 96
6.4.3 Topologias Ótimas por Classes de Fones 96
6.4.4 Resultados Comparativos: QMTI/Cl + LDMGA × QMTI/Cl + MANOVA × RT/Cl 97
6.5 Análise dos Efeitos dos *Fatores* Lingüísticos 97
6.6 Análise de Desempenho dos Modelos: QMTI/Ph + LDM-GA 100
6.7 Considerações Finais 107
6.7.1 Considerações Sobre os Resultados 107

7 Algoritmo OPWI: Fundamentos Teóricos **109**

7.1 Introdução 109

7.2 Formulação do Problema 112

7.2.1 Etapas de Análise e Síntese 113

7.3 Estimativa dos Instantes de Análise e Segmentação Sonoro/Não-Sonoro 114

7.4 Decomposição CEL × CER 115

7.4.1 Etapa de Análise da Decomposição CEL/CER 117

7.4.2 Etapa de Síntese da Decomposição CEL/CER 118

7.4.3 Freqüências de Corte do Filtro de Decomposição CEL/CER 118

7.5 Estimativa do Nível de Estacionariedade do Sinal de Fala 121

7.5.1 Primeiro Critério, C^1 121

7.5.2 Segundo Critério, C^2 122

7.5.3 Terceiro Critério, C^3 122

7.6 Protótipo Ótimo e sua Representação Temporal 123

7.7 Estimativa dos Protótipos Ótimos: Método I 124

7.7.1 Normalização de Protótipos 124

7.7.2 Critério de Otimização no Domínio Temporal 124

7.7.3 Critério de Otimização no Domínio Espectral 125

7.7.4 Solução do Critério de Otimização no Domínio Espectral 127

7.7.5 Aproximação para a Estimativa do Protótipo Seguinte 129

7.8 Estimativa dos Protótipos Ótimos: Método II 131

7.9 Análise da Componente CER: Modelo Auto-regressivo 132

7.10 Síntese da Componente CEL: Interpolação Tempo-Freqüência 134

7.11 Modificações Prosódicas na Componente CEL 137

7.11.1 Modificações na Taxa de Articulação: TSM 137

7.11.2 Modificações do Contorno da Freqüência Fundamental: PSM 137

7.11.3 Modificações Conjuntas de PSM e TSM 138

7.12 Síntese da Componente CER 138

7.13 Suavização Espectral 141

7.13.1 Suavização da Componente CEL 142

7.13.2 Suavização da Componente CER 143

7.14 Aspectos de Implementação 144
7.14.1 Etapa de Análise 144
7.14.2 Etapa de Síntese 144
7.15 Considerações Finais 145

8 Algoritmo OPWI: Resultados Experimentais e Análises **147**
8.1 Introdução 147
8.2 Sinais de Fala a Serem Utilizados 148
8.3 Decomposição CEL/CER 151
8.3.1 Filtros para Decomposição CEL/CER 151
8.3.2 Avaliação Espectral 154
8.3.3 Avaliação Temporal 157
8.4 Estimativa do Nível de Estacionariedade 160
8.5 Estimativa dos Protótipos Ótimos 163
8.6 Análise e Ressíntese 167
8.7 Modificações Prosódicas 168
8.7.1 TSM por um Fator Constante 168
8.7.2 PSM por um Fator Constante 173
8.7.3 PSM e TSM por Fatores Variáveis 177
8.8 Exemplos para Avaliação Audível 182
8.8.1 Sinais Originais 182
8.8.2 Sinais Sintetizados Sem Modificações Prosódicas 182
8.8.3 Sinais Sintetizados Com Modificações Prosódicas de TSM 183
8.8.4 Sinais Sintetizados Com Modificações Prosódicas de PSM 183
8.8.5 Sinais Sintetizados Com TSM e PSM Cossenoidais 183
8.9 Considerações Finais 183

9 Conclusões **185**
9.1 Principais Contribuições 185
9.1.1 Ampla Revisão sobre Sistemas CTF-SCAUS 185
9.1.2 Algoritmo LDM-GA 186
9.1.3 Algoritmo OPWI 187

9.2 Sugestões para Trabalhos Futuros 188
9.2.1 Algoritmo LDM-GA 188
9.2.2 Algoritmo OPWI 188
9.3 Considerações Finais 189

Referências bibliográficas **190**

A Histogramas das Durações dos Fones para Avaliação do Algoritmo LDM-GA **201**

B Normalização dos Protótipos **207**

Lista de Figuras

1.1 Diagrama de blocos com os principais componentes de um sistema CTF-SCAUS. . . . 4

2.1 Diagrama de blocos das principais operações do módulo de *Front-End*. 12

2.2 Análise sintática da frase em inglês "*The astronomers saw stars with ears*". As Figuras (a) e (b) mostram duas possíveis análises sintáticas desta mesma sentença. 14

2.3 Exemplo de um etiquetador morfossintático (*Part-of-Speech Tagger*) para a frase em francês, "*J' entre par la porte*". A linha contínua (em negrito) indica as etiquetas selecionadas, (PN-Pess., V., Prep., Art., Subst.). 15

3.1 Diagrama de blocos das principais operações do módulo prosódico. 22

3.2 Gramática do sistema de Janet Pierrehumbert (Pierrehumbert, 1980). 27

3.3 Componentes do modelo superposicional de Fujisaki. 29

3.4 Ilustração do posicionamento e do formato dos *eventos entoacionais*. O símbolo (s) indica a posição dos núcleos das sílabas e os símbolos (a) e (b) representam os *eventos tonais* e *eventos de fronteira*, respectivamente. 31

3.5 Parâmetros RFC. 32

3.6 Comportamentos dos eventos de acordo com os valores de *tilt*. 33

3.7 Geração da curva de F_0 a partir do texto analisado lingüisticamente. 34

3.8 Modelo entoacional de Kagoshima para geração automática do contorno de F_0. 37

3.9 Diagrama de blocos do modelo entoacional de Kagoshima para geração automática do contorno de F_0. 37

4.1 Processo de seleção automática de unidades de síntese. 44

4.2 Estimativa do custo de concatenação. 46

4.3 Processo de clusterização de unidades de síntese. 48

4.4 Processo de seleção de unidade de síntese utilizando-se árvores de clusterização e algoritmos de programação dinâmica (DTW ou algoritmo de Viterbi). 50
4.5 Processo de alinhamento para o cálculo do custo de concatenação. 51
4.6 Projeto e aquisição da base de dados de unidades de síntese. 54

5.1 Freqüência dos 45 fones presentes na base de dados. Os fones seguem a mesma ordenação da Tabela 5.5 65
5.2 Diagrama de blocos das etapas do algoritmo LDM-GA. 69
5.3 Algoritmo para estimação das topologias ótimas dos modelos QMTI. 70
5.4 Rotina p/ estimação da $i-ésima$ topologia intermediária utilizando: $DCl_i^{T_j}$ e $DCl_i^{V_j}$. . 70
5.5 Algoritmo para Clusterização dos Fones 73
5.6 Rotina para partição do $i-ésimo$ cluster do $j-ésimo$ nível da árvore. 74

6.1 Dados para o fone [aa] no formato binário (quantificado). Os primeiros 45 níveis (ao longo do eixo horizontal) representam a identidade do fone [aa]. 80
6.2 Análise de Componentes Principais sobre os dados do fone [aa]. Esta análise permite reduzir a dimensão original do espaço de 168 a 77, 55 e 27, com perdas de representação dos dados, respectivamente, iguais a 0%, 2% e 18%. 81
6.3 Topologias ótimas selecionadas pelo método LDM-GA para cada um dos 45 fones da tabela 5.4. 82
6.4 Frequência de ocorrência de cada um dos 13 fatores linguísticos ao longo das 45 topologias ótimas da Figura 6.3. 83
6.5 Topologias ótimas selecionadas pelo método MANOVA para cada um dos 45 fones da tabela 5.4. 83
6.6 Frequência de ocorrência de cada um dos 13 fatores linguísticos ao longo das topologias ótimas das 45 topologias ótimas da Figura 6.5. 84
6.7 50 topologias intermediárias obtidas para o fone [AR]. 85
6.8 50 topologias intermediárias obtidas para o fone [b]. 85
6.9 50 topologias intermediárias obtidas para o fone [@]. 86
6.10 QMTI/Ph + LDM-GA (indicado por ∘) versus QMTI/Ph cheio (indicado por ∗). . . . 87
6.11 QMTI/Ph + LDM-GA (indicado por ∘) versus QMTI/Ph cheio (indicado por ∗). . . . 88
6.12 QMTI/Ph + LDM-GA (indicado por ∘) versus QMTI/Ph cheio (indicado por ∗). . . . 88

6.13 QMTI + Regra Majoritária (indicado por ○) versus QMTI + MANOVA (indicado por □). 89
6.14 QMTI/Ph + LDM-GA (indicado por ○) versus QMTI/Ph + LDM-GA (indicado por □). 90
6.15 QMTI/Ph + LDM-GA (indicado por ○) versus QMTI/Ph + MANOVA (indicado por □). 90
6.16 QMTI/Ph + LDM-GA (indicado por ○) versus QMTI/Ph + RT (indicado por ▽). . . 91
6.17 QMTI/Ph + LDM-GA (indicado por ○) versus QMTI/Ph + RT (indicado por ▽). . . 92
6.18 QMTI/Ph + LDM-GA (indicado por ○) versus QMTI/Ph + RT (indicado por ▽). . . 92
6.19 Árvore de clusterização: Primeira Parte. 95
6.20 Árvore de clusterização: Segunda Parte. 95
6.21 Árvore de clusterização: Terceira Parte. 96
6.22 Classes de fones selecionadas da árvore de clusterização. O símbolo (○) indica os resultados dos modelos QMTI/Cl + LDM-GA. O símbolo (∗) indica a média dos modelos QMTI/Ph + LDM-GA (para os fones contidos em cada classe). 97
6.23 Topologias para as classes de fones selecionadas. 99
6.24 Comparação de desempenho dos modelos QMTI/Cl + LDM-GA (indicado por ○), QMTI/Cl + MANOVA (indicado por ◇) e RT/Cl (indicado por □). 99
6.25 Efeitos dos *fatores* lingüísticos associados ao fone [e]. 100
6.26 Erro em RMS dos modelos QMTI/Ph + LDM-GA para todos os 45 fones da Tabela 5.4 101
6.27 Percentual de Erro RMS dos modelos QMTI/Ph + LDM-GA para todos os 45 fones da Tabela 5.4 101
6.28 Desempenho do modelo QMTI/Ph + LDM-GA para o fone [ae] 103
6.29 Desempenho do modelo QMTI/Ph + LDM-GA para o fone [b] 103
6.30 Desempenho do modelo QMTI/Ph + LDM-GA para o fone [d] 104
6.31 Desempenho do modelo QMTI/Ph + LDM-GA para o fone [e] 104
6.32 Desempenho do modelo QMTI/Ph + LDM-GA para o fone [g] 105
6.33 Desempenho do modelo QMTI/Ph + LDM-GA para o fone [H] 105
6.34 Desempenho do modelo QMTI/Ph + LDM-GA para o fone [p] 106
6.35 Desempenho do modelo QMTI/Ph + LDM-GA para o fone [r] 106
6.36 Desempenho do modelo QMTI/Ph + LDM-GA para o fone [z] 107

7.1 Diagrama de blocos das etapas de análise e síntese do algoritmo OPWI. 115
7.2 Estimativa dos instantes de análise nos segmentos sonoros (localizados próximos aos IFGs). 116

7.3 Estimativa dos instantes de análise nos segmentos não-sonoros, por interpolação linear dos períodos fundamentais (T_0) dos segmentos sonoros adjacentes (anterior e seguinte). 116

7.4 Processo de decomposição CEL X CER: Etapa de análise. 119

7.5 Processo de decomposição CEL X CER: Etapa de síntese. 120

7.6 (a) Segmento sonoro do sinal de fala. (b) Componente CER, correspondente ao sinal em (a). (c) Resíduo de predição correspondente a componente CER em (b). 133

7.7 (a) Segmento da componente CEL. (b) Protótipos ótimos. (c) Extensão periódica (dois períodos fundamentais) das respectivas representações temporais dos protótipos apresentados em (b). 135

7.8 (a) Processo de *Overlap and Add* utilizado para síntese do resíduo de predição da componente CER. 142

8.1 Forma de onda e espectograma do sinal SF_PB. Taxa de amostragem 22050 Hz. . . . 149

8.2 Forma de onda e espectograma do sinal SM_PB. Taxa de amostragem 22050 Hz. . . . 149

8.3 Forma de onda e espectograma do sinal SF_US. Taxa de amostragem 16000 Hz 150

8.4 Forma de onda e espectograma do sinal SM_GER. Taxa de amostragem 16000 Hz . . 150

8.5 Freqüências de corte dos filtros de decomposição CEL/CER, $f_c(k)$, para o sinal SF_PB (linha contínua). Contorno estilizado de $f_c(k)$ (linha pontilhada). 152

8.6 Freqüências de corte dos filtros de decomposição CEL/CER, $f_c(k)$, para o sinal SM_PB (linha contínua). Contorno estilizado de $f_c(k)$ (linha pontilhada). 152

8.7 Freqüências de corte dos filtros de decomposição CEL/CER, $f_c(k)$, para o sinal SF_US (linha contínua). Contorno estilizado de $f_c(k)$ (linha pontilhada). 153

8.8 Freqüências de corte dos filtros de decomposição CEL/CER, $f_c(k)$, para o sinal SM_GER (linha contínua). Contorno estilizado de $f_c(k)$ (linha pontilhada). 153

8.9 Forma de onda e espectograma da componente CEL de SF_PB. Há uma ênfase na estrutura harmônica do sinal, os segmentos não-sonoros são eliminados e a componente de ruído dos segmentos sonoros é fortemente atenuada. 155

8.10 Forma de onda e espectograma da componente CER de SF_PB. Esta componente contém toda a energia dos segmentos não-sonoros e também as componentes de ruído que se misturavam (ao longo de toda a banda de freqüências) à componente harmônica (CEL) 155

8.11 Espectro de magnitude de um segmento janelado do sinal SF_PB, centrado no instante de tempo 3,8 segundos. 156

8.12 Espectro de magnitude de um segmento janelado da componente CEL, do sinal SF_PB, centrado no instante de tempo 3,8 segundos. 156

8.13 Espectro de magnitude de um segmento janelado da componente CER, do sinal SF_PB, centrado no instante de tempo 3,8 segundos. 157

8.14 Espectro de magnitude de um segmento janelado do sinal SF_PB, centrado no instante de tempo 6,06 segundos. 157

8.15 Espectro de magnitude de um segmento janelado da componente CEL, do sinal SF_PB, centrado no instante de tempo 6,06 segundos. 158

8.16 Espectro de magnitude de um segmento janelado da componente CER, do sinal SF_PB, centrado no instante de tempo 6,06 segundos. 158

8.17 Segmento do sinal SF_PB submetido ao processo de decomposição CEL/CER. (a) Sinal original, (b) componente CEL e (c) componente CER. (Destaque para o trecho fricativo sonoro) 159

8.18 Segmento do sinal SF_PB submetido ao processo de decomposição CEL/CER. (a) Sinal original, (b) componente CEL e (c) componente CER. 159

8.19 Segmento do sinal SF_PB submetido ao processo de decomposição CEL/CER. (a) Sinal original, (b) componente CEL e (c) componente CER. 160

8.20 Análise e detecção dos níveis de estacionariedade do sinal SM_GER. (a) As faixas verticais (cinza) indicam os segmentos estacionários. (b) Critério C3 que combina os critérios C1 e C2. (c) Critério C2, mede variações na estrutura dos formantes do sinal ao longo do tempo. (d) Critério C1, mede variações de energia ao longo do tempo. . . 161

8.21 Classificação estacionário × não-estacionário ao longo de um trecho da componente CEL do sinal SM_GER. (Segmentos estacionários são indicados pelas faixas cinzas). 162

8.22 Classificação estacionário × não-estacionário ao longo de um trecho da componente CEL do sinal SM_GER. (Segmentos estacionários são indicados pelas faixas cinzas). 162

8.23 Classificação estacionário × não-estacionário ao longo de um trecho da componente CEL do sinal SM_GER. (Segmentos estacionários são indicados pelas faixas cinzas). 162

8.24 Análise da representação temporal dos protótipos ótimos estimados. (a) sinal original (linha tracejada); extensão periódica da representação temporal do protótipo estimado ao longo de dois períodos fundamentais (linha contínua). (b) Espectro de magnitude correspondente à extensão periódica (2 períodos) da representação temporal do protótipo em (a). (c) Espectro de magnitude correspondente ao sinal original em (a) (ao longo dos 2 períodos). 164

8.25 Análise da representação temporal dos protótipos ótimos estimados. (a) sinal original (linha tracejada); extensão periódica da representação temporal do protótipo estimado ao longo de dois período fundamentais (linha contínua). (b) Espectro de magnitude correspondente à extensão periódica (2 períodos) da representação temporal do protótipo em (a). (c) Espectro de magnitude correspondente ao sinal original em (a) (ao longo dos 2 períodos). 165

8.26 Análise da representação temporal dos protótipos ótimos estimados. (a) sinal original (linha tracejada); extensão periódica da representação temporal do protótipo estimado ao longo de dois períodos fundamentais (linha continua). (b) Espectro de magnitude correspondente à extensão periódica (2 períodos) da representação temporal do protótipo em (a). (c) Espectro de magnitude correspondente ao sinal original em (a) (ao longo dos 2 períodos). 166

8.27 Segmento original do sinal SF_PB no intervalo de 0,57 seg. a 4,88 seg. 169

8.28 Segmento da Figura 8.27 submetido a uma TSM igual a 2,4. 169

8.29 Segmento da Figura 8.27 submetido a uma TSM igual a 0,7. 170

8.30 Segmento do sinal SF_PB submetido a uma TSM igual a 0,7. (a) Sinal original; (b) Componente CEL sintetizada; (c) Componente CER sintetizada; (d) Sinal sintetizado (CEL + CER) 170

8.31 Segmento do sinal SF_PB submetido a uma TSM igual a 2,4. (a) Sinal original; (b) Componente CEL sintetizada; (c) Componente CER sintetizada; (d) Sinal sintetizado (CEL + CER) 171

8.32 Segmento do sinal SF_PB submetido a uma TSM igual a 2,4. (a) Sinal original; (b) Componente CEL sintetizada; (c) Componente CER sintentizada; (d) Sinal sintetizado (CEL + CER) 171

8.33 Segmento do sinal SF_PB submetido a uma TSM igual a 2,4. (a) Sinal original; (b) Componente CEL sintetizada; (c) Componente CER sintentizada; (d) Sinal sintetizado (CEL + CER) 172

8.34 Segmento do sinal SM_GER submetido a uma TSM = 1,6. (a) Sinal original; (b) Componente CEL sintetizada; (c) Componente CER sintentizada; (d) Sinal sintetizado (CEL + CER) 172

8.35 Segmento do sinal SF_PB submetido a uma PSM igual a $\frac{1}{2,2}$. A estrutura harmônica se encontra altamente comprimida, porém sem nenhum prejuizo para a estrutura formântica do sinal. 174

8.36 Segmento do sinal SF_PB submetido a uma PSM igual a $\frac{1}{0,7}$. A estrutura harmônica se encontra expandida, porém a estrutura formântica do sinal encontra-se inalterada. . 174

8.37 Segmento do sinal SF_PB submetido a uma PSM igual a $\frac{1}{2,2}$. (a) sinal original, (b) componente CEL, (c) componente CER e (d) componentes CEL + CER. 175

8.38 Segmento do sinal SF_PB submetido a uma PSM igual a $\frac{1}{1,5}$. (a) sinal original, (b) componente CEL, (c) componente CER e (d) componentes CEL + CER. 175

8.39 Segmento do sinal SF_PB submetido a uma PSM igual a $\frac{1}{1,5}$. (a) sinal original, (b) componente CEL, (c) componente CER e (d) componentes CEL + CER. Destaque para o trecho fricativo sonoro. 176

8.40 Segmento do sinal SF_PB submetido a uma PSM $\frac{1}{0,7}$. (a) sinal original, (b) componente CEL, (c) componente CER e (d) componentes CEL + CER. Destaque para a presença de uma leve oscilação na componente DC, no final da componente CER. 176

8.41 Segmento do sinal SF_PB (CEL + CER) submetido a uma PSM com variação cossenoidal. 178

8.42 Segmento da componente CEL do sinal SF_PB submetido a uma PSM com variação cossenoidal. 178

8.43 Segmento da componente CER do sinal SF_PB submetido a uma PSM com variação cossenoidal. 179

8.44 Segmento do sinal SF_PB (CEL + CER) submetido a uma TSM com variação cossenoidal. 179

8.45 Segmento da componente CEL do sinal SF_PB submetido a TSM com variação cossenoidal. 180

8.46 Segmento da componente CER do sinal SF_PB submetido a TSM com variação cossenoidal. 180
8.47 Segmento do sinal SF_PB (CEL + CER) submetido a TSM e PSM com variação cossenoidal. 181
8.48 Segmento da componente CEL do sinal SF_PB, submetido a TSM e PSM com variação cossenoidal. 181
8.49 Segmento da componente CER do sinal SF_PB, submetido a TSM e PSM com variação cossenoidal. 182

A.1 Histogramas dos fones [@], [AR], [ER], [H], [OR], [Q]. 201
A.2 Histogramas dos fones [aa], [ae], [ai], [au], [b], [ccc], [ch], [d], [dh], [dx]. 202
A.3 Histogramas dos fones [e], [ei], [f], [g], [i],[ii], [jh], [k], [l], [m]. 203
A.4 Histogramas dos fones [n], [ng], [oi], [oo], [ou], [p], [r], [s], [sh], [t]. 204
A.5 Histogramas dos fones [th], [u], [uh], [uu], [v], [w], [y], [z], [zh]. 205

Lista de Tabelas

5.1 Descrição dos 14 *fatores* lingüísticos utilizados nos modelos de predição da duração segmental dos fones. 62

5.2 Descrição dos níveis associados aos 14 *fatores* lingüísticos utilizados. 63

5.3 Exemplos de *fatores* lingüísticos associados ao fone [@] (to allow). 63

5.4 Descrição dos 45 fones utilizados. Notações no formato da Toshiba e do *International Phonetic Alphabet* - IPA. 64

5.5 Número de exemplos na base de dados para cada um dos 45 fones 64

6.1 Testes estatísticos de Mann-Whitney: QMTI/Ph + LDM-GA versus QMTI/Ph Cheio 89

6.2 Testes estatísticos de Mann-Whitney: QMTI/Ph + LDM-GA versus QMTI/Ph + MANOVA 91

6.3 Testes estatísticos de Mann-Whitney: QMTI/Ph + LDM-GA versus QMTI/Ph + RT 92

6.4 Clusters da árvore de clusterização: Primeira parte 93

6.5 Classes da árvore de clusterização: Segunda e Terceira partes. 94

6.6 Testes estatísticos não paramétricos de Mann-Whitney: QMTI/Cl + LDM-GA versus QMTI/Cl + MANOVA e versus QMTI/Cl + RT 98

8.1 Valores para o fator ξ. 151

8.2 Valores de SNR obtidos pelo algoritmo OPWI no processo de análise/ressíntese. 167

Lista de Símbolos

$\lfloor \cdot \rfloor$	-	Menor inteiro maior ou igual a
$\lceil \cdot \rceil$	-	Maior inteiro menor ou igual a
$\langle \cdot \rangle$	-	Inteiro mais próximo
$mod(A, B)$	-	Resto da divisão inteira entre A e B
$s(n)$	-	Sinal de fala
$x(n)$	-	Componente de Evolução Lenta - CEL - do sinal de fala
$y(n)$	-	Componente de Evolução Rápida - CER - do sinal de fala
$e(n)$	-	Resíduo LPC da componente CER
$X_j^P(k)$	-	Protótipo ótimo associado ao $j-$ésimo instante de análise
$x_j^P(n)$	-	Representação temporal de $X_j^P(k)$
$we_j^d(n)$	-	Segmento de *pitch* a direita
$we_j^e(n)$	-	Segmento de *pitch* a esquerda
n_j^a	-	$j-$ésimo instante de análise
n_j^s	-	$j-$ésimo instante de síntese
F_s	-	Freqüência de amostragem
F_{0_i}	-	Freqüência fundamental associada a n_i^a
T_{0_i}	-	Período fundamental associado a n_i^a
N_{ij}^a	-	Número de amostras entre n_i^a e n_j^a
N_{ij}^s	-	Número de amostras entre n_i^s e n_j^s
$C^t(U, T_i)$	-	Custo fonético-prosódico entre a unidade U e o alvo T_i
$C^c(U_{i-1}, U_i)$	-	Custo de concatenação entre as unidades U_{i-1} e U_i
$\Re(A)$	-	Parte real de A
$\Im(A)$	-	Parte imaginária de A

s - Segundos
ms - Milissegundos

Lista de Acrônimos

ANN	-	*Artificial Neural Network*
ANOVA	-	*Analysis of Variance*
C4.5	-	Algoritmo para Treinamento de Árvores de Classificação
CART	-	*Classification And Regression Tree*
CEL	-	Componente de Evolução Lenta
CELP	-	*Coded Excitation Linear Prediction*
CER	-	Componente de Evolução Rápida
CTF	-	Conversão de Texto em Fala / Conversor de Texto em Fala
CTF-SCAUS	-	CTF empregando a Tecnologia SCAUS
CTF-SCAUS-DI	-	CTF empregando a Tecnologia SCAUS em Domínio Irrestrito
CTF-SCAUS-DR	-	CTF empregando a Tecnologia SCAUS em Domínio Restrito
CT	-	*Classification Tree*
DFT	-	*Discrete Fourier Transform*
DTW	-	*Dinamic Time Warpping*
FST	-	*Finite State Transducer*
GA	-	*Genetic Algorithm*
HMM	-	*Hidden Markov Models*
HNM	-	*Harmonic plus Noise Model*
IDFT	-	*Inverse Discrete Fourier Transform*
IFG	-	Instante de Fechamento da Glote
IPA	-	*International Phonetic Alphabet*
LDM-GA	-	*Linguistic Data Mining Using Genetic Algorithm*
LNRE	-	*Large Number of Rare Events*
LPC	-	*Linear Prediction Coefficients/Coding*
LP-PSOLA	-	*Linear Prediction Pitch Synchronous Overlap And Add*
LSF	-	*Line Spectral Coefficients*
MAP	-	*Maximization a Posteriori*
MBE	-	*Multiband Excitation Coding*
MBROLA	-	*Multiband Resynthesis Overlap and Add*

MFCC	-	*Mel Frequency Cepstral Coefficients*
MLLR	-	*Maximum Likelihood Linear Regression*
OPWI	-	*Optimized Prototype Waveform Interpolation*
PCA	-	*Principal Component Analysis*
PoS	-	*Part-of-Speech*
PSM	-	*Pitch Scale Modification*
QMTI	-	*Quantification Method Type I*
RAF	-	Reconhecimento Automático de Fala
RELP	-	*Residual-Excited Linear Prediciton*
RFC	-	*Rise and Fall Conection*
RLMV	-	Regressão Linear Multivariável
RMS	-	*Root Mean Square*
RNLMV	-	Regressão Não-Linear Multivariável
RT	-	*Regression Tree*
SCAUS	-	Seleção e Concatenação Automática de Unidades de Síntese
SVM	-	*Support Vector Machine*
TD-PSOLA	-	*Time Domain Pitch Synchronous Overlap And Add*
ToBI	-	*Tons and Break Indexes*
TSM	-	*Time Scale Modification*
VMM	-	*Visible Markov Models*

Trabalhos Publicados Pelo Autor

1. E. Morais, F. Violaro. "Exploratory Analysis of Linguistic Data Based on Genetic Algorithm for Robust Modeling of the Segmental Duration of Speech". *Interspeech - International Congress on Speech Processing* (Interspeech'2005), Lisboa, Portugal, Setembro 2005.

2. E. Morais, F. Violaro. "Tutorial Estendido: Data-Driven Text-to-Speech Synthesis". *XXII Simpósio Brasileiro de Telecomunicações* (SBrT'2005), Campinas, SP, Brasil, Setembro 2005.

3. E. Morais, F. Violaro. "Exploratory Analysis of Linguistic Data and Its Application to Speech Segmental Duration". *XXII Simpósio Brasileiro de Telecomunicações* (SBrT'2005), Campinas, SP, Brasil, Setembro 2005.

4. E. Morais, J. Vieira, P. Arantes e A. Matte. "Metodologias para Projeto e Aquisição de uma Base de Dados Lingüísticos para Treinamentos e Avaliações de Sistemas de Reconhecimento de Fala". *III TIL - Workshop em Tecnologia da Informação e da Linguagem* (TIL'2005), São Leopoldo, RS, Brasil, Julho 2005.

5. E. Morais, F. Violaro. "Análise Exploratória de Dados Lingüísticos para uma Modelagem Robusta da Duração Segmental da Fala". *III TIL - Workshop em Tecnologia da Informação e da Linguagem* (TIL'2005), São Leopoldo, RS, Brasil, Julho 2005.

6. L. Barbosa, J. Gonçalves, E. Morais. "Uma Proposta de Canal de Interatividade para o SBTV através de Comunicação sem Fio em RF Intrabanda". *Workshop de TV Digial e Interativa* (SIBGRAPI 2005), Natal, RN, Brasil, Outubro, 2005.

7. E. Morais, F. Violaro. "Mini Curso: Corpus-Based Concatenative Text-to-Speech Synthesis". *IWT - International Workshop on Telecommunication* (IWT 2004), Santa Rita do Sapucaí, MG, Brasil, Agosto 2004.

8. A. Schweitzer, N. Braunschweiler, E. Morais, et al. Chapter "*SmartKom - Foundations of Multimodal Dialogue Systems*", chapter Multimodal Speech Synthesis, In Wolfgang Wahlster (Ed.), Spring-Verlag, 2004.

9. K. Knill, E. Morais, T. Burrows, P. Jackson. "Toshiba Technical Report'2003". *Speech Technology Group, Cambridge Research Laboratory*, Cambridge, Inglaterra, 2003.

10. K. Knill, E. Morais, T. Burrows, P. Jackson. "Toshiba Technical Report'2002". *Speech Technology Group, Cambridge Research Laboratory*, Cambridge, Inglaterra, 2002.

11. A. Schweitzer, N. Braunschweiler, E. Morais. "Prosody Generation in the SmartKom Project". *International Congress of Prosody Generation* (Speech Prosody 2002), Aix-en-Provence, França, Maio 2002.

12. E. Morais, G. Dogil. "Concatenative Text-to-Speech Synthesis Based on Waveform Interpolation (A Time Frequency Approach)". *The Journal of the Acoustic Society of America* (2001), Volume 110, Issue 5, pp 2775, Lauderdale, Flórida, USA, 2001.

13. E. Morais, P. Taylor, F. Violaro. "Prototype Waveform Interpolation Applied for Concatenative Speech Synthesis (A Time Frequency Approach)". *ICSLP - International Congress of Speech and Language Processing*, Beijin, China, 2000.

14. E. Morais, F. Violaro, C. Ynoguti. "A Comparative Study Between a Hybrid System and a Traditional HMM System for Continous Speech Recognition". *International Telecommunication Symposium*, São Paulo, SP, Brasil, 1998.

15. E. Morais, F. Violaro, P. A. Barbosa. "Prosodic Speech Modification Using TFI (Time Frequency Interpolation)". *International Telecommunication Symposium*, São Paulo, SP, Brasil, 1998.

16. E. Morais. "Reconhecimento Automático de Fala Contínua Empregando Modelos Híbridos ANN + HMM". Tese de Mestrado, *Universidade Estadual de Campinas*, Campinas, SP, Brasil, Dezembro de 1997.

17. E. Morais, F. Violaro. "Sistemas Híbridos ANN-HMM para Reconhecimento Automático de Fala Contínua". *XV SBT - Symposium Brasileiro de Telecomunicações*, Recife, PE, Brasil, 1997.

18. E. Morais, F. Violaro. "Modelos Ocultos de Markov Treinados a Partir de ANN e sua Aplicação em Reconhecimento Automático de Fala". *III CBRN - III Congresso Brasileiro de Redes Neurais Artificiais*, Florianópolis, SC, Brasil, 1997.

19. E. Morais, F. Violaro. "Sistemas Híbridos ANN-HMM Baseados nos Critérios ML e MAP e sua Aplicação ao Reconhecimento de Séries Temporais". 3° *Simpósio Brasileiro de Automação Inteligente, 1997*, pp. 406-411, Vitória, ES, Brasil, 1997.

20. E. Morais, L. M. Silva. "Um Estudo Comparativo de Taxas de Transmissão de Protótipos para a Codificação TFI de Sinais de Voz". TELEMO'96, 1997, pp. 337-342, Curitiba, PR, Brasil, 1996.

21. E. Morais. "Sistema CELP/TFI para Codificação de Voz e Sua Aplicação na Síntese de Fala a Partir de Texto". Dissertação Final de Graduação, *Universidade de Brasília*, Brasília, DF, Brasil, Dezembro de 1995.

Capítulo 1

Introdução

1.1 Considerações Iniciais

A área de Conversão de Texto em Fala (CTF) passou por uma ruptura de paradigma na última década (Ostendorf and Bulyko, 2002). Os sistemas, desenvolvidos nos laboratórios de pesquisa e/ou comerciais, deixaram de ser quase que puramente "artesanais"e passaram a ser predominantemente treinados, automaticamente, a partir de corpora de fala (devidamente segmentados e etiquetados lingüisticamente) (Hunt and Black, 1996), (Eide, 2003), (Huang and Acero, 1998), (Black, 1996), (Quazza et al., 2001). Hoje em dia, a ênfase tem sido dada, primordialmente, a técnicas de engenharia tais como: otimização de funções de custo, modelagem estatística e processamento digital de sinais, muito mais do que à elaboração de regras lingüísticas *ad hoc*. Como conseqüência, os modernos sistemas CTF passaram a ser capazes de sintetizar fala com uma qualidade muito próxima à da fala natural.

Durante vários anos as áreas de Reconhecimento Automático de Fala (RAF) e CTF foram consideradas conceitualmente diferentes. O objetivo principal dos sistemas de RAF sempre foi a modelagem das variabilidades entre-locutores, enquanto os sistemas CTF sempre se concentraram nas variabilidades intra-locutores (variabilidades na fala de um locutor específico). Sistemas RAF, em geral, não se preocupavam com aspectos prosódicos da fala. Por outro lado, aspectos prosódicos tais como ritmo e entoação, sempre foram considerados de fundamental importância para os sistemas CTF. Entretanto, apesar destas diferenças conceituais entre as áreas de CTF e RAF, convergências significativas têm ocorrido entre estas duas áreas. Muitas das técnicas atualmente utilizadas em sistemas CTF foram trazidas da área de RAF. Por outro lado, aspectos prosódicos vêm gradativamente ganhando importância na área de RAF, principalmente em sistemas de diálogo, os quais envolvem reconhecimento e compreensão de fala. Um bom exemplo desta convergência entre as áreas de RAF e CTF é a nova tecnologia para sistemas CTF introduzida por K. Tokuda, (Tokuda et al., 2002), (Yoshimura et al., 1999), (Shichiri et al., 2002), (Tokuda et al., 1999), que emprega Modelos Ocultos de Markov - HMM (Hidden Markov Models) para converter o texto de entrada em uma seqüência de parâmetros acústicos e prosódicos (coeficientes *Mel Cepstrais*, freqüência fundamental e duração) e em seguida sintetiza o sinal de fala a partir desta seqüência de parâmetros.

Algumas das técnicas atualmente utilizadas em sistemas CTF que foram trazidas da área de RAF são: Árvores de decisão baseadas em entropia (Breiman et al., 1993), (Quinlan, 1993), HMM (*Hidden Markov Models*) (Rabiner, 1989), programação dinâmica (Rabiner, 1989), métodos de busca baseados no algoritmo de Viterbi (Rabiner, 1989), coeficientes Mel Ceptrais (Rabiner, 1989) e adaptação de locutor utilizando MAP (*Maximum a posteriori*) (Rabiner, 1989) e/ou MLLR (*Maximum Likelihood Linear Regression*) (Leggetter and Woodland, 1995). Além destas técnicas trazidas da área de RAF, outros métodos também largamente utilizados nos modernos sistemas CTF são: regressão linear e não-linear de múltiplas variáveis (Jobson, 1991), árvores de regressão (Breiman et al., 1993), (Quinlan, 1993), redes neurais (Haykin, 1994), (Bishop, 1995), transdutores de estado-finito (Bulyko, 2002), algoritmos gulosos (Zhu and Zhang, 2002), (Möbius, 2000), algoritmos genéticos (Bäck et al., 2000a), (Bäck et al., 2000b), e várias técnicas de processamento digital de sinais (Quatieri, 2002), (Dutoit, 1997) e (Stylianou, 1996).

1.2 Tecnologia SCAUS

A tecnologia que tem dominado o cenário de sistemas CTF nos últimos anos é a de Seleção e Concatenação Automática de Unidades de Síntese (SCAUS) (Hunt and Black, 1996), (Black, 1996). Os sistemas CTF que empregam a tecnologia SCAUS (sistemas CTF-SCAUS) formulam o paradigma de conversão de texto em fala como um problema não-paramétrico no qual a fala sintetizada pode ser gerada a partir de um processo de seleção e concatenação de unidades de síntese previamente gravadas (extraídas de um sinal de fala natural).

Os primeiros sistemas CTF a obterem sucesso com a tecnologia SCAUS utilizavam um único exemplar de cada unidade de síntese. Como conseqüência, suas bases de dados eram consideradas compactas e o processo de seleção de unidades de síntese era imediato. Entretanto, a existência de um único exemplar para cada unidade de síntese não era suficiente para garantir toda a variabilidade fonético/prosódica exigida pelo processo de síntese. Como resultado, a simples concatenação destas unidades levava, quase que invariavelmente, a descontinuidades espectrais e a contornos prosódicos inadequados. Para minimizar estes problemas, esta tecnologia confiava fortemente em técnicas de processamento digital de sinais para suavizar as fronteiras espectrais (entre unidades adjacentes) e modificar prosodicamente estas unidades.

Apesar dos enormes esforços/avanços realizados na área de processamento digital de sinais de fala, considera-se que nenhuma das técnicas de processamento de sinais, até hoje desenvolvida, seja capaz de realizar suavizações espectrais e modificações prosódicas sem introduzir degradações no sinal de fala (Stylianou, 1996), (Dutoit, 1997), (Chappell and Hansen, 1998). Por esta razão, somados aos avanços tecnológicos na área de informática (capacidade de processamento e disponibilidade de memória), os sistemas CTF-SCAUS passaram a empregar grandes inventários de unidades de síntese (com vários exemplares de cada unidade), devidamente projetados para maximizar a cobertura das variabilidades fonético-prosódicas associadas à lingua que se deseja modelar. Segundo o estado-da-arte dos sistemas CTF-SCAUS, o processo de síntese de uma sentença consiste na seleção *on-fly* (em tempo de execução)

da seqüência de unidades de síntese que melhor satisfaz as seguintes propriedades:

- A seqüência de unidades de síntese deve apresentar um contorno prosódico o mais próximo possível do desejado (critério paradogmático);
- As descontinuidades espectrais, ao longo das fronteiras entre as unidades de síntese selecionadas, devem ser minimizadas (critério sintagmático).

Analisando-se as duas propriedades acima, pode-se concluir que um dos principais objetivos da tecnologia SCAUS é minimizar a necessidade de suavizações espectrais e modificações prosódicas das unidades de síntese por meio de técnicas de processamento digital de sinais. Em sistemas CTF-SCAUS que empregam inventários de unidades de síntese extremamente extensos, e que apresentam um processo de seleção de unidades de síntese altamente robusto, as operações de processamento digital de sinais são normalmente limitadas apenas à concatenação (com leves suavizações espectrais nas fronteiras) das unidades de síntese, eliminando-se, por completo, qualquer tipo de modificação prosódica. A não utilização de modificações prosódicas nesses sistemas tem como objetivo preservar o *nível de qualidade vocal* das unidades de síntese, dado que qualquer tipo de modificação prosódica, introduz, em maior ou menor grau, algum tipo de distorção na qualidade vocal das unidades de síntese (Dutoit, 1997).

1.2.1 Arquitetura de um Sistema CTF-SCAUS

O diagrama da Figura 1.1 mostra os quatro módulos principais de um sistema CTF-SCAUS: módulo de *Front-End* (modelagem lingüística), módulo prosódico, módulo de seleção de unidades de síntese e módulo de *Back-End* (síntese do sinal de fala) (Huang et al., 2001). Além destes quatro módulos principais, a Figura 1.1 também mostra o corpus de fala empregado na modelagem/treinamento do sistema e o inventário de unidades de síntese utilizado pelo módulo de *Back-End* para a síntese de sinal de fala. A seguir, é apresentada uma breve descrição de cada um destes componentes:

- ***Front-End*: Módulo Lingüístico**. Responsável por várias das análises lingüísticas a serem realizadas no texto a ser convertido em fala. Algumas destas análises são: Determinação da estrutura do texto (por exemplo, divisão do texto em frases), normalização de texto (expansão de abreviaturas, siglas e números), análises lingüísticas (sintática, semântica), desambiguação homográfica, análise morfológica e conversão de grafemas em fonemas.
- **Módulo Prosódico**. Responsável pela modelagem dos eventos prosódicos do sinal de fala. Em sistemas CTF, as manifestações fonético-acústicas mais importantes da prosódia são: o contorno de *pitch* (correspondente perceptivo da freqüência fundamental - F_0 - da fala), o posicionamento e a duração de pausas, a duração das sílabas/segmentos fonéticos e o contorno da intensidade do sinal de fala.
- **Módulo de Seleção Automática de Unidades de Síntese**. Dadas todas as informações provenientes dos módulos de *Front-End* e de modelagem prosódica, o módulo de seleção de

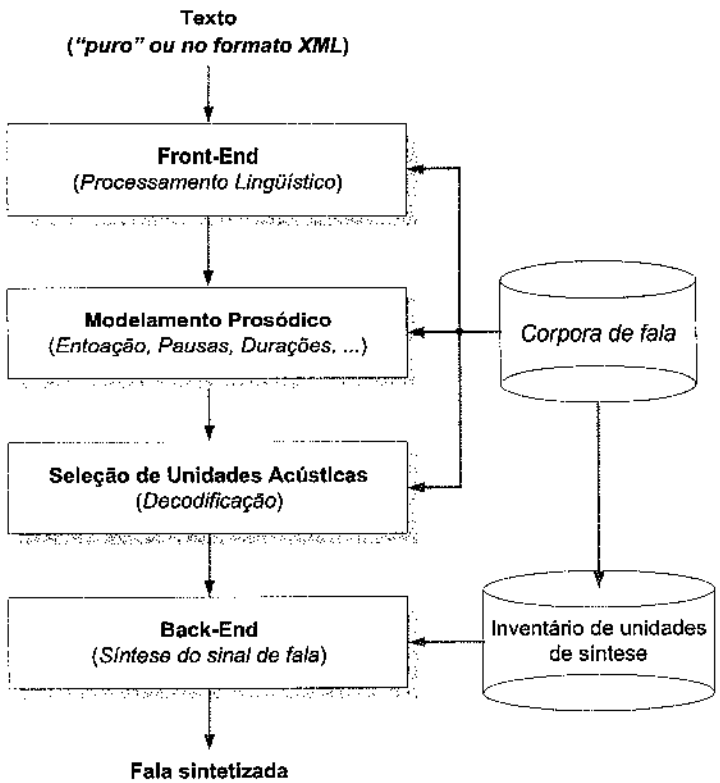


Figura. 1.1: Diagrama de blocos com os principais componentes de um sistema CTF-SCAUS.

unidades de síntese responsabiliza-se por pesquisar o inventário de unidades de síntese e selecionar *on-fly* a melhor seqüência de unidades a ser enviada ao módulo de *Back-End* para síntese do sinal de fala.

- ***Back-End*: Geração de Forma de Onda.** Dadas as unidades de síntese selecionadas pelo módulo de seleção de unidades, o módulo de *Back-End* é responsável por concatená-las e, se for necessário, realizar as devidas suavizações espectrais e modificações prosódicas estimadas pelo módulo prosódico.
- **Corpora de Fala.** Corpora de fala devidamente segmentados e etiquetados têm sido largamente utilizados na modelagem/treinamento de sistemas CTF-SCAUS.

  - O processo de segmentação destes corpora consiste na identificação (ao longo do sinal acústico) das fronteiras entre fones, unidades de síntese, palavras, frases e unidades entoacionais (sílabas, grupos acentuais, grupos entoacionais, etc). Além disso, também são identificadas as fronteiras entre sílabas e palavras com diferentes níveis de acentuação ao longo das sentenças.
  - O processo de etiquetagem destes corpora consiste na atribuição de rótulos a cada um dos itens segmentados. Estes rótulos, geralmente, são de natureza sintática, semântica,

fonológica e fonética.

- **Inventário de Unidades de Síntese.** Todas as unidades de síntese a serem utilizadas, bem como as informações sintáticas, semânticas, fonológicas e prosódicas associadas a cada uma delas podem ser extraídas diretamente do corpus de fala (que se encontra devidamente segmentado e etiquetado). Algumas das principais unidades de síntese empregadas em sistemas CTF-SCAUS são: difones, fones, metades-de-fones e senones. Uma descrição detalhada sobre cada uma destas unidades de síntese será apresentada na seção 4.2.

1.2.2 Síntese em Domínio Irrestrito × Síntese em Domínios Restritos

Sistemas CTF-SCAUS, normalmente, são classificados quanto ao seu domínio de aplicação em: sistemas CTF-SCAUS para Domínio Irrestrito (CTF-SCAUS-DI) e sistemas CTF-SCAUS para Domínio Restrito (CTF-SCAUS-DR).

- Sistemas CTF-SCAUS-DI são projetados e treinados para sintetizar qualquer palavra, sentença ou texto de uma dada língua. Portanto, sistemas CTF-SCAUS-DI empregam extensos inventários de unidades de síntese devidamente projetados para apresentar uma cobertura balanceada de todos os possíveis eventos fonético-prosódicos, que possam ocorrer na língua a ser modelada.
- Sistemas CTF-SCAUS-DR são projetados e treinados para apresentar um excelente desempenho em um determinado domínio específico, como por exemplo, notícias sobre a previsão do tempo ou horóscopo. Por conseguinte, o inventário de unidades de síntese de sistemas CTF-SCAUS-DR deve ser projetado para maximizar a cobertura dos eventos fonético-prosódicos com maior probabilidade de ocorrerem neste domínio específico. Além disso o processo de seleção de unidades de sistemas CTF-SCAUS-DR deve ser otimizado para maximizar a robustez do sistema no domínio especificado.

1.3 Treinamento de Sistemas CTF-SCAUS

Vários dos componentes que fazem parte do estado-da-arte de sistemas CTF-SCAUS são treinados automaticamente a partir de corpora empregando métodos estatísticos. No módulo de *Front-End*, o estado-da-arte dos analisadores sintáticos (*syntatic parsers*) emprega técnicas estatísticas tais como PCFG (*Probabilistic Context Free Grammar*), treinadas a partir do algoritmo *Inside-Outside* (Manning and Schutze, 1999), e *Shallow parsers* treinados com o uso do algoritmo CART (*Classification and Regression Trees*), (Manning and Schutze, 1999). Etiquetadores morfossintáticos - PoS - (*Part-of-Speech Taggers*) e conversores grafema-fonema são normalmente contruídos e treinados empregando-se HMM (*Hidden Markov Models*) (Manning and Schutze, 1999) e/ou CART (Schmid, 1994).

No módulo prosódico, o algoritmo CART tem sido largamente utilizado para predizer o posicionamento e a duração de pausas ao longo das sentenças (Atterer, 2002), (Atterer and Klein, 2002).

Modelos de Soma de Produtos - SoP - (*Sum-of-Product Models*), métodos de Regressão Linear empregando Variáveis Nominais (RLVN) (Morais et al., 2005),(Morais and Violaro, 2005a),(Morais and Violaro, 2005b), Redes Neurais Artificiais (RNA) e árvores de regressão - RT - (*Regression Trees*) (Breiman et al., 1993), têm sido amplamente utilizados na predição da duração de fones ou de sílabas. Em modelos de entoação, como por exemplo os modelos Tilt (Taylor, 2000), algoritmos como HMM, RLVN e CART são, geralmente, utilizados para a modelagem e síntese do contorno de F_0 (freqüência fundamental).

No módulo de seleção de unidades de síntese, técnicas como RLVN, CART ou RNA têm sido largamente utilizadas nas estimativas dos custos prosódico-fonético e de concantenação entre as unidades de síntese (Hunt and Black, 1996), (Donovan, 2003). A clusterização das unidades de síntese, normalmente, é realizada empregando-se o algoritmo CART (Donovan, 2003), (Donovan, 2000). Algoritmos de programação dinâmica como DTW (*Dinamic Time Warpping*) ou algoritmo de Viterbi (Rabiner, 1989) são utilizados no processo de seleção de unidades de síntese.

O processo de seleção das sentenças que devem fazer parte do corpus de fala geralmente utiliza algoritmos gulosos ou algoritmos genéticos (Zhu and Zhang, 2002), para garantir que o corpus de fala seja balanceado fonética e prosodicamente. Além disso, métodos para adaptação de locutor, MAP (*Maximum A Posteriori*) (Rabiner, 1989) e MLLR (*Maximum Likelihood Linear Regression*) (Leggetter and Woodland, 1995) têm sido empregados para adaptar/alterar as características vocais do corpus de fala original, nas características vocais de um outro locutor; empregando para isto apenas um número limitado de sentenças. Este procedimento, comumente, denominado de *Voice Conversion* (Ye and Young, 2003), (Ye and Young, 2004), permite que sistemas CTF-SCAUS sejam capazes de sintetizar novas vozes, a partir deste corpus adaptado/alterado, eliminando, portanto, a necessidade de gravação de novos corpora.

1.4 Problemas com Sistemas CTF-SCAUS

1.4.1 Distribuições LNRE

Os módulos prosódico e de seleção de unidades de síntese de um sistema CTF-SCAUS são treinados a partir dos atributos lingüísticos (sintáticos, semânticos, fonológicos e fonéticos) presentes em um corpus de fala. Estes atributos são simbólicos (não-numéricos) e o produto fatorial (combinatorial) de todos eles define um espaço (domínio) de elevada dimensão. Além disso, por mais extensos e bem projetados que sejam estes corpora, este espaço combinatorial de atributos normalmente apresenta distribuições de probabilidades desbalanceadas, com problemas de falta de dados e, na maioria das vezes, com comportamentos não-gaussianos. Möbius denomina este tipo de distribuição de LNRE (*Large Number of Rare Events*) (Möbius, 2001). Este caráter LNRE destas distribuições é considerado um dos principais obstáculos na obtenção de modelos robustos para os módulos prosódicos e de seleção de unidades de síntese.

Um dos possíveis procedimentos para minimizar problemas com distribuições LNRE é a realização

de análises exploratórias prévias dos dados lingüísticos disponíveis no corpus. Esta análise exploratória deve ser capaz de apontar não somente a melhor técnica de treinamento a ser utilizada (por exemplo, regressão linear, regressão não-linear ou redes neurais), mas também de indicar quais são as topologias ótimas para os modelos a serem treinados, e também propor possíveis clusterizações de dados lingüísticos com o objetivo de minimizar os problemas de desbalanceamento e falta de dados das distribuições LNRE.

1.4.2 Nível de Qualidade Vocal × Nível de Estabilidade Vocal

Apesar de as unidades de síntese apresentarem um excelente *nível de qualidade vocal* (unidades de síntese correspondem a segmentos de um sinal de fala natural), a qualidade vocal da fala sintetizada pela tecnologia SCAUS (especialmente se não forem realizadas modificações prosódicas pelo módulo de *Back-End*) pode, eventualmente, apresentar problemas de falta de homogeneidade no *nível de qualidade vocal* ao longo da fala. Segundo Kagoshima (Mizutani and Kagoshima, 2005), esta falta de homogeneidade é uma conseqüência do *fraco nível de estabilidade vocal* da tecnologia SCAUS. Este *fraco nível de estabilidade vocal* se manifesta principalmente nos sistemas CTF-SCAUS-DI. Os dois principais fatores apontados por Kagoshima, para este *fraco nível de estabilidade vocal* dos sistemas CTF-SCAUS, são:

- Durante o processo de síntese, uma ou mais das unidades de síntese necessárias para gerar o sinal de fala sintetizado com um nível de homogeneidade vocal adequado podem não estar presentes no inventário de unidades de síntese.
- O processo de seleção de unidades de síntese pode cometer erros. Ajustar adequadamente as funções de custo fonético-prosódico e de concatenação utilizadas no processo de seleção de unidades, para que elas apresentem uma elevada correlação com aspectos auditivo-perceptivos, é uma tarefa extremamente difícil (basta lembrar do caráter LNRE das distribuições dos atributos lingüísticos).

Com o objetivo de aumentar o *nível de estabilidade vocal* dos sistemas CTF-SCAUS, Kagoshima desenvolveu, recentemente, um método denominado seleção e fusão de unidades de síntese. Neste método, as unidades de síntese a serem utilizadas na síntese do sinal de fala são versões modificadas, em maior ou menor grau (por meio de processamento digital de sinais), das unidades de síntese originais. Por conseguinte, os *níveis de qualidade vocais* destas unidades de síntese modificadas são, necessariamente, inferiores aos níveis das unidades de síntese originais. A principal novidade/contribuição deste processo de seleção e fusão apresentado por Kagohsima é a flexibilização do controle da relação entre o *nível de estabilidade vocal* dos sistemas CTF-SCAUS e o *nível de qualidade vocal* das unidades de síntese. Em (Mizutani and Kagoshima, 2005), Kagoshima mostra que aumentos significativos no *nível de estabilidade vocal* dos sistemas CTF-SCAUS podem ser obtidos às custas de uma redução relativamente pequena no *nível de qualidade vocal* das unidades de síntese (unidades modificadas pelo processo de fusão). Entretanto, é importante enfatizar que o desempenho do algoritmo de fusão e

seleção proposto por Kagoshima é fortemente dependente da flexibilidade e robustez do algoritmo de *Back-End* utilizado pelo sistema CTF.

1.5 Motivações e Objetivos

Entre as principais motivações para a realização desta Tese de Doutorado destacam-se: (1) A ausência de trabalhos de Mestrado ou Doutorado no Brasil sobre sistemas CTF-SCAUS treinados a partir de extensos corpora de fala. (2) O reduzido número de trabalhos realizados no Brasil sobre modelagem prosódica (no contexto de sistemas CTF-SCAUS). (3) O reduzido número de trabalhos (no Brasil e no exterior) sobre problemas causados a sistemas CTF-SCAUS por distribuições LNRE. (4) O problema, ainda não solucionado, do *fraco nível de estabilidade vocal* dos sistemas CTF-SCAUS. (5) Uma demanda sempre constante, pela comunidade que trabalha com a tecnologia CTF-SCAUS, por novas técnicas de *Back-End* que sejam flexíveis e capazes de realizar suavizações espectrais e modificações prosódicas de alta qualidade. Diante destas motivações, os principais objetivos desta Tese são:

- Apresentar uma ampla revisão bibliográfica sobre os principais componentes de um sistema CTF-SCAUS: módulo de *Front-End*, módulo prosódico, módulo de seleção automática de unidades de síntese, módulo de *Back-End*, corpora de fala e inventário de unidades de síntese.
- Descrever detalhes sobre algumas das principais técnicas estatísticas (técnicas *Data-Driven*) utilizadas no treinamento de sistemas CTF-SCAUS a partir de corpora.
- Introduzir um novo algoritmo para análise exploratória de dados lingüísticos, LDM-GA (*Linguistic Data Mining Using Genetic Algorithm*) e aplicá-lo à modelagem da duração segmental da fala, por meio de técnicas de regressão linear de múltiplas variáveis nominais.
- Introduzir um novo algoritmo de *Back-End* denominado OPWI (*Optimized Prototype Waveform Interpolation*), o qual deve possuir flexibilidade e robustez suficientes para realizar as seguintes operações:
  - Suavizações espectrais de alta qualidade na fronteira entre unidades de síntese;
  - Modificações de alta qualidade na Taxa de Articulação (TSM - *Time Scale Modification*) e na Freqüência Fundamental (PSM - *Pitch Scale Modification*). Estas operações de TSM e PSM são de fundamental importância para a realização de modificações prosódicas, para a clusterização de unidades de síntese e para a estimativa das funções de custo fonético-prosódico do módulo de seleção de unidades de síntese.
  - Fusão de unidades de síntese através do método proposto por Kagoshima (Mizutani and Kagoshima, 2005). Conforme mencionado anteriormente, esta fusão tem com objetivo aumentar o *nível de estabilidade vocal* dos sistemas CTF-SCAUS.

1.6 Estrutura da Tese

O capítulo 2 realiza uma breve revisão sobre o módulo de *Front-End* de sistemas CTF-SCAUS, discutindo aspectos tais como: estrutura de texto, normalização de texto, análise lingüística (análise sintática e semântica), análise fonética, análise morfológica, conversão grafema-fonema e projeto e compressão de léxico. Além disso, dá um destaque especial à construção de etiquetadores morfossintáticos empregando VMM (*Visible Markov Models*) e HMM (*Hidden Markov Models*).

O capítulo 3 apresenta uma breve revisão sobre diferentes aspectos relacionados à modelagem e à geração da prosódia (no contexto de sistemas CTF-SCAUS), tais como: aspectos paralingüísticos (estilo de elocução), aspectos fonológicos (unidades entoacionais/segmentação prosódica) e aspectos prosódicos (modelagem de duração, freqüência fundamental e contorno de intensidade). Apresenta uma revisão sobre as quatro principais classes de modelos entoacionais abordados na literatura: modelos baseados em tons de Pierrehumbert (Silverman et al., 1992), modelos perceptivos (Santen et al., 1997), modelos superposicionais (Santen et al., 1997) e modelos de estilização acústica (Santen et al., 1997). Além disso, apresenta detalhes sobre os modelos entoacionais de Paul Taylor (*Tilt model of intonation*) (Taylor, 2000) e Takehiko Kagoshima (Kagoshima et al., 1998).

O capítulo 4 apresenta uma breve revisão sobre o módulo de seleção automática de unidades de síntese. Realiza uma discussão sobre as principais unidades de síntese utilizadas pela tecnologia SCAUS. Descreve o processo de seleção de unidades em detalhes, dando ênfase à estimativa das funções de custo fonético-prosódico e de concatenação e também ao processo de busca empregando programação dinâmica, DTW (*Dynamic Time Warping*) ou algoritmo de Viterbi. Discute técnicas para clusterização e poda de unidades de síntese, dando ênfase à definição de métricas para medir as distâncias acústico-prosódicas entre unidades de síntese. Detalha técnicas para projeto, segmentação, etiquetagem e compressão de corpora de fala.

O capítulo 5 apresenta os fundamentos teóricos de um novo modelo para análise exploratória de dados lingüísticos denominado LDM-GA (*Linguistic Data Mining using Genetic Algorithm*), o qual pode ser aplicado para minimizar problemas relacionados a distribuições LNRE. Formula o algoritmo LDM-GA sob o contexto da modelagem e predição da duração segmental da fala empregando modelos de regressão linear. Apresenta conceitos e definições sobre o problema de modelagem e predição da duração segmental da fala. Apresenta em detalhes a estrutura e os princípios de operação do algoritmo LDM-GA, dando destaque para os processos de clusterização e seleção automática de topologias ótimas.

O capítulo 6 apresenta resultados e análises dos experimentos sobre a aplicação do algoritmo LDM-GA ao problema de modelagem e predição da duração segmental da fala empregando modelos QMTI. Compara os resultados experimentais com árvores de regressão, RT (*Regression Trees*), e modelos de regressão linear com topologias selecionadas empregando ANOVA (*Analysis of Variance*). Realiza considerações sobre problemas relacionados à esparsidade do espaço de fatores (atributos) lingüísticos. Apresenta resultados e análises sobre o processo de seleção de topologias ótimas e sobre o processo de clusterização de fones. Analisa o desempenho dos modelos QMTI otimizados pelo algoritmo LDM-GA na predição da duração segmental de fones. Apresenta algumas considerações finais e sugestões para

trabalhos futuros.

O capítulo 7 apresenta os fundamentos teóricos do algoritmo OPWI. Apresenta inicialmente uma breve revisão sobre os principais algoritmos de Back-End empregados pela tecnologia SCAUS. Apresenta uma formulação detalhada dos módulos de análise e síntese do algoritmo OPWI. No módulo de análise, destaca as operações de decomposição CEL/CER (*Componente de Evolução Lenta / Componente de Evolução Rápida*), estimativa do nível de estacionaridade e estimativa dos protótipos ótimos. No módulo de síntese, destaca a síntese da componente CEL por interpolação tempo-freqüência dos protótipos e a síntese da componente CER por modelagem autoregressiva. Apresenta detalhes sobre as operações de modificação prosódica e de suavização espectral empregando o algoritmo OPWI. Discute aspectos de implementação capazes de reduzir significativamente o custo computacional, tanto da estimativa dos protótipos ótimos quanto na operação de interpolação tempo-freqüência. Apresenta também algumas considerações finais sobre o algoritmo OPWI.

O capítulo 8 apresenta os resultados e análises dos experimentos realizados com o algoritmo OPWI. Analisa os resultados do processo de decomposição CEL/CER, o processo de estimativa do nível de estacionaridade e a estimativa dos protótipos ótimos. Apresenta resultados e análises sobre diversos tipos de modificações prosódicas. Disponibiliza exemplos audíveis para avaliação do algoritmo. Conclui com algumas considerações finais sobre o desempenho do algoritmo OPWI.

O capítulo 9 realiza algumas considerações finais, conclui a Tese e apresenta sugestões para trabalhos futuros.

Capítulo 2

Front-End: Módulo Lingüístico

2.1 Introdução

A tarefa do módulo de *Front-End* de um sistema de conversão texto-fala (sistema CTF) consiste na transformação do texto de entrada em uma representação lingüística interna. O texto de entrada pode ser um texto "puro" (uma seqüência de caracteres, ASCII ou UNICODE), ou um texto formatado através de *Markup Languages*, por exemplo, XML (Vlist, 2002). A representação lingüística interna consiste em um conjunto de estruturas que representam, por exemplo, as palavras que ocorrem no texto, suas categorias gramaticais e semânticas, análises morfológicas, transcrições fonéticas, segmentações prosódicas e propriedades acentuais. O módulo de *Front-End* é normalmente subdividido, conforme ilustrado na Figura 2.1, em três sub-módulos principais: análise de texto, análise fonética e léxico. O sub-módulo de análise de texto é responsável pela determinação da estrutura do texto (frases, parágrafos, turnos de diálogos, etc), pela conversão de símbolos não-ortográficos/abreviaturas/siglas em palavras e pela análise sintática e semântica do texto. O sub-módulo de análise fonética é responsável por derivar uma transcrição fonética detalhada para o texto a ser processado (eliminando ambigüidades e lidando com efeitos coarticulatórios). O léxico contém várias informações úteis aos sub-módulos de análise de texto e análise prosódica, tais como possíveis classes morfossintáticas das palavras e regras para expansão e conversão grafema-fonema de abreviaturas e siglas.

O restante deste capítulo é composto por quatro seções. A seção 2.2 descreve as principais operações envolvidas no sub-módulo de análise de texto, com destaque para os etiquetadores morfossintáticos. A seção 2.3 apresenta o sub-módulo de análise fonética. A seção 2.4 realiza algumas considerações sobre o léxico e a seção 2.5 conclui este capítulo com algumas considerações finais.

2.2 Análise do Texto

Este sub-módulo é responsável por explicitar todas as informações lingüísticas sobre o texto a ser processado, que não sejam de caráter fonético ou prosódico. Implementações simplificadas deste sub-módulo limitam-se a converter itens não-ortográficos (números, siglas ou abreviaturas) em palavras.

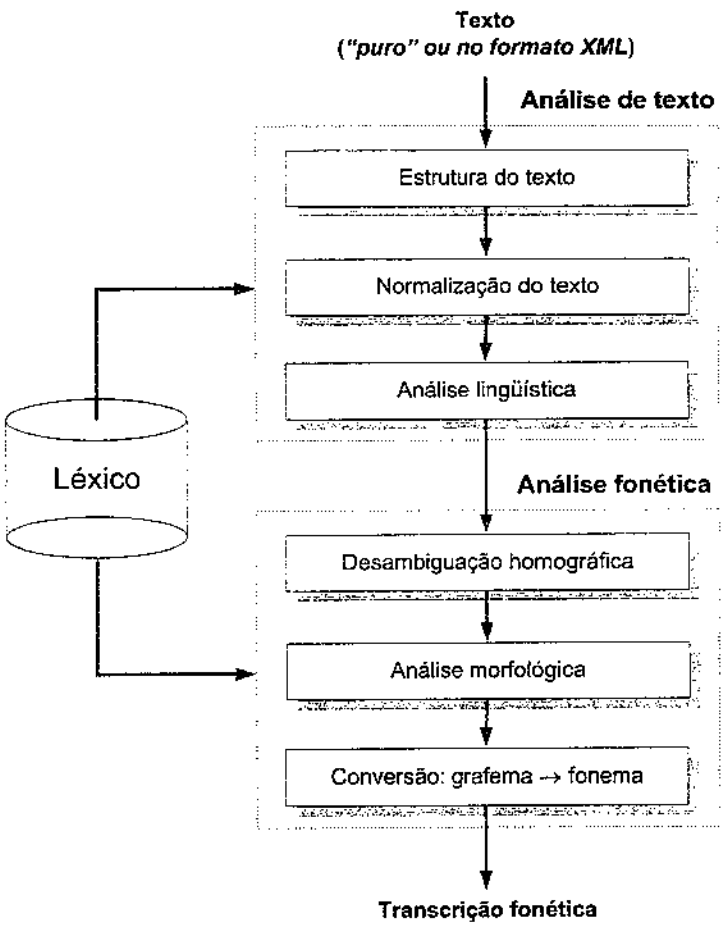


Figura. 2.1: Diagrama de blocos das principais operações do módulo de *Front-End*.

Por outro lado, implementações mais complexas, além de lidarem com itens não-ortográficos, realizam elaboradas análises sobre a estrutura do texto (para identificação de parágrafos, sentenças, frases e turnos de diálogos), bem como sofisticadas análises sintáticas e semânticas das frases/sentenças (Sproat, 1998), (Huang et al., 2001), (Dutoit, 1997). Todas estas análises visam a determinação de atributos lingüísticos necessários para que o sub-módulo de análise fonética realize conversões grafema-fonema detalhadas e também para que o módulo prosódico possa gerar modelos duracionais (predição de pausas e duração de sílabas/fones) e entoacionais elaborados. A seguir é apresentada uma descrição dos três componentes envolvidos no sub-módulo de análise de texto.

2.2.1 Determinação da Estrutura do Texto

Este sub-módulo é responsável por determinar estruturas do texto, tais como fronteiras de parágrafos, sentenças, frases e turnos de diálogos. A determinação destas fronteiras é de fundamental importância para o desempenho do módulo prosódico. Entretanto, se o texto a ser processado for

um documento formatado segundo uma *Markup Language* como XML (*Extended Markup Language*), então várias informações sobre a estrutura do texto já poderão estar disponíveis. Por exemplo, um livro formatado em XML poderá conter informações sobre a localização de títulos, capítulos, seções, referência, notas de rodapé, parágrafos, etc. Esta estruturação permitirá, por exemplo, que diferentes partes do texto possam ser convertidas em fala com diferentes entoações ou estilos de elocução.

2.2.2 Normalização do Texto

Sistemas CTF podem ter que lidar com textos que incluem abreviaturas (Apto. - Apartamento) e siglas (DF - Distrito Federal); com manuais técnicos que podem conter gráficos e tabelas seguidos de números; com *e-mails* que geralmente apresentam vocabulários próprios (quanto -qto-, você -vc-, não -naun-, abraço -abc-, é -eh-) e sinais de emoção (*emoticons*); e, também, com endereços de Internet, que podem apresentar vários símbolos especiais. Para poder lidar com estas abreviaturas, siglas, números, expressões de e-mail e símbolos não-ortográficos, todo sistema CTF deve possuir um sub-módulo para normalização de texto. A seguir são apresentados alguns exemplos de normalização de texto:

Abreviaturas e Siglas

- Dr. - Doutor;
- Sr. - Senhor;
- kHz - Quilohertz;
- cm - Centímetro;
- SP - São Paulo;
- Unicamp - Universidade Estadual de Campinas.

Datas e Horas

- 14/12/1967 - Quatorze de dezembro de mil novecentos e sessenta e sete;
- 14/12/1967 - Quatorze do doze de mil novecentos e sessenta e sete;
- 11:45 - Onze horas e quarenta e cinco minutos;
- 11:45 - Quinze para as doze;
- 11:45 - Quinze para o meio-dia;
- 02:30 - Duas e trinta;
- 02:30 - Duas e meia.

Número de Telefone

- 015 61 33415897 - zero-quinze, meia-um, três-três, quatro-um, cinco-oito, nove-sete;
- 015 61 33415897 - zero-quinze, sessenta-e-um, trinta-e-três, quarenta-e-um, cinqüenta-e-oito, noventa-e-sete.

2.2.3 Análise Lingüística

Este sub-módulo é responsável pela estimação das classes morfossintáticas (PoS - *Part-of-Speech*) de cada palavra e também pela estimação/análise (*parsing*) das estruturas sintáticas e semânticas de cada frase do texto. Tais informações são de extrema importância para o sub-módulo de análise fonética e também para os subseqüentes módulos prosódicos e de seleção automática de unidades de síntese.

O *estado-da-arte* dos etiquetadores morfossintáticos (PoS) e dos analisadores (*parsing*) sintáticos e semânticos é dominado por métodos estatísticos baseados em corpora. Quando treinados com uma elevada quantidade de dados, estes etiquetadores e *parsers* são capazes de atingir excelentes desempenhos (Schmid, 1994), (Manning and Schutze, 1999).

No contexto de sistema CTF a tarefa de um *parser* sintático consiste não apenas em estimar as possíveis estruturas sintáticas de uma frase, mas também de indicar (se houver mais de uma estrutura correta, como no caso da Figura 2.2), qual delas é a mais apropriada para os módulos subseqüentes do sistema, módulo prosódico e módulo de seleção automática de undidades de síntese.

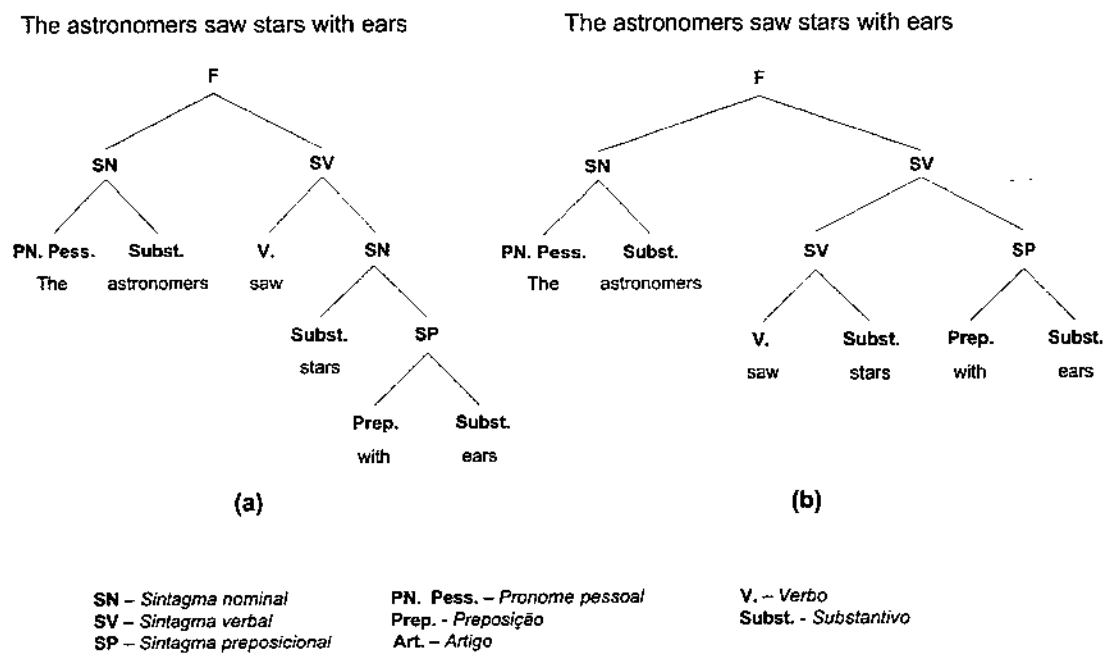


Figura. 2.2: Análise sintática da frase em inglês "*The astronomers saw stars with ears*". As Figuras (a) e (b) mostram duas possíveis análises sintáticas desta mesma sentença.

A Figura 2.3 ilustra o processo de etiquetagem morfossintática da sentença em francês "*J' entre par la porte*". Segundo a Figura 2.3, o processo de etiquetagem morfossintática pode ser interpretado como um procedimento de busca ao longo de uma rede de estados-finitos (sendo cada estado desta rede representado por uma das possíveis classes morfossintáticas).

Algumas das técnicas mais comumente utilizadas na construção de etiquetadores morfossintáticos

são VMM (*Visible Markov Models*), HMM (*Hidden Markov Models*) e CAT (*Classification Trees*) (Kepler, 2005). A construção de etiquetadores morfossintáticos através das técnicas VMM e CAT assume a existência de uma base de dados previamente segmentada e classificada morfossintaticamente. Por esta razão, VMM e CAT são denominadas técnicas supervisionadas. Por outro lado, a técnica HMM permite a construção de etiquetadores morfossintáticos a partir de bases de dados sem qualquer segmentação e classificação morfossintática prévia. Por conseguinte, a técnica HMM é denominada não-supervisionada. A seguir, detalhes serão apresentados sobre os processos de modelagem/operação de etiquetadores morfossintático baseados em VMM e HMM.

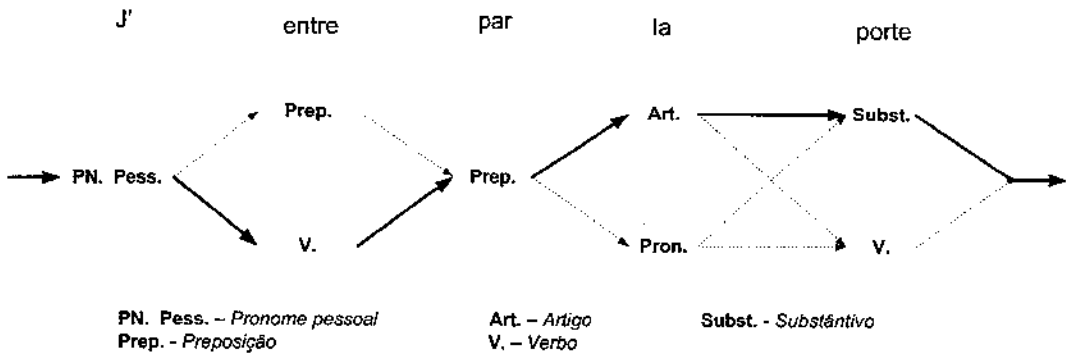


Figura 2.3: Exemplo de um etiquetador morfossintático (*Part-of-Speech Tagger*) para a frase em francês, "*J' entre par la porte*". A linha contínua (em negrito) indica as etiquetas selecionadas, (PN-Pess., V., Prep., Art., Subst.).

Etiquetagem Morfossintática empregando VMM

Várias abordagens determinísticas e estatísticas têm sido propostas para o problema de etiquetagem morfossintática, PoS (*Part-of-Speech Tagging*). Uma das abordagens estatísticas mais largamente utilizadas para estimar PoS emprega Cadeias de Markov Visíveis VMM (*Visible Markov Models*), as quais são treinadas de maneira supervisionada a partir de corpora devidamente etiquetados morfossintaticamente (Manning and Schutze, 1999). No contexto de cadeias de Markov visíveis, o problema de PoS apresenta a seguinte formulação: dada uma sentença composta por N palavras $\mathbf{W} = (w_1, w_2, \cdots, w_N)$ e o conjunto de todas as possíveis seqüências de etiquetas morfossintáticas, $\mathbf{T} = (t_1, t_2, \cdots, t_N)$ de duração N, com $t_i \in C_T = \{c_1, c_2, \cdots, c_M\}$ (sendo C_T o conjunto de todas as M possíveis classes morfossintáticas, verbo, substantivo, adjetivo, etc), então o problema de *Part-of-Speech* consiste em determinar a seqüência de etiquetas $\hat{\mathbf{T}} \in \mathbf{T}$ que "morfossintaticamente" melhor representa a sentença $\mathbf{W}$. Probabilisticamente esta formulação pode ser escrita como:

$$\hat{\mathbf{T}} = \underset{\mathbf{T}}{ArgMax}\, P(\mathbf{T}|\mathbf{W}) \tag{2.1}$$

Segundo a regra de *Bayes* a equação 2.1 pode ser reescrita como:

$$\hat{\mathbf{T}} = \underset{\mathbf{T}}{ArgMax}\, \frac{P(\mathbf{T},\mathbf{W})}{P(\mathbf{W})} = \underset{\mathbf{T}}{ArgMax}\, \frac{P(\mathbf{W}|\mathbf{T}) \cdot P(\mathbf{T})}{P(\mathbf{W})} \tag{2.2}$$

O denominador em 2.2 é independente de $\mathbf{T}$ e pode ser ignorado durante o processo de busca pela seqüência $\hat{\mathbf{T}}$, portanto:

$$\hat{\mathbf{T}} = \underset{\mathbf{T}}{ArgMax}\, P(\mathbf{W}|\mathbf{T}) \cdot P(\mathbf{T}) \tag{2.3}$$

Os modelos de Markov visíveis aplicados à estimação de PoS fazem uso de duas suposições significativamente fortes:

- A probabilidade de uma palavra w_i está condicionada apenas a sua respectiva etiqueta morfossintática t_i;
- A probabilidade de uma etiqueta morfossintática t_i está condicionada apenas às n etiquetas anteriores a ela $t_{i-1}, t_{i-2}, \cdots, t_{i-n}$

Como resultado destas duas suposições,

$$P(\mathbf{W}|\mathbf{T}) = P(w_1, w_2, \cdots, w_N|t_1, t_2, , t_N) = \prod_{i=1}^{N} P(w_i|t_i) \tag{2.4}$$

$$P(\mathbf{T}) = P(t_1, t_2, \cdots, t_N) = \prod_{i=1}^{N} P(t_i|t_{i-1}, t_{i-2}, \cdots, t_{i-n}) \tag{2.5}$$

Substituindo 2.4 e 2.5 na equação 2.3, tem-se:

$$\hat{\mathbf{T}} = \underset{T}{ArgMax} \prod_{i=1}^{N} P(w_i|t_i) \cdot P(t_i|t_{i-1}, \cdots, t_{i-n}) \tag{2.6}$$

Portanto, o problema de etiquetagem morfossintática (PoS) consiste na busca pela seqüência de etiquetas $\hat{\mathbf{T}}$ que maximiza a equação 2.6. Este problema de busca é geralmente solucionado utilizando-se técnicas de programação dinâmica, como por exemplo o algoritmo de Viterbi.

Na prática o valor de n na equação 2.5 é normalmente limitado a 2. Com isto, reduz-se o termo $P(t_i|t_{i-1}, \cdots, t_{i-n})$ a uma gramática *tri-gram* $P(t_i|t_{i-1}, t_{i-2})$ (Manning and Schutze, 1999), (Dutoit, 1997). Além disso, com o objetivo de lidar com problemas de esparsidade de dados, o termo $P(t_i|t_{i-1}, t_{i-2})$ é submetido a um processo de suavização por meio de interpolação linear:

$$P_I\left(t_i|t_{i-1},t_{i-2}\right) = \lambda_1 \cdot P\left(t_i\right) + \lambda_2 \cdot P\left(t_i|t_{i-1}\right) + \lambda_3 \cdot P\left(t_i|t_{i-1},t_{i-2}\right) \tag{2.7}$$

Uma solução elegante para a estimativa dos valores de λ_1, λ_2 e λ_3 pode ser encontrada em (Jelinek, 1997).

Substituindo 2.7 em 2.6, tem-se:

$$\hat{\mathbf{T}} = \underset{T}{ArgMax} \prod_{i=1}^{N} P\left(w_i|t_i\right) \cdot P_I\left(t_i|t_{i-1},t_{i-2}\right) \tag{2.8}$$

Como está sendo assumida a existência de um corpora devidamente etiquetado, então os parâmetros da equação 2.8 podem ser estimados, segundo o método da maximização da verossimilhança, através das equações 2.9, 2.10, 2.11 e 2.12,

$$P\left(w_i|t_j\right) = \frac{C\left(w_i,t_j\right)}{C\left(t_j\right)} \tag{2.9}$$

$$P\left(t_i\right)) = \frac{C\left(t_i\right)}{NTags} \tag{2.10}$$

$$P\left(t_i|t_j\right) = \frac{C\left(t_i,t_j\right)}{C\left(t_j\right)} \tag{2.11}$$

$$P\left(t_i|t_j,t_k\right) = \frac{C\left(t_i,t_j,t_k\right)}{C\left(t_j,t_k\right)} \tag{2.12}$$

sendo que $C(w_i,t_j)$ representa o número de vezes que a palavra w_i ocorre conjuntamente com a etiqueta t_j. $C(t_i)$, $C(t_i,t_j)$ e $C(t_i,t_j,t_k)$ indicam, respectivamente, o número de vezes que as seqüências de etiquetas t_i, (t_i,t_j) e (t_i,t_j,t_k) ocorrem no corpus. $NTags$ representa o número total de etiquetas morfossintáticas existentes no corpus.

Etiquetagem Morfossintática empregando HMM

O método de etiquetagem morfossintática empregando VMM assume a existência de uma base de dados devidamente etiquetada. Entretanto, estas bases de dados nem sempre se encontram disponíveis e, portanto, técnicas de etiquetagem não-supervisionada (que possam ser treinadas a partir de corpora sem etiquetagem prévia), são de extrema importância. Uma das técnicas que tem sido largamente

utilizada na etiquetagem morfossintática não-supervisionada é a dos Modelos Ocultos de Markov, HMM (*Hidden Markov Models*). A formulação probabilística da etiquetagem morfossintática utilizando HMM é semelhante ao do método VMM. A única diferença reside na estimativa dos parâmetros que modelam a probabilidade $P(\mathbf{W}|\mathbf{T})$ (presente na equação 2.3) os quais, em HMM, são estimados iterativamente utilizando-se o algoritmo *Baum-Welch* (Manning and Schutze, 1999), (Rabiner, 1989).

2.3 Análise Fonética

O objetivo principal da análise fonética é converter símbolos ortográficos em sua correspondente representação fonética. Além disso, também cabe ao módulo de análise fonética estimar/produzir outras informações como separação silábica e sílabas acentuadas. Este processo de análise fonética é considerado relativamente simples para línguas como, por exemplo, português, espanhol e finlandês. Para estas línguas é possível estimar um conjunto de regras lingüísticas *ad hoc* capaz de lidar, com precisão, com mais de 90% das análises fonéticas. O mesmo não acontece, por exemplo, com a língua inglesa, principalmente por ser morfologicamente mais complexa e também por possuir palavras oriundas de várias outras línguas. Algumas das três operações mais utilizadas em um bom sistema para análise fonética são: desambiguação homográfica, análise morfológica e conversão grafema-fonema.

2.3.1 Desambiguação Homográfica

Um bom conversor grafema-fonema requer que palavras com grafias idênticas, porém com pronúnicas distintas, sejam desambiguadas. Por exemplo: sede (verbo) e sede (substantivo). A maioria das desambiguações homográficas podem ser resolvidas utilizando-se apenas as etiquetas morfossintáticas, entretanto, em alguns casos, análises semânticas e de discurso também são necessárias.

2.3.2 Análise Morfológica

Este módulo analisa relações entre ortografia e pronúncia através da análise de componentes morfológicos tais como prefixos, sufixos e radicais das palavras. A análise destes componentes morfológicos providencia importantes informações para se obter pronúncias precisas para palavras flexionadas e/ou palavras compostas.

2.3.3 Conversão Grafema-Fonema

O último estágio do módulo de análise fonética, normalmente, inclui regras (lingüísticas do tipo *ad hoc* ou derivadas através de métodos estatísticos), para conversão de grafemas em fonemas e um dicionário de pronúncias para as palavras que sejam exceções às regras de conversão. As técnicas estatísticas mais comumente utilizadas para realizar conversões grafema-fonema são HMM (*Hidden Markov Models*) e CART (*Classification and Regression Trees*). É importante enfatizar que este módulo deve realizar uma conversão grafema-fonema detalhada do texto a ser convertido em fala, contemplando, inclusive, variações fonéticas decorrentes da coarticulação entre palavras.

Este processo de conversão grafema-fonema é geralmente realizado em duas etapas. A primeira consiste na derivação de uma transcrição fonética lexical (transcrição fonética canônica das palavras). A segunda consiste na modificação desta transcrição lexical para levar em consideração possíveis variações fonéticas decorrentes dos processos coarticulatórios. Por exemplo, no português brasileiro, o fone /s/, que em uma transcrição lexical normalmente seria fricativo não-sonoro, deve ser modificado para fricativo sonoro quando diante de palavras iniciadas por vogais ou consoantes sonoras ("*muitas emoções*", "*vários dias*").

2.4 Léxico

O léxico consiste em uma das fontes de informação mais importantes para os módulos de análise de texto e análise fonética. O léxico é também referido como dicionário e, em geral, também é de fundamental importância para os módulos prosódico e de seleção de unidades de síntese. Um bom léxico deve conter, entre outras, as seguintes informações:

- Formas flexionadas das unidades lexicais;
- Transcrição fonética (incluindo múltiplas pronúncias) para todas as unidades do léxico;
- Separação silábica e identificação de sílabas acentuadas para todas as unidades do léxico;
- Informações para análises morfológicas. Identificação de radicais, prefixos, sufixos, etc;
- Informações para expansão e conversão grafema-fonema de abreviaturas e siglas;
- Informações para gerar pronúncias adequadas para nomes próprios e palavras estrangeiras;
- Pronúncia de todos os caracteres permitidos. Isto garantirá que o sistema seja capaz de soletrar qualquer caracter;
- *Part-of-Speech* (PoS) para todas as unidades do léxico;
- Identificação de atributos sintáticos e semânticos para todas as unidades do léxico.

2.4.1 Compressão do Léxico

Duas das principais vantagens do uso de um léxico extenso e elaborado são:

- Pode garantir uma maior qualidade ao conversor texto-fala. Isto acontece porque todas as informações presentes no léxico (transcrição fonética, derivações morfológicas, etc) podem ser previamente analisadas para confirmar a sua precisão.
- Pode representar uma redução significativa no tempo de processamento do conversor texto-fala, dado que várias das informações que a princípio deveriam ser estimadas *on-fly*, a partir do texto, já se encontram explicitadas no léxico.

Por outro lado, a grande desvantagem de um léxico extenso é, obviamente, o aumento no *footprint* (tamanho) do sistema. A utilização de sistemas CTF em ambientes embarcados (de reduzido *footprint*), quase sempre exige que eficientes técnicas de compressão sejam aplicadas ao léxico. A seguir, é

apresentado um algoritmo relativamente simplicado para a compressão de léxicos voltados a sistemas CTF.

- Se uma unidade lexical possui uma freqüência de ocorrência extremamente elevada, então esta unidade lexical, juntamente com todos os seus campos (transcrição fonética, PoS, sílabas, acentuação, classe semântica,...), deve permanecer no léxico.
- Se uma unidade lexical não possuir uma freqüência de ocorrência extremamente elevada e seus campos (transcrição fonética, PoS, sílabas, acentuação, classe semântica,...) puderem ser perfeitamente preditos a um baixo custo computacional, a partir do texto de entrada, então todos os campos desta unidade devem ser eliminados do léxico. Entretanto, se algum campo desta unidade lexical não puder ser predito a partir do texto de entrada e/ou qualquer outro campo associado a esta unidade lexical, então este campo, bem como sua palavra lexical, deverá permanecer no léxico.
- Todas unidades lexicais remanescentes (que não puderem ser removidas), devem ser submetidas a algum tipo de compactação, como por exemplo empregando códigos de *Huffman* (Cover and Thomas, 1991).

2.5 Considerações Finais

Este capítulo apresentou uma breve revisão sobre as principais operações envolvidas no módulo de *Front-End*, de sistemas CTF-SCAUS. Por se tratar de uma breve revisão algumas operações importantes foram apenas mencionadas e merecem ser tratadas em mais detalhes em trabalhos futuros. Entre essas operações destacam-se:

- O uso de *Voice Markup Languages* para a formatação do texto de entrada (texto a ser convertido em fala), permitindo a utilização de informações adicionais àquelas que podem ser extraídas diretamente do texto por meio de análises lingüísticas.
- A construção de analisadores (*parses*) sintáticos e semânticos (elaborados) empregando métodos probabilísticos.

Capítulo 3

Módulo Prosódico

3.1 Introdução

No contexto de sistemas para conversão texto-fala (sistemas CTF), o termo prosódia, geralmente, refere-se a variações audíveis na freqüencia fundamental (F_0), na intensidade, na duração de sílabas ou segmentos fonéticos e no posicionamento e duração de pausas ao longo da fala. Certos autores também atribuem à prosódia aspectos relacionados à estrutura rítmica e à taxa de elocução da fala. Análises mais detalhadas sobre ritmo da fala e taxa de elocução podem ser encontradas em (Barbosa, 1994), (Barbosa, 2006), (Barbosa, 2002) (Keller, Forthcoming).

A entoação, representada pela variação sistemática do contorno de F_0, é considerada o fenômeno prosódico mais importante em sistemas CTF. Em sistemas CTF a modelagem do contorno de F_0 tem sido utilizada para caracterizar uma sentença como declarativa, interrogativa ou exclamativa; para direcionar (focalizar) a atenção do ouvinte para aspectos específicos da mensagem que está sendo dita; e, mais recentemente, para expressar sentimentos e emoções (Bailly et al., 2003). Se uma sentença é pronunciada com um contorno de F_0 constante, sem pausas ou com pausas uniformes entre palavras, soará extremamente não-natural.

Esta Tese irá se limitar a descrever apenas três dos fenômenos prosódicos considerados mais importantes em sistemas CTF: contorno de F_0, duração de segmentos fonéticos e localização/duração de pausas. Apesar de reconhecer a inter-relação existente entre estes fenômentos prosódicos, todas as discussões sobre contorno de F_0, pausas e durações segmentais, realizadas nesta Tese, assumirão, por uma questão de simplicidade, a independência entre estes fenômenos.

O restante deste capítulo é dividido em quatro seções. A seção 3.2 apresenta os principais componentes do módulo prosódico, desde aspectos paralingüísticos e fonológicos até aspectos fonéticos relacionados à modelagem prosódica. A seção 3.3 descreve as quatro classes principais de modelos de entoação: modelos baseados em tons, modelos perceptivos, modelos superposicionais e modelos de estilização acústica. A seção 3.4 apresenta o modelo entoacional de Paul Taylor (modelo Tilt). Detalhes sobre o modelo entoacional de Takeiko Kagoshima são apresentados na seção 3.5. A seção 3.6 conclui o capítulo com algumas discussões finais.

3.2 Módulo Prosódico

A Figura 3.1 mostra um diagrama de blocos com os principais módulos envolvidos na modelagem prosódica de sistemas CTF. Este diagrama inclui desde aspectos pragmáticos e fonológicos até a realização fonético-acústica dos eventos prosódicos. A entrada do módulo prosódico descrito na Figura 3.1 consiste no texto analisado lingüisticamente (informações lingüísticas extraídas do sub-módulo de análise de texto) e a sua respectiva transcrição fonética (acrescida de outras informações, como separação silábica e acento lexical).

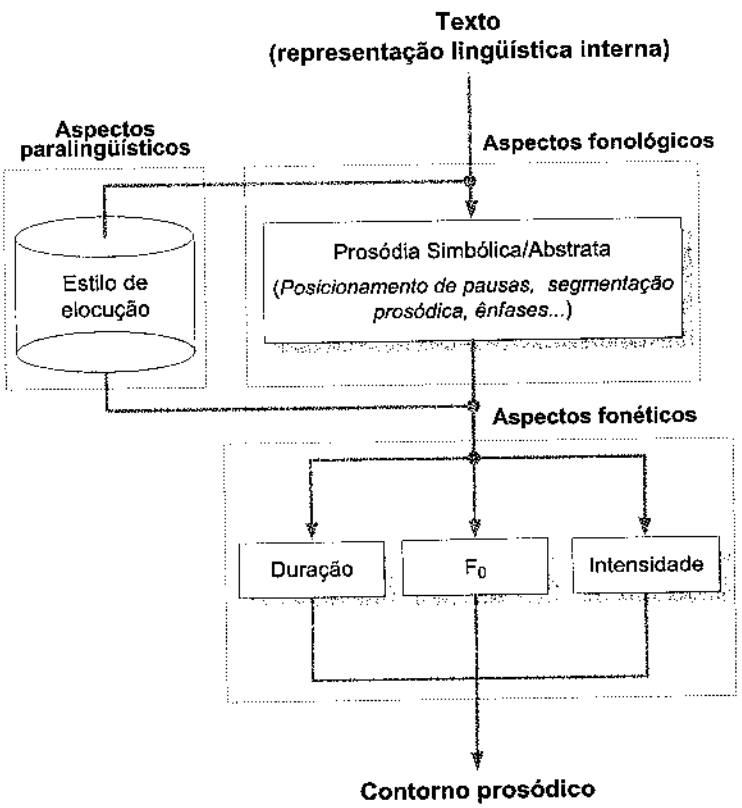


Figura. 3.1: Diagrama de blocos das principais operações do módulo prosódico.

A Figura 3.1 considera que a modelagem e a geração prosódica se manifestam em três áreas lingüísticas bem delineadas:

- Paralingüística. Principalmente através de aspectos relacionados ao estilo de elocução.
- Fonologia (prosódia simbólica/abstrata). Por meio da segmentação prosódica (identificação de unidades entoacionais), caracterização de fronteiras prosódicas (coesão forte/fraca entre palavras), localização de pausas silenciosas e atribuições de ênfase (foco) às unidades entoacionais.
- Fonética. Através da realização física de aspectos prosódicos, tais como: duração segmental dos fones/sílabas e pausas silenciosas, contorno da freqüência fundamental (F_0) e evolução temporal

da intensidade da fala.

As subseções 3.2.1, 3.2.2 e 3.2.3 apresentam uma breve discussão sobre cada um destes itens.

3.2.1 Aspectos Paralingüísticos: Estilo de Elocução

Por mais completas que sejam as informações lingüísticas extraídas de uma sentença, elas geralmente não serão suficientes para explicar, em definitivo, seus aspectos prosódicos. Diferentes pessoas podem gerar diferentes realizações prosódicas para uma mesma sentença. Além disso, um mesmo locutor também pode gerar diferentes contornos prosódicos para uma mesma sentença, dependendo, por exemplo, do seu estado de humor ou de saúde. Entre os aspectos paralingüísticos mais comumente estudados no contexto de sistemas CTF, destaca-se o estilo de elocução. Alguns dos principais efeitos prosódicos associados ao estilo de elocução são:

- Qualidade vocal (falseto, voz crepitante, voz soprosa, voz rouca, etc);
- Taxa de elocução. Por exemplo, uma alta taxa de elocução pode indicar que o locutor esteja muito excitado;
- Fala sub ou sobre-articulada. Muitas vezes utilizada por locutores profissionais;
- Gama tonal. Indica a faixa de variação de F_0 ao longo da fala (Huang et al., 2001). Por exemplo, uma reduzida excursão do contorno de F_0 pode indicar que o locutor (sem patologia) esteja aborrecido, deprimido ou controlando um sentimento de raiva.

A emoção é outro aspecto importante relacionado com estilo de elocução. Além disso, o estudo da emoção em sistemas CTF é um tema emergente na área de ciência e tecnologia da fala (Huang et al., 2001). Algumas emoções simples que têm sido estudadas no contexto de sistemas CTF são: raiva/cólera, alegria e tristeza (Bailly et al., 2003).

3.2.2 Aspectos Fonológicos: Prosódia Simbólica/Abstrata

A prosódia simbólica/abstrata é a responsável pela ligação entre as características pragmáticas, semânticas e sintáticas de uma sentença (ou texto) com seus respectivos contornos de F_0, duração segmental e intensidade. Algumas das principais operações do sub-módulo de prosódia simbólica são:

- Segmentação prosódica: identificação de unidades entoacionais.
- Caracterização das fronteiras prosódicas. Identificação do nível de coesão forte/fraca entre palavras que se encontram nas fronteiras prosódicas.
- Localização de pausas silenciosas e atribuições de ênfase (foco) a sílabas ao longo das unidades entoacionais.

Apesar de não haver um consenso entre os estudiosos de prosódia, algumas das unidades prosódicas mais comumente utilizadas na modelagem do contorno de F_0, na duração segmental dos fones e na

localização/duração de pausas silenciosas são: a sílaba, o grupo acentual, o grupo entoacional e outras unidades de ordem superior. A terminologia utilizada neste trabalho, para se referir às unidades entoacionais, é a mesma adotada por (López, 1993) e (Mancebo, 2002), em estudos sobre entoação para o espanhol europeu. Apesar de todos os autores reconhecerem a existência destas unidades, ainda não existe um consenso na hora de estabelecer qual delas é a mais indicada para descrever este ou aquele fenômeno prosódico.

Unidades Entoacionais

- **A sílaba.** Possui um papel importante na prosódia, uma vez que ela está associada ao acento lexical e/ou frasal. O acento se manisfesta como a proeminência de uma sílaba específica, sílaba acentuada, frente a outras sílabas da mesma palavra ou grupo acentual. Esta proeminência geralmente se manifesta através de movimentos significativos em F_0, em combinação com variações de outras propriedades como a duração, a intensidade e a qualidade vocal.
- **O grupo acentual.** Também referido como *grupo tônico*, *grupo rítmico-semântico*, *grupo rítmico* ou *stress group*, apresenta, entre outras, as seguintes definições:
  - Conjunto delimitado por uma palavra acentuada e todas as palavras não-acentuadas que a precedem,
  - Conjunto dado por uma sílaba acentuada seguida de qualquer sílaba não-acentuada até a sílaba acentuada seguinte (Santen, 2002).
- **O grupo entoacional.** Também referido em espanhol como *unidad melódica*, *grupo melódico*, *grupo fónico* ou em inglês como *intonational phrase, intonational unit, prosodic sentence* ou *breath group*, é definido como uma estrutura entoacional coerente, que não inclui nenhuma "ruptura prosódica importante" (*major prosodic break*). As rupturas prosódicas são as fronteiras que delimitam os grupos entoacionais e são, geralmente, representadas por pausas e inflexões significativas no contorno de F_0. Um grupo entoacional pode ser constituído por uma única sílaba, um sintagma (nominal, verbal...), ou uma frase inteira.

  Estudos têm indicado (porém, de forma ainda não conclusiva) que não existe uma relação direta entre grupos entoacionais e unidades sintáticas. Por esta razão Botinis (Botinis et al., 2001) afirma que "*a correspondência entre grupos de entoação e sintaxe é muito mais casual do que causal*".

  Apesar de os grupos acentuais e das sílabas poderem ser identificados de forma determinística ao longo de um texto, estimar os grupos entoacionais é um problema muito mais complexo. Dada uma locução, somente será possível determinar precisamente onde estão localizadas as fronteiras entre os grupos entoacionais (segundo a definição de grupo entoacional apresentada nesta Tese), se for realizada uma análise cuidadosa (e na maioria das vezes, complexa) do contorno F_0 desta locução.

- **Unidades Superiores.** A entoação possui um papel importante na organização do discurso e do diálogo. Este fato é conseqüência de estruturas prosódicas que superam o nível da frase e que se estendam ao longo de turnos de diálogo ou de parágrafos. Apesar destas unidades superiores de entoação serem de grande importância para a naturalidade de sistemas CTF, elas ainda não se encontram devidamente caracterizadas e formalizadas (no contexto de sistemas CTF), constituindo, portanto, uma área de pesquisa ainda a ser explorada.

3.2.3 Aspectos Prosódicos: Realização Acústica

Modelagem da Entoação

Segundo uma classificação dada por Botinis et al (Botinis et al., 2001), os modelos de entoação se distinguem em: modelos baseados em tons, modelos perceptivos, modelos superposicionais e modelos de estilização acústica. A seção 3.3 apresenta uma breve revisão sobre estes modelos entoacionais. Em seguida as seções 3.4 e 3.5 apresentam detalhes sobre dois modelos entoacionais de estilização acústica largamente utilizados em sistemas CTF, o modelo Tilt de Paul Taylor (Taylor, 2000) e o modelo de Takeiko Kagoshima (Kagoshima et al., 1998).

Modelagem da Estrutura Duracional da Fala

O objetivo da modelagem da duração é estimar a estrutura duracional da fala a partir de atributos simbólicos de ordem pragmática, semântica, sintática, fonológica e prosódica, derivados no módulo lingüístico e nos sub-módulos de estilo de elocução (pragmática) e prosódia simbólica (fonologia entoacional). Alguns dos fatores que, geralmente, dificultam a estimação da estrutura duracional da fala para sistemas CTF são:

- Elevado número de atributos lingüísticos, bem como complexas interações entre eles;
- Elevado número de possíveis combinações entre os atributos lingüísticos;
- Ausência de teorias lingüísticas elaboradas que sejam possíveis de serem implementadas computacionalmente e que tenham sido devidamente avaliadas e testadas no contexto de sistemas CTF.

Apesar de existirem várias abordagens para a modelagem da estrutura duracional da fala (Morais and Violaro, 2005b), (Möbius and Santen, 1996), (Barbosa, 1994), esta Tese se limitará a analisar apenas a modelagem da duração segmental de fones. Várias técnicas têm sido apresentadas para a predição da duração segmental de fones. Algumas das mais usuais são: Árvores de Regressão, RT (*Regression Trees*) (Breiman et al., 1993), Soma de Produtos, SoP (*Sum-of-Products*) (Sproat, 1998) e modelos de regressão linear de múltiplas variáveis nominais (por exemplo, QMTI - *Quantification Method Type I*) (Morais et al., 2005).

A seção 5.2 apresenta uma descrição detalhada sobre o problema de predição da duração segmental da fala empregando modelos de regressão linear de múltiplas variáveis nominais.

Modelagem do Contorno de Intensidade

A intensidade é um parâmetro proporcional à energia do sinal de fala. Intuitivamente, poder-se-ia acreditar que a intensidade é um parâmetro prosódico tão importante quanto a duração e a entoação. Entretanto, na prática isto não se confirma. Durante vários anos acreditou-se que os segmentos fonéticos acentuados se destacavam dos demais por meio de um padrão de energia mais elevado; no entanto, o que se observou é que os acentos (ao longo das sentenças) são caracterizados principalmente por variações de duração e de F_0. Obviamente, a importância relativa de cada um destes parâmetros (duração, F_0 e intensidade) pode apresentar variação (de maior ou menor grau) de língua para língua.

Devido a esta importância "secundária" do contorno de intensidade na prosódia da fala, a maioria dos sistemas CTF-SCAUS não o modelam explicitamente, atentando-se exclusivamente à modelagem da duração e do contorno de F_0.

3.3 Modelagem da Entoação: Contorno de F_0

3.3.1 Modelos Baseados em Tons

O trabalho de maior influência na área de modelos entoacionais baseados em tons é, sem dúvida, a Tese de Doutorado de Janet Pierrehumbert (Pierrehumbert, 1980) *apud* (Sproat, 1998). O modelo de Pierrehumbert foi desenvolvido utilizando conceitos da fonologia métrica (Liberman and Prince, 1977) *apud* (Sproat, 1998) e autosegmental (Leben, 1976) *apud* (Sproat, 1998). Pierrehumbert descreve a entoação do inglês americano empregando apenas duas categorias de tons: alto (H) e baixo (L). Segundo Pierrehumbert estes tons não interagem entre si, apenas se concatenam seqüencialmente no tempo. A Figura 3.2 ilustra a gramática de estados-finitos do modelo de Pierrebumbert. Esta gramática estabelece as regras de concatenação dos vários símbolos que representam os tons alto (H) e baixo (L) ao longo de um *grupo entoacional* (*intonational phrase*). Os principais eventos do modelo de Pierrehumbert são:

- Acentos tonais (*Pitch accent*): São identificados pelo símbolo "$*$", (H$*$, L$*$). Acentos tonais também podem ser bi-tonais (H$*$+L, H+L$*$, L$*$+H, L+H$*$).
- Acentos frasais (*Phrase accent*): São identificados pelos símbolo "$-$", (H $-$, L $-$). São utilizados para representar movimentos de *pitch* entre os acentos tonais e os tons de fronteira.
- Tons de fronteira (*Boundary tones*): São representados pelo símbolo "%", (H%, L%) para tons de fronteira finais e (%H, %L) para tons de fronteira iniciais. Estes tons são alinhados com as fronteiras dos grupos entoacionais (*intonational phrases*).

O modelo de Pierrehumbert é a base do sistema de transcrição prosódica ToBI (*Tones and Break Indexes*) apresentado por Silverman et al (Silverman et al., 1992).

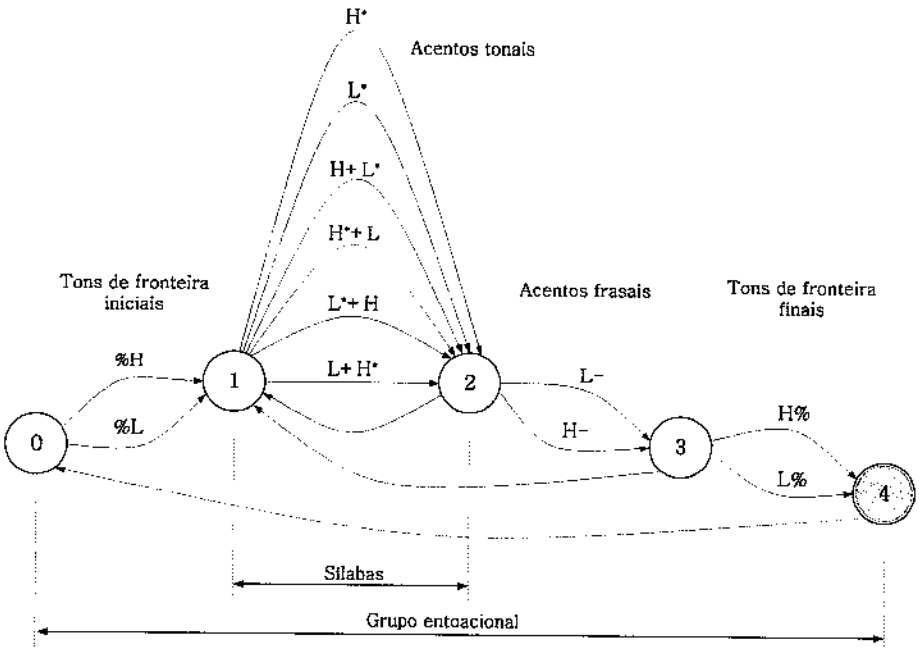


Figura. 3.2: Gramática do sistema de Janet Pierrehumbert (Pierrehumbert, 1980).

Aplicações

Graças à disponibilidade de corpora devidamente etiquetados com *acentos tonais*, *acentos frasais* e *tons de fronteira*, Ostendorf et al (Wightman and Ostendorf, 1994) realizaram trabalhos de posicionamento automático dos símbolos ToBI ao longo de sentenças. Em (Ross, 1994) e (Ross and Ostendorf, 1996), Ross gerou, automaticamente, contornos de entoação (curva de F_0) para sistemas de CTF posicionando, inicialmente, símbolos ToBI ao longo das sentenças e, em seguida, gerando a curva de F_0 a partir desta seqüência de símbolos ToBI.

Vantagens e Desvantagens

O principal argumento a favor do uso do modelo de Pierrehumbert (através da notação ToBI) é a sua utilidade como sistema de representação fonológica da entoação.

A desvantagem deste modelo é a dificuldade de se sintetizar a curva de F_0 a partir dos símbolos ToBI. Esta síntese somente é possível através de um sistema de regras sofisticado ou através de métodos estatísticos treinados a partir de extensos corpora devidamente etiquetados segundo a transcrição ToBI.

3.3.2 Modelos Perceptivos

Os modelos perceptivos baseiam-se no fato de que nem todos os movimentos observados na curva de F_0 são audíveis. Em outras palavras, parte-se do presuposto que somente um número limitado de movimentos característicos na curva de F_0 provocam a sensação audível de entoação e, portanto,

somente estes movimentos necessitam ser representados e modelados. O processo de análise e síntese dos modelos perceptivos consiste, inicialmente, em identificar movimentos característicos na curva de F_0 e, em seguida, utilizar estes movimentos característicos (segundo uma gramática previamente definida), na modelagem da curva entoacional.

O modelo perceptivo mais conhecido é o desenvolvido pelo IPO (*Instituut voor Perceptieonderzoek*). A etapa de análise do modelo IPO consiste em três passos. Inicialmente, movimentos no contorno de F_0, que sejam perceptivamente relevantes, são estilizados através de segmentos de reta. Este procedimento resulta em uma seqüência de segmentos de reta (*close copy contours*), que é perceptivamente indistinguível do contorno de F_0 original. Em outras palavras, os contornos de F_0 originais e estilizados são *perceptivamente equivalentes*. A razão para a estilização do contorno de F_0, de acordo com os pesquisadores do IPO, reside no fato da enorme variabilidade presente na curva de F_0 original representar um sério obstáculo para a observação de regularidades.

O segundo passo consiste no estudo das regularidades presentes nestes segmentos de reta (*close copy contours*), caracterizados por amplitudes, inclinações, durações e suas respectivas posições ao longo das sílabas, às quais se encontram alinhados. A idéia é que nem todos os *movimentos-estilizados* de F_0 são distintos perceptivamente, portanto, é possível o estabelecimento de classes de equivalência, sendo cada classe representada por um *movimento-estilizado-padrão* de F_0.

O terceiro passo consiste na definição de uma gramática que estabelece os possíveis seqüenciamentos de *movimentos-estilizados-padrões* de F_0.

Aplicação

O modelo IPO tem sido aplicado para modelar a entoação e reproduzi-la em sistemas CTF para vários idiomas, especialmente holandês, inglês e alemão (Sproat, 1998). Entretanto, uma limitação importante deste modelo é que seu processo de estilização não é automático. Versões mais modernas do modelo IPO (Mertens et al., 1996) descrevem um método para automatizar este processo de estilização.

Problemas

Alguns autores (Bellegarda and Silverman, 2001) argumentam que o processo de estilização da curva de F_0, apesar de ser guiado por aspectos perceptivos, não é capaz de modelar importantes variações entoaciocionais relativas à microprosódia.

3.3.3 Modelos Superposicionais

Estes métodos modelam a entoação de forma hierárquica, incluindo vários componentes simultâneos com alcances temporais distintos. A curva de F_0 final é obtida pela superposição aditiva de todos estes componentes. Um dos primeiros trabalhos sobre modelagem superposicional da entoação foi realizado para a língua Sueca e pode ser encontrado em (Öhman, 1967) *apud* (Huang et al., 2001). Uma versão refinada e elaborada deste modelo foi proposta por Fujisaki (Fujisaki, 1988) *apud* (Sproat, 1998).

O modelo de Fujisaki, ilustrado na Figura 3.3, distingue dois tipos de eventos discretos denominados comandos frasais e comandos acentuais, os quais são responsáveis, respectivamente, pelas componentes frasais e acentuais do contorno de F_0. Os comandos frasais são modelados como funções impulsos e os comandos acentuais são modelados como funções pulsos. Estes comandos funcionam como excitações de filtros sub-amortecidos de segunda ordem, cujas saídas são somadas para gerar a curva de F_0 final. Em (Fujisaki, 1992) *apud* (Dutoit, 1997), Fujisaki argumenta que a escolha das funções de transferência dos filtros e dos comandos acentuais e frasais possui uma correspondência direta com aspectos fisiológicos do processo de produção da fala.

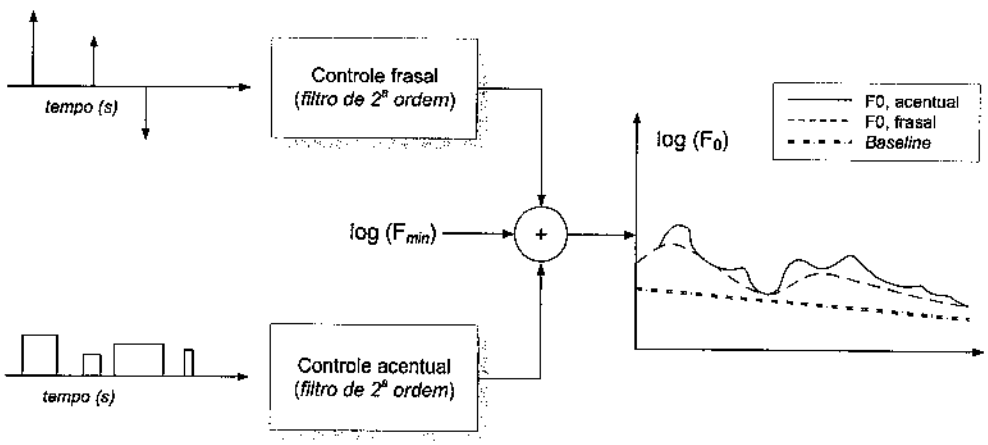


Figura. 3.3: Componentes do modelo superposicional de Fujisaki.

Aplicação

Este modelo é amplamente utilizado na geração de entoação para múltiplas línguas (Sproat, 1998),(Hirai et al., 1996). Isto faz com que o modelo superposicional seja referência obrigatória em qualquer estudo sobre entoação.

Problemas

O principal problema do modelo de Fujisaki é o seu treinamento. É complicado determinar se uma variação de F_0 é provocada pela componente acentual ou pela componente frasal. Geralmente, a estimativa dos parâmetros do modelo de Fujisaki é realizada de forma iterativa, minimizando-se a distância entre o contorno de F_0 sintético e o contorno analisado e impondo-se restrições aos valores dos parâmetros (Möebius, 1996), (Mancebo, 2002). Por uma questão de simplicidade, as componentes frasais e acentuais são, em geral, otimizadas separadamente.

3.3.4 Modelos de Estilização Acústica

Estes modelos possuem em comum o fato de se apoiarem, exclusivamente, nos perfis de F_0, sem fazerem qualquer consideração sobre a natureza fonológica da entoação ou sobre processos de produção/geração da mesma. Métodos de estilização acústica se apóiam na modelagem dos perfis de F_0, mediante alguma representação paramétrica que seja capaz de modelar a forma e a evolução temporal destes perfis. Estes métodos normalmente empregam técnicas estatísticas, e são capazes de modelar a entoação de um determinado locutor ou corpus, utilizando um número reduzido de parâmetros. As seções 3.4, 3.5 apresentam detalhes sobre dois modelos de estilização acústicas denominados Tilt, de autoria de Paul Taylor (Taylor, 2000) e modelo de Kagoshima, de autoria de Kagoshima (Kagoshima et al., 1998).

Vantagens

No contexto de sistemas CTF, a principal vantagem destes modelos é que eles se baseiam, exclusivamente, na curva de F_0. Seu objetivo é capturar as inflexões significativas da entoação, para depois reproduzi-las. Deste ponto de vista, estes modelos apresentam a vantagem de não estabelecer suposições *a priori*, tais como superposição de níveis acentuais, existência de uma estrutura de tons subjacente e aparecimento de movimentos de F_0 imperceptíveis; suposições estas que, apesar de estarem carregadas de fundamentos (lingüísticos, perceptivos, etc), não têm conseguido solucionar, inteiramente, o problema de modelagem da entoação para sistemas CTF.

3.4 O Modelo Entoacional de Paul Taylor: Modelo Tilt

3.4.1 Processo de Estilização

O modelo Tilt, introduzido por Paul Taylor em sua Tese de Doutorado (Taylor, 1992), é, talvez, o mais característico dos modelos de estilização acústica. A unidade básica do modelo Tilt é o *evento entoacional*. Os tipos básicos de *eventos entoacionais* são: *acentos tonais* e *acentos de fronteira*. Os *acentos tonais* são *gestos* em F_0 (movimentos em F_0) associados a sílabas específicas ao longo da elocução. Estes *acentos tonais* são os responsáveis por atribuir diferentes graus de ênfase às palavras ou às sílabas. Os *acentos de fronteira* são *gestos* de F_0 específicos, que ocorrem nas fronteiras dos grupos entoacionais. A Figura 3.4 ilustra uma seqüência de *eventos entoacionais*, juntamente com a indicação das sílabas às quais os *eventos entoacionais* se encontram alinhados.

Os *eventos entoacionais* são representados, inicialmente, em termos de um modelo denominado RFC (*Rise and Fall Connection Model*). No modelo RFC os *eventos entoacionais* são caracterizados por uma subida contínua até uma região de máximo e uma posterior descida. O modelo RFC modela estes *eventos entoacionais* utizando curvas paramétricas e contínuas, com formatos próximos ao de uma parábola, conforme descrito na Figura 3.5. Os parâmetos de um *evento entoacional*, segundo o modelo RFC, são:

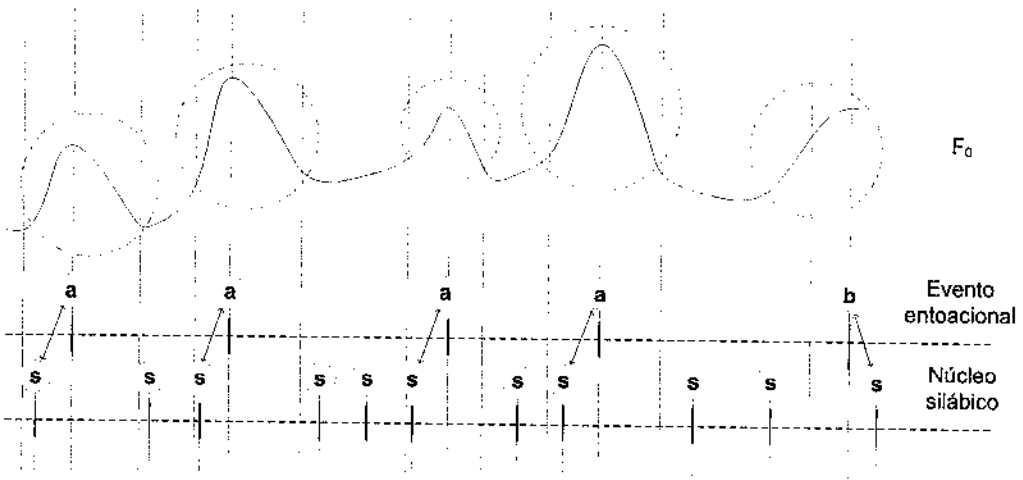


Figura. 3.4: Ilustração do posicionamento e do formato dos *eventos entoacionais*. O símbolo (s) indica a posição dos núcleos das sílabas e os símbolos (a) e (b) representam os *eventos tonais* e *eventos de fronteira*, respectivamente.

- A_{sub} - amplitude de subida (Hz)
- D_{sub} - duração de subida (segundos)
- A_{desc} - amplitude de descida (Hz)
- D_{desc} - duração de descida (segundos)
- P_{silaba} - posição na sílaba (segundos)
- A_{abs} - altura de F_0 (Hz)

Apesar de o modelo RFC ser capaz de modelar adequadamente os *eventos entoacionais*, ele não é muito fácil de ser manipulado matematicamente e, também, de ser interpretado lingüisticamente. A representação Tilt surgiu como uma tentativa para solucionar/minimizar estes dois problemas. Segundo a representação Tilt, os dois parâmetros de amplitude, juntamente com os dois parâmetros de duração do modelo RFC, são transformados em apenas 3 parâmetros:

- Amplitude *tilt* (Hz): Soma das amplitudes de subida e descida do modelo RFC;
- Duração *tilt* (segundos): Soma das durações de subida e descida do modelo RFC;
- *tilt* (sem dimensão): Número que expressa o formato do *evento entoacional*, independentemente de sua amplitude ou duração.

Este valor de *tilt* é um valor contínuo entre -1 e 1, que estabelece a proporção entre os movimentos de subida e de descida do *evento entoacional*: 1 indica apenas subida, 0 indica subida e descida na mesma proporção e -1 indica apenas descida. A Figura 3.6 ilustra o formato dos *eventos entoacionais* para diferentes valores de *tilt*.

Os processos de análise e síntese do modelo Tilt envolvem as seguintes operações:

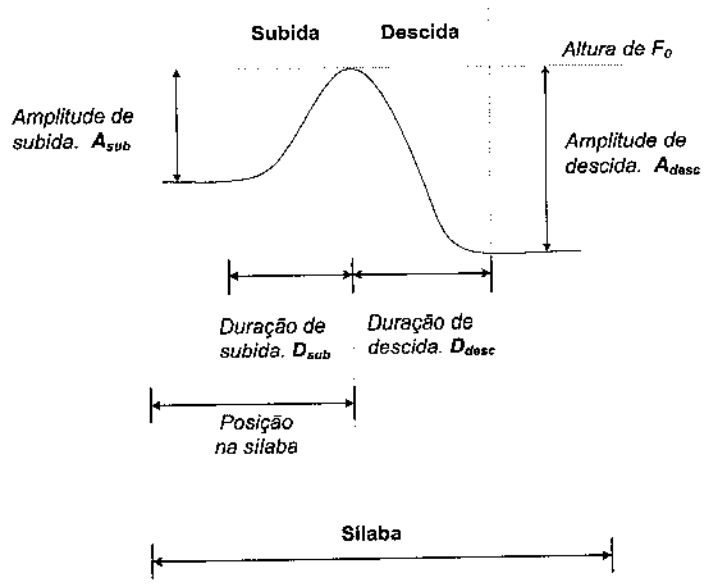


Figura. 3.5: Parâmetros RFC.

- Análise
  - Suavização do contorno de F_0 utilizando-se funções *splines*,
  - Segmentação silábica do contorno de F_0,
  - Localização das sílabas que devem receber *eventos entoacionais*,
  - Extração dos parâmetros RFC,
  - Transformação dos parâmetros RFC em parâmetros Tilt.

- Síntese
  - Predição das sílabas que devem receber os *eventos entoacionais*,
  - Predição dos parâmetros Tilt,
  - Transformação dos parâmetros Tilt em parâmetros RFC,
  - Síntese do contorno de F_0 a partir dos parâmetros RFC.

As subseções 3.4.2 e 3.4.3 descrevem em detalhes algumas das etapas de análise e síntese do modelo Tilt.

3.4.2 Análise

Estimação da Localização de Eventos Entoacionais

O primeiro passo do processo de análise consiste no treinamento de modelos/algoritmos/regras capazes de segmentar a curva de F_0 em termos de sílabas (e seus elementos: núcleo, onset, coda)

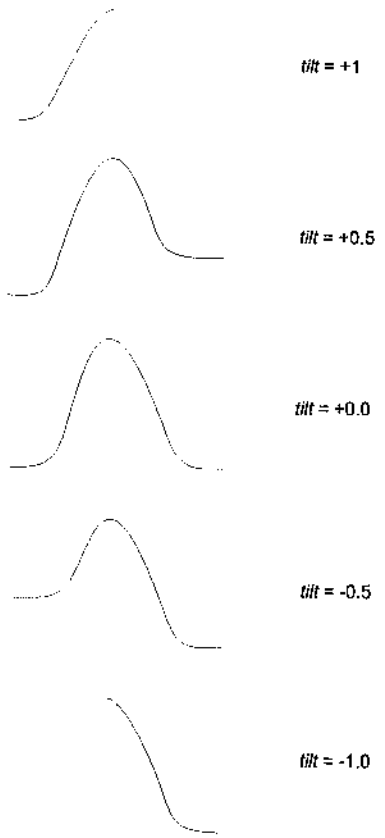


Figura. 3.6: Comportamentos dos eventos de acordo com os valores de *tilt*.

e indicar quais são as sílabas (ao longo da sentença) que devem receber *eventos entoacionais*. Este procedimento pode ser realizado utilizando-se modelos ocultos de Markov - HMM - (*Hidden Markov Model*) (Taylor, 2000) ou CAT (*Classification Tree*) (Dusterhoff and Black, 1997). Entretanto, tanto o treinamento de modelos HMM quanto o de CAT requer a existência de uma base de dados devidamente etiquetada com as segmentações silábicas e com as indicações de quais sílabas devem receber *eventos entoacionais*.

Estimação dos parâmetros RFC

Uma vez identificadas as sílabas que devem receber *eventos entoacionais*, o próximo passo consiste na determinação dos parâmetros RFC (destes *eventos entoacionais*), diretamente da curva de F_0. Para isto, a curva de F_0 é, inicialmente, interpolada ao longo dos segmentos não-sonoros e depois suavizada para eliminar possíveis variações abruptas devidas a erros na estimação de F_0. Em seguida empregam-se algoritmos de aproximação de curva (Taylor, 1995), para estimar os parâmetros RFC.

Conversão de parâmetros RFC em parâmetros Tilt

Os parâmetros RFC podem ser convertidos em parâmetros Tilt empregando-se as seguintes equações:

$$tilt = \frac{|A_{sub}| - |A_{desc}|}{2 \cdot (|A_{sub}| + |A_{desc}|)} + \frac{|D_{sub}| - |D_{desc}|}{2 \cdot (|D_{sub}| + |D_{desc}|)} \tag{3.1}$$

$$A_{evento} = |A_{sub}| + |A_{desc}| \tag{3.2}$$

$$D_{evento} = |D_{sub}| + |D_{desc}| \tag{3.3}$$

3.4.3 Síntese

A Figura 3.7 apresenta um diagrama de blocos com as principais operações da etapa de síntese do modelo Tilt. A seguir são apresentados detalhes sobre cada uma destas operações.

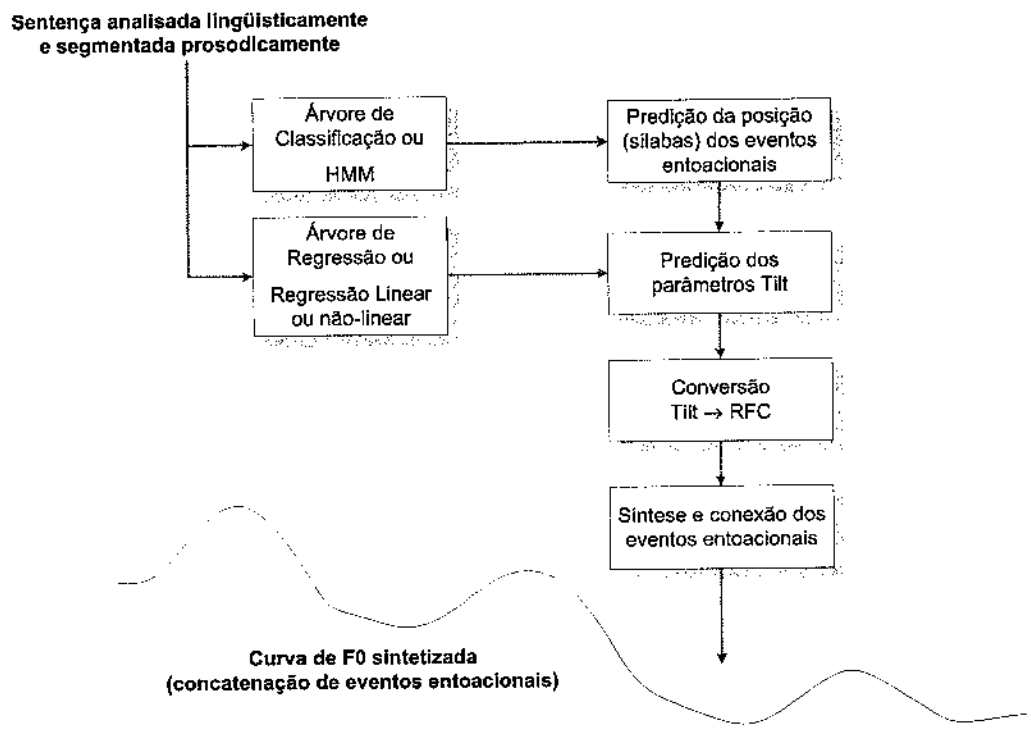


Figura. 3.7: Geração da curva de F_0 a partir do texto analisado lingüisticamente.

Estimativa das Posições dos Eventos Entoacionais

O primeiro passo do processo de síntese da curva de F_0 consiste na predição das sílabas que devem receber *eventos entoacionais* (*acentos tonais* e *acentos de fronteira*). Este procedimento pode ser realizado utilizando-se HMM (Taylor, 1992) ou CART (Dusterhoff and Black, 1997).

Estimativa dos Parâmetros Tilt a partir de Parâmetros Lingüísticos

Durante o processo de conversão de texto em fala, os parâmetros dos *eventos entoacionais* devem ser estimados a partir de atributos lingüísticos extraídos da sentença a ser sintetizada (atributos provenientes do módulo lingüístico e do módulo de prosódia simbólica). A estimativa destas funções capazes de mapear atributos lingüísticos em parâmetros dos *eventos entoacionais* pode ser obtida a partir de métodos estatísticos tais como árvore de regressão RT (*Regression Trees*) (Dusterhoff and Black, 1997) ou regressão linear de múltiplas variáveis nominais (Jobson, 1991).

Conversão de Parâmetros Tilt em Parâmetros RFC

Os parâmetros Tilt podem ser convertidos em parâmetros RFC empregando-se as seguintes equações:

$$A_{sub} = \frac{A_{evento} \cdot (1 + tilt)}{2} \tag{3.4}$$

$$A_{desc} = \frac{A_{evento} \cdot (1 - tilt)}{2} \tag{3.5}$$

$$D_{sub} = \frac{D_{evento} \cdot (1 + tilt)}{2} \tag{3.6}$$

$$D_{desc} = \frac{D_{evento} \cdot (1 - tilt)}{2} \tag{3.7}$$

Síntese do Contorno de F_0 a Partir dos Parâmetros RFC

O contorno de F_0 pode ser gerado a partir dos parâmetros RFC, segundo a função quadrática por partes descrita na equação 3.8.

$$F_0(n) = \begin{cases} A_{abs} + A - 2 \cdot (\frac{n}{D})^2 & \text{se} \quad 0 < n < \frac{D}{2} \\ A_{abs} + 2 \cdot (1 - \frac{n}{D})^2 & \text{se} \quad \frac{D}{2} < n < D \end{cases} \tag{3.8}$$

sendo $A = A_{desc}$, $D = D_{desc}$ e A_{abs} igual ao valor de F_0 no início do evento entoacional corrente, o qual é igual ao valor de F_0 no final do evento entoacional ou conexão precedente.

As regiões que não apresentam *eventos entoacionais* são representadas por conexões (linhas retas), segundo a equação 3.9

$$F_0(n) = A_{abs} + A \cdot (\frac{n}{D}) \quad \text{, para} \quad 0 < n < D \tag{3.9}$$

sendo A, D e A_{abs} os mesmos da equação 3.8.

3.5 O Modelo Entoacional de Kagoshima

3.5.1 Processo de Estilização

Assim como o modelo Tilt, o modelo de Kagoshima (Kagoshima et al., 1998) também pode ser classificado como pertencente à classe dos modelos de estilização acústica. Segundo Kagoshima, o contorno entoacional de uma sentença pode ser descrito com uma seqüência de gestos de F_0, somados a níveis de *offset*, seguido de algumas operações de pós-processamento. Em (Kagoshima et al., 1998) Kagoshima aplica este modelo ao japonês associando estes gestos de F_0 a *grupos acentuais* medidos em números de *mora* (unidade lingüística básica para a língua japonesa). O modelo de Kagoshima foi aplicado por Morais (Knill et al., 2002) (Knill et al., 2003), às línguas inglesa (americano e britânico), alemão e espanhol (europeu), associando cada gesto de F_0 a *unidades entoacionais* medidas em número de sílabas. As unidades entoacionais utilizadas por Morais foram as próprias palavras do léxico. Os resultados obtidos por Morais foram considerados de boa qualidade para o inglês e o alemão e excelentes para o espanhol (Knill et al., 2003). De acordo com o modelo de Kagoshima, uma sentença composta por N palavras terá seu contorno de F_0 escrito como:

$$\mathbf{F_0} = \{\mathbf{F}_g^1 + \mathit{Offset}^1, \mathbf{F}_g^2 + \mathit{Offset}^2, \cdots, \mathbf{F}_g^N + \mathit{Offset}^N\} \tag{3.10}$$

sendo $\mathbf{F_0}$ o contorno de F_0 para toda a sentença, $\mathbf{F}_g^i$ o gesto de F_0 para a *i-ésima* palavra na sentença e Offset^i o nível de *offset* a ser aplicado a $\mathbf{F}_g^i$. A Figura 3.8 ilustra um exemplo para o caso de $N = 10$.

A Figura 3.9 apresenta um diagrama de blocos do modelo de Kagoshima. As seções 3.5.2 e 3.5.3 apresentam uma breve descrição das operações envolvidas nos módulos de análise de síntese deste modelo.

3.5.2 Análise

O processo de análise do modelo de Kagoshima consiste na criação de inventários de *gestos* de F_0. Estes inventários são obtidos através de um processo de quantização vetorial em malha fechada,

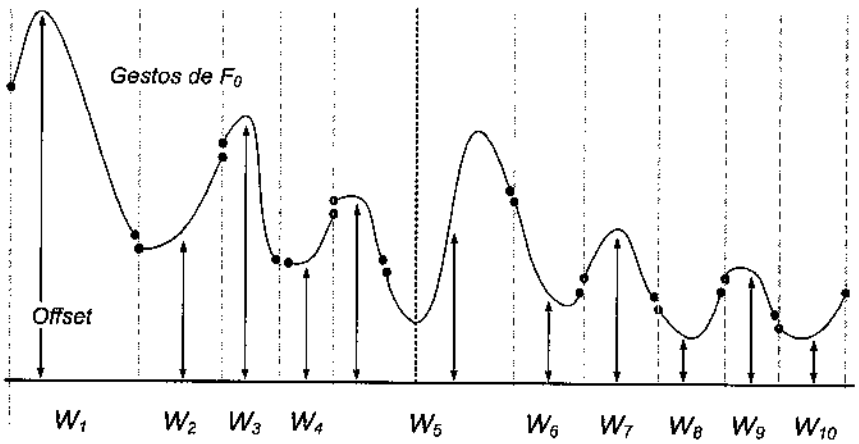


Figura. 3.8: Modelo entoacional de Kagoshima para geração automática do contorno de F_0.

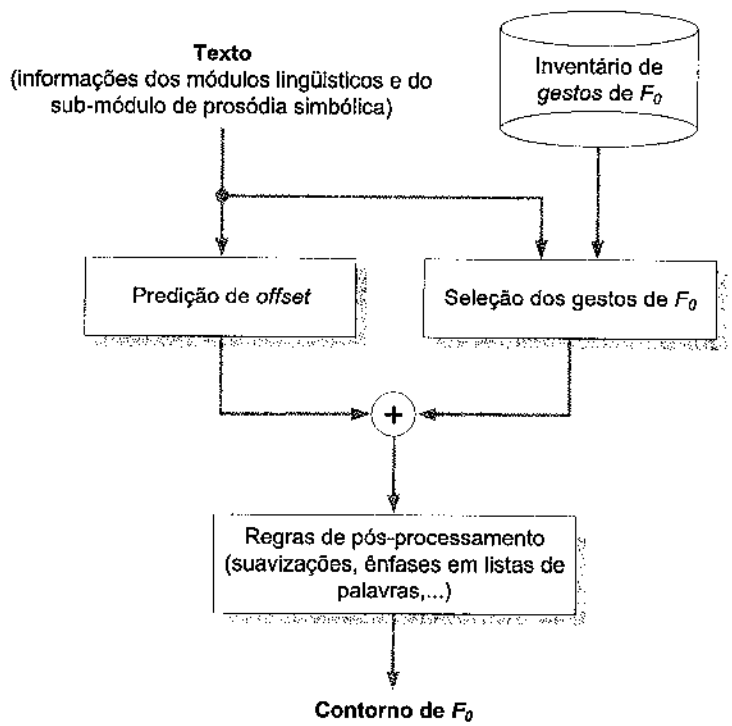


Figura. 3.9: Diagrama de blocos do modelo entoacional de Kagoshima para geração automática do contorno de F_0.

conforme descrito em (Akamine and Kagoshima, 1998), (Kagoshima et al., 1998), e resumido a seguir.

Dicionário de Gestos de F_0

O procedimento de quantização vetorial pode ser dividido em quatro passos principais:

- Os contornos de F_0 de todas as sentenças presentes na base de dados são suavidados utilizando-se funções *splines*;
- Estes contornos suavizados são então segmentados, nos respectivos gestos de F_0, delimitados pelas fronteiras das unidades entoacionais (grupos acentuais ou palavras);
- Os gestos de F_0 são agrupados em classes. Cada classe irá conter somente gestos de F_0 associados a unidades entoacionais com o mesmo número de sílabas e com sílaba acentuada em uma posição específica. Por exemplo, todos os gestos de F_0 associados a unidades entoacionais de 3 sílabas e acento na última sílaba serão agrupados na mesma classe;
- Para cada classe, um número representativo de gestos de F_0 são selecionados/sintetizados através do procedimento em malha fechada descrito em (Akamine and Kagoshima, 1998). O número de gestos de F_0 selecionado/sintetizado por classe irá depender da variabilidade do espaço definido pelos gestos de F_0 associados a esta classe.

3.5.3 Síntese

A etapa de síntese do modelo de Kagoshima consiste, inicialmente, na seleção dos gestos de F_0 (associados às unidades entoacioniais) e na predição dos valores de *offsets*. Em seguida, concatena-se os gestos de F_0 já somados aos seus respectivos valores de *offsets* e finaliza-se o processo com a aplicação de algumas regras de pós-processamento (suavização das fronteiras entre os gestos de F_0 e intensificação/atenuação de algumas ênfases).

Seleção dos gestos de F_0

A identificação de qual inventário de *gestos* de F_0 deve ser utilizado para cada unidade entoacional (grupo acentual ou palavra) dependerá da combinação entre o número de sílabas da unidade entoacional *versus* a posição da sílaba acentuada na unidade entoacional.

O próximo passo consiste na seleção de qual gesto de F_0, presente no inventário selecionado, deverá ser utilizado. Este problema foi resolvido por Kagoshima e Morais, utilizando-se análise discriminante linear (Knill et al., 2002), (Knill et al., 2003), a partir dos atributos lingüísticos (provenientes do módulo de *Front-End* e do sub-módulo de prosódia simbólica). Outras técnicas tais com CART ou ANN também podem ser aplicadas a este problema de seleção/classificação.

Predição dos Offsets

A função para mapear atributos lingüísticos nos valores de *offsets* de cada *gesto* de F_0 deve ser projetada e treinada utilizando-se métodos estatísticos treinados a partir de corpora de fala (do locutor que se deseja modelar). Kagoshima e Morais utilizaram modelos de regressão linear a partir de variáveis nominais (método QMTI - *Quantification Method Type I* - (Hayashi, 1950)) para realizar esta tarefa. Entretanto, técnicas como RT (Regression Trees) e ANN (*Artificial Neural Networks*) também poderiam ser utilizadas para a predição destes *offsets*.

Pós-Processamento Utilizando Regras Lingüísticas do tipo Ad Hoc

Após a estimação da sequência de gestos de F_0 e seus correspondentes valores de *offset*, um procedimento de suavização deve ser aplicado nas fronteiras entre os *gestos* de F_0. Além disso, algumas regras *ad hoc* devem ser utilizadas, caso necessário, para enfatizar alguns movimentos na curva de F_0 (Kagoshima et al., 1998).

3.6 Considerações Finais

Este capítulo apresentou uma revisão sobre as principais operações envolvidas no módulo prosódico de sistemas CTF-SCAUS. Foram discutidos desde aspectos paralingüísticos e fonológicos até aspectos fonéticos relacionados à modelagem prosódica. Abordou-se brevemente a modelagem da duração segmental (de fones) e do contorno de intensidade da fala. Um destaque especial foi dado à modelagem da entoação, discutindo-se as principais características dos modelos baseados em tons, dos modelos perceptivos, dos modelos superposicionais e dos modelos de estilização de F_0. Além disso foram apresentados detalhes sobre dois modelos entoacionais baseados em estilização de F_0, que têm sido amplamente utilizados em sistemas CTF-SCAUS comerciais: o modelo Tilt de Paul Taylor e o modelo de Kagohisma.

É importante ressaltar que o levantamento bibliográfico realizado durante esta Tese, revelou que existem poucos trabalhos sobre modelagem prosódica para o português brasileiro (PB) no contexto de sistemas CTF. Além disso, dos poucos trabalhos encontrados, a grande maioria versa apenas sobre modelagem da duração e do ritmo da fala. São raros (e muitas vezes incipientes) os trabalhos sobre entoação para o PB.

Capítulo 4

Módulo de Seleção Automática de Unidades de Síntese

4.1 Introdução

Sistemas de conversão texto-fala que empregam a tecnologia SCAUS (sistemas CTF-SCAUS), formulam o processo de síntese de fala como um procedimento de Seleção e Concatenação Automática de Unidades de Síntese previamente gravadas (Hunt and Black, 1996), (Black, 1996), (Quazza et al., 2001). Os modernos sistemas CTF-SCAUS empregam grandes inventários de unidades de síntese, os quais são devidamente projetados para garantir um alto grau de variabilidade acústica (fonética) e prosódica (Zhu and Zhang, 2002), (Eide, 2003). A principal motivação para o uso de grandes inventários de unidades de síntese reside na possibilidade de que o processo de seleção possa encontrar unidades de síntese próximas das "ideais", minimizando, portanto, possíveis descontinuidades espectrais entre unidades adjacentes e garantindo, por conseguinte, o contorno prosódico adequado ao sinal sintetizado. Entretanto, o uso de extensas bases de unidades de síntese requer um processo de busca altamente eficiente, capaz não somente de selecionar unidades de síntese próximas das "ideais", mas também de apresentar um custo computacional compatível com a aplicação desejada.

Este capítulo discute algumas das principais operações e cuidados a serem tomados pelos sistemas CTF-SCAUS a partir de grandes bases de dados. Algumas destas operações e cuidados são:

- Qual é a melhor unidade de síntese a ser utilizada?
- Como projetar, adquirir e etiquetar um corpus de fala que seja "ótimo" para este tipo de tecnologia?
- Como organizar as unidades de síntese em estruturas que facilitem o processo de busca e ao mesmo tempo eliminem redundâncias (acústicas e prosódicas) existentes no conjunto inicial de unidades de síntese?
- Como implementar um processo de seleção de unidades de síntese eficiente: rápido, preciso e robusto (que seja capaz de operar em tempo real e minimize erros de seleção durante o processo

de busca pelas unidades ótimas)?

- Quais são as melhores técnicas para compactar (comprimir) a base de unidades de síntese?

O restante deste capítulo é dedicado a descrever detalhes sobre cada uma das operações listadas acima. A seção 4.2 discute algumas das unidades de síntese mais comumente utilizadas em sistemas que empregam a tecnologia SCAUS. A seção 4.3 apresenta fundamentos sobre o processo de seleção de unidades de síntese, discutindo questões como funções de custo fonético-prosódico, custo concatenativo e custo total. A seção 4.4 discute técnicas de clusterização hierárquica binária como uma maneira eficiente de organizar as unidades de síntese, apresenta algumas das principais métricas utilizadas para medir distâncias entre unidades de síntese e discute algumas das técnicas de poda para eliminar redundâncias presentes nas árvores de clusterização. A seção 4.5 apresenta procedimentos para projeto, aquisição e etiquetagem do corpus de fala, e também técnicas para compressão do inventário de unidades de síntese. Por último, a seção 4.6 encerra este capítulo apresentando algumas considerações finais.

4.2 Unidades de Síntese

A escolha das unidades de síntese mais adequadas a sistemas CTF-SCAUS continua a ser uma importante área de pesquisa. Entre as várias unidades de síntese, utilizadas ao longo da última década, destacam-se: difones, fones, metade-de-fones e senones (unidades sub-fonéticas). A seguir são apresentadas algumas considerações sobre cada uma destas unidades.

Difones

Sistemas que empregam difones independentes de contexto (um único exemplar de cada difone) têm sido estudados há vários anos (Klatt, 1987). Difones são unidades de síntese que incluem as transições entre os pares de fones que podem existir em uma dada língua. Por exemplo, para o Inglês que possui, aproximadamente, 50 fones, o número total de difones encontra-se entre 1500 e 2000 (dado que várias combinações de pares de fones nunca ocorrem). Sistemas baseados em difones independentes de contexto são capazes de produzir fala sintética com alta inteligibilidade, entretanto, para serem capazes de expressar diversidades contextuais (variações prosódicas e efeitos coarticulatórios), os difones devem ser submetidos a elevados níveis de modificações prosódicas e suavizações espectrais. Infelizmente, conforme colocado por vários autores (Dutoit, 1997), (Stylianou, 1996) e (Quatieri, 2002), todas as técnicas de processamento digital de sinais desenvolvidas para manipulação dos sinais de fala introduzem, em maior ou menor grau, distorções indesejadas no sinal sintetizado.

Fones

Mais recentemente, pesquisadores atacaram o problema de seleção de unidades a partir de grandes bases de dados de fones dependentes de contexto (vários exemplares de cada fone) (Hunt and

Black, 1996), (Hon et al., 1998). O sistema CHATR (Black, 1996) do ATR (*Advanced Telecommunication Research*) é um exemplo bem conhecido do uso deste tipo de unidade de síntese. Contudo, em (Bulyko, 2002), Bulyko relata que o sistema CHATR apresenta um comportamento não muito estável, oscilando entre falas sintetizadas com um alto grau de naturalidade e falas sintetizadas com distorções desagradáveis. Uma das razões para este desempenho inconsistente do sistema CHATR é que, ao empregar fones como unidade de síntese, todas as concatenações, necessariamente, passam a ser realizadas nas fronteiras entre fones. Para os segmentos fricativos este problema não é tão grave, dado que os efeitos de coarticulação em suas fronteiras são mínimos (Sproat, 1998). Porém, o mesmo não acontece com segmentos vocálicos, os quais, em geral, encontram-se sujeitos a fortes efeitos de coarticulação. Portanto, mesmo utilizando fones dependentes de contexto, o sistema CHART, muitas vezes, defronta-se com o problema de concatenar segmentos vocálicos que apresentam descasamentos espectrais (descasamentos de maior ou menor grau na estrutura formântica e no contorno de F_0). O efeito cumulativo destes descasamentos espectrais é, geralmente, a principal fonte de distorções do sistema CHATR (Klabbers and Veldhuis, 2001).

Metades-de-Fones

O sistema ATT *Next-Generation* (Beutnagel et al., 1999a) minimiza os problemas enfrentados pelo sistema CHART através do uso de unidades de síntese dependentes de contexto e com extensões iguais à metade de um fone (*metade-de-fone*). Estas unidades agregam maior flexibilidade ao sistema, permitindo que as concatenações possam ser realizadas nas fronteiras entre os fones ou no meio dos fones. O sistema ATT *Next-Generation* utiliza elaboradas funções de custo de concatenação, os quais são dependentes de contexto e diferenciados para concatenações entre fronteiras de fones ou entre regiões centrais de fones.

Senones

Utilizando-se de conceitos da área de reconhecimento de fala, vários autores têm empregado estados de HMM dependentes de contextos para definir o inventário de unidades de síntese. Em (Donovan, 1996), (Huang and Acero, 1998), HMM com três estados são utilizados para produzir, para cada fone, três unidades de síntese denominadas *senones*. Apesar de *senones* serem capazes de modelar variações espectrais em um nível sub-fonético refinado, elas introduzem muitas junções durante a concatenação, as quais podem resultar em uma potencial degradação na qualidade do sinal e um aumento no tempo de busca pelas unidades de síntese.

4.3 Seleção de Unidades de Síntese

O método de seleção automática de unidades de síntese formula o paradigma de conversão texto-fala como um problema não-paramétrico no qual a fala sintetizada pode ser gerada a partir de grandes bases de fala, através de um procedimento de busca. O objetivo deste processo de busca é selecionar

a seqüência de unidades de síntese que melhor satisfaz as seguintes propriedades: (1) a seqüência de unidades de síntese deve apresentar um contorno prosódico o mais próximo possível do desejado (critério paradogmático) e (2) as unidades de síntese selecionadas devem minimizar possíveis descontinuidades espectrais ao longo das junções entre elas (critério sintagmático). Em outras palavras, o principal objetivo do processo de seleção automática de unidades de síntese é evitar, o máximo possível, qualquer tipo de modificação prosódica ou suavização espectral nas unidades de síntese selecionadas. A Figura 4.1 ilustra o processo de seleção de unidades de síntese empregado por sistemas CTF-SCAUS.

O método de seleção de unidades, em geral, emprega duas funções de custo principais: (1) função de custo fonético-prosódico e (2) função de custo de concatenação. A função de custo fonético-prosódico, $C^t(U_i, T_i)$, estima a diferença entre o contexto fonético-prosódico da unidade de síntese, U_i, presente na base de dados, e o contexto fonético-prosódico desejado, T_i. A função custo de concatenação, $C^c(U_{i-1}, U_i)$, estima o grau de descontinuidades entre duas unidades de síntese consecutivas U_{i-1} e U_i.

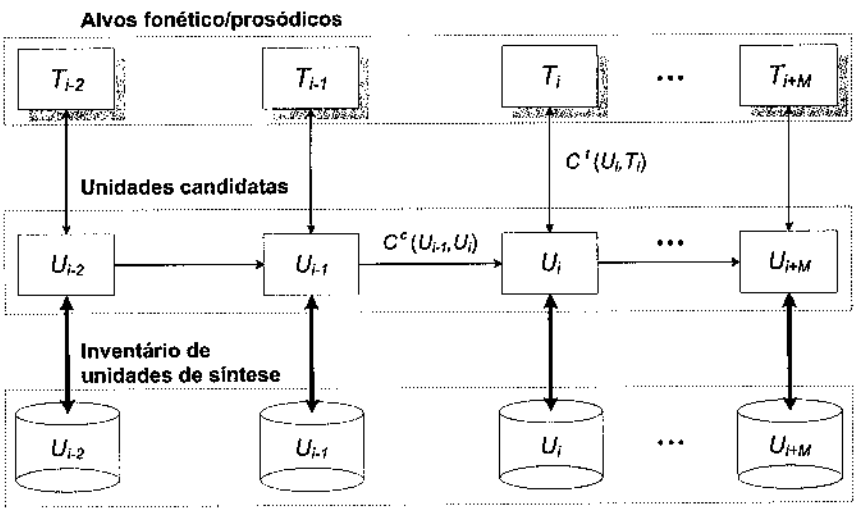


Figura. 4.1: Processo de seleção automática de unidades de síntese.

O primeiro estágio do processo de busca avalia o custo fonético-prosódico $C^t(U_i, T_i)$ de todas as unidades de síntese candidatas. Em outras palavras, para cada alvo fonético-prosódico T_i da sentença a ser sintetizada, as respectivas unidades de síntese candidatas são avaliadas segundo o seu custo fonético-prosódico. No segundo estágio de busca, um custo de concatenação $C^c(U_{i-1}, U_i)$ é associado a cada possível transição entre duas unidades de síntese candidatas. Por último, um algoritmo de programação dinâmica (ou busca síncrona baseada no algoritmo de *Viterbi*) é utilizado para encontrar a seqüência ótima de unidades de síntese que minimiza a soma acumulada dos custos fonético-prosódicos e de concatenação. Um aspecto muito importante deste procedimento de busca é que as unidades de síntese que forem consecutivas na base de dados e apresentarem um reduzido custo fonético-prosódico, terão alta probabilidade de serem selecionadas. Este fenômeno permite ao sistema selecionar "unidades

de síntese" de comprimentos não-uniformes. Por exemplo, se a unidade de síntese básica for a *metade-de-fones*, então este procedimento de seleção de unidades permitirá a seleção de *metades-de-fones*, difones, sílabas, palavras, segmentos de sentenças e até mesmo sentenças completas presentes na base de dados.

4.3.1 Funções de Custo Fonético-Prosódico

As funções de custo fonético-prosódico incluem, em geral, rótulos fonéticos, fonológicos, duração de segmentos/sílabas e valores de F_0. A utilização de atributos simbólicos e contínuos no cálculo da distância entre as unidades de síntese e seus respectivos alvos fonético-prosódicos, faz da estimação destas funções de custo uma tarefa relativamente complexa. No caso de distâncias simbólicas, uma simples comparação binária não é uma medida adequada. Em geral, estas distâncias simbólicas são estimadas empregando-se métodos de regressão linear/não-linear de múltiplas variáveis nominais (Hunt and Black, 1996). Distâncias entre atributos contínuos são, em geral, calculadas utilizando-se métricas específicas capazes de explorar características perceptivas, principalmente nos casos de F_0 e duração segmental (Bulyko, 2002).

Formalmente, o custo fonético-prosódico $C^t(T_i, U_i)$ é dividido em p subcustos $C_j^t(T_i, U_i)$ $(j = 1, \cdots, p)$, associados a cada um dos parâmetros simbólicos (parâmetro fonético, fonológico, etc) e numéricos (duração segmental e valor de F_0). Alguns autores (Hunt and Black, 1996) reportam o uso de 20 a 30 subcustos fonético-prosódicos. O custo fonético-prosódico total é estimado como uma soma ponderada de seus respectivos subcustos:

$$C^t(T_i, U_i) = \sum_{j=1}^{p} w_j^t \cdot C_j^t(T_i, U_i) \tag{4.1}$$

sendo w_j^t os pesos associados a cada subcusto $C_j^t(T_i, U_i)$.

4.3.2 Funções de Custo Concatenativo

O custo de concatenação é calculado na fronteira entre as unidades de síntese a serem concantenadas. Estes custos devem medir não somente o grau de descontinuidade no ponto de junção entre as unidades, mas também o descasamento de características espectrais em quadros de análise adjacentes ao ponto de junção entre as unidades de síntese. A Figura 4.2 ilustra o caso em que o custo de concatenação, entre as unidades U_{i-1} e U_i, é estimado ao longo de dois quadros de análise adjacentes ao ponto de junção. Neste exemplo, o custo de concatenação é dado por:

$$C^c(U_{i-1}, U_i) = C^{ce}(U_{i-1}, U_i) + C^{cd}(U_{i-1}, U_i) \tag{4.2}$$

sendo $C^{ce}(U_{i-1}, U_i)$ definido como o custo de concatenação à esquerda e mede a distância entre o

último quadro de análise da unidade U_{i-1} e o quadro de análise que precede à unidade U_i (é importante lembrar que U_{i-1} e U_i foram extraídos de uma elocução contínua). $C^{cd}(U_{i-1}, U_i)$ é definido como o custo de concatenação à direita e mede a distância entre o primeiro quadro de análise da unidade U_i e o quadro de análise que segue a unidade U_{i-1}.

De forma semelhante aos custos fonético-prosódicos, os custos de concatenação $C^{c}(U_{i-1}, U_i)$, os custos de concatenação à esquerda $C^{ce}(U_{i-1}, U_i)$ e à direita $C^{cd}(U_{i-1}, U_i)$ são estimados como uma soma ponderada de q subcustos:

$$C^{c}(U_{i-1}, U_i) = \sum_{j=1}^{q} w_j^{c} \cdot C_j^{c}(U_{i-1}, U_i) \tag{4.3}$$

$$C^{ce}(U_{i-1}, U_i) = \sum_{j=1}^{q} w_j^{ce} \cdot C_j^{ce}(U_{i-1}, U_i) \tag{4.4}$$

$$C^{cd}(U_{i-1}, U_i) = \sum_{j=1}^{q} w_j^{cd} \cdot C_j^{cd}(U_{i-1}, U_i) \tag{4.5}$$

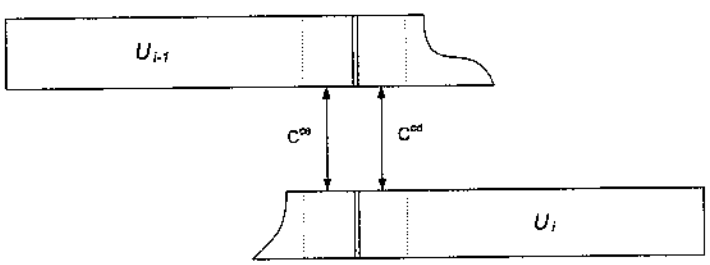


Figura. 4.2: Estimativa do custo de concatenação.

sendo w_j^{c}, w_j^{ce} e w_j^{cd} os pesos de cada um dos subcustos das equações 4.3, 4.4 e 4.5, respectivamente.

Vários estudos têm sido realizados em busca de métricas para estimar subcustos de concatenação que apresentem uma forte correlação com aspectos auditivo-perceptivos (Wouters and Macon, 1998), (Donovan, 2001). Entretanto, até o momento, não existe um consenso sobre qual seja a melhor métrica para o cálculo destes subcustos. Algumas das distâncias mais utilizadas são: distância euclidiana ou distância de Mahalonabis (Bulyko, 2002) entre coeficientes MFCC (*Mel Frequency Cepstral Coefficientes*) (Wouters and Macon, 1998), distância de Kullback-Leibler entre formantes (Klabbers and Veldhuis, 2001) e distâncias euclidianas entre coeficientes LSF (*Line Spectral Frequency Coeficients*) (Wouters, 2001), (Hon et al., 1998). A diferença do logaritmo da potência de F_0 também tem sido utilizada como um dos subcustos de concatenação (Hunt and Black, 1996).

4.3.3 Custo Total

Sendo $T = T_1, T_2, \cdots, T_N$ a seqüência de alvos prosódico-fonéticos desejados para uma sentença a ser sintetizada, e sendo $U = U_1, U_2, \cdots, U_N$ uma seqüência de unidades de síntese candidata à seqüência ótima, então o custo total da seqüência U é estimado como a soma acumulada de seus custos fonético-prosódico e de concatenação:

$$C(T,U) = \sum_{i=1}^{N} C^t(T_i, U_i) + \sum_{i=2}^{N-1} C^c(U_{i-1}, U_i) + C^c(S, U_1) + C^c(U_N, S) \tag{4.6}$$

sendo que S denota silêncio, $C^c(S, U_1)$ e $C^c(U_N, S)$ definem as condições iniciais e finais (concatenação de silêncio para U_1 e de U_N para silêncio). Expandindo a equação 4.6 em termos de seus subcustos (equações 4.1 e 4.2), tem-se:

$$C(T,U) = \sum_{i=1}^{N} \sum_{j=1}^{p} w_j^t \cdot C_j^t(T_i, U_i) + \sum_{i=1}^{N-1} \sum_{j=2}^{q} w_j^c \left(C_j^c(U_{i-1}, U_i) + C_j^c(S, U_1) + C_j^c(U_N, S) \right) \tag{4.7}$$

Formalmente, o processo de seleção de unidades de síntese consiste na determinação da seqüência de unidades $\bar{U}$, de extensão N, que minimiza o custo total definido pela equação 4.8

$$\bar{U} = \underset{\bar{U}_1, \bar{U}_2, \cdots, \bar{U}_N}{ArgMin}\ C(T,U) \tag{4.8}$$

Como já foi mencionado anteriormente, este procedimento de busca pode ser solucionado utilizando-se o algoritmo de Viterbi (Rabiner, 1989). Contudo, para grandes bases de unidades de síntese, esta busca pode demandar um alto custo computacional. Para reduzir este custo computacional, técnicas de clusterização de unidades de síntese, seguidas por operações de poda, têm sido largamente utilizadas (Donovan, 2000), (Donovan, 2003), (Black and Taylor, 1997). A próxima seção será dedicada a apresentar alguns destes procedimentos.

4.4 Clusterização de Unidades de Síntese

Armazenar uma unidade de síntese (fone, metade-de-fone ou *senone*), para cada possível contexto lingüístico, não é um procedimento prático. Por exemplo, para a língua inglesa, a representação de todas as combinações de contextos fonéticos imediatamente à esquerda e à direita, requer mais de 50.000 unidades de síntese (Bulyko, 2002). Além disso, este número pode crescer significativamente se mais informações lingüísticas tais com sílabas e informações prosódicas forem levadas em consideração.

Árvores de clusterização, largamente utilizadas em sistemas para reconhecimento de fala, permitem controlar a relação entre o tamanho da base de dados e o seu grau de variabilidade acústica. Adicional-

mente, este procedimento de clusterização também permite acelerar o processo de seleção de unidades de síntese, reduzindo-se o número de unidades candidatas por alvo fonético-prosódico.

Em geral, este procedimento de clusterização é realizado empregando-se árvores de classificação hierárquica binária. Segundo este procedimento, uma árvore de classificação é gerada para cada unidade de síntese (por exemplo, se forem utilizados fones como unidades de síntese, e se existirem 50 fones distintos, então devem ser geradas 50 árvores de classificação). Dois dos algoritmos geralmente utilizados para o crescimento (construção) destas árvores são o CART (*Classification and Regression Trees*) (Breiman et al., 1993) e o C4.5 (Quinlan, 1993). O uso do algoritmo CART requer o estabelecimento prévio dos seguintes itens:

- Um método para medir o grau de impuridade em um cluster. O grau de impuridade de um cluster é, geralmente, medido como a distância média entre as unidades de síntese pertencentes a este cluster. Portanto, a medida do nível de impuridade requer o estabelecimento de métricas para medir a distância entre pares de unidades de síntese.
- Uma lista de características fonéticas, lexicais e fonológicas. A partir desta lista são geradas várias perguntas sobre possíveis contextos fonéticos, lexicais e/ou fonológicos. Cada pergunta define uma possível divisão de um *cluster* em dois *cluster-filhos*. A pergunta ótima será aquela que gerar *cluster-filhos* cuja soma de suas impuridades seja mínima. O processo de seleção da pergunta ótima é realizado utilizando-se um algoritmo guloso (*greedy algorithm*).

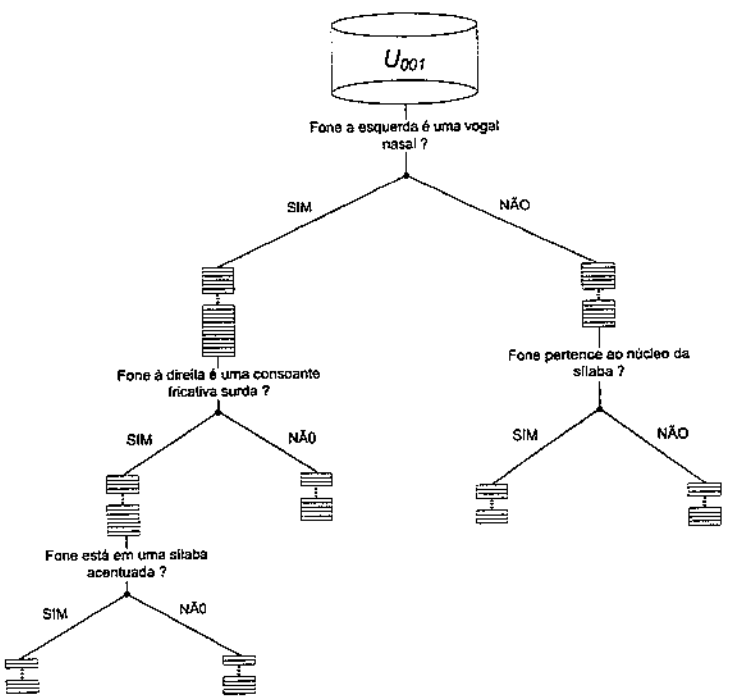


Figura. 4.3: Processo de clusterização de unidades de síntese.

Após a construção das árvores de clusterização (uma para cada unidade de síntese), o processo de seleção de unidades de síntese pode ser realizado, conforme ilustrado na Figura 4.4, através dos seguintes passos:

- Primeiro: Dada a transcrição da sentença a ser sintetizada (em termos de suas respectivas unidades de síntese) e todas as informações lingüísticas e prosódicas a ela associada, descer ao longo das correspondentes árvores de clusterização (uma para cada unidade de síntese) até encontrar os respectivos *clusters* terminais ótimos.
- Segundo: Utilizando-se apenas as unidades de síntese pertencentes aos *clusters* terminais selecionados, empregar a equação 4.8 para selecionar as unidades de síntese ótimas.

Dado que a busca pelos *clusters* terminais ótimos consiste em um procedimento extremamente rápido (basta descer ao longo das árvores realizando as devidas perguntas lingüísticas e prosódicas) e que o número de exemplares de unidades de síntese por *cluster* terminal é relativamente reduzido, então pode-se verificar que o procedimento mostrado na Figura 4.4 pode reduzir drasticamente o tempo de busca pela seqüência ótima de unidades de síntese.

O principal argumento desfavorável com relação ao procedimento da Figura 4.4 consiste no fato que as unidades ótimas não necessariamente estarão presentes nos *clusters* terminais selecionados. Ou seja, a robustez do procedimento descrito pela Figura 4.4 é extremamente dependente da qualidade do processo de clusterização das unidades de síntese. Um dos itens fundamentais para que este processo de clusterização seja de alta qualidade, é a definição de métricas eficientes para o cálculo das distâncias acústicas e prosódicas entre unidades de síntese. Por esta razão, a subseção 4.4.1 será dedicada à apresentação de algumas das métricas mais comumente utilizadas durante o processo de clusterização de unidades de síntese.

4.4.1 Métricas para Estimar a Distância Entre Unidades de Síntese

Bulyko (Bulyko, 2002) utiliza os seguintes parâmetros acústicos para calcular a distância entre unidades de síntese: MFCC (*Mel Frequency Coefficientes*) de ordem 12 e sua derivada de primeira ordem (Delta MFCC), as durações destas unidades e seus valores de F_0. Cada vetor de parâmetros acústicos (contendo MFCC, Delta MFCC e F_0) é estimado para quadros de análise de 5 *ms* com superposição de 50%.

O cálculo da distância entre duas unidades de síntese U e V requer, inicialmente, um alinhamento entre os quadros de análise destas unidades. Um alinhamento extremamente simples é o mostrado na Figura 4.5. Entretanto alinhamentos mais complexos que façam uso de *Time Scale Modifications* e *Pitch Scale Modifications*, como será apresentado no capítulo 7, sobre o módulo *Back-End*, também podem ser empregados. Em seguida, utilizando-se métricas específicas para os parâmetros *Cepstrais* (MFCC em conjunto com os Delta MFCC), as durações e os valores de F_0, estima-se a distância entre as unidades de síntese. Formalmente, a distância $D(U,V)$, entre as unidades U e V (assumindo U mais extensa que V), é estimada pela equação 4.9.

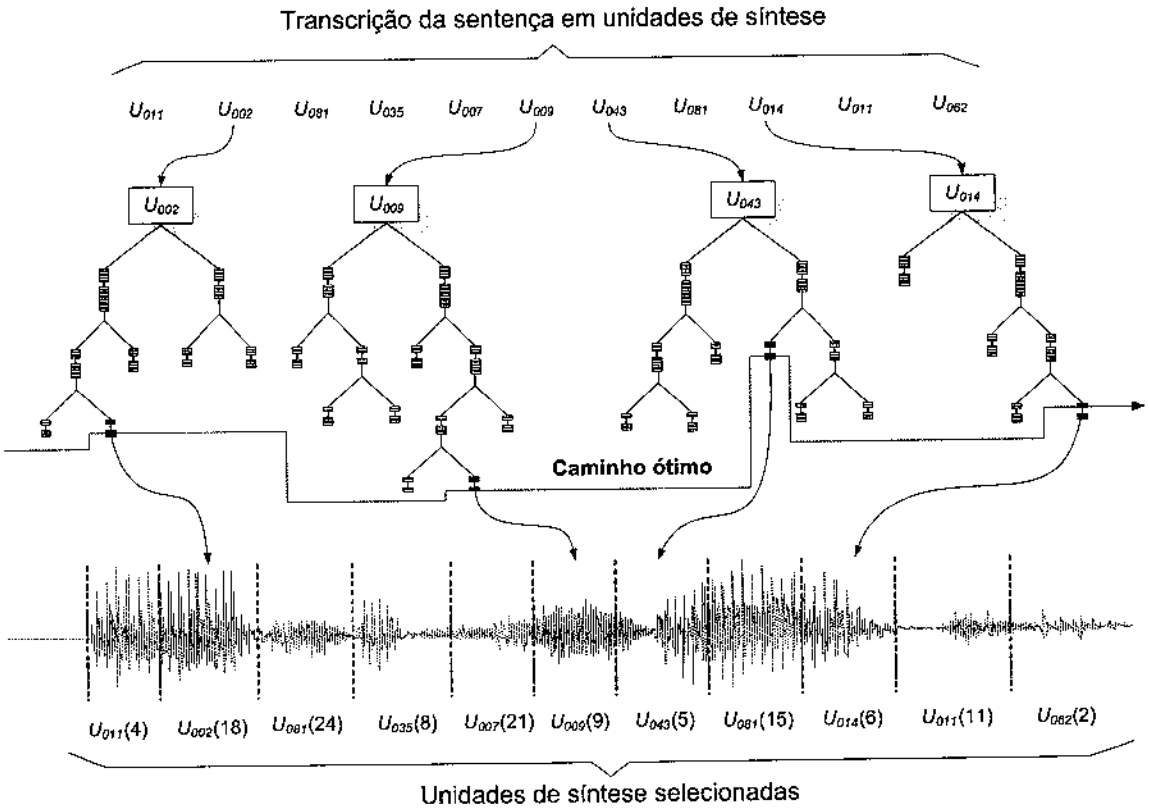


Figura. 4.4: Processo de seleção de unidade de síntese utilizando-se árvores de clusterização e algoritmos de programação dinâmica (DTW ou algoritmo de Viterbi).

$$D\left(U,V\right)_{N_U>N_V} = w_D \cdot d_D + \frac{1}{N_U} \cdot \sum_{i=1}^{N_U} \left[w_C \cdot d_C\left(u_i, v_{j(i)}\right) + w_{F_0} \cdot d_{F_0}\left(u_i, v_{j(i)}\right)\right], \tag{4.9}$$

sendo N_U e N_V os números de quadros de análise das unidades U e V, respectivamente; d_D é um fator de penalidade para diferenças de duração entre as unidades; $d_C\left(u_i, v_{j(i)}\right)$ e $d_{F_0}\left(u_i, v_{j(i)}\right)$ são as distâncias *cepstrais* e de F_0 entre o *i-ésimo* quadro de análise da unidade U e o $j(i)$-*ésimo* quadro de análise da unidade V; e $j(i) = \langle i \cdot \frac{N_V}{N_U} \rangle$ (a expressão $\langle \cdot \rangle$ significa o inteiro mais próximo de). As quantidades w_D, w_C e w_{F_0} correspondem, respectivamente, aos pesos a serem aplicados às penalidades de duração, e às distâncias *cepstrais* e de F_0. A manipulação dos valores destes pesos permite controlar a contribuição relativa das diferenças entre as unidades de síntese em termos de: duração, parâmetros *Cepstrais* e valor de F_0.

O fator de penalidade d_D é estimado pela equação 4.10.

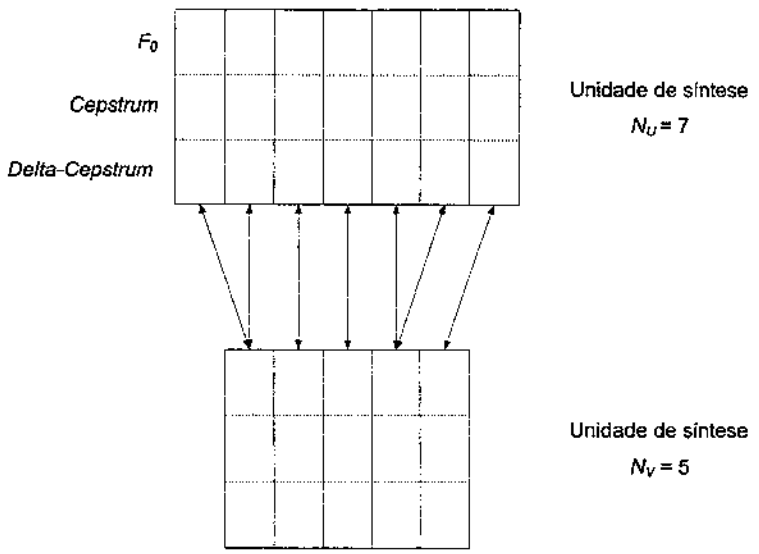


Figura. 4.5: Processo de alinhamento para o cálculo do custo de concatenação.

$$d_D = \frac{N_U}{N_V} - 1, \tag{4.10}$$

No exemplo da Figura 4.5 considera-se $N_U > N_V$.

A distância *cepstral* d_C entre os quadros de análise u_i e $v_{j(i)}$ é estimada utilizando-se a distância de Mahalanobis normalizada entre os vetores *cepstrais* (MFCC e Delta MFCC):

$$d_C\left(u_i, v_{j(i)}\right) = \frac{1}{N} \cdot \sum_{k=1}^{N} \frac{\left(f_k(u_i) - f_k(v_{j(i)})\right)^2}{\sigma_k^2} \tag{4.11}$$

sendo N o número de coeficientes *cepstrais* (12 MFCC + 12 Delta MFCC) e f_k o k-ésimo componente *cepstral*. A variância σ_k^2 deve ser calculada ao longo dos valores de f_k para todas as unidades de síntese a serem clusterizadas. Por exemplo, assumindo que a unidade de síntese a ser clusterizada seja a unidade *metade-de-fone* /aa/, então σ_k^2 deve ser calculada ao longo dos valores de f_k para todos os exemplos da unidade *metade-de-fone* /aa/ pertencentes ao inventário de unidades de síntese.

A distância entre valores de F_0 para as unidades U e V é estimada utilizando-se dois termos: a diferença em valores absolutos de F_0, $d_{|F_0|}$, e a diferença em valores da derivada de F_0, $d_{\Delta F_0}$:

$$d_{F_0}\left(u_i, v_{j(i)}\right) = d_{|F_0|}\left(u_i, v_{j(i)}\right) + d_{\Delta F_0}\left(u_i, v_{j(i)}\right). \tag{4.12}$$

O valor de $d_{|F_0|}$ é estimado em função do nível de sonoridade dos quadros de análise u_i e $v_j(i)$,

segundo a equação 4.13 a seguir:

$$d_{|F_0|}(u_i, v_j(i)) = \begin{cases} \frac{(F_0(u_i) - F_0(v_{j(i)}))}{\sigma^2_{F_0}} & \text{, se ambos os quadros de análise são sonoros} \\ 0 & \text{, se ambos os quadros de análise são não-sonoros} \\ 1 & \text{, caso contrário} \end{cases} \quad (4.13)$$

sendo $\sigma^2_{F_0}$ a variância de F_0 nos quadros de análise sonoros de todos os exemplos da unidade de síntese que está sendo clusterizada (semelhantemente a σ^2_k da equação 4.11).

O valor de $d_{\Delta F_0}$ é estimado pela equação 4.14 a seguir:

$$d_{\Delta F_0}(u_i, v_{j(i)}) = \frac{\Delta F_0(u_i) - \Delta F_0(v_{j(i)})}{\sigma^2_{\Delta F_0}} \quad (4.14)$$

sendo ΔF_0 a diferença entre valores de F_0 associados a quadros de análise consecutivos e sonoros:

$$\Delta F_0(u_i) = \begin{cases} F_0(u_i) - F_0(v_{j(i)}) & \text{, se ambos os quadros são sonoros} \\ 0 & \text{, caso contrário} \end{cases} \quad (4.15)$$

sendo $\sigma^2_{\Delta F_0}$ a variância de ΔF_0, a qual deve ser estimada de forma semelhante à variância $\sigma^2_{F_0}$ da equação 4.13.

4.4.2 Técnicas de Poda

Como a complexidade computacional do processo de seleção de unidades de síntese pode crescer quadraticamente (dependendo do algoritmo de busca utilizado), com o número de unidades de síntese candidatas por alvo fonético-prosódico, então algumas medidas devem ser tomadas para tornar este procedimento de busca possível de ser realizado na prática. Uma maneira de reduzir o custo computacional é fazer com que a base de dados seja a menor possível, através da poda de unidades ao longo dos *clusters* da árvore de clusterização. Diferentes técnicas de poda têm sido investigadas: algumas concentram-se na eliminação de unidades *outliers* e também unidades que apresentam alvos fonético-prosódicos muito próximos entre si (Black and Taylor, 1997). Outras técnicas procuram diversificar as características espectrais nas fronteiras entre as unidades de síntese, com o objetivo de minimizar os custos de concatenação (Hon et al., 1998).

4.4.3 *Lookup Tables*

Em um sistema CTF-SCAUS os custos de concatenação podem, em princípio, ser estimados *offline* e armazenados em tabelas para posterior consulta. Entretanto, se o número de possíveis pares de unidades de síntese for demasiadamente elevado, então o armazenamento de todos os possíveis custos de concatenação pode ser proibitivo. Diante destes fatos, duas estratégias têm sido largamente utilizadas para reduzir o custo computacional relativo ao cálculo dos custos de concatenação durante o processo de seleção de unidades de síntese:

- Os custos mais freqüentes (mais prováveis de ocorrerem durante a síntese) são calculados *offline* e armazenados em tabelas para posterior consulta (Beutnagel et al., 1999b);
- O espaço de unidades de síntese é quantizado vetorialmente gerando classes de unidades de síntese. Em seguida, uma tabela com todas as distâncias entre estas classes de unidades de síntese é armazenada para posterior consulta (Beutnagel et al., 1999a), (Coorman et al., 2000).

4.5 Corpus e Inventário de Unidades de Síntese

A Figura 4.6 ilustra os principais passos para a construção de um corpus de fala voltado a sistema CTF-SCAUS. Além disso, mostra a derivação do inventário de unidades de síntese a partir do corpus de fala devidamente segmentado e etiquetado lingüisticamente.

4.5.1 Definição da Aplicação

Se o corpus a ser projetado será utilizado no treinamento de um sistema CTF de domínio restrito, sistema CTF-SCAUS-DR, então este corpus deve ser cuidadosamente projetado para maximizar a cobertura dos eventos fonéticos e prosódicos mais prováveis de ocorrerem neste domínio especificado.

Por outro lado, se o corpus a ser projetado será utilizado no treinamento de um sistema CTF de domínio irrestrito, sistema CTF-SCAUS-DI, então o corpus deve ser cuidadosamente projetado para apresentar um balanceamento de todos os eventos fonéticos e prosódicos pertencentes à língua que se deseja modelar.

4.5.2 Projeto

O conjunto de sentenças a ser gravado deve contemplar o maior número possível de eventos fonéticos e prosódicos, que apresentem uma alta probabilidade de ocorrerem durante o processo de síntese (conforme o domínio previamente definido na sub-seção 4.5.1). Para obter um corpus de fala que apresente esta característica, algoritmos gulosos e algoritmos genéticos (Monteiro et al., 2005) têm sido utilizados para selecionar um conjunto de sentenças (em torno de 1500 a 5000 sentenças, dependendo do tipo de unidade de síntese a ser utilizada e também do domínio - restrito ou irrestrito - a ser modelado) de um corpus de milhares de sentenças (por exemplo, várias edições do jornal Folha de São Paulo), que maximize a cobertura dos eventos fonético-acústicos e prosódicos desejados.

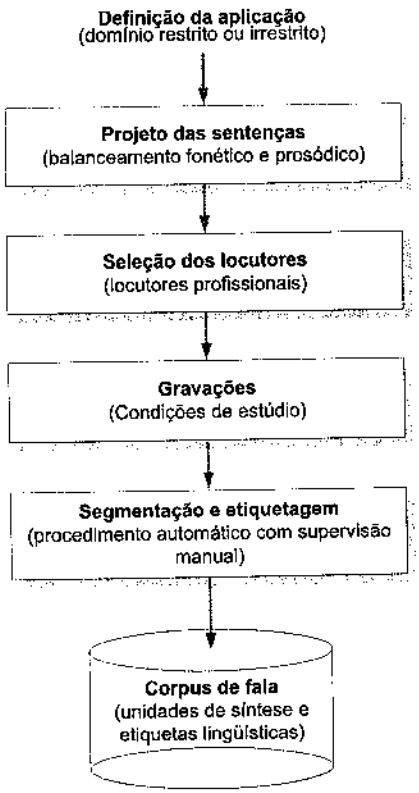


Figura. 4.6: Projeto e aquisição da base de dados de unidades de síntese.

4.5.3 Seleção de Locutores

A seleção dos locutores deve contemplar, primordialmente, aspectos como qualidade de voz e estilo de elocução. Entretanto, testes rigorosos devem ser realizados para garantir que as vozes dos locutores escolhidos sejam adequadamente modeladas/manipuladas pelas técnicas de processamento de sinais a serem utilizadas no módulo de *Back-End*.

4.5.4 Gravações

As gravações devem ser realizadas com equipamentos profissionais e em condições de estúdio. Elas devem ser realizadas no modo estéreo, adquirindo-se em um canal a fala do locutor e no outro a dinâmica dos pulsos glotais (sinal laringográfico (Huang et al., 2001)). A aquisição da dinâmica dos pulsos glotais pode ser realizada através do uso de um eletroglotógrafo digital. A informação sobre a dinâmica dos pulsos glotais é de fundamental importância para algoritmos de síntese, do módulo de *Back-End*, que operam de forma síncrona com o período fundamental (inverso de F_0) e que também façam uso dos instantes de fechamento e/ou abertura da glote.

4.5.5 Segmentação e Etiquetagem

As segmentações fonéticas das bases de unidades de síntese são geralmente realizadas utilizado-se segmentadores automáticos, baseados na tecnologia de HMM (*Hidden Markov Models*) (Adell and Bonafonte, 2004), seguido de uma inspeção manual para verificar a qualidade da segmentação e corrigir erros sistemáticos.

A etiquetagem lingüística (etiquetas sintáticas, semânticas, fonológicas e prosódicas) é, normalmente, realizada utilizando-se procedimentos semi-automáticos, baseados em métodos estatísticos tais como CART (*Classification and Regression Trees*), QMTI (*Quantification Method Type I*), HMM e ANN, seguidos de uma inspeção/correção manual por lingüistas especializados.

4.5.6 Inventário de Unidades de Síntese

Montagem

A montagem do inventário de unidades de síntese se inicia com a extração das unidades de síntese bem como das informações sintáticas, semânticas, fonológicas e prosódicas associadas a cada uma delas, do corpus de fala (que se encontra devidamente segmentado e etiquetado). Em seguida os vários exemplares de cada unidade de síntese são submetidos ao processo de clusterização e poda descritos na seção 4.4. Por último as unidades de síntese que se encontram nos clusters terminais das árvores de clusterização são compactadas utilizando técnicas de codificação de fala.

Compactação

A área de codificação de fala a baixas taxas, para sistemas de telefonia fixa, telefonia móvel ou internet, pode ser considerada uma área já bem estabelecida e dominada. Uma divisão clássica da área de codificação de fala, segundo Quatieri (Quatieri, 2002), consiste em:

- Codificação no Domínio da Freqüência
  - *Subband Coding*,
  - *Sinusoidal Coding*,
  - *Multi-Band Excitation Vocoder* (MBE).
- Codificação Paramétrica
  - LPC *Vocoder* (*Basic Linear Predictive Coding*),
  - *Mixed Excitation* LPC (MELP).
- Codificação do Resíduo LPC
  - *Multi-Pulse Linear Prediction*,
  - *Code-Excited Linear Prediction* (CELP).

Técnicas para compressão de bases de unidades de síntese, geralmente, baseiam-se em uma ou mais das técnicas apresentadas acima e, além disso, procuram explorar algumas das especificidades associadas ao inventário de unidades de síntese:

- O inventário de unidades de síntese é proveniente de um único locutor.
- O inventário de unidades de síntese encontra-se completamente etiquetado foneticamente.
- Alguns algoritmos de *Back-End* como HNM (*Harmonic + Noise Model*), LPC (*Linear Prediction Coding*) e Modelos Senoidais são paramétricos e, portanto, a parametrização intrínseca destes algoritmos deve ser levada em consideração durante o procedimento de codificação.
- O sinal gravado é de altíssima qualidade (locutor e estúdio profissionais, etc).
- O sinal pode ser gravado a taxas de amostragem acima do necessário, operado digitalmente para a extração dos parâmetros necessários e, somente após isto, ser reamostrado digitalmente para a taxa de amostragem desejada.
- O sinal laringográfico normalmente encontra-se disponível (com indicações precisas sobre os instantes de abertura e fechamento da glote).

4.6 Considerações Finais

Este capítulo apresentou um revisão sobre as principais operações envolvidas no módulo de seleção automática de unidades de síntese de sistemas CTF-SCAUS. Discutiram-se os fundamentos do processo de busca/seleção de unidades de síntese ótimas através da minimização de funções de custo fonético-prosódicas e funções de custo de concatenação entre unidades de síntese. Apresentou-se o método de clusterização de unidades de síntese como uma alternativa para reduzir o tamanho do inventário de unidades e também para minimizar o tempo de busca pelas unidades de síntese ótimas. Algumas das métricas mais comumente utilizadas para medir a distância entre unidades de síntese foram descritas em detalhes. Para finalizar, foram levantadas algumas considerações sobre o projeto, aquisição, segmentação e etiquetagem lingüística de corpora de fala para sistemas CTF-SCAUS.

Apesar do vasto levantamento bibliográfico realizado nesta Tese não foi possível encontrar, até o presente momento, nenhum trabalho publicado (artigos, livros, Teses, etc) sobre seleção automática de unidades de síntese para o português brasileiro (PB). Este mesmo levantamento bibliográfico também revelou poucas iniciativas para a construção de corpora de fala para o PB, voltados ao desenvolvimento de trabalhos na área de CTF empregando a tecnologia SCAUS (Cirigliano et al., 2005). Além disso, dos poucos corpora de fala encontrados, para o PB, nenhum deles se encontra, até o presente momento, disponível para uso irrestrito pela comunidade acadêmica brasileira.

Capítulo 5

Algoritmo LDM-GA: Formulação Teórica

5.1 Introdução

Um dos maiores desafios no treinamento de sistemas CTF-SCAUS a partir de corpora de fala (empregando técnicas estatísticas - *Data-Driven*), é a modelagem robusta de eventos sintáticos, semânticos, fonológicos e fonético-acústicos que apresentam baixa freqüência de ocorrência na fala natural (Möbius, 2001). Os problemas causados por eventos lingüísticos e da fala que possuem distribuições de freqüências extremamente desbalanceadas são freqüentemente subestimados ou mesmo desconhecidos (Möbius, 2001).

Vários fenômenos lingüísticos e da fala são caracterizados segundo (Möbius, 2001), como pertencentes à classe de distribuições denominada LNRE, *Large Number of Rare Events*. Segundo (Möbius, 2001) a classe LNRE é caracterizada por distribuições extremamente desbalanceadas: enquanto alguns membros da classe apresentam uma alta freqüência de ocorrência, a sua grande maioria apresenta freqüências de ocorrência extremamente baixas. Como o número de eventos com baixíssima freqüência de ocorrência é extremamente elevado, então Möbius argumenta que a probabilidade de que pelo menos um destes eventos possa ocorrer ao longo de uma sentença é relativamente alta. Möbius relata ter observado, em trabalhos sobre síntese de fala, para várias línguas, distribuições LNRE em três diferentes contextos: em análises lingüísticas (módulo de *Front-End*), na modelagem da duração segmental e do contorno entoacional (módulo prosódico) e no projeto de corpora de fala para treinamento de sistemas baseados na tecnologia SCAUS.

No contexto de análises lingüísticas, Möbius argumenta que a distribuição do número de ocorrências de sílabas em línguas com estruturas silábicas complexas (como por exemplo inglês e alemão) apresenta características LNRE. Nestas línguas, apenas algumas centenas de sílabas ocorrem muito freqüentemente na fala natural. O restante das sílabas (sua grande maioria), é raramente utilizado na fala natural.

No tocante à modelagem prosódica, Möbius reporta que distribuições com características LNRE também são observadas na modelagem da duração segmental da fala. Möbius e Santen (Möbius and Santen, 1996), (Santen, 1994), verificaram que o conjunto de atributos lingüísticos, que têm algum efeito

sobre a duração dos segmentos fonéticos, define um espaço fatorial de elevada dimensão e que a grande maioria destes atributos lingüísticos geralmente apresenta uma freqüência de ocorrência extremamente reduzida.

Segundo Möbius, o projeto de corpora de fala para treinamento de sistemas CTF-SCAUS, também defronta-se com o problema de distribuições LNRE. Möbius e colegas (Möbius, 2001) observaram que o espaço fatorial de atributos lingüísticos utilizados na seleção das unidades de síntese é extremente elevado, e a grande maioria destes atributos geralmente apresenta uma reduzida freqüência de ocorrência.

Com o objetivo de lidar com o problema de distribuições LNRE este capítulo apresenta um novo método para análise exploratória de dados lingüísticos. Este novo método é baseado em Algoritmos Genéticos - AG, e será denominado LDM-GA (*Linguistic Data Mining using Genetic Algorithm*). A motivação inicial para a criação do algoritmo LDM-GA surgiu de uma necessidade prática dos Laboratórios de Tecnologia da Fala da Toshiba em Cambridge, na Inglaterra, em desenvolver um algoritmo eficiente para análise exploratória de dados lingüísticos (dados simbólicos/categorias), que pudesse ser utilizado tanto em problemas de regressão ou classificação quanto em análises puramente lingüísticas. Além disto, este novo método não deveria fazer qualquer tipo de suposição quanto às características dos dados a serem analisados, diferentemente, por exemplo, do método de Análise de Variância - ANOVA (Jobson, 1991). Apesar de o método LDM-GA poder ser aplicado a problemas de classificação, o que englobaria problemas de projeto de corpora e seleção automática de unidades de síntese, esta Tese se restringirá a apresentar o algoritmo LDM-GA apenas no contexto da modelagem linear robusta da duração segmental da fala.

Este capítulo concentra-se apenas na formulação teórica do algoritmo LDM-GA. Análises e avaliações dos resultados da aplicação deste algoritmo ao problema de predição da duração segmental da fala serão apresentados no capítulo 6. Outro aspecto importante sobre este capítulo é que não será realizada nenhuma revisão sobre fundamentos em Algoritmos Genéticos - GA. Assume-se que todas as operações envolvendo GA, a serem utilizadas neste capítulo, são clássicas e que existe uma vasta gama de livros tratando deste assunto, como por exemplo (Bäck et al., 2000a), (Bäck et al., 2000b) e (Mitchell, 1998).

Para uma melhor compreensão do problema de predição da duração segmental da fala, a seção 5.2 deste capítulo é dedicada à formulação deste problema, introduzindo definições e conceitos fundamentais, bem como discutindo alguns dos principais problemas encontrados na modelagem da duração segmental da fala. A seção 5.3 descreve em detalhes o corpus utilizado nos experimentos realizados, dando ênfase aos *fatores* lingüísticos empregados na predição da duração dos segmentos fonéticos e à elevada dimensionalidade do espaço fatorial associado a todas as possíveis combinações destes *fatores*. O método utilizado para estimar os modelos de regressão, QMTI - *Quantification Method Type I* - é apresentado na seção 5.4. A seção 5.5 apresenta uma descrição detalhada das duas principais etapas do algoritmo LDM-GA: (1) Estimação das classes de fones por *clusterização* hierárquica e binária e (2) estimação das topologias ótimas (fatores lingüísticos ótimos) para os respectivos modelos de regressão. Finalmente, a seção 5.6 encerra o capítulo com algumas considerações finais.

5.2 Formulação do Problema de Predição da Duração Segmental da Fala

5.2.1 Conceitos Fundamentais e Definições

O problema de predição da duração segmental de fones pode ser compreendido como um mapeamento do espaço de *fatores* lingüísticos, extraídos da sentença que se deseja converter em fala, no intervalo de durações segmentais dos fones.

Nesta Tese cada *fator* lingüístico F_i é representado por meio de um vetor cujas respectivas posições são denomidadas *níveis* e cada um destes *níveis* corresponderá uma determinada categoria lingüística. Por exemplo, se F_i representa o *fator Categoria Gramatical*, então:

$$Categoria\ Gramatical \Leftrightarrow F_i = [Verbo, Substantivo, Adjetivo, \cdots] \tag{5.1}$$

O espaço fatorial de todos os *fatores* lingüísticos disponíveis $F_1, F_2, F_3, \cdots, F_N$ será denominado $\mathbf{F}$, e dado pelo seguinte produto cartesiano:

$$\mathbf{F} = F_1 \times F_2 \times F_3 \times \cdots \times F_N \tag{5.2}$$

O intervalo duracional, que compreende todos os possíveis valores de duração assumidos pelos segmentos fonéticos, será representado por $\mathbf{D}$. O intervalo $\mathbf{D}$ estará contido no eixo dos números reais $\Re$ e limitado por um valor mínimo e máximo de duração. Em outras palavras, para todo valor de duração $dur \in \mathbf{D}$:

$$dur_{min} < dur \leq dur_{max} \tag{5.3}$$

O objetivo da modelagem da duração segmental da fala consiste na estimativa de uma função DUR capaz de mapear qualquer valor do *espaço fatorial* $\mathbf{F}$ no *intervalo duracional* $\mathbf{D}$.

$$DUR : \mathbf{F} \rightarrow \mathbf{D} \tag{5.4}$$

Removendo-se o valor médio do intervalo $\mathbf{D}$, pode-se definir um novo intervalo $\mathbf{ZD}$ dado por:

$$\mathbf{ZD} = \mathbf{D} - mean\,(\mathbf{D}) \tag{5.5}$$

Utilizando-se este novo intervalo $\mathbf{ZD}$, pode-se definir o mapeamento $ZDUR$ a seguir:

$$ZDUR : \mathbf{F} \rightarrow \mathbf{ZD} \tag{5.6}$$

No caso de um modelo de regressão aditivo simples, o mapeamento $ZDUR$ pode ser decomposto em N termos A_i, um para cada um dos N fatores F_i:

$$A_i : F_i \rightarrow \mathbf{ZD} \tag{5.7}$$

Combinando aditivamente todos os termos A_i e utilizando-se a relação entre **D** e **ZD** dada por 5.5, têm-se:

$$DUR : \mathbf{F} \rightarrow \mathbf{D} \Leftrightarrow mean\,(\mathbf{D}) + \sum_{i=1}^{N} A_i\,(F_i) \tag{5.8}$$

A contribuição de cada termo A_i para o modelo $DUR(\mathbf{F})$ será denominada *efeito* do fator F_i.

5.2.2 Problemas na Modelagem da Duração Segmental da Fala

Algumas das principais razões que dificultam o processo de modelagem e estimação da duração segmental da fala são:

1. Desbalanceamento da base de dados (*Database Imbalance and Missing Data Problems*);
2. Interações complexas entre os fatores lingüísticos;
3. Ausência de uma teoria lingüística elaborada que seja plausível de ser modelada matematicamente, e que já tenha sido rigorosamente avaliada e testada no contexto de sistemas CTF.

O algoritmo LDM-GA, introduzido neste capítulo, será utilizado para lidar com o primeiro problema listado acima, isto é, com o problema de desbalanceamento de base de dados. O método LDM-GA será empregado para melhorar a robustez de modelos de regressão linear de múltiplas variáveis. Foram escolhidos modelos de regressão linear de múltiplas variáveis devido à simplicidade destes modelos e também por serem considerados modelos totalmente conectados (Hansa and Sagisaka, 2004), diferentemente, por exemplo, de árvores de regressão - *Regression Trees* - *RT* (Breiman et al., 1993). Apesar de ser aplicado a modelos de regressão linear, o método LDM-GA pode ser facilmente adaptado para operar com outros modelos de regressão, como por exemplo *Sum-of-Products* - SoP, (Santen, 1994), (Sproat, 1998) e *Artificial Neural Networks* - ANN, (Haykin, 1994).

5.3 Descrição e Análise do Corpus Utilizado

A base de dados utilizada para avaliar o algoritmo LDM-GA foi cedida pelo STG-CRL (*Speech Technology Group - Cambridge Research Laboratory*) da Toshiba em Cambridge, na Inglaterra. Trata-se de uma base de dados de um único locutor masculino e falante do inglês americano e compreende 1474 sentenças, 18172 palavras e 60136 fones. Cada fone da base de dados possui uma descrição com formato dado pela equação 5.9

$$O(i) = \langle F_1, F_2, F_3, F_4, \cdots, F_{14}, DurPh \rangle \tag{5.9}$$

O primeiro fator lingüístico F_1 indica a identidade do fone a ser modelado - phID. Os *fatores* F_i com $i = 2, 3, \cdots, 14$, representam os parâmetros *fatores* lingüísticos (sintáticos, fonológicos, morfológicos,$\cdots$) a serem utilizados. O termo *DurPh* representa a duração em milissegundos do fone a ser modelado. Nesta base de dados, todos os *fatores* lingüísticos foram derivados automaticamente e verificados manualmente.

Uma descrição detalhada dos *fatores* lingüísticos e seus respectivos níveis é apresentada na subseção 5.3.1. O conjunto de fones utilizado, segundo a notação do IPA (*International Phonetic Alphabet*) e dos Laboratórios da Toshiba, é apresentado na subseção 5.3.2. Uma análise sobre o balanceamento da base de dados é apresentada na subseção 5.3.3. Finalmente, na subseção 5.3.4 é apresentada uma análise sobre a dimensionalidade do espaço fatorial dos *fatores* lingüísticos.

5.3.1 Fatores e Níveis

A lista contendo a descrição dos 14 *fatores* lingüísticos utilizados é apresentada na Tabela 5.1. É importante notar que o primeiro *fator* F_1 é associado à identidade do fone. Conforme será apresentado no capítulo 6 sobre os resultados experimentais, o termo F_1 não fará parte dos modelos de regressão que forem estimados exclusivamente para um único fone, isto é, no caso em que houver uma relação biunívoca entre os modelos de regressão e os fones, o termo F_1 se fará desnecessário. Por outro lado, quando forem estimados modelos de regressão para classes de fones (um único modelo de regressão para classes com mais de um fone) o termo F_1 necessariamente terá que fazer parte deste modelo.

A Tabela 5.2 apresenta os níveis assumidos pelos *fatores* lingüísticos descritos na Tabela 5.1. Pode ser verificado da Tabela 5.2 que o número total de níveis associados aos 14 *fatores* lingüísticos (incluindo a identidade dos phones - phID) é igual a 168. No caso do termo F_1 não ser considerado, então o número total de níveis é reduzido a 123. Estas dimensões de 168 ou 123 irão definir o número máximo de coeficientes de regressão por modelo (modelo por fones ou classes-de-fones), conforme será discutido ao longo deste capítulo.

Com o objetivo de ilustrar o formato dos dados presentes na base de dados, a Tabela 5.3 apresenta alguns exemplos do fone [@].

Tabela. 5.1: Descrição dos 14 *fatores* lingüísticos utilizados nos modelos de predição da duração segmental dos fones.

Fatores	Descrição dos fatores
F1 : phID	Identidade do fone corrente
F2: PosInSyll	Posição do fone corrente em relação à sílaba acentuada da palavra corrente
F3: PrevPh	Classe do fone anterior
F4: NextPh	Classe do fone seguinte
F5: NNextPh	Classe do fone após o fone seguinte
F6: PoS	Etiqueta morfossintática
F7: ACC	Nível de ênfase da palavra corrente
F8: NSyll	Número de sílabas na palavra corrente
F9: DistEnd	Distância até o final da palavra (em número de sílabas)
F10: DistStress	Distância até a sílaba acentuada da próxima palavra (em número de sílabas)
F11: NextPause	Distância até a próxima pausa (em número de sílabas)
F12: PrevPause	Distância a partir da pausa anterior (em número de sílabas)
F13: Chunk	Distância até o final do grupo acentual corrente (em número de sílabas)
F14: PosInWord	Posição da sílaba que contém o fone em análise dentro da palavra corrente

5.3.2 Conjunto de Fones Utilizados

O conjunto dos 45 fones utilizados na avaliação do algoritmo LDM-GA é apresentado na Tabela 5.4. São utilizadas duas notações fonéticas na Tabela 5.4, a notação fonética internacional - IPA e a notação adotada pelos Laboratórios da Toshiba em Cambridge, Inglaterra. Como todas as análises e resultados apresentados neste capítulo fazem uso da notação da Toshiba, então a Tabela 5.4 será de extrema importância para a identificação dos símbolos fonéticos utilizados.

5.3.3 Freqüência de Ocorrência dos Fones na Base de Dados

A Tabela 5.5 apresenta a quantidade de exemplos, na base de dados, de cada um dos 45 fones da Tabela 5.4. Pode ser verificado que, enquanto os fones [i] e [n] apresentam em torno de 4000 exemplos, os fones [oi] e [zh] ocorrem, aproximadamente, apenas 100 vezes ao longo de toda a base de dados. Levando-se em consideração a dimensionalidade do espaço de *fatores* lingüísticos (o número total de níveis pode alcançar 123 ou 168), pode-se verificar o elevado grau de esparsidade dos dados associados aos fones [zh]e [oi].

A Figura 5.1 apresenta as mesmas informações da Tabela 5.5, porém em termos de freqüência relativa de ocorrência de cada um dos 45 fones da Tabela 5.4. Pode ser verificado na Figura 5.1 que o número de exemplos para o fone [@] corresponde a 6.2% de todos os exemplos contidos na base de dados. Por outro lado, o número de exemplos do fone [oi] corresponde a apenas 0.22% de toda a base de dados.

Tabela. 5.2: Descrição dos níveis associados aos 14 *fatores* lingüísticos utilizados.

Fatores	Níveis Assumidos
F1: phID	@, AR, ER, H, OR, Q, aa, ae, ai, au, b, ccc, ch, d, dh, dx, e, ei, f, g, i, ii, jh, k, l, m, n, ng, oi, oo, ou, p, r, s, sh, t, th, u, uh, uu, v, w, y, z, zh
F2: PosInSyll	pre (antes da sílaba lexical acentuada), aft (após a sílaba lexical acentuada), mid (na sílaba lexical acentuada), none (no caso da palavra não ser acentuada)
F3: PrevPh	ShortVowel, LongVowel, Diphthong, VC1 (fricativas sonoras), VC2 (nasais), VPlosive (plosiva sonora), UPlosive (plosiva não-sonora), Closure, UC (consoante não-sonora), Sil, none
F4: NextPh	ShortVowel, LongVowel, Diphthong, VC1, VC2, VPlosive, UPlosive, Closure, UC, Sil, none
F5: NNextPh	ShortVowel, LongVowel, Diphthong, VC1, VC2, VPlosive, UPlosive, Closure, UC, Sil, none
F6: PoS	n (substantivo), nam, adj (adjetivo), adv (advérbio), itf, deny, dig, pron2 (pronome relativo), vi (verbo intransitivo), vs, vt (verbo transitivo), bv, NULL, w, pnc, nud, int (interjeição), prep (preposição), freq
F7: ACC	deacc (sem ênfase), acc (com ênfase), high (fortemente enfatizada)
F8: NSyll	0,1,2,3,4,5,6,7,8,9
F9: DistEnd	0,1,2,3,4,5,6,7,8,9
F10: DistStress	0,1,2,3,4,5,6,7,8,9,none
F11: NextPause	0,1,2,3,4,5,6,7,8,9
F12: PrevPause	0,1,2,3,4,5,6,7,8,9
F13: Chunk	0,1,2,3,4,5,6,7,8,9
F14: PosInWord	start (primeira sílaba da palavra), end (última sílaba da palavra), middle (qualquer outra sílaba na palavra)

Tabela. 5.3: Exemplos de *fatores* lingüísticos associados ao fone [@] (to allow).

F_1	F_2	F_3	F_4	F_5	F_6	F_7	F_8	F_9	F_{10}	F_{11}	F_{12}	F_{13}	F_{14}	DurPh (ms)
@	mid	UC	UC	SV	prep	deacc	1	1	0	4	9	4	start	34
@	aft	VC1	UC	UC	N	high	2	0	2	9	3	9	end	64
@	pre	UC	UC	SV	N	acc	4	4	4	9	4	9	start	45
@	mid	UC	UP	SV	freq	deacc	1	1	1	4	9	4	start	48
@	aft	UP	none	none	N	acc	2	0	none	1	9	0	end	124
@	aft	VC1	VC1	UP	adv	acc	3	2	2	5	2	5	mid	30
@	pre	UP	UC	SV	N	acc	2	2	2	9	3	2	start	30
@	mid	UC	VC1	SV	freq	deacc	1	1	0	9	1	4	start	24
@	aft	UP	VC1	none	nam	acc	2	1	none	0	9	1	end	46

Tabela. 5.4: Descrição dos 45 fones utilizados. Notações no formato da Toshiba e do *International Phonetic Alphabet* - IPA.

Toshiba	IPA	Exemplo		Toshiba	IPA	Exemplo
ii	iː	ease		ai	aɪ	rise
i	ɪ	pit		au	aʊ	house
e	ɛ	pet		oi	ɔɪ	noise
ae	æ	pat		ei	eɪ	raise
aa	ɑː	calm		ou	oʊ	nouse
uh	ʌ	cut		AR	ɑ(ɾ)	far
oo	ɔː	cause		OR	ɔ(ɾ)	port
uu	uː	lose		p	p	pin
u	ʊ	put		t	t	tin
ER	ɚ	mother		k	k	kin
@	ə	allow		b	b	bin
d	d	din		jh	ʤ	gin
g	g	give		H	h	hit
f	f	fin		m	m	mock
v	v	van		n	n	not
s	s	sir		ng	ŋ	doing
z	z	zoo		l	l	left
sh	ʃ	shin		r	r	right
zh	ʒ	measure		dx	ɾ	writer,rider
th	θ	thin		w	w	wasp
dh	ð	this		y	j	yes
ch	ʧ	chin		Q	glotal stop	
ccc	Intervalo de oclusão					

Tabela. 5.5: Número de exemplos na base de dados para cada um dos 45 fones

#	Fone	$N°$ ex.	#	Fone	$N°$ ex.	#	Fone	$N°$ ex.	#	Fone	$N°$ ex.
1	i	4326	13	r	1667	25	v	1107	37	sh	483
2	n	3858	14	k	1622	26	H	1034	38	th	405
3	@	3730	15	e	1592	27	ei	987	39	au	387
4	t	3400	16	ER	1581	28	u	981	40	jh	331
5	s	2771	17	ccc	1519	29	uu	815	41	ch	323
6	l	2287	18	Q	1459	30	ou	772	42	AR	320
7	ii	1941	19	uh	1360	31	aa	676	43	dx	316
8	d	1940	20	w	1353	32	ng	653	44	oi	132
9	z	1860	21	ai	1242	33	y	570	45	zh	89
10	dh	1831	22	p	1153	34	OR	515			
11	ae	1770	23	b	1139	35	oo	497			
12	m	1711	24	f	1137	36	g	494			

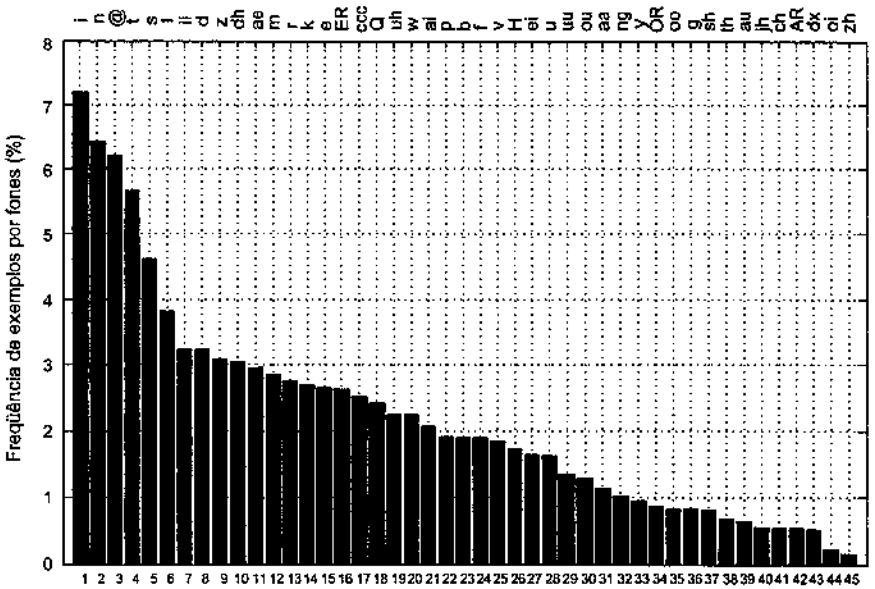


Figura. 5.1: Freqüência dos 45 fones presentes na base de dados. Os fones seguem a mesma ordenação da Tabela 5.5

5.3.4 Dimensionalidade do Espaço de *Fatores* Lingüísticos

A partir da Tabela 5.2, que descreve o número de níveis associados a cada um dos *fatores* lingüísticos, pode ser verificado que o espaço combinatorial $\mathbf{F} = F_1 \times F_2 \times F_3 \times \cdots \times F_{14}$ terá uma dimensão da ordem de 10^{13}.

$$Dim\,(\mathbf{F}) = 45 \cdot 4 \cdot 11 \cdot 11 \cdot 11 \cdot 19 \cdot 3 \cdot 10 \cdot 10 \cdot 11 \cdot 10 \cdot 10 \cdot 10 \cdot 3 = 4.5 \cdot 10^{13} \tag{5.10}$$

Entretanto, se as identidades dos fones *phID* não forem consideradas como um dos *fatores* pertencentes ao espaço combinatorial $\mathbf{F}$, então a dimensão de $\mathbf{F}$ será da ordem de 10^{12}.

$$Dim\,(\mathbf{F}) = 4 \cdot 11 \cdot 11 \cdot 11 \cdot 19 \cdot 3 \cdot 10 \cdot 10 \cdot 11 \cdot 10 \cdot 10 \cdot 10 \cdot 3 = 1.0 \cdot 10^{12} \tag{5.11}$$

Este elevado número de possíveis combinações de *fatores* (considerando-se ou não a identidade fonética como um *fator*) é um dos grandes desafios na modelagem robusta da estrutura duracional da fala.

5.3.5 Histogramas dos Fones da Base de Dados

As Figuras A.1, A.2, A.3, A.4 e A.5 do Apêndice apresentam os histogramas das durações dos 45 fones da Tabela 5.4. Algumas observações que podem ser inferidas, a partir das distribuições dos fones

associados a estes histogramas, são:

- As distribuições dos fones geralmente são assimétricas, com uma tendência a terem longas caudas à direita, aproximando-se muitas vezes de distribuições *gamma*.
- Algumas distribuições apresentam um comportamento bimodal (ou mesmo trimodal) como, por exemplo, as distribuições dos fones [r] e [p].
- Parece ser evidente que algumas distribuições apresentam problemas de falta de dados. Isto ocorre, por exemplo, com os fones [ah], [oi], [jh] e [AR].

5.4 Regressão Linear Multivariável Empregando QMTI

5.4.1 Quantificação e Modelagem

Neste trabalho foi utilizado o método de regressão linear a partir de variáveis nominais, proposto por Hayashi e denominado *Quantification Method Type I* - QMTI (Hayashi, 1950). Quando aplicado à modelagem da duração segmental da fala, este método prediz estatisticamente a duração desejada de um determinado fone (variável contínua), a partir de *fatores/atributos* (variáveis categoriais/simbólicas) associados ao contexto lingüístico em que este fone se encontra. O método QMTI prediz a duração do $k-$*ésimo* fone quando inserido no contexto lingüístico Ctx_Ling, a partir da equação 5.12, a seguir:

$$\widehat{durPh_k}\left(Ctx\_Ling\right) = \overline{durPh_k} + \sum_{i=1}^{NF_k} \sum_{l=1}^{NL(i)} a_k\left(i,l\right) \cdot \delta\left(i,l,Ctx\_Ling\right) \tag{5.12}$$

sendo NF_k o número total de *fatores* lingüísticos para o $k-$*ésimo* fone; $NL(i)$ o número total de *níveis* lingüísticos associados ao $i-$*ésimo fator* lingüístico; $\widehat{durPh_k}\left(Ctx\_Ling\right)$ o valor da duração predito para o $k-$*ésimo* fone, no contexto lingüístico Ctx_Ling; $\overline{durPh_k}$ o valor médio da duração do $k-$*ésimo* fone (estimado a partir de todos os exemplos da base de dados associados a este fone); $a_k(i,l)$ representa os coeficientes de regressão para o $k-$*ésimo* fone; $\delta(i,l,Ctx\_Ling)$ é uma função característica associada ao contexto lingüístico Ctx_Ling e representada pela equação 5.13 a seguir:

$$\delta(i,l,Ctx\_Ling) = \begin{cases} 1: & \text{Se o } l-\textit{ésimo} \text{ nível do } i-\textit{ésimo fator} \text{ lingüístico,} \\ & \text{associado ao contexto } Ctx\_Ling, \text{ estiver presente.} \\ 0: & \text{Caso contrário.} \end{cases} \tag{5.13}$$

A função característica $\delta(i,l,Ctx\_Ling)$ é a responsável pela *quantificação* dos *fatores* lingüísticos, originalmente representados por variáveis simbólicas, em valores binários (0 e 1).

O algoritmo QMTI estima os coeficientes de regressão $a_k(i,l)$ através da minimização do erro quadrático ϵ, definido na equação 5.14:

$$\epsilon = \sum_{j=1}^{NeT_k} \left(\widehat{durPh_k}(j) - durPh_k(j) \right)^2 \tag{5.14}$$

sendo NeT_k o número de exemplos de treinamento do $k-ésimo$ fone e sendo $durPh_k(j)$ e $\widehat{durPh_k}(j)$, respectivamente, os valores de duração esperado e predito (segundo a equação 5.12), para o $j-ésimo$ exemplo de treinamento do $k-ésimo$ fone. É importante enfatizar que cada exemplo presente na base de dados corresponde a um determinado contexto lingüístico Ctx_Ling.

Os coeficientes $a_k(i,l)$ serão os responsáveis pelo mapeamento das informações presentes no espaço dos atributos lingüísticos $\mathbf{F}$ (após sua quantificação através da função $\delta(i,l,Ctx\_Ling)$), nos valores de duração do $k-ésimo$ fone.

5.4.2 Efeito dos Fatores Lingüísticos

Comparando-se a equação 5.8 com a equação 5.12, verifica-se que:

$$\begin{aligned} \widehat{durPh_k}\left(Ctx\_Ling\right) &= mean\left(\mathbf{D}\right) + \sum_{i=1}^{NF_k} A_i\left(F_i\right) \\ &= \overline{durPh_k} + \sum_{i=1}^{NF_k} \sum_{l=1}^{NL(i)} a_k\left(i,l\right) \cdot \delta_{Ctx\_Ling}\left(i,l\right) \end{aligned} \tag{5.15}$$

Logo, pode-se verificar que, para o $k-ésimo$ fone, o efeito do $i-ésimo\ fator$ lingüístico A_i será dado por:

$$A_i(F_i) = \sum_{l=1}^{NL(i)} a_k\left(i,l\right) \cdot \delta\left(i,l,Ctx\_Ling\right) \tag{5.16}$$

5.5 Algoritmo LDM-GA

5.5.1 Princípios de Operação do Algoritmo LDM-GA

A abordagem empregada pelo algoritmo LDM-GA para lidar com o problema de desbalanceamento de base de dados, para modelagem da duração segmental da fala, pode ser dividida em duas etapas.

A primeira etapa consiste na busca por fones que apresentem "características duracionais" semelhantes e que possam ser agrupados em classes e modelados conjuntamente em um único modelo de regressão. O principal objetivo deste agrupamento de fones em classes é o aumento da quantidade de exemplos de treinamento disponíveis, por modelo de regressão. A solução empregada pelo algoritmo

LDM-GA, para esta questão, utiliza métodos de busca baseados em Algoritmos Genéticos para analisar, de maneira eficiente, "todo o espaço" de fatores lingüísticos (para todos os fones), e construir automaticamente uma árvore de *clusterização* binária, indicando classes de fones (contendo um ou mais fones) que apresentam características duracionais semelhantes. Em seguida, realiza-se uma busca *Bottom-Up* nesta árvore para selecionar os clusters de fones que maximizam um critério de otimização global previamente definido.

A segunda etapa do algoritmo LDM-GA diz respeito a quais atributos lingüísticos (e interações entre atributos lingüísticos, no caso de modelos SoP) devem ser utilizados em cada modelo de regressão (classes de fones). A solução empregada pelo algoritmo LDM-GA para solucionar este problema também utiliza Algoritmo Genético. Inicialmente, para cada classe de fones, são estimadas várias topologias intermediárias. Cada uma destas topologias intermediárias é estimada a partir de uma determinada partição da base de dados em dados de treinamento e dados de validação. O objetivo da utilização de diferentes partições (aleatórias) da base de dados é verificar se as distribuições dos dados resultantes destas partições serão polarizadas (com diferentes *viezes* e variâncias). Quanto maior for a polarização destas distribuições, resultantes das partições, maior será o grau de diversidade das soluções intermediárias (número de soluções intermediárias distintas). Em seguida, através da aplicação de uma operação denominada Regra Majoritária - RM, realiza-se uma espécie de *ensamble* destas topologias intermediárias, gerando uma única topologia ótima para a respectiva classe de fones.

Estas duas etapas do algoritmo LDM-GA serão denominadas, ao longo deste capítulo, como:

- *Estimação de Classes de Fones por Clusterização Hierárquica Binária*: Identificação de fones com características duracionais semelhantes, para serem modelados conjuntamente em um único modelo de regressão;
- *Estimação das Topologias Ótimas dos Modelos de Regressão*: Quais *fatores* lingüísticos devem ser empregados em cada modelo de regressão.

A Figura 5.2 apresenta um diagrama de blocos descrevendo as duas etapas do algoritmo LDM-GA. As subseções a seguir dedicam-se a descrever detalhadamente cada uma das operações envolvidas nas etapas de estimação de classes de fones por *clusterização* hierárquica binária e estimação das topologias ótimas dos modelos de regressão. Por uma questão didática as operações relacionadas à etapa de estimação das topologias ótimas serão apresentadas primeiro.

5.5.2 Estimação das Topologias Ótimas dos Modelos de Regressão

O algoritmo LDM-GA estima a topologia ótima, de cada classe de fones, através da aplicação da operação de regra majoritária, descrita na subseção 5.5.4, ao conjunto de soluções intermediárias associadas a estas classes. A Figura 5.3 apresenta o *pseudo-código* utilizado para a estimação das topologias ótimas.

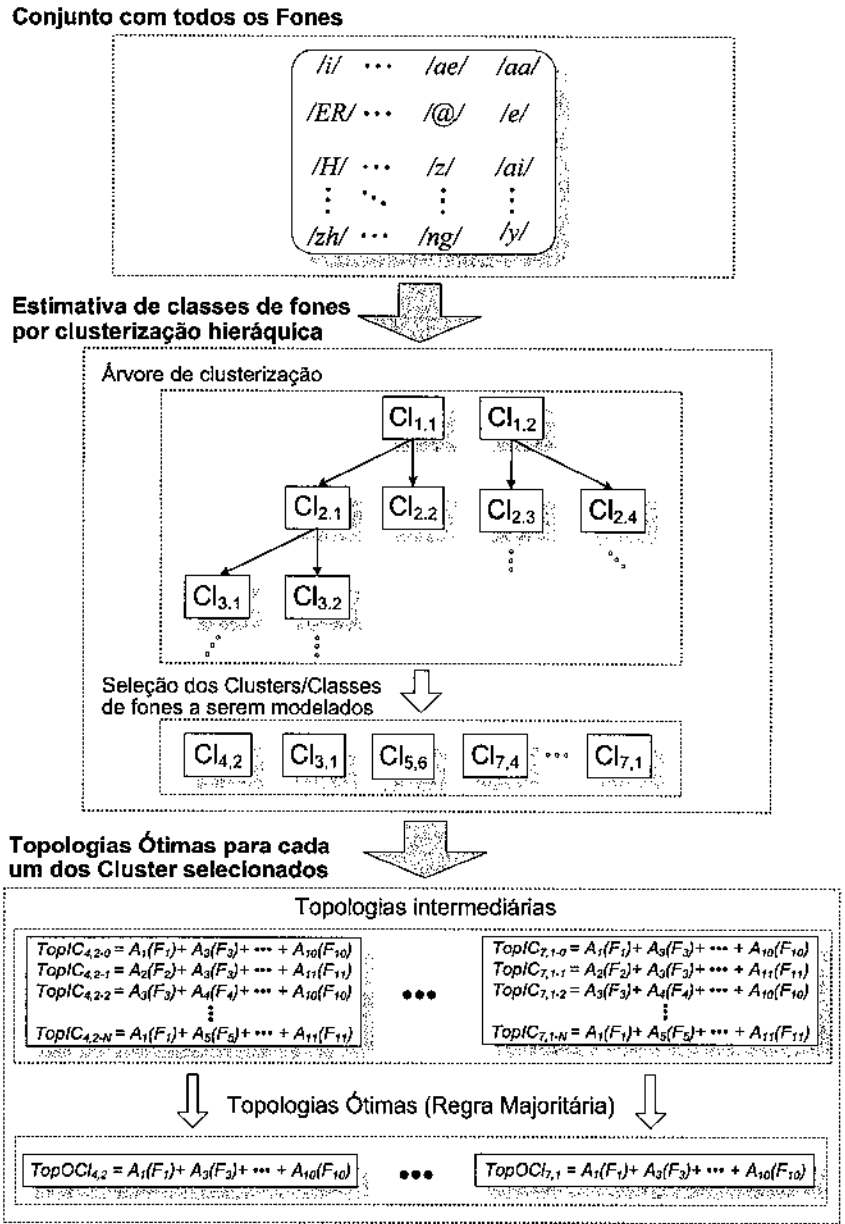


Figura. 5.2: Diagrama de blocos das etapas do algoritmo LDM-GA.

5.5.3 Estimação das Topologias Intermediárias dos Modelos de Regressão

Segundo (Jobson, 1991), a solução ótima para estimar a melhor topologia de um modelo de regressão linear consiste na busca linear, ao longo de todo o espaço de possíveis combinações de parâmetros, pelo subconjunto de parâmetros que maximiza o critério de otimização desejado.

Como mencionado anteriormente, as soluções intermediárias podem ser compreendidas como as melhores topologias, especificamente estimadas para uma determinada partição da base de dados em

dados de treinamento e dados de validação. Portanto, a solução ótima para determinar uma topologia intermediária, consiste na solução dada por (Jobson, 1991) aplicada a uma partição específica da base de dados. Entretanto, para o caso particular da predição da duração segmental da fala, utilizando a base de dados descrita na seção 5.3, cujo espaço *fatorial* dos atributos lingüísticos é da ordem de 10^{13}, a aplicação direta da solução apresentada por (Jobson, 1991), por meio de uma busca linear ao longo de todo o espaco de *fatores* lingüísticos, é absolutamente inviável computacionalmente. A solução empregada pelo algoritmo LDM-GA para lidar com este problema combinatorial, de elevada dimensão, consiste na utilização de Algoritmos Genéticos para pesquisar (explorar) o espaço de *fatores* lingüísticos de uma maneira mais eficiente. A estimação das topologias intermediárias, pelo algoritmo LDM-GA, segue o procedimento descrito pelo *pseudo-código* da Figura 5.4.

Figura. 5.3: Algoritmo para estimação das topologias ótimas dos modelos QMTI.

Algoritmo para Estimação das Topologias Ótimas dos Modelos QMTI:
begin
  for $i := 1$ **to** $N_{classes}$ **do** ($N_{classes}$: *número de classes de fones*)
    $DCl_i :=$ *Conjunto de todos os dados disponíveis para os fones da classe* Cl_i
    for $j := 1$ **to** N_j **do** (N_j : *número de topologias intermediárias*)
      Partição de DCl_i *nos conjuntos de: Treinamento -* $DCl_i^{T_j}$ *e Validação -* $DCl_i^{V_j}$
      Estimação da $i-$*ésima topologia intermediária utilizando:* $DCl_i^{T_j}$ *e* $DCl_i^{V_j}$
    end
    $TopologiaOtima :=$ *Regra Majoritária das* N_j *topologias intermediárias*
  end
end

Figura. 5.4: Rotina p/ estimação da $i-$*ésima* topologia intermediária utilizando: $DCl_i^{T_j}$ e $DCl_i^{V_j}$.

Rotina para estimação da $i-$***ésima*** **topologia intermediária utilizando:** $DCl_i^{T_j}$ **e** $DCl_i^{V_j}$
begin
  for $k := 1$ **to** N_e **do** (N_e : *Número de épocas*)
    Geração de população inicial
    Estimação dos modelos QMTI utilizando apenas $DCl_i^{T_j}$
    Avaliação da função de fitness utilizando apenas $DCl_i^{V_j}$
    Reprodução e seleção dos indivíduos
    for $l := 1$ **to** N_g **do** (N_g : *Número de gerações*)
      Estimação dos modelos QMTI utilizando apenas $DCl_i^{T_j}$
      Avaliação da função de fitness utilizando apenas $DCl_i^{V_j}$
      Reprodução e seleção dos indivíduos
    end
    $TopologiaIntermEpoca(k) :=$ *Indivíduo com o melhor fitness da* k*-ésima época*
  end
  $TopologiaInterm :=$ *Indivíduo com o melhor fitness das* N_e *épocas*
end

5.5.4 Operações Envolvidas na Estimação das Topologias Intermediárias e Ótimas

As subseções a seguir apresentam um descrição detalhada de todas as operações empregadas nos *pseudo-códigos* das Figuras 5.3 e 5.4.

Partição da Base de Dados: Dados de Treinamento × Dados de Validação

Cada partição da base de dados em dados de treinamento e dados de validação é realizada de forma aleatória e nas proporções de 80% para dados de treinamento e 20% para dados de validação.

Geração de População Inicial

A geração de populações iniciais, com topologias candidatas à topologia ótima, é realizada através da amostragem aleatória do espaço combinatorial de todas as possíveis combinações entre os 14 *fatores* lingüísticos descritos na Tabela 5.1. Conforme indicado no *pseudo-código* da rotina 5.4, cada época desta rotina requer uma nova população inicial de indivíduos.

Representação Cromossômica

Cada indivíduo da população inicial deve ser representado através de uma "codificação cromossômica". Na representação cromossômica utilizada neste trabalho, o $k-$*ésimo* indivíduo da população inicial Top_k será representado por um vetor binário de dimensão 14 (número de *fatores* lingüísticos) dado pela expressão a seguir:

$$CR\_Top_k(i) = \begin{cases} 1: & \text{Se } F_i \in Top_k \\ 0: & \text{Caso contrário.} \end{cases} \tag{5.17}$$

Em outras palavras, a $i-$*ésima* posição do $k-$*ésimo* cromossomo será igual a 1 se o $i-$ *ésimo fator* lingüístico estiver presente no $k-$*ésimo* indivíduo da população inicial. Se o *i-ésimo fator* lingüístico não estiver presente no $k-$*ésimo* indivíduo da população inicial, então a $i-$*ésima* posição do $k-$*ésimo* cromossomo será igual a 0.

Estimação dos Modelos de Regressão

Para cada indivíduo da população, deve ser estimado um modelo de regressão utilizando o método QMTI da equação 5.12. A estimação dos parâmetros deste modelo de regressão deve ser realizada utilizando-se apenas os dados de treinamento.

Função Objetivo e Função de Fitness

A função objetivo para o $k-$*ésimo* fone é dada pela equação 5.18:

$$FunObj\left(Cl_k\right) = R\left(Cl_k\right) = \frac{\left(\sum_{i=1}^{NeV_k} durCl_k(i) \cdot \widehat{durCl_k(i)}\right)^2}{\left(\sum_{i=1}^{NeV_k} durCl_k(i)^2\right) \cdot \left(\sum_{i=1}^{NeV_k} \widehat{durCl_k(i)}^2\right)} \tag{5.18}$$

sendo NeV_k o número de exemplos de validação para a $k-ésima$ classe de fones e $durCl_k$ e $\widehat{durCl_k}$ as durações original e predita para a $k-ésima$ classe de fones, respectivamente. O termo $R\left(Cl_k\right)$ também é conhecido como coeficiente de correlação de Pearson (Jobson, 1991), entre as durações originais e as durações estimadas para a $k-ésima$ classe de fones.

A função de *fitness* é obtida a partir de uma suavização linear da função objetivo (Mitchell, 1998). Esta suavização tem como objetivo reduzir a pressão seletiva dos indivíduos a cada geração.

Operações Evolutivas

As operações evolutivas utilizadas foram:

- *Crossover*: utilizou-se operações de *Crossover* de um único ponto e a uma taxa fixa de 80%;
- *Mutação*: utilizou-se operações de Mutação do tipo binária e a uma taxa fixa de 0.01%;
- *Seleção*: empregou-se o método da "Roleta" para seleção dos indivíduos que deveriam passar direto para a próxima geração. A taxa de Seleção utilizada foi de 40%;
- *Reprodução*: a taxa de Reprodução utilizada foi de 60%.

Regra Majoritária

A Regra Majoritária foi a operação utilizada neste trabalho para derivar a topologia ótima, de cada classe de fones, a partir das respectivas soluções intermediárias. A operação de Regra Majoritária pode ser interpretada como uma espécie de *ensamble* (Duffy and Helmbold, 2000), (Meir and Rätsch, 2003), (Valentini and Masulli, 2002), (Breiman, 1994), das soluções intermediárias. Segundo a operação da Regra Majoritária, a topologia ótima (em representação cromossômica), para a $k-ésima$ classe de fones, será dada por:

$$\mathbf{TopOCl_k} = [TopOCl_k(1)\ TopOCl_k(2) \cdots TopOCl_k(14)] \tag{5.19}$$

sendo:

$$TopOCl_k(i) = \lfloor \left(\frac{1}{NTI_k} \sum_{j=1}^{NIT_k} TopICl_k(j,i) \right) + 0.5 \rfloor \tag{5.20}$$

Na equação 5.20 o operador $\lfloor \cdot \rfloor$ representa o maior inteiro menor que. A variável $TopICl_k(j,i)$ é definida como:

$$TopICl_k(j,i) = \begin{cases} 1: & \text{Se a j-}\textit{ésima}\text{ solução intermediária da k-}\textit{ésima}\text{ classe de fones} \\ & \text{contém o i-}\textit{ésimo}\text{ fator lingüístico} \\ 0: & \text{Caso contrário.} \end{cases} \quad (5.21)$$

e NTI_k representa o número de topologias intermediárias para a $k-$*ésima* classe de fones.

Analisando-se as equações 5.19 e 5.20, verifica-se que a operação da Regra Majoritária simplesmente observa a freqüência de ocorrência de cada *fator* lingüístico ao longo das soluções intermediárias e seleciona para a solução do problema (topologia ótima) somente aqueles *fatores* lingüísticos que ocorrerem em mais de 50% das soluções intermediárias.

5.5.5 Construção da Árvore de *Clusterização*

O algoritmo LDM-GA gera a árvore de clusterização dos fones empregando um procedimento de clusterização do tipo *Top Down*, hierárquico e binário. Inicialmente todos os fones são agrupados em um único *cluster* no primeiro nível da árvore. Em seguida utilizando-se algoritmos genéticos, realiza-se uma busca pela melhor partição (divisão) deste *cluster* inicial em dois clusters filhos. O critério utilizado para a escolha da melhor partição foi a maximização do valor de *fitness* dos *cluster* filhos (o cálculo do valor de fitness será discutido na subseção sobre "Função Objetivo e Função de Fitness", seção 5.5.6). De posse destes dois novos clusters, do segundo nível da árvore, realiza-se o mesmo procedimento de partição aplicado ao cluster inicial, gerando 4 novos *clusters* filhos, no terceiro nível da árvore. Este procedimento de partição/duplicação é aplicado iterativamente até que todos os *clusters* terminais da árvore (*folhas* da árvore) possuam apenas um único fone por *cluster*. As Figuras 5.5 e 5.6, a seguir, apresentam os *pseudo-códigos* dos algoritmos para construção da árvore de clusterização.

Figura. 5.5: Algoritmo para Clusterização dos Fones

Algoritmo para Clusterização dos Fones:
begin
Inicie a árvore com apenas um único cluster no primeiro nível e contendo todos os fones
$j := 1$ (j : *Índice para indicar o nível da árvore a ser processado*)
do *Até cada cluster de fone ser reduzida a um único fone*
for $i := 1$ **to** N_j **do** (N_j : *Número de clusters no* $j-$ *esimo nível da árvore*)
Particionar o $i-$ *ésimo cluster do* $j-$ *ésimo nível da árvore*
end
$j := j + 1$ (*Siga para o próximo nível da árvore*)
end
end

Figura. 5.6: Rotina para partição do $i-$*ésimo* cluster do $j-$*ésimo* nível da árvore.

Rotina para partição do $i-$ésimo cluster do $j-$ésimo nível da árvore
begin
DCl_{ij} = Dados associados aos fones do $i-$ ésimo cluster do $j-$ ésimo nível da árvore
for $k := 1$ **to** N_e **do** *(N_e : Número de épocas)*
Geração de população inicial
Estimação dos modelos QMTI utilizando: DCl_{ij}
Avaliação da função de fitness utilizando: DCl_{ij}
Reprodução e seleção dos indivíduos
for $l := 1$ **to** N_g **do** *(N_g : Número de gerações)*
Estimação dos modelos QMTI utilizando: DCl_{ij}
Avaliação da função de fitness utilizando: DCl_{ij}
Reprodução e seleção dos indivíduos
end
$MelhorParticaoDaEpoca(i)$:= Indivíduo com melhor fitness na $k-$ ésima época
end
$MelhorParticaoDoCluster$:= Indivíduo com melhor fitness em todas as N_e épocas
end

5.5.6 Principais Operações na Construção da Árvore de Clusterização

As principais operações envolvidas na construção da árvore de clusterização dos fones são:

Geração de População Inicial

A geração de populações iniciais, com indivíduos candidatos à partição ótima de cada *cluster*, é realizada através da amostragem aleatória do espaço combinatorial de todas as possíveis partições de cada *cluster*. Conforme indicado no *pseudo-código* da Figura 5.6, a cada época da rotina para partição do $i-$*ésimo cluster* do $j-$*ésimo* nível da árvore de clusterização, uma nova população inicial é gerada.

Representação Cromossômica

Na representação cromossomômica utilizada neste trabalho, o $k-$*ésimo* indivíduo da população inicial Top_k é representado por um vetor binário com dimensão igual ao número de fones no *cluster* a ser particionado. Estes cromossomos são representados por:

$$CR\_Cl_k(i) = \begin{cases} 1: & \text{Se o i-}ésimo\text{ fone fará parte do } cluster \text{ filho à direita} \\ 0: & \text{Se o i-}ésimo\text{ fone fará parte do } cluster \text{ filho à esquerda} \end{cases} \quad (5.22)$$

Em outras palavras, a $i-$*ésima* posição do $k-$*ésimo* cromossomo será igual a 1, se o $i-$*ésimo fone* pertencente ao *cluster* for selecionado para fazer parte do *cluster* filho à direita. Se o $i-$*ésimo* fone

for selecionado para fazer parte do *cluster* filho à esquerda, então a $i-ésima$ posição do $k-ésimo$ cromossomo será igual a 0.

Estimação dos Modelos QMTI

Um modelo de regressão linear deve ser estimado para cada indivíduo da população, utilizando-se o método QMTI da equação 5.12. A estimação dos parâmetros deste modelo de regressão deve ser realizada empregando-se **todos** os dados disponíveis (*Treinamento* + *Validação*).

Função Objetivo e Função de Fitness

Definindo Cl_{il}^{L} e Cl_{il}^{R}, respectivamente, como os clusters filhos à esquerda e à direita do $i-ésimo$ cluster do $j-ésimo$ nível da árvore, então a função objetivo associada à partição que gerou Cl_{il}^{L} e Cl_{il}^{R} será dada pela equação 5.23:

$$FunObj\left(Cl_{il}^{L}, Cl_{il}^{R}\right) = R\left(Cl_{il}^{L}, Cl_{il}^{R}\right) = \frac{R\left(Cl_{il}^{L}\right) + R\left(Cl_{il}^{R}\right)}{2} \tag{5.23}$$

sendo:

$$R\left(Cl_{il}^{L}\right) = \frac{\left(\sum_{k=1}^{NE\_Cl_{il}^{L}} durCl_{il}^{L}(k) \cdot \widehat{durCl_{il}^{L}}(k)\right)^{2}}{\left(\sum_{k=1}^{NE\_Cl_{il}^{L}} durCl_{il}^{L}(k)^{2}\right) \cdot \left(\sum_{k=1}^{NE\_Cl_{il}^{L}} \widehat{durCl_{il}^{L}}(k)^{2}\right)} \tag{5.24}$$

e sendo $NE\_Cl_{il}^{L}$ o número de exemplos na base de dados para a classe Cl_{il}^{L} e $durCl_{il}^{L}(k)$ e $\widehat{durCl_{il}^{L}}(k)$ as durações originais e preditas para a classe Cl_{il}^{L}, respectivamente. O termo $R\left(Cl_{il}^{L}\right)$ também é conhecido como coeficiente de correlação de Pearson (Jobson, 1991) entre as durações originais e estimadas para a classe Cl_{il}^{L}. O termo $R\left(Cl_{il}^{R}\right)$, associado ao *cluster* filho a direita, é definido de forma semelhante a $R\left(Cl_{il}^{L}\right)$.

A função de *fitness* é obtida a partir de uma suavização da função objetivo (Mitchell, 1998). Esta suavização tem como objetivo reduzir a pressão seletiva dos indivíduos a cada geração.

Operações Evolutivas

As operações evolutivas utilizadas foram:

- *Crossover*: utilizou-se opearações de *Crossover* de um único ponto e a uma taxa fixa de 80%;
- *Mutação*: utilizou-se operações de Mutação do tipo binária e a uma taxa fixa de 0.01%;

- *Seleção*: empregou-se o método da "Roleta" para seleção dos indivíduos que deveriam passar direto para a próxima geração. A taxa de Seleção utilizada foi de 40%;
- *Reprodução*: a taxa de Reprodução utilizada foi de 60%.

5.5.7 Seleção das Classes de Fones Ótimas

O passo seguinte à geração da árvore de *clusterização* consiste na análise e seleção das classes/*clusters* de fones "ótimos" a serem modeladas pelo método QMTI. Este processo de análise e seleção das classes "ótimas" (ao longo da árvore de clusterização), ocorre de maneira *Bottom Up*, a partir do último nível da árvore. O algoritmo a seguir descreve o processo de seleção de *clusters* "ótimos", utilizado neste trabalho:

1. Definições e Inicializações;
   (a) Define-se Cl_{il} como o $i-ésimo$ cluster (classe de fones) do $l-ésimo$ nível da árvore de clusterização;
   (b) Define-se Cl_List como a lista contendo os *clusters*, da árvore de clusterização, candidatos a *clusters* ótimos;
   (c) Inicializa-se $l = (Número\ de\ níveis\ da\ árvore\ de\ clusterização) - 1$;
   (d) Inicializa-se Cl_List com todos os *clusters* do último nível da árvore de clusterização;
   (e) Estima-se modelos de regressão, com topologias ótimas selecionadas pelo método da Regra Majoritária, para cada um dos *clusters* do último nível da árvore de *clusterização*. Em seguida, calcula-se o valor de *fitness* de cada um destes modelos;
2. Para todo *cluster*, presente no $l-ésimo$ nível da árvore de clusterização, estima-se um modelo de regressão, com topologia ótima selecionada pelo método da Regra Majoritária. Em seguida, calcula-se o valor de *fitness* de cada um destes modelos;
3. Se Cl_{il} apresentar maior valor de *fitness* que a média dos valores de *fitness* de todos os elementos em Cl_List que forem dependentes (filhos, netos, ...) de Cl_{il}, então substituir todos estes elementos de Cl_List (descendentes de Cl_{il}) por Cl_{il};
4. Se $l = 1$ então *ENCERRAR* o algoritmo. Senão, fazer $l = l - 1$;
5. Retonar ao passo 2;

No final deste algoritmo, a variável Cl_List irá conter a lista das classes ótimas de fones a serem modelados.

5.6 Considerações Finais

Este capítulo apresentou os fundamentos teóricos do algoritmo LDM-GA (*Linguistic Data Mining using Genetic Algorithm*), o qual foi desenvolvido para minimizar os problemas associados às distribuições com características LNRE (*Large Number of Rare Events*). Apesar de o algoritmo LDM-GA

ter sido formulado sob o contexto de otimização de modelos de regressão linear, ele pode, facilmente, ser adaptado para otimizar modelos de regressão empregando SoP (*Sum of Product Models*) ou ANN (*Artificial Neural Networks*). Além disso, o algoritmo LDM-GA também pode ser adaptado para otimizar problemas de classificação, por meio de análise discrimitativa linear ou não-linear. Por exemplo, o algoritmo LDM-GA pode ser facilmente adaptado para otimizar os classificadores a serem utilizados no processo de seleção de *gestos* de F_0 do modelo entoacional de Kagoshima (Kagoshima et al., 1998) (descrito na seção 3.5). É importante relembrar que o processo de seleção de *gestos* de F_0 do modelo entoacional de Kagoshima se faz a partir de um espaço combinatorial de atributos lingüísticos (simbólicos) que normalmente apresenta distribuições com características LNRE.

Duas outras importantes propriedades do algoritmo LDM-GA são: (1) Ele não faz qualquer suposição sobre as características das distribuições dos dados lingüísticos a serem modelados. (2) Apesar do critério de otimização empregado pelo algoritmo LDM-GA neste capítulo (e no capítulo 6) ser baseado no coeficiente de correlação, nada impede que qualquer outro critério de otimização (que contemple aspectos auditivo-perceptivos, por exemplo) seja utilizado no algoritmo LDM-GA.

As representações cromossômicas utilizadas pelo algoritmo LDM-GA, neste capítulo, tanto na seleção das topologias ótimas quanto na clusterização dos fones são extremamente simples, e portanto, sugere-se a investigação e o uso de representações cromossômicas mais elaboradas em trabalhos futuros.

O capítulo 6 a seguir, apresenta os resultados de uma série de experimentos realizados com o algoritmo LDM-GA no contexto da predição da duração segmental da fala. Estes experimentos serão úteis não somente para avaliar o desempenho do algoritmo LDM-GA, mas também para tornar mais claros vários dos fundamentos teóricos apresentados neste capítulo.

Capítulo 6

Algoritmo LDM-GA: Resultados Experimentais e Análises

6.1 Introdução

Este Capítulo apresenta o resultado de vários experimentos para avaliar o desempenho do algoritmo LDM-GA na otimização de modelos de regressão linear aplicados à modelagem da duração segmental da fala. A base de dados utilizada neste Capítulo, é a mesma descrita na seção 5.3. Os modelos de regressão otimizados pelo algoritmo LDM-GA foram comparados com modelos de regressão otimizados pelo método MANOVA (Multivariate Analysis of Variance) (Jobson, 1991) e também com modelos de regressão estimados por árvores de regressão - RT (*Regression Trees*) (Breiman et al., 1993); (Quinlan, 1993). Este Capítulo é dividido em 6 seções principais. A seção 6.2 faz uso do método de análise de componentes principais - PCA (*Principal Component Analysis*) para analisar o grau de esparsidade dos dados lingüísticos presentes na base de dados. A seção 6.3 apresenta resultados referentes à modelagem individual de cada um dos 45 fones da Tabela 5.4. Em outras palavras, os resultados da subseção 6.3 não fazem nenhuma referência ao processo de clusterização de fones do algoritmo LDM-GA e, portanto, todos os fones são analisados/modelados individualmente. A seção 6.4.3 apresenta resultados referentes à modelagem de classes de fones, obtidas pelo processo de clusterização hierárquica binária do algoritmo LDM-GA descrito na seção 5.5.5 do Capítulo 5. A seção 6.5 analisa o efeito de cada um dos *fatores* linguísticos na modelagem da duração segmental dos fones. A seção 6.6 analisa algumas das limitações apresentadas pelos modelos de regressão linear, na modelagem da duração segmental da fala. Finalmente a seção 6.7 apresenta algumas considerações finais sobre o algoritmo LDM-GA.

Para facilitar a identificação dos resultados apresentados neste Capítulo, serão utilizadas as seguintes notações:

- QMTI/Ph Cheios - Modelos de regressão estimados por fone, empregando o método QMTI e com topologias cheias (utilizando todos os *fatores* lingüísticos disponíveis).

- QMTI/Ph + LDM-GA - Modelos de regressão estimados por fone, empregando o método QMTI e com topologias estimadas segundo o algoritmo LDM-GA.
- QMTI/Cl + LDM-GA - Modelos de regressão estimados por classes de fones, empregando o método QMTI e com topologias e classes de fones estimadas segundo o algoritmo LDM-GA.
- QMTI/Ph + MANOVA - Modelos de regressão estimados por fones, empregando o método QMTI e com topologias estimadas segundo a técnica MANOVA.
- RT/Ph - Modelos de regressão estimados por fones, empregando árvores de regressão.
- RT/Cl - Modelos de regressão estimados por classes de fones, empregando árvores de regressão.

6.2 Esparsidade do Espaço de *Fatores* Lingüísticos

A Figura 6.1 mostra os dados do fone [aa] após o processo de quantificação descrito pela equação 5.13. Observe que foram utilizados 14 *fatores* lingüísticos, incluindo a identidade dos fones, phID, e que a dimensão de cada vetor binário é igual à soma do número de níveis destes *fatores*, isto é, 168. Analisando a Figura 6.1 pode-se verificar que o grau de esparsidade dos dados, após o processo de quantificação apresentado em 5.13, é extremamente elevado.

Com o objetivo de quantificar o grau de esparsidade dos dados da Figura 6.1, o método de Análise de Componentes Principais - PCA (*Principal Componente Analysis*) foi aplicado a estes dados, gerando os dados apresentados na Figura 6.2. Segundo o método PCA, a dimensão original dos dados, 168, pode ser reduzida para 77 sem nenhuma perda na representação dos dados. Nos casos de serem toleradas perdas na representação dos dados, iguais a 2% e 18%, então, a dimensão original dos dados, 168, pode ser reduzida, respectivamente, para 55 e 27.

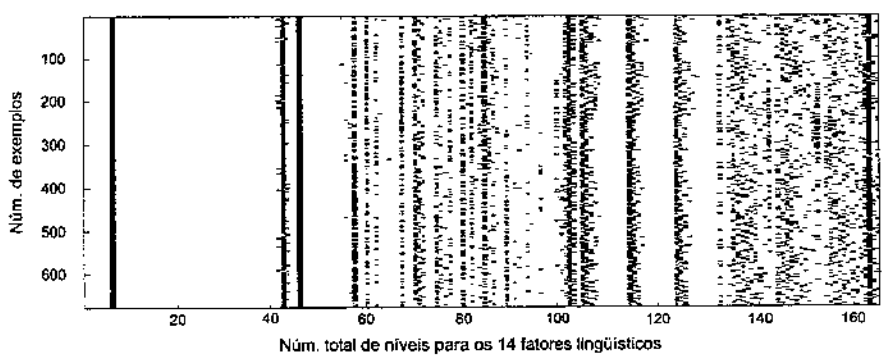


Figura. 6.1: Dados para o fone [aa] no formato binário (quantificado). Os primeiros 45 níveis (ao longo do eixo horizontal) representam a identidade do fone [aa].

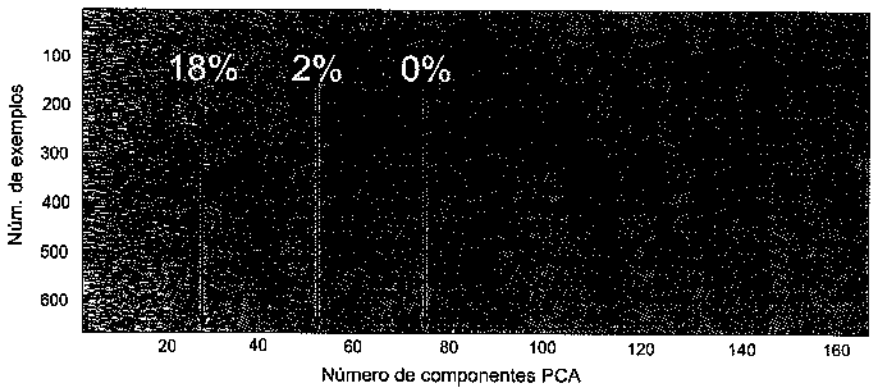


Figura. 6.2: Análise de Componentes Principais sobre os dados do fone [aa]. Esta análise permite reduzir a dimensão original do espaço de 168 a 77, 55 e 27, com perdas de representação dos dados, respectivamente, iguais a 0%, 2% e 18%.

6.3 Modelagem por Fones

6.3.1 Topologias Ótimas Segundo o Método LDM-GA

A Figura 6.3 mostra as 45 topologias ótimas, obtidas para cada um dos fones, através do uso do método LDM-GA. Nesta Figura o eixo horizontal indica os 45 fones modelados e o eixo vertical indica os 13 *fatores* lingüísticos utilizados. Observe que nesta seção cada modelo de regressão corresponde especificamente a um determinado fone, portanto nestas condições o *fator* lingüístico phID torna-se desnecessário. A presença de um retângulo de cor preta na coordenada referente ao *i-ésimo* fone e ao *j-ésimo* fator lingüístico, indica que este *j-ésimo* fator lingüístico deve fazer parte da topologia ótima do *i-ésimo* fone. Por outro lado a presença de um retângulo de cor branca indica que o respectivo *fator* lingüístico não faz parte da topologia ótima do fone em questão. Analisando-se a Figura 6.3 pode-se verificar que a topologia selecionada para o fone [zh] é a dada pela equação 6.1

$$DUR\{[zh]\} = média([zh]) + A_1\,(PosSyll) + A_3\,(NextPh) \tag{6.1}$$

As topologias ótimas da Figura 6.3 foram obtidas utilizando-se os algoritmos descritos nos pseudo-códigos das Figuras 5.3 e 5.4. Os parâmetos utilizados na estimativas destas topologias ótimas foram:

- População inicial: 50 indivíduos.
- Número de classes: 45 (igual ao número de fones).
- Número de topologias intermediárias: 50.
- Número de épocas: 10.
- Número de gerações: 15.
- Partição dos dados: 80% Treinamento, e 20% Validação.

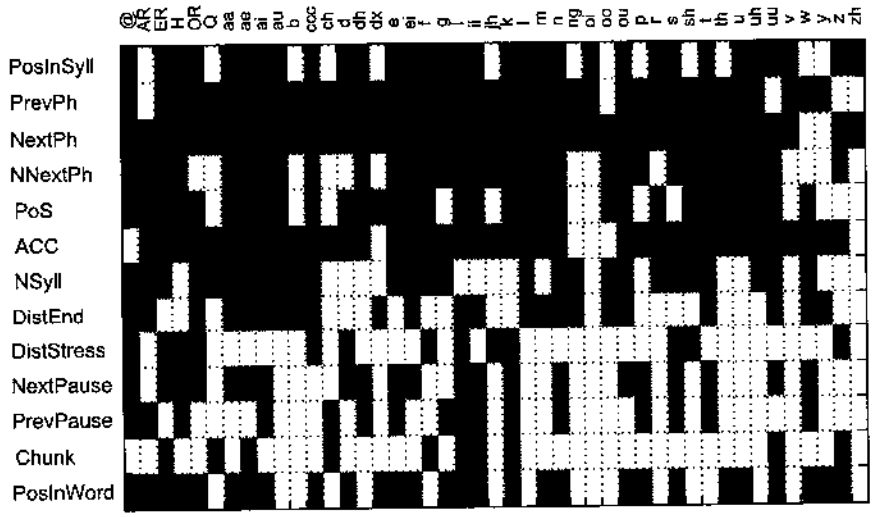


Figura. 6.3: Topologias ótimas selecionadas pelo método LDM-GA para cada um dos 45 fones da tabela 5.4.

- Operações evolutivas:
  - *Crossover*: utilizou-se operações de *Crossover* de um único ponto e a uma taxa fixa de 80%.
  - *Mutação*: utilizou-se operações de Mutação do tipo binária e a uma taxa fixa de 0,01%
  - *Seleção*: empregou-se o método da "Roleta" para seleção dos indivíduos que deveriam passar direto para a próxima geração. A taxa de Seleção utilizada foi de 40%
  - *Reprodução*: a taxa de Reprodução utilizada foi de 60%

A Figura 6.4 apresenta a frequência de ocorrência de cada um dos fatores linguísticos ao longo de todas as 45 topologias ótimas da Figura 6.3

6.3.2 Topologias Ótimas Segundo o Método MANOVA

A Figura 6.5 mostra as topologias ótimas selecionadas com o uso do método MANOVA. De 6.5 observa-se que para o fone [zh] a topologia ótima selecionada foi a representada pela equação 6.3.

$$DUR\{[zh]\} = média([zh]) + A_1\,(PosSyll) + A_2\,(PrevPh) + A_3\,(NextPh) \tag{6.2}$$

As análises de variâncias - MANOVA, empregadas na estimação das topologias mostradas na Figura 6.5 foram realizadas utilizando-se o software Matlab. Os *fatores* lingüísticos selecionados foram aqueles que apresentaram um nível de significância estatística $\rho \leq 0,001$ (sendo ρ também conhecido como $p-value$).

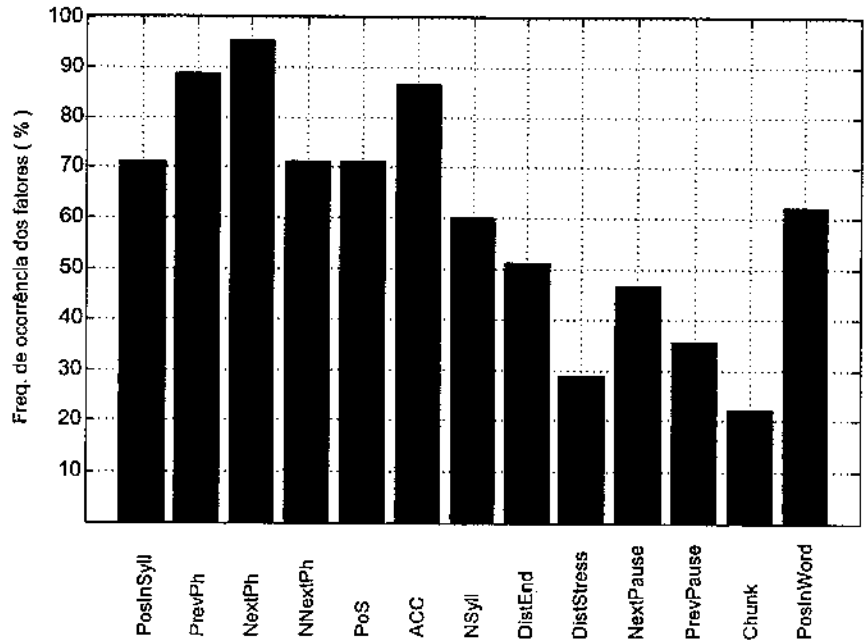


Figura. 6.4: Frequência de ocorrência de cada um dos 13 fatores linguísticos ao longo das 45 topologias ótimas da Figura 6.3.

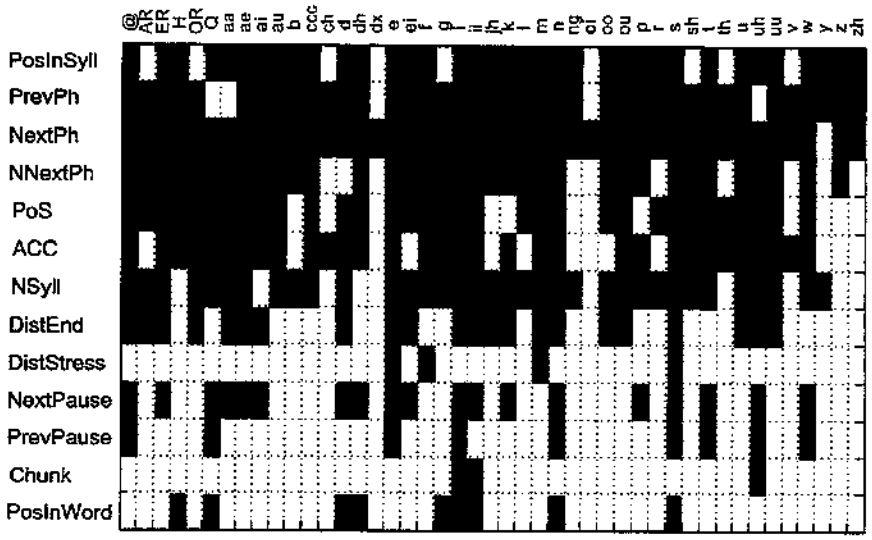


Figura. 6.5: Topologias ótimas selecionadas pelo método MANOVA para cada um dos 45 fones da tabela 5.4.

A Figura 6.6 apresenta a frequência de ocorrência de cada um dos *fatores* linguísticos ao longo de todas as 45 topologias ótimas da Figura 6.5. Pode-ser observar da Figura 6.6 que os *fatores* mais freqüentes, ao longo das 45 topologias, são as identidades dos fones vizinhos (PrevPh e NextPh) ao fone em análise. Outras duas influências importantes a serem observadas são as dos *fatores* ACC (nível

acentual da palavra que contém o fone em análise) e PoS (*Part-of-Speech*).

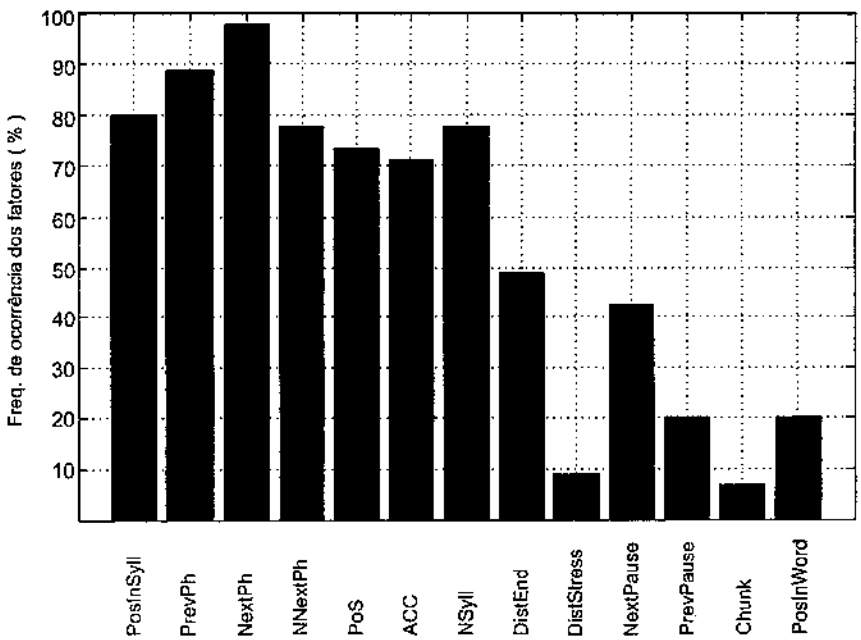


Figura. 6.6: Frequência de ocorrência de cada um dos 13 fatores linguísticos ao longo das topologias ótimas das 45 topologias ótimas da Figura 6.5.

6.3.3 Análise do Operador Regra Majoritária

As Figuras 6.7, 6.8 e 6.9 apresentam as 50 topologias intermediárias obtidas, respectivamente, para os fones [AR], [b] e [@]. Cada uma das 50 topologias intermediárias foi obtida para diferentes partições da base de dados em dados de treinamento (80%) e validação (20%). O objetivo da estimação de 50 partições (aleatórias) da base de dados em dados de treinamento e validação foi o de verificar se estas partições poderiam gerar distribuições polarizadas (com diferentes *viezes* e variâncias) e, portanto, conduzirem a diferentes topologias intermediárias. Pode-se verificar das Figuras 6.7, 6.8 e 6.9 que o fone [AR] apresenta uma maior diversidade de soluções intermediárias distintas que os fones [@] e [b]. Analisando-se as distribuições de durações associadas aos fones [AR], [b] e [@], Figura A.1, verifica-se que o fone [@] é aquele que apresenta a maior quantidade de dados e também a distribuição mais próxima a uma distribuição gaussiana. Esta relação entre o grau de diversidade das soluções intermediárias (número de soluções intermediárias diferentes) e o nível de balanceamento da base de dados se mostrou uma regra para todos os fones.

Aplicando-se a regra majoritária às topologias intermediárias associadas ao fones [AR], [b] e [@], tem-se:

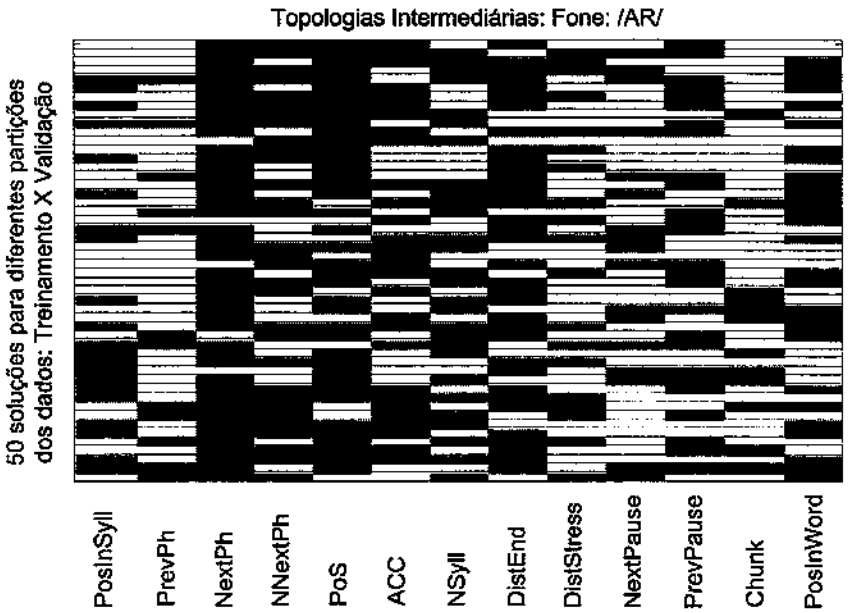


Figura. 6.7: 50 topologias intermediárias obtidas para o fone [AR].

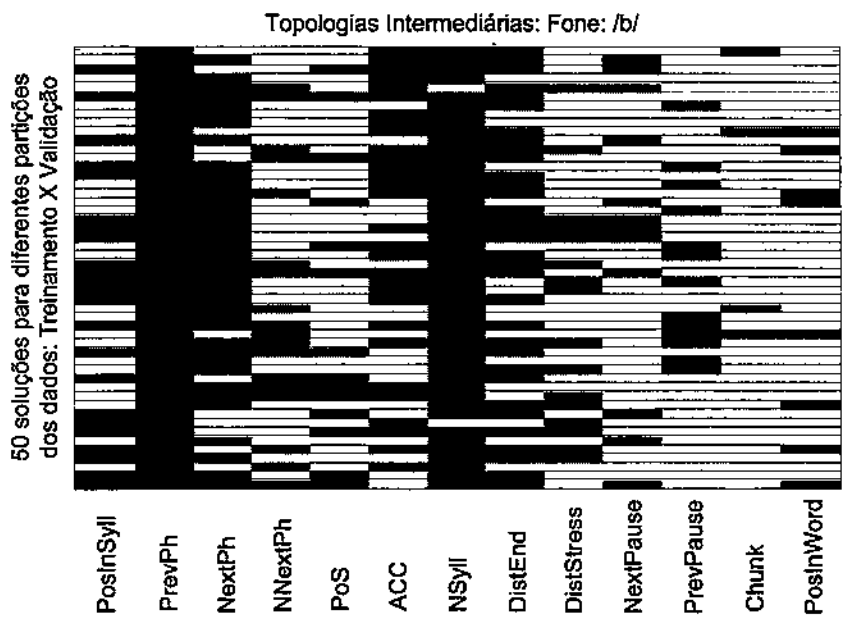


Figura. 6.8: 50 topologias intermediárias obtidas para o fone [b].

$$\begin{aligned} DUR\{[AR]\} \;=\; & \textit{média}([AR]) + A_3\,(NextPh) + A_4\,(NNextPh) + A_5\,(PoS) + \\ & A_6\,(ACC) + A_7\,(NSyll) + A_8\,(DistEnd) + \\ & A_{11}\,(PrevPause) + A_{13}\,(PosInWord) \end{aligned} \tag{6.3}$$

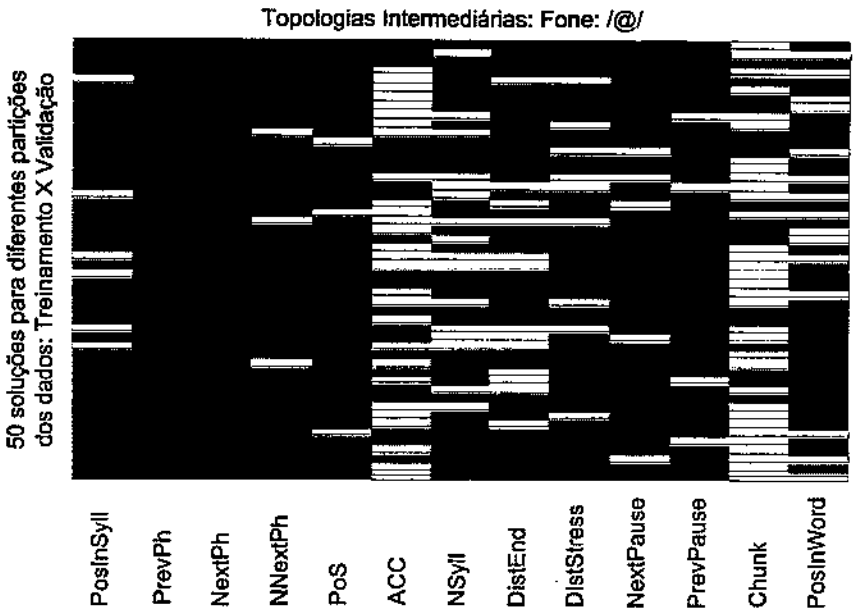


Figura. 6.9: 50 topologias intermediárias obtidas para o fone [@].

$$DUR\{[b]\} = média([b]) + A_2\,(PrevPh) + A_3\,(NextPh) + A_6\,(ACC) + A_7\,(NSyll) + A_8\,(DistEnd) \quad (6.4)$$

$$DUR\{[@]\} = média([@]) + A_2\,(PosInSyll) + A_2\,(PrevPh) + A_3\,(nextPh) + A_4\,(NNextPh) + A_5\,(PoS) + A_7\,(NSyll) + A_8\,(DistEnd) + A_9\,(DistStress) + A_{10}\,(NextPause) + A_{11}\,(PrevPause) + A_{13}\,(PosInWord) \quad (6.5)$$

6.3.4 Resultados Comparativos: QMTI/Ph + LDM-GA × QMTI/Ph Cheio

As Figuras 6.10, 6.11 e 6.12 mostram o desempenho dos modelos de regressão QMTI/Ph + LDM-GA para os fones [AR], [g] e [r] (representados por ○). Nesta Figuras os modelos QMTI/Ph + LDM-GA são comparados com modelos QMTI/Ph Cheio (representados por ∗), isto é, modelos utilizando todos os *fatores* linguísticos disponíveis. Em cada uma das Figuras 6.10, 6.11 e 6.12 são mostrados resultados correspondentes a 25 diferentes partições da base de dados em dados de treinamento (80% dos dados) e validação (20% dos dados). Para cada uma das 25 partições as 50 topologias intermediárias são treinadas e avaliadas utilizando-se os respectivos dados de treinamento e validação correspondentes à

partição em questão. Cada barra vertical (uma para cada partição), indica a faixa de variação dos coeficientes de correlação associados às 50 topologias intermediárias. Em outras palavras (para cada partição da base de dados), o limite superior de cada barra vertical indica o maior coeficiente de correlação alcançado pelas 50 topologias intermediárias (para a correspondente partição dos dados), enquanto o limite inferior indica o menor coeficiente de correlação atingido por estas mesmas 50 topologias intermediárias. Assumindo que os resultados das soluções intermediárias se distribuem uniformemente ao longo das barras verticais, então, o ponto médio, associado a cada uma destas barras, indica o coeficiente de correlação médio para as 50 topologias intermediárias.

As Figuras 6.10, 6.11 e 6.12 mostram que na média os modelos QMTI/Ph + LDM-GA alcançam desempenhos superiores aos modelos QMTI/Ph Cheios. Outro resultado importante, mostrado nas Figuras 6.10, 6.11 e 6.12, é que, na média, o resultado obtido pelos modelos QMTI/Ph + LDM-GA são superiores ao valores médios obtidos pelas 50 topologias intermediárias. Este resultado pode ser observado verficando-se que o coeficiente de correlação dos modelos QMTI/Ph + LDM-GA, são na média, superiores ao valor médio das barras verticais.

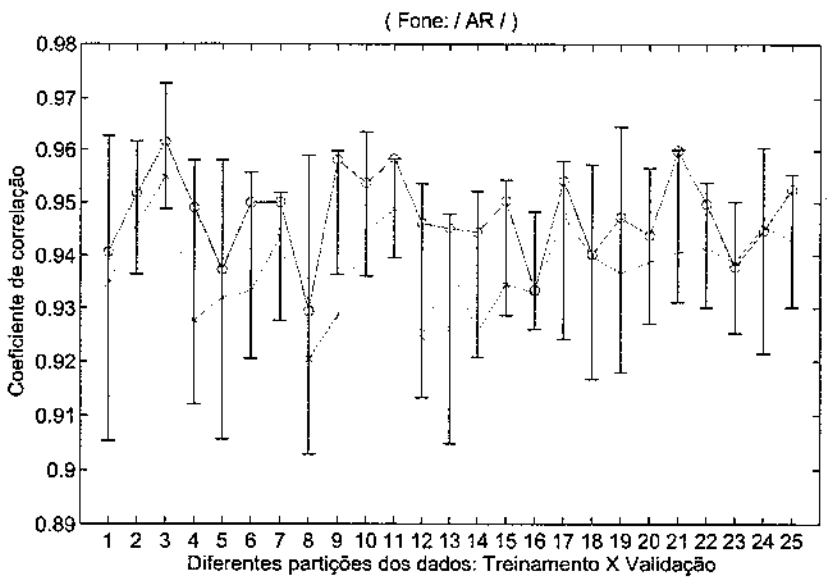


Figura. 6.10: QMTI/Ph + LDM-GA (indicado por o) versus QMTI/Ph cheio (indicado por $*$).

Testes estatísticos não paramétricos de Mann-Whitney (Gibbons, 1985) (a um nível de significância de $0,05$, isto é, valores de $p-value$ inferiores a $0,05$ indicam significância estatística) foram aplicados aos resultados das Figuras 6.10, 6.11 e 6.12 e os respectivos *p-value* encontram-se na Tabela 6.1.

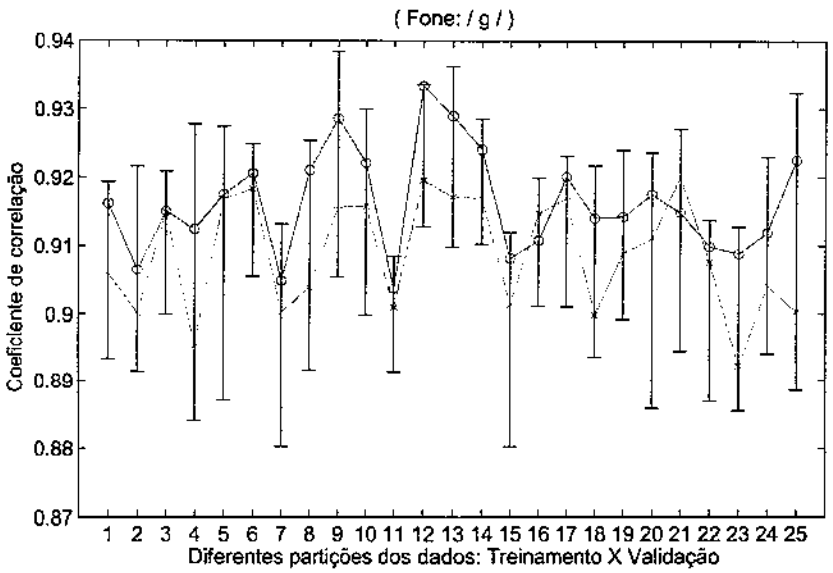


Figura. 6.11: QMTI/Ph + LDM-GA (indicado por ○) versus QMTI/Ph cheio (indicado por ∗).

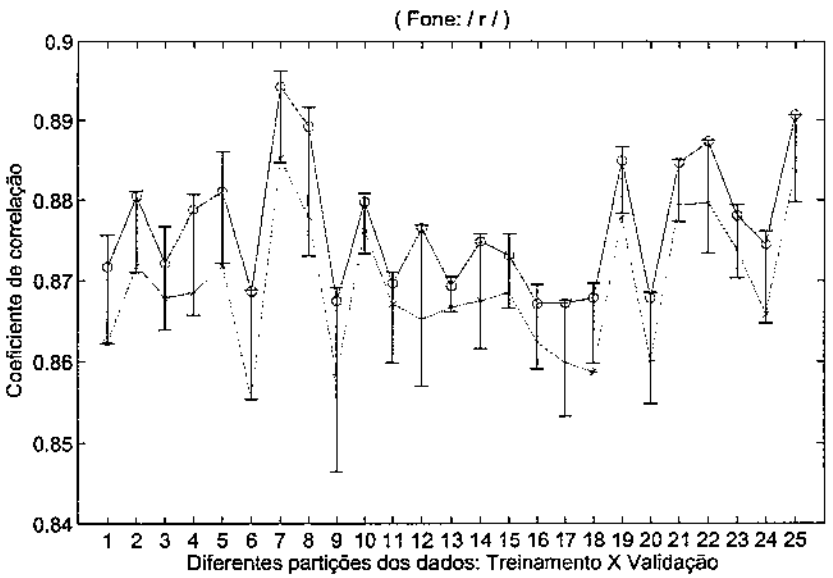


Figura. 6.12: QMTI/Ph + LDM-GA (indicado por ○) versus QMTI/Ph cheio (indicado por ∗).

6.3.5 Resultados Comparativos: QMTI/Ph + LDM-GA × QMTI/Ph + MANOVA

As Figuras 6.13, 6.14 e 6.15 comparam o desempenho dos modelos de regressão QMTI/Ph + LDM-GA (representados por ○) com os modelos de regressão QMTI/Ph + MANOVA (representados por □), para os fones [AR], [g] e [r] . Os resultados mostrados nas Figuras 6.13, 6.14 e 6.15 correspondem

Tabela. 6.1: Testes estatísticos de Mann-Whitney: QMTI/Ph + LDM-GA versus QMTI/Ph Cheio

Fone	*p-value*
[AR]	$9,14 \cdot 10^{-14}$
[g]	$7,81 \cdot 10^{-2}$
[r]	$5,88 \cdot 10^{-3}$

a 25 diferentes partições da base de dados em dados de treinamento (80% dos dados) e validação.(20% dos dados). Estes 25 experimentos (para cada um dos três fones) mostram que na média os modelos QMTI/Ph + LDM-GA alcançam desempenhos ligeiramente superiores aos modelos QMTI/Ph + MANOVA para os fones [AR] e [r]. Entretanto, para o fone [g] os resultados obtidos pelos algoritmos QMTI/Ph + LDM-GA não se mostraram significativamente superiores aos resultados obtidos pelo método QMTI/Ph + MANOVA.

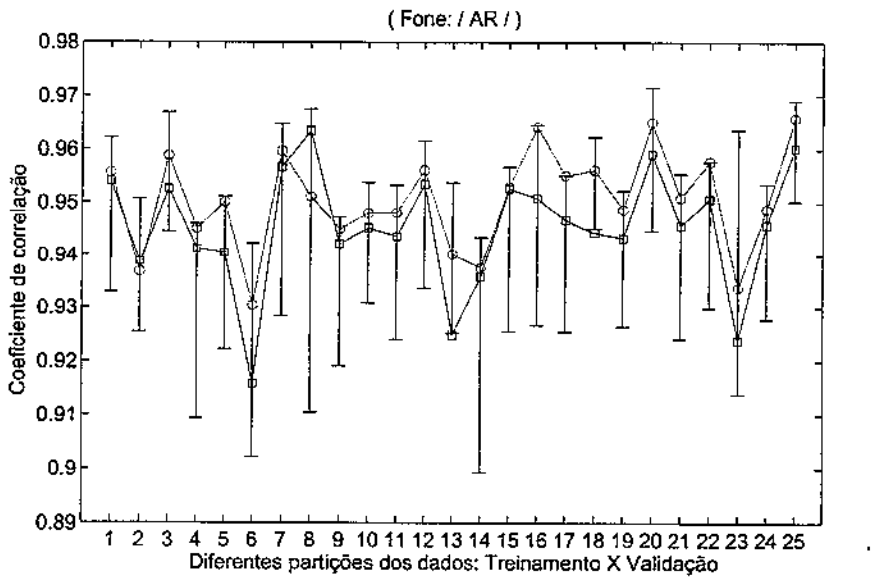


Figura. 6.13: QMTI + Regra Majoritária (indicado por ○) versus QMTI + MANOVA (indicado por □).

Testes estatísticos não paramétricos de Mann-Whitney (Gibbons, 1985) (a um nível de significância de 0,05) foram aplicados aos resultados das Figuras 6.13, 6.14 e 6.15 e os respectivos *p-value* encontram-se na Tabela 6.2.

6.3.6 Resultados Comparativos: QMTI/Ph + LDM-GA × RT/Ph

As Figuras 6.16, 6.17 e 6.18 comparam o desempenho dos modelos de regressão QMTI/Ph + LDM-GA (representados por ○) com os modelos de árvore de regressão RT/Ph (representados por ▽), para

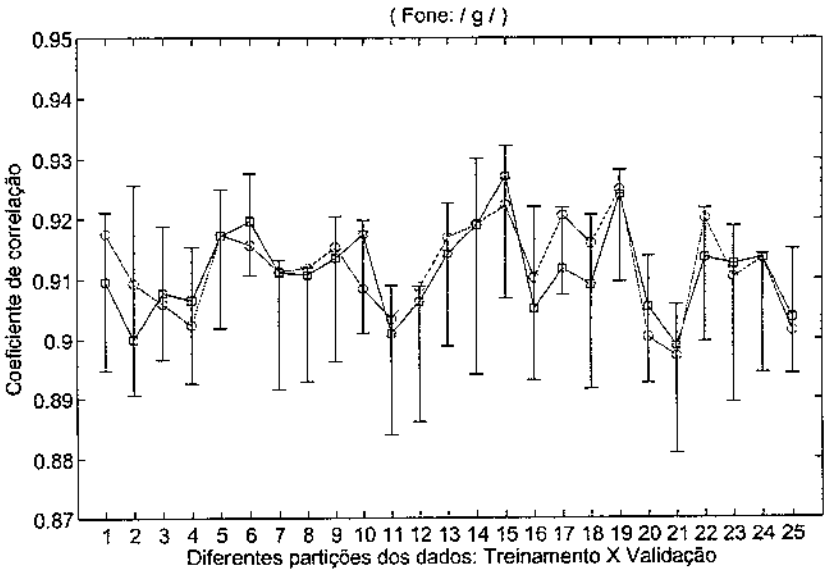


Figura. 6.14: QMTI/Ph + LDM-GA (indicado por o) versus QMTI/Ph + LDM-GA (indicado por □).

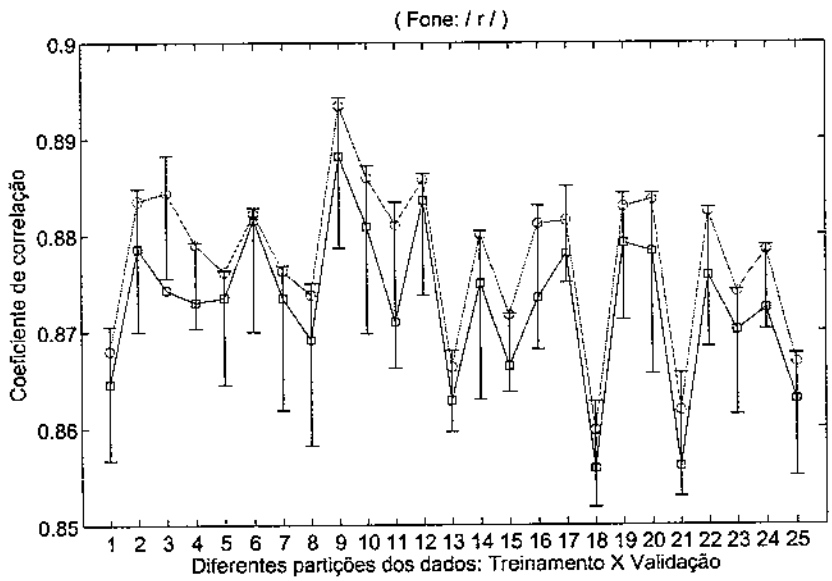


Figura. 6.15: QMTI/Ph + LDM-GA (indicado por o) versus QMTI/Ph + MANOVA (indicado por □).

os fones [AR], [g] e [r]. Os resultados mostrados nas Figuras 6.16, 6.17 e 6.18 correspondem a 25 diferentes partições da base de dados em dados de treinamento (80% dos dados) e validação (20% dos dados). Estes 25 experimentos (para cada um dos três fones) mostram que na média os modelos QMTI/Ph + LDM-GA alcançam desempenhos ligeiramente superiores que os modelos RT/Ph para os fones [AR] e [r]. Entretanto, para o fone [g] os resultados obtidos pelos algoritmos RT/Ph se mostraram

Tabela. 6.2: Testes estatísticos de Mann-Whitney: QMTI/Ph + LDM-GA versus QMTI/Ph + MANOVA

Fone	*p-value*
[AR]	$7,04 \cdot 10^{-2}$
[g]	$7,03 \cdot 10^{-1}$
[r]	$7,93 \cdot 10^{-2}$

equiparáveis aos resultados obtidos pelos método QMTI/Ph + LDM-GA.

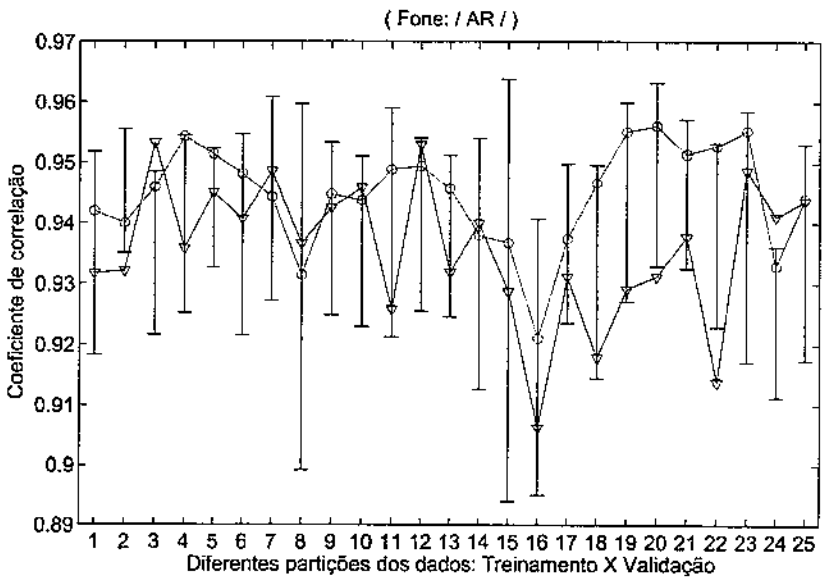


Figura. 6.16: QMTI/Ph + LDM-GA (indicado por $\circ$) versus QMTI/Ph + RT (indicado por $\triangledown$).

Testes estatísticos não paramétricos de Mann-Whitney (Gibbons, 1985) (a um nível de significância de 0,05) foram aplicados aos resultados das Figuras 6.16, 6.17 e 6.18 e os respectivos *p-value* encontram-se na Tabela 6.3.

6.4 Modelagem por Classes de Fones

6.4.1 Árvore de Clusterização

A aplicação do algoritmo de clusterização do método LDM-GA à base de dados descrita na Seção 5.5.5, resultou em uma árvore de clusterização binária (assimétrica) com 10 níveis e 89 clusters. Os clusters desta árvore se encontram descritos nas Tabelas 6.4 e 6.5.

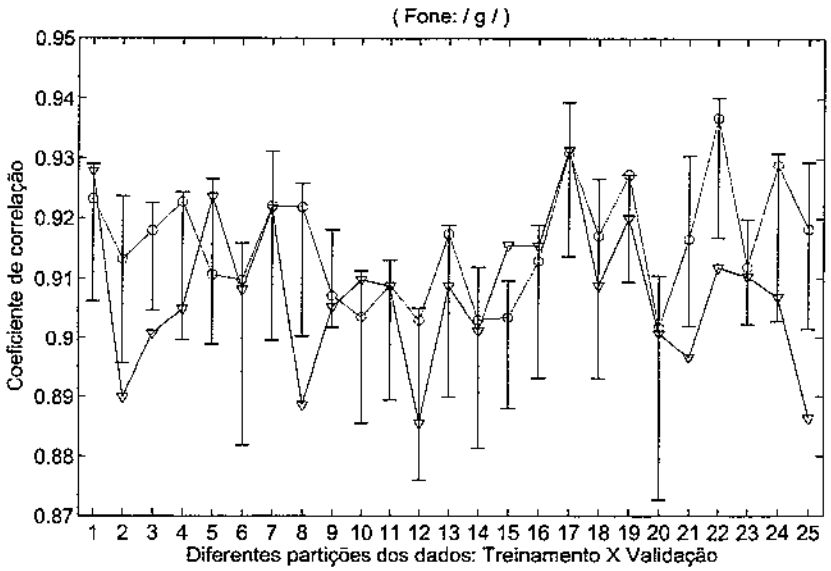


Figura. 6.17: QMTI/Ph + LDM-GA (indicado por ∘) versus QMTI/Ph + RT (indicado por ∇).

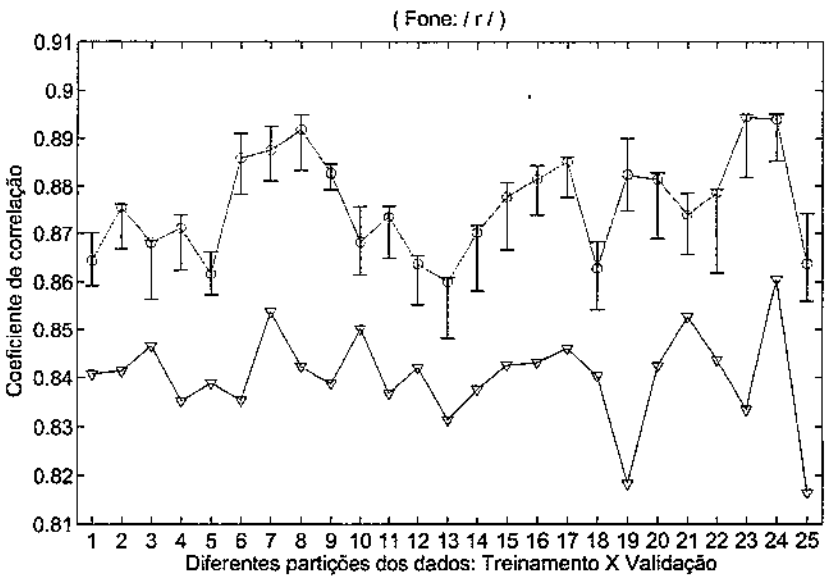


Figura. 6.18: QMTI/Ph + LDM-GA (indicado por ∘) versus QMTI/Ph + RT (indicado por ∇).

Tabela. 6.3: Testes estatísticos de Mann-Whitney: QMTI/Ph + LDM-GA versus QMTI/Ph + RT

Fone	*p-value*
[AR]	$3,64 \cdot 10^{-2}$
[g]	$1,59 \cdot 10^{-2}$
[r]	$1,45 \cdot 10^{-9}$

Clusters da Árvore de Clusterização: Primeira parte
Classes no primeiro nível
Cl.1.1: @,H,Q,b,ccc,d,dh,dx,e,g,i,l,m,n,ng,r,t,u,uh,v,w,y,z,zh,AR, ER,OR,aa,ae,ai,au,ch,ei,f,ii,jh,k,oi,oo,ou,p,s,sh,th,uu
Classes no segundo nível
Cl.2.1: @,H,Q,b,ccc,d,dh,dx,e,g,i,l,m,n,ng,r,t,u,uh,v,w,y,z,zh,
Cl.2.2: AR,ER,OR,aa,ae,ai,au,ch,ei,f,ii,jh,k,oi,oo,ou,p,s,sh,th,uu
Classes no terceiro nível
Cl.3.1: H,Q,b,e,g,i,l,m,n,ng,t,uh,z,zh
Cl.3.2: @,ccc,d,dh,dx,r,u,v,w,y
Cl.3.3: ER,aa,ae,f,ii,jh,k,oo,p,s,sh,th
Cl.3.4: AR,OR,ai,au,ch,ei,oi,ou,uu
Classes no quarto nível
Cl.4.1: H,Q,b,n,ng,t,zh
Cl.4.2: e,g,i,l,m,uh,zh
Cl.4.3: @,ccc,d,u,v,w,y
Cl.4.4: dh,dx,r
Cl.4.5: ae,ii,s,th
Cl.4.6: ER,aa,f,jh,k,oo,p,sh
Cl.4.7: au,ch,ei,ou,uu
Cl.4.8: AR,OR,ai,oi
Classes no quinto nível
Cl.5.1: H,t,zh
Cl.5.2: Q,b,n,ng
Cl.5.3: g,i,zh
Cl.5.4: e,l,m,uh
Cl.5.5: @,ccc,u,v,w,y
Cl.5.6: d
Cl.5.7: dh,dx
Cl.5.8: r
Cl.5.9: ae,ii
Cl.5.10: s,th
Cl.5.11: ER,oo
Cl.5.12: aa,f,jh,k,p,sh
Cl.5.13: au,ch,ei,ou
Cl.5.14: uu
Cl.5.15: AR,OR
Cl.5.16: ai,oi

Tabela. 6.4: Clusters da árvore de clusterização: Primeira parte

Clusters da Árvore de Clusterização	
Segunda Parte	**Terceira Parte**
Classes no sexto nível	
Cl.6.1: H,zh	Cl.6.13: ae
Cl.6.2: t	Cl.6.14: ii
Cl.6.3: Q,b,ng	Cl.6.15: s
Cl.6.4: n	Cl.6.16: th
Cl.6.5: i,zh	Cl.6.17: ER
Cl.6.6: g	Cl.6.18: oo
Cl.6.7: e,m,uh	Cl.6.19: aa,jh,k,p,sh
Cl.6.8: l	Cl.6.20: f
Cl.6.9: ccc,u,v,w,y	Cl.6.21: ch,ei,ou
Cl.6.10: @	Cl.6.22: au
Cl.6.11: dh	Cl.6.23: AR
Cl.6.12: dx	Cl.6.24: OR
	Cl.6.25: oi
	Cl.6.26: ai
Classes no sétimo nível	
Cl.7.1: H	Cl.7.9: u,v,w,y
Cl.7.2: Z	Cl.7.10: ccc
Cl.7.3: Q	Cl.7.11: aa,jh,p,sh
Cl.7.4: b,ng	Cl.7.12: k
Cl.7.5: i	Cl.7.13: ch,ou
Cl.7.6: zh	Cl.7.14: ei
Cl.7.7: m,uh	
Cl.7.8: e	
Classes no oitavo nível	
Cl.8.1: b	Cl.8.7: a,jh,sh
Cl.8.2 ng	Cl.8.8: p
Cl.8.3: m	Cl.8.9: ch
Cl.8.4: uh	Cl.8.10: ou
Cl.8.5: v,w,y	
Cl.8.6: u	
Classes no nono nível	
Cl.9.1 v,y	Cl.9.3: aa,sh
Cl.9.2: w	Cl.9.4: jh
Classes no décimo nível	
Cl.10.1: v	Cl.10.3: aa
Cl.10.2: y	Cl.10.4: sh

Tabela. 6.5: Classes da árvore de clusterização: Segunda e Terceira partes.

A Figuras 6.19, 6.20 e 6.21 mostram a árvore de clusterização com a identificação das respectivas classes de fones descritas nas Tabelas 6.4 e 6.5. As Figuras 6.19, 6.20 também apresentam os valores médios e desvios padrões de duração para cada um das classes/clusters de fones em milissegundos.

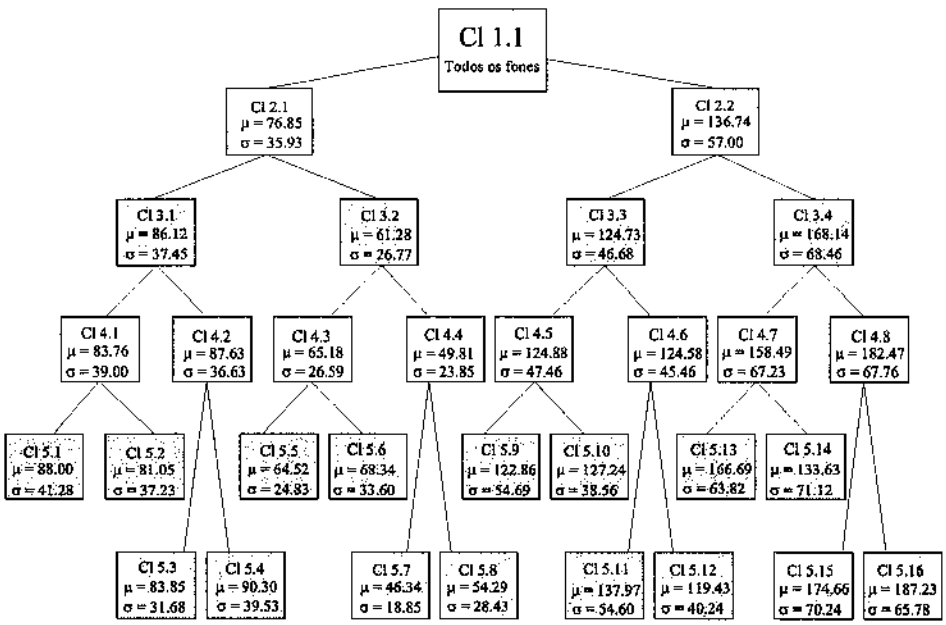


Figura. 6.19: Árvore de clusterização: Primeira Parte.

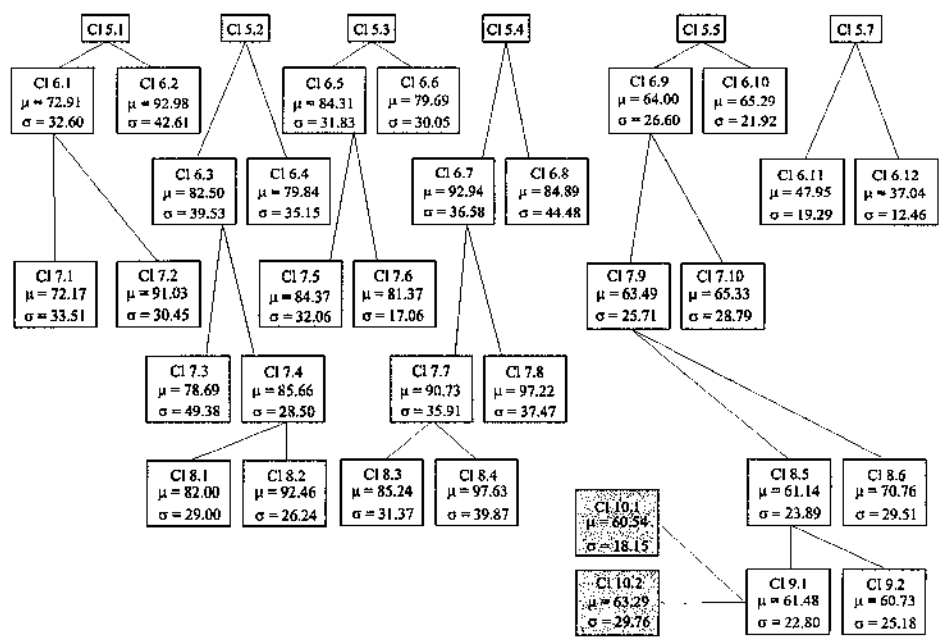


Figura. 6.20: Árvore de clusterização: Segunda Parte.

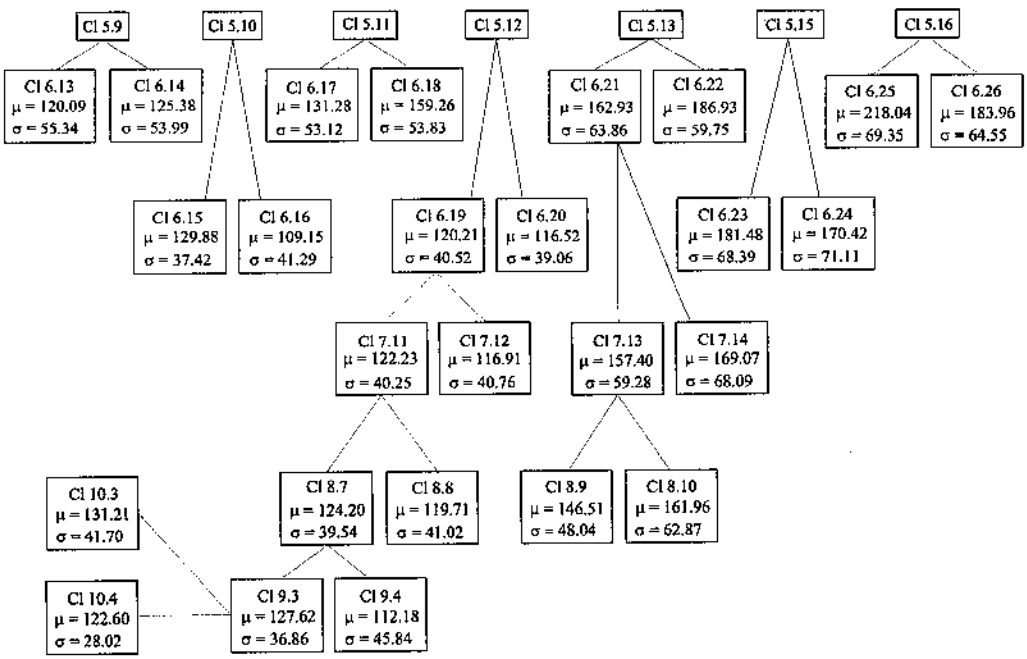


Figura. 6.21: Árvore de clusterização: Terceira Parte.

6.4.2 Seleção das Classes de Fones Ótimas

A seleção das *classes-de-fones* ótimas foi realizada aplicando-se o algoritmo descrito na Seção 5.5.7 na árvore de clusterização, Figuras 6.19, 6.20 e 6.21. Foram selecionadas 32 classes de fones, sendo 23 destas classes possuem um único fone, 7 classes possuem 2 fones, 1 classes possui 3 fones e a última classe possui 5 fones. As classes com mais de um fone são: C01 ([aa],[sh]); C02 ([Q],[b]); C03 ([H],[z]); C04 ([AR],[OR]); C05 ([dh],[dx]); C06 ([ccc],[u],[v],[w],[y]); C07 ([g],[i],[zh]); C08 ([ae],[ii]); C09 ([ai],[oi]). A Figura 6.22 mostra os coeficientes de correlação obtidos para cada uma das classes de fones com mais de um fone (C01, C02, C03, C04, C05, C05, C07, C08 e C09) e os compara com o valor médio dos coeficientes de correlação estimados para os fones contidos em cada uma destas classes.

6.4.3 Topologias Ótimas por Classes de Fones

A Figura 6.23 mostra as topologias ótimas obtidas para cada uma das 32 classes de fones selecionadas. É importante observar que para as classes de fones com mais de um fone (C01, C02, C03, C04, C05, C05, C07, C08 e C09), o *fator* lingüístico phID se fez extremamente necessário.

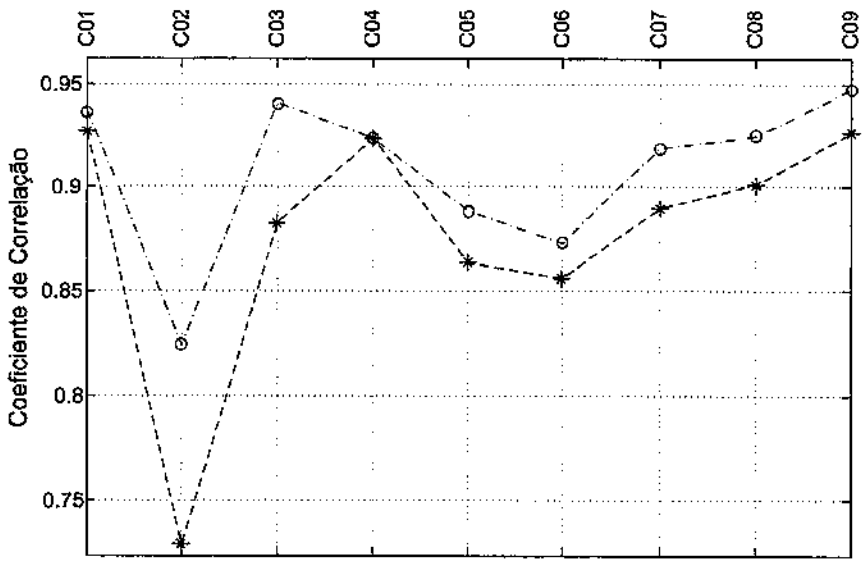


Figura. 6.22: Classes de fones selecionadas da árvore de clusterização. O símbolo (o) indica os resultados dos modelos QMTI/Cl + LDM-GA. O símbolo (*) indica a média dos modelos QMTI/Ph + LDM-GA (para os fones contidos em cada classe).

6.4.4 Resultados Comparativos: QMTI/Cl + LDMGA × QMTI/Cl + MANOVA × RT/Cl

A Figura 6.24 apresenta o resultado da avaliação dos modelos QMTI/Cl + LDM-GA quando comparados com os modelos QMTI/Cl + MANOVA e RT/Cl. Os coeficientes de correlação mostrados na Figura 6.24 são o resultado da média de 50 experimentos. Cada um destes 50 experimentos corresponde a uma diferente partição dos dados em dados de treinamento (80%) e validação (20%). Em cada um dos 50 experimentos os modelos QMTI/Cl + LDM-GA, QMTI/Cl + MANOVA e RT/Cl foram treinados utilizando-se apenas os dados de treinamento e avaliados utilizando-se os dados de validação. O objetivo de tomar a média destes 50 experimentos foi o de apresentar resultados estatisticamente mais confiáveis.

Testes estatísticos não paramétricos de Mann-Whitney (Gibbons, 1985) (a um nível de significância de 0,05) foram aplicados aos resultados das Figuras 6.10, 6.11 e 6.12 e os respectivos *p-value* encontram-se na Tabela 6.6.

6.5 Análise dos Efeitos dos *Fatores* Lingüísticos

Os modelos QMTI/Ph + LDM-GA ou QMTI/Cl + LDM-GA também podem ser úteis para análises puramente lingüísticas, como por exemplo para medir o efeito (influência) dos *fatores* lingüísticos na composição da duração segmental de fones ou classes-de-fone. A Figura 6.25 apresenta os efeitos dos *fatores* lingüísticos associados ao fone [e]. Os efeitos de cada *fator* são representados pelas

Tabela. 6.6: Testes estatísticos não paramétricos de Mann-Whitney: QMTI/Cl + LDM-GA versus QMTI/Cl + MANOVA e versus QMTI/Cl + RT

Fone	*p-value*: LDM-GA versus MANOVA	*p-value*: LDM-GA versus RT
[@]	$7,59 \cdot 10^{-1}$	$< 10,0 \cdot 10^{-7}$
[ER]	$5,46 \cdot 10^{-1}$	$< 10,0 \cdot 10^{-7}$
[au]	$7,96 \cdot 10^{-4}$	$< 10,0 \cdot 10^{-7}$
[ch]	$7,93 \cdot 10^{-2}$	$< 10,0 \cdot 10^{-7}$
[d]	$3,03 \cdot 10^{-1}$	$< 10,0 \cdot 10^{-7}$
[e]	$1,37 \cdot 10^{-1}$	$< 10,0 \cdot 10^{-7}$
[ei]	$4,22 \cdot 10^{-1}$	$< 10,0 \cdot 10^{-7}$
[f]	$2,37 \cdot 10^{-1}$	$< 10,0 \cdot 10^{-7}$
[jh]	$2,00 \cdot 10^{-1}$	$< 10,0 \cdot 10^{-7}$
[k]	$4,58 \cdot 10^{-1}$	$4,13 \cdot 10^{-1}$
[l]	$5,55 \cdot 10^{-1}$	$1,50 \cdot 10^{-2}$
[m]	$9,88 \cdot 10^{-1}$	$3,76 \cdot 10^{-2}$
[n]	$8,63 \cdot 10^{-1}$	$5,78 \cdot 10^{-2}$
[ng]	$5,88 \cdot 10^{-5}$	$< 10,0 \cdot 10^{-7}$
[oo]	$3,96 \cdot 10^{-7}$	$< 10,0 \cdot 10^{-7}$
[ou]	$4,54 \cdot 10^{-1}$	$5,06 \cdot 10^{-2}$
[p]	$1,69 \cdot 10^{-1}$	$2,03 \cdot 10^{-2}$
[r]	$7,64 \cdot 10^{-1}$	$1,67 \cdot 10^{-7}$
[s]	$5,01 \cdot 10^{-1}$	$2,96 \cdot 10^{-1}$
[t]	$9,53 \cdot 10^{-1}$	$3,19 \cdot 10^{-1}$
[th]	$9,75 \cdot 10^{-1}$	$< 10,0 \cdot 10^{-7}$
[uh]	$7,43 \cdot 10^{-1}$	$1,28 \cdot 10^{-1}$
[uu]	$2,13 \cdot 10^{-1}$	$< 10,0 \cdot 10^{-7}$
C01	$1,26 \cdot 10^{-3}$	$< 10,0 \cdot 10^{-7}$
C02	$3,15 \cdot 10^{-1}$	$< 10,0 \cdot 10^{-7}$
C03	$2,00 \cdot 10^{-1}$	$2,27 \cdot 10^{-2}$
C04	$2,31 \cdot 10^{-10}$	$< 10,0 \cdot 10^{-7}$
C05	$8,92 \cdot 10^{-2}$	$1,35 \cdot 10^{-4}$
C06	$1,97 \cdot 10^{-7}$	$1,77 \cdot 10^{-1}$
C07	$2,71 \cdot 10^{-1}$	$8,65 \cdot 10^{-1}$
C08	$7,38 \cdot 10^{-1}$	$< 10,0 \cdot 10^{-7}$
C09	$9,97 \cdot 10^{-1}$	$< 10,0 \cdot 10^{-7}$

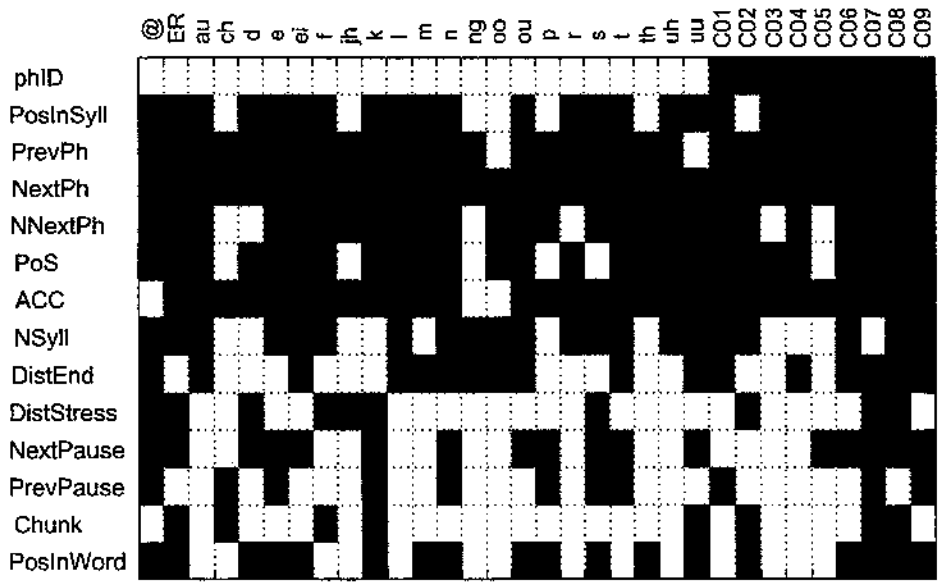


Figura. 6.23: Topologias para as classes de fones selecionadas.

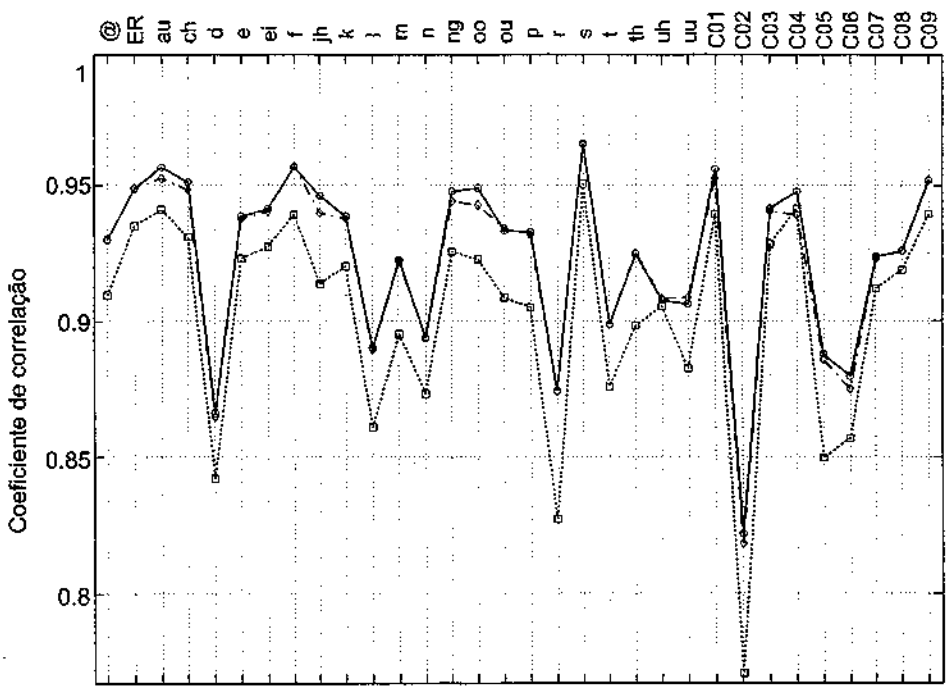


Figura. 6.24: Comparação de desempenho dos modelos QMTI/Cl + LDM-GA (indicado por ○), QMTI/Cl + MANOVA (indicado por ◇) e RT/Cl (indicado por □).

barras verticais (associadas aos níveis associados ao respectivo *fator*) e são medidos em milissegundos. Conforme apresentado na equação 5.16 do Capítulo 5, a soma dos efeitos causados pelos *fatores* lingüísticos pode ser interpretada como uma perturbação sobre o valor médio de duração associada ao

fone/classe-de-fone em análise.

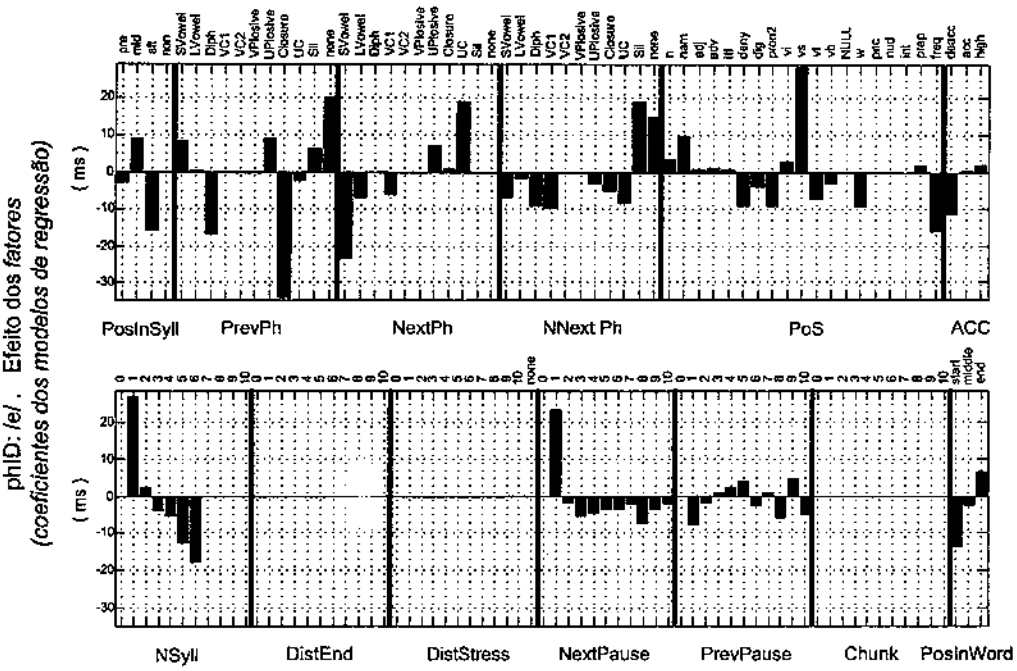


Figura. 6.25: Efeitos dos *fatores* lingüísticos associados ao fone [e].

6.6 Análise de Desempenho dos Modelos: QMTI/Ph + LDM-GA

Conforme dito na seção 5.2.2, a escolha de modelos de regressão linear para a modelagem da duração segmental da fala, deveu-se, principalmente, à simplicidade destes modelos e também por se tratarem de modelos totalmente conectados (Hansa and Sagisaka, 2004). Esta simplicidade dos modelos lineares foi extremamente importante para facilitar todo o processo de análise do algoritmo LDM-GA e, também, para gerar resultados mais fáceis de serem interpretados e analisados. Entretanto, conforme será mostrado nesta seção, os modelos de regressão linear, otimizados pelo método LDM-GA e estimados pelo método QMTI, ainda apresentam várias limitações. Estas limitações se devem principalmente à ausência de interações mais complexas entre os *fatores* lingüísticos (diferentemente, por exemplo, dos modelos *Sum-of-Products* - SoP, (Santen, 1994)) ou ANN (*Artificial Neural Network*).

A Figura 6.26 mostra os erros em valores RMS (*Root Mean Squared Errors*) para todos os 45 fones utilizados (Tabela 5.4).

Pode-se verificar na Figura 6.26 que para alguns fones, por exemplo, [ai], [ei] e [uu], este erro RMS de predição chega a quase $30ms$. Para se poder ter uma melhor percepção dos erros em valores RMS apresentados na Figura 6.26, a Figura 6.27 apresenta uma versão normalizada destes erros RMS. Esta normalização é realizada segundo a equação 6.6:

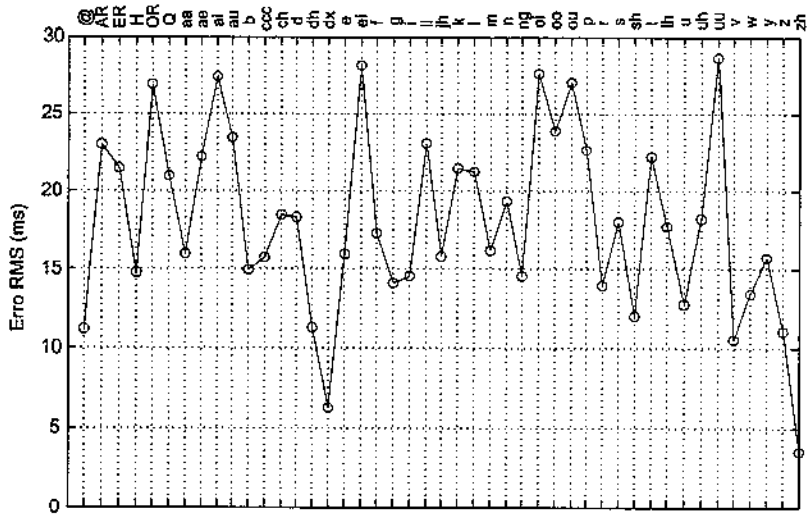


Figura. 6.26: Erro em RMS dos modelos QMTI/Ph + LDM-GA para todos os 45 fones da Tabela 5.4

$$\epsilon(Ph_k) = \frac{RMS\,(Ph_k)}{média\,(dur\,Ph_k)} \cdot 100\% \tag{6.6}$$

A equação 6.6 divide os erros cometidos na predição da duração de cada fone (erros em valores RMS) pela sua respectiva duração média. Os resultados são multiplicados por 100 e interpretados como percentuais de erro de duração cada fone (em valores RMS) em relação ao seus respetivos valores médios de duração.

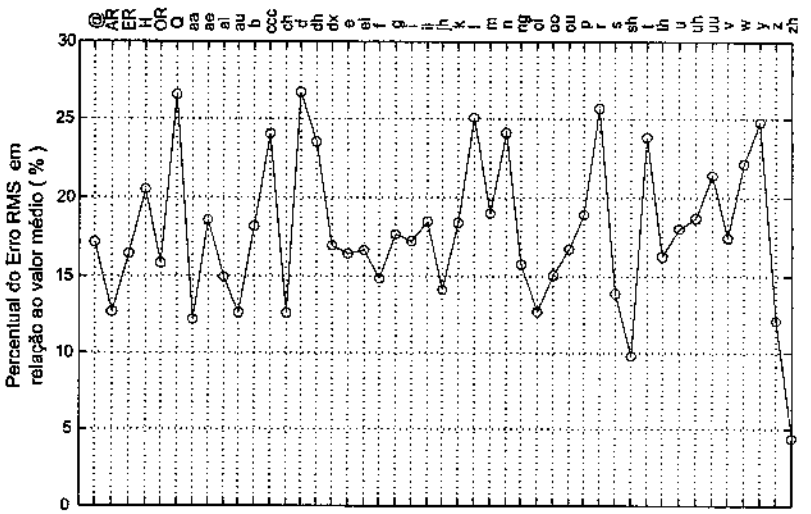


Figura. 6.27: Percentual de Erro RMS dos modelos QMTI/Ph + LDM-GA para todos os 45 fones da Tabela 5.4

A Figura 6.26 mostra que os fones [ei] e [oi] possuem erros RMS próximos a 30 milisegundos. Por outro, lado a Figura 6.27 mostra que o percentual de erro RMS dos fones [ei] e [oi] (conforme definido em 6.6) é de apenas 15%.

As Figuras 6.28, 6.29, 6.30, 6.31, 6.32, 6.33, 6.34, 6.35 e 6.36 mostram resultados da avaliação dos modelos QMTI/Ph + LDM-GA para os fones [ae], [b], [d], [e], [g], [H], [p], [r] e [z], respectivamente. Cada uma destas Figuras é composta por dois gráficos, alinhados em relação ao eixo das durações segmentais. O primeiro apresenta o histograma da duração do fone em análise (durações presentes na base de dados), e o segundo apresenta a relação entre as durações esperadas e preditas. No segundo gráfico as durações esperadas (durações presentes na base de dados) são indicadas pela legenda (o), e os valores de duração preditos, pela legenda ($\times$). A linha horizontal, representada pelos símbolos ($*$) indica o valor médio do fone em análise. Algumas observações importantes sobre estas figuras são:

- Os modelos QMTI + LDM-GA não são capazes de realizar boas predições para valores de duração que se desviam muito do valor médio de duração.
- Nem todas as distribuições possuem longas caudas à direita. Por exemplo, as distribuições dos fones [b] e [g] (Figuras 6.29 e 6.32) podem ser consideradas quase simétricas ou com cauda ligeiramente maior à esquerda. Portanto, conforme amplamente discutido por (Bellegarda and Silverman, 2001), a utilização de transformações logarítmicas para modificar as distribuições originais (na tentativa de aproximá-las de uma distribuição normal), nem sempre é uma solução apropriada.
- Distribuições com caráter bimodal, como por exemplo o fone [r], Figura 6.35, tendem a aglomerar as durações preditas em torno de regiões da distribuição com maior concentração de dados.
- As durações preditas para o fone [z], 6.36, apresentaram um comportamento até então não muito bem explicado, concentrando-se em três regiões bem definidas, apesar de suas distribuições não apresentarem um caráter explicitamente trimodal.

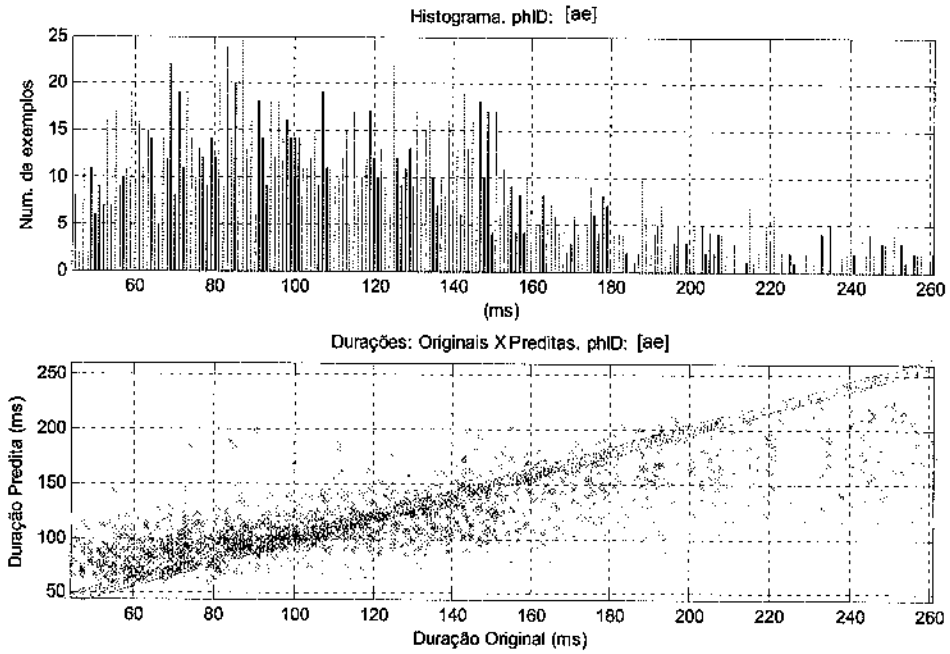


Figura. 6.28: Desempenho do modelo QMTI/Ph + LDM-GA para o fone [ae]

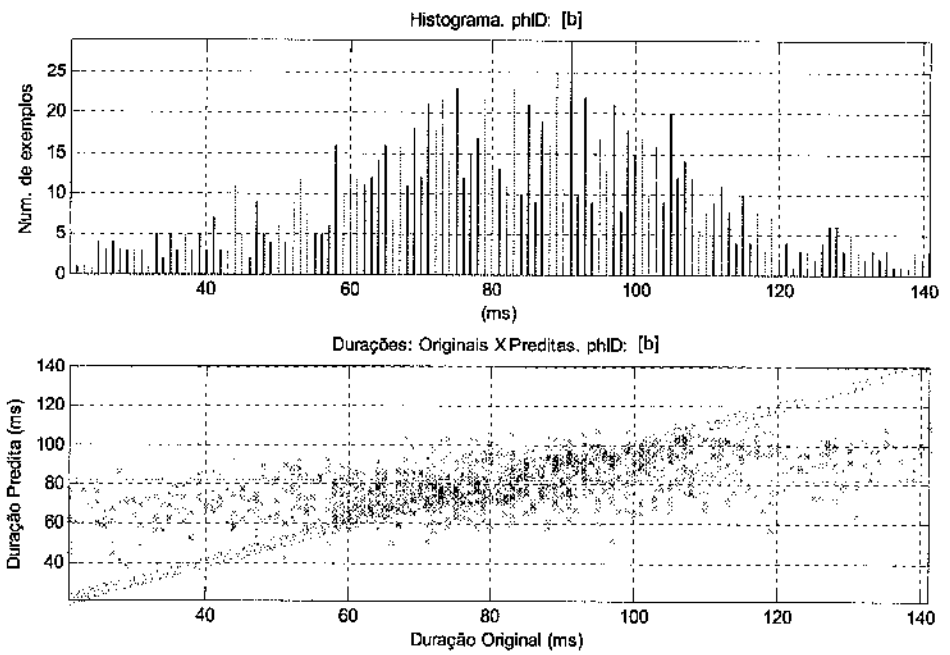


Figura. 6.29: Desempenho do modelo QMTI/Ph + LDM-GA para o fone [b]

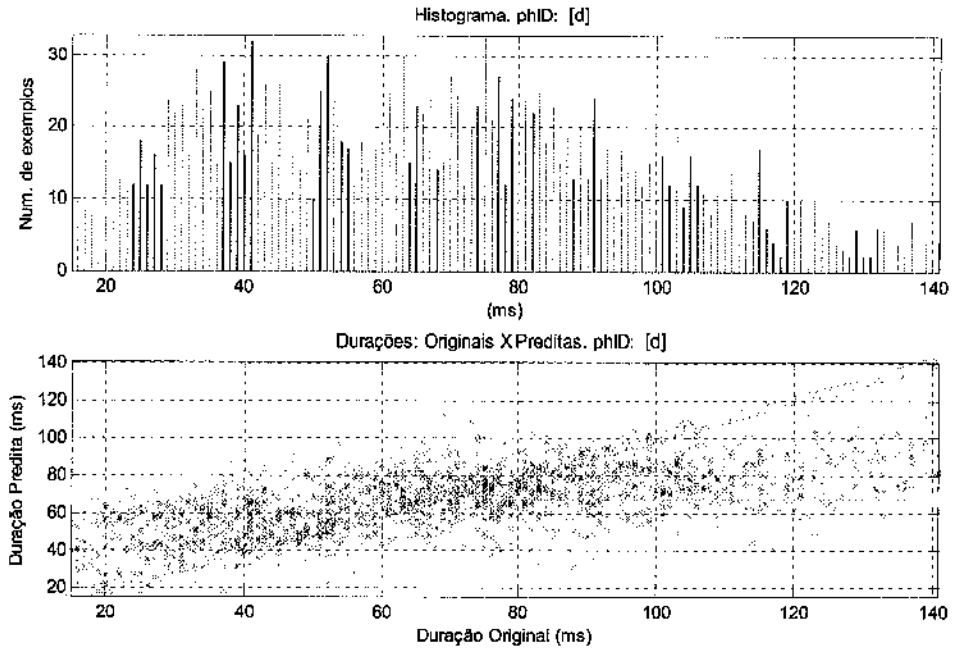


Figura. 6.30: Desempenho do modelo QMTI/Ph + LDM-GA para o fone [d]

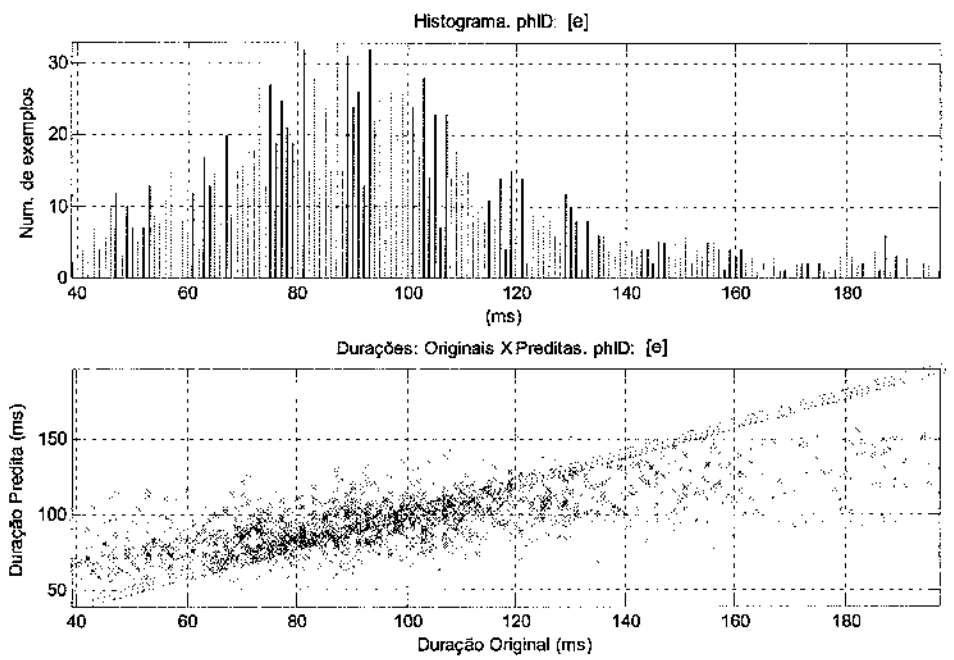


Figura. 6.31: Desempenho do modelo QMTI/Ph + LDM-GA para o fone [e]

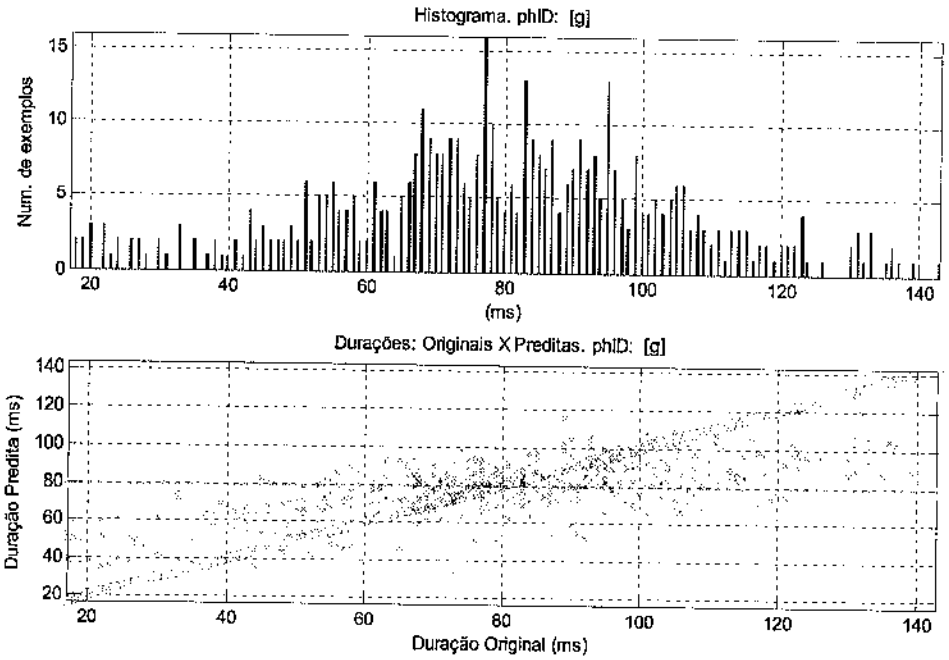


Figura. 6.32: Desempenho do modelo QMTI/Ph + LDM-GA para o fone [g]

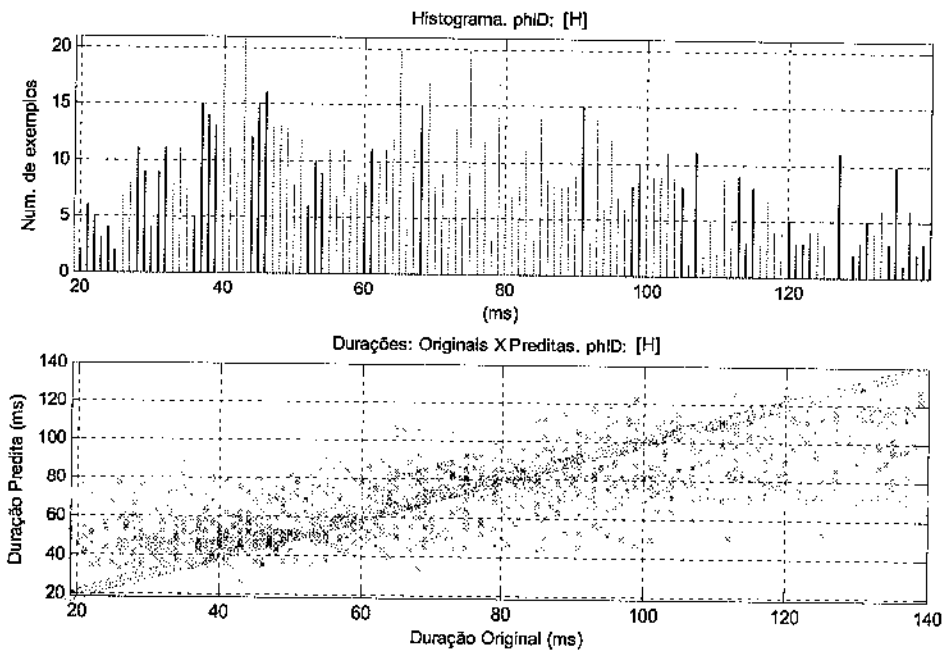


Figura. 6.33: Desempenho do modelo QMTI/Ph + LDM-GA para o fone [H]

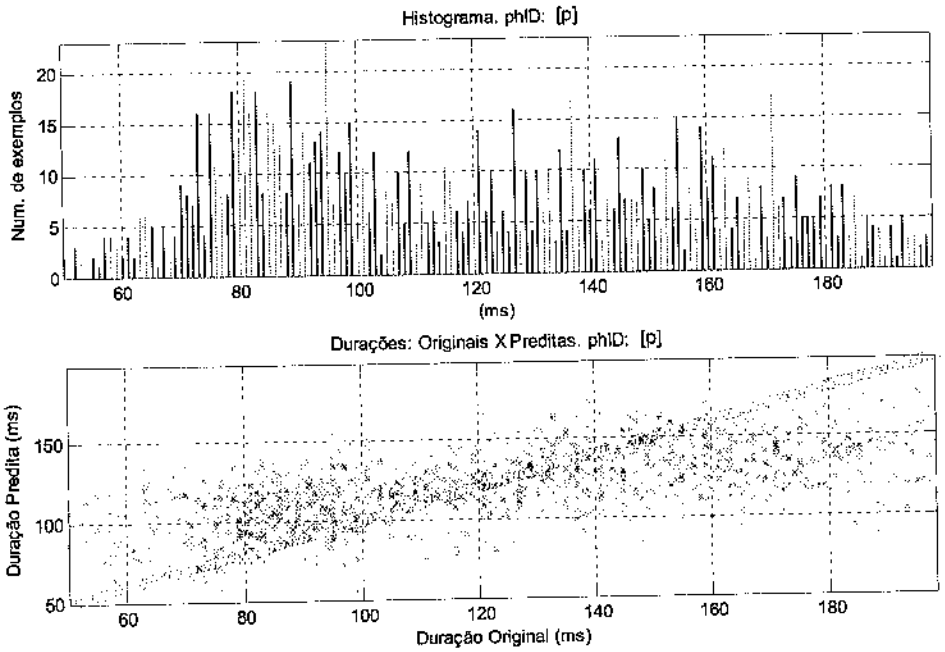


Figura. 6.34: Desempenho do modelo QMTI/Ph + LDM-GA para o fone [p]

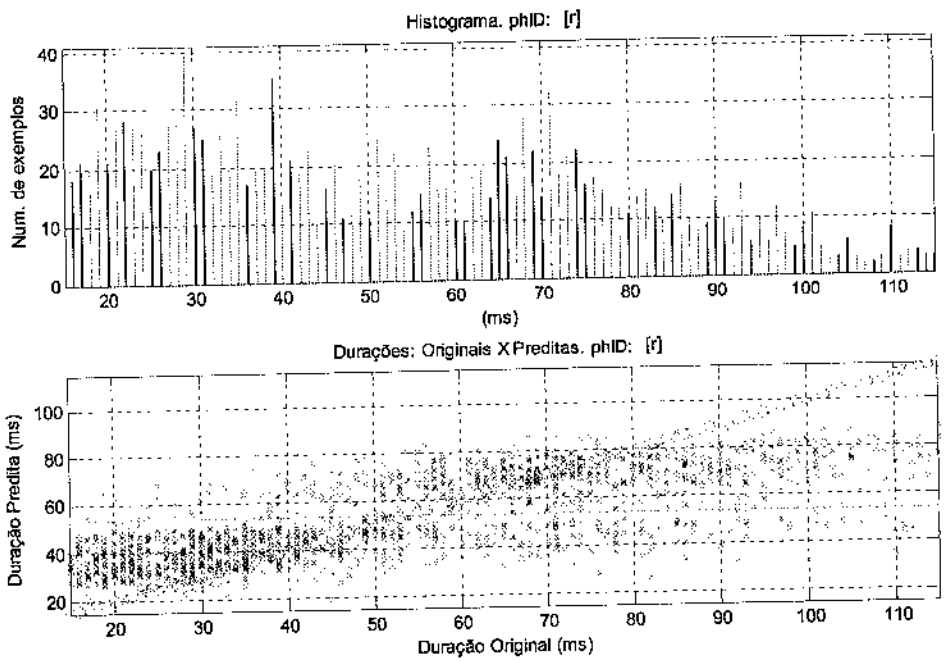


Figura. 6.35: Desempenho do modelo QMTI/Ph + LDM-GA para o fone [r]

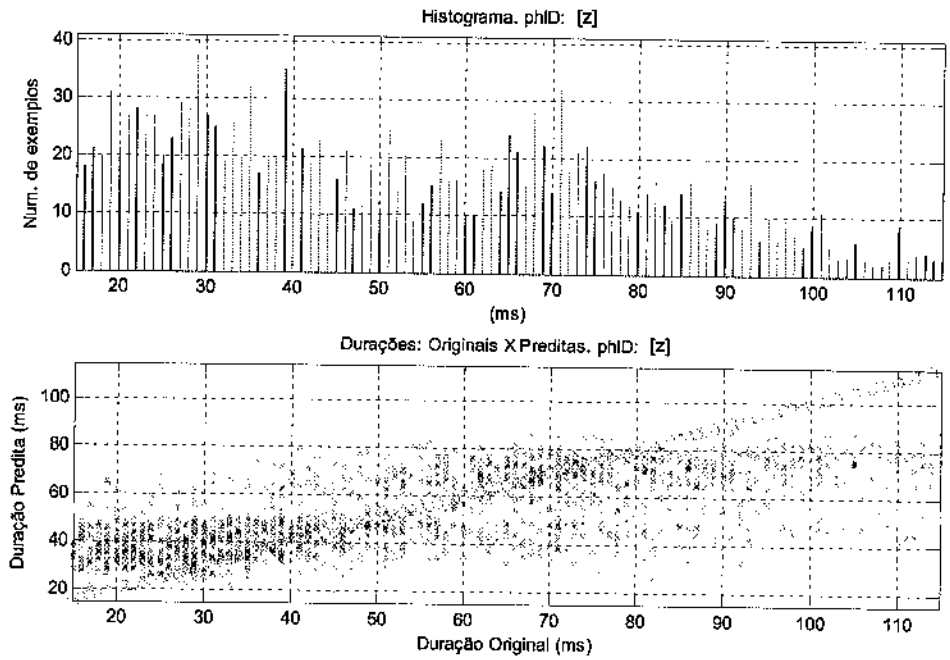


Figura. 6.36: Desempenho do modelo QMTI/Ph + LDM-GA para o fone [z]

6.7 Considerações Finais

6.7.1 Considerações Sobre os Resultados

Este Capítulo apresentou vários experimentos que avaliam o desempenho do algoritmo LDM-GA na otimização de modelos de regressão linear (modelos QMTI) aplicados à modelagem da duração segmental da fala. Modelos QMTI + LDM-GA foram comparados a modelos QMTI + MANOVA e também a modelos de regressão estimados utilizando árvores de regressão - RT (*Regression Trees*). Os resultados obtidos confirmam as seguintes suposições iniciais:

- Os modelos QMTI/Ph + LDM-GA apresentam uma capacidade de generalização significativamente superior aos modelos QMTI Cheios.
- Os modelos QMTI/Ph + LDM-GA apresentam uma capacidade de generalização ligeiramente superior aos modelos QMTI/Ph + MANOVA.
- De 1 e 2 pode-se concluir que o processo de seleção de topologias ótimas do algoritmo LDM-GA, mesmo sem qualquer tipo de clusterização dos fones, se mostrou eficiente.
- Os modelos QMTI/Ph + LDM-GA apresentam uma melhor capacidade de generalização que as árvores de regressão. Este resultado mostra que, apesar de as árvores de regressão serem métodos não-lineares (de fato, lineares por partes), elas não apresentam um boa capacidade de generalização, mesmo quando submetidas a técnicas de poda (*pruning*).

- O método de clusterização hierárquica binária do algoritmo LDM-GA se mostrou eficiente na seleção de classes de fones que maximizam o desempenho global dos modelos de regressão.
- Das nove classes de fones selecionadas, que possuem mais de um fone, pode-se verificar que os fones agrupados (clusterizados) não necessariamente possuem qualquer similaridade fonética, quanto a aspectos duracionais, entre si.
- Apesar de os modelos de regressão lineares QMTI terem se mostrado úteis na análise e validação do algoritmo LDM-GA, eles se mostraram demasiadamente simples para modelar adequadamente a duração segmental da fala. Muito provavelmente, esta limitação dos modelos de regressão lineares QMTI deve-se ao fato de estes modelos não contemplarem interações mais complexas entre fatores lingüísticos. Esta limitação dos modelos QMTI/Ph + LDM-GA pode ser observada nos elevados erros de modelagem para os valores de duração que se encontram distantes do valor médio, conforme mostrado nas Figuras 6.28, 6.29, 6.30, 6.31, 6.32, 6.33, 6.34, 6.35 e 6.36.

Capítulo 7

Algoritmo OPWI: Fundamentos Teóricos

7.1 Introdução

Em um sistema CTF-SCAUS (Hunt and Black, 1996), (Morais and Violaro, 2005c), a síntese do sinal de fala é de responsabilidade do módulo de *Back-End*. Este módulo recebe como entrada uma seqüência de unidades de síntese (selecionada pelo módulo de seleção automática de unidades), e os contornos prosódicos (gerados pelo módulo prosódico), e fornece como saída o sinal de fala sintetizado. Entre as principais operações realizadas pelo módulo de *Back-End* destacam-se: (1) Concatenação das unidades de síntese; (2) Normalização do contorno de energia na fronteira entre as unidades de síntese concatenadas; (3) Suavizações espectrais nas fronteiras entre as unidades de síntese concatenadas (por exemplo: suavizações das freqüências centrais e larguras de banda dos formantes); (4) Modificações prosódicas, tais como modificações na taxa de articulação - TSM (*Time Scale Modifications*) e modificações no contorno da freqüência fundamental - PSM (*Pitch Scale Modifications*).

Algumas das técnicas de *Back-End* mais comumente utilizadas no estado-da-arte de sistemas CTF-SCAUS são: TD-PSOLA (*Time-Domain Pitch Synchronous Overlap and Add*) (Moulines and Charpentier, 1990), MBROLA (*Multiband Resynthesis Overlap and Add*) (Dutoit and Leich, 1993), HNM (*Harmonic plus Noise Model*) (Stylianou, 1996) e modelos baseados em análise preditiva linear - LPC (*Linear Prediction Coding*) (Edgington et al., 1998), (Pacheco and Seara, 2002).

A técnica TD-PSOLA tem sido a mais largamente empregada nos módulos de *Back-End* de sistemas CTF-SCAUS comerciais. O sucesso deste algoritmo deve-se, principalmente, ao seu baixo custo computacional associado a uma excelente qualidade segmental (na síntese do sinal de fala sem modificação prosódica) e um bom desempenho na realização de modificações prosódicas. O algoritmo TD-PSOLA opera sincronamente com os pulsos glotais (mais especificamente com os Instantes de Fechamento da Glote - IFGs) e diretamente sobre o sinal de fala, sem parametrizá-lo explicitamente; por esta razão ele é classificado como um algoritmo não-paramétrico. O elemento básico do algoritmo TD-PSOLA é o que será denominado *segmento de pitch*, o qual é estimado a cada IFG. O *i−ésimo segmento de pitch* do algoritmo TD-PSOLA, é definido como um segmento do sinal de fala (centrado no *i−ésimo* IFG e extendendo-se do IFG anterior até o IFG seguinte) multiplicado por uma janela de Hanning

assimétrica (centrada no *i−ésimo* IFG e extendendo-se do IFG anterior até o IFG seguinte). O algoritmo TD-PSOLA ressintetiza o sinal de fala através de uma mera operação de *Overlap and Add* dos *segmentos de pitch*, garantindo uma perfeita reconstrução do sinal de fala. As modificações de TSM são realizadas através de repetições e/ou eliminações de *segmentos de pitch* seguidas pelo processo de *Overlap and Add*. As modificações de PSM são realizadas aumentando-se ou diminuindo-se o graus de superposição entre os *segmentos de pitch* antes do processo de *Overlap and Add*.

Se por um lado a característica não-paramétrica da técnica TD-PSOLA garante a ela um reduzido custo computacional nas operações de ressíntese e de modificações prosódicas (TSM e PSM), por outro, ela é considerada a principal responsável por alguns dos problemas e limitações do algoritmo TD-PSOLA na realização de suavizações espectrais entre unidades de síntese. Entre esses problemas/limitações destacam-se:

- *Descasamento de fase.* O mau posicionamente dos instantes de análise (estimativas imprecisas dos IFGs), poderá causar descontinuidades na concatenação de unidades de síntese;
- *Descasamento de freqüência fundamental* (F_0). A concatenação de duas unidades de síntese com envelopes espectrais semelhantes e com IFGs corretamente posicionados, porém com diferentes valores de F_0, poderá ocasionar descontinuidades espectrais;
- *Envelopes espectrais distintos.* Por ser uma técnica não-paramética operando no domínio do tempo, o algoritmo TD-PSOLA não apresenta nenhuma maneira imediata para ajustar (suavizar/interpolar) os envelopes espectrais presentes na fronteira entre as unidades de síntese a serem concatenadas.

Outro problema tipicamente associado a esta característica não-paramétrica do algoritmo TD-PSOLA é a sua incapacidade de modelar adequadamente os segmentos mistos da fala (segmentos compostos por uma componente quase-periódica e uma componente ruidosa), como por exemplo, fricativas sonoras. A redução na taxa de articulação de sons mistos através do algoritmo TD-PSOLA normalmente introduz um caráter metálico ao sinal sintetizado. Isto ocorre porque o algoritmo TD-PSOLA reduz a taxa de articulação da fala através da simples repetição dos *segmentos de pitch*, o que, por conseguinte, introduz uma periodicidade na componente ruidosa que poderia não existir originalmente.

A técnica MBROLA procura contornar os problemas apresentados pelo algoritmo TD-PSOLA, no processo de concatenação de unidades de síntese, ressintetizando os *segmentos de pitch* do algoritmo TD-PSOLA com uma fase modificada (para garantir o perfeito alinhamento entre estes *segmentos de pitch*) e F_0 constantes. Esta ressíntese destes *segmentos de pitch* é feita utilizando a técnica MBE (*Multiband Excitation Vocoder*) (Griffin and Lim, 1988). Durante a concatenação das unidades de síntese, os *segmentos de pitch* (com F_0 constantes e fases modificadas) ao longo da fronteira de concatenação são interpolados para garantir uma melhor suavização espectral (Dutoit, 1997). Apesar de a técnica MBROLA garantir uma boa suavização espectral na fronteira entre as unidades de síntese, as modificações de fase e de F_0, realizadas com a técnica MBE, normalmente introduzem algumas degradações na qualidade vocal do sinal sintetizado.

Vários métodos que empregam o modelo fonte-filtro (Quatieri, 2002) (geralmente utilizando análise preditiva linear - LPC - *Linear Predictive Coding*), têm sido propostos para operarem como módulo de *Back-End*. Entre eles destacam-se *Vocoders* LPC com excitações mistas (Huang and Acero, 1998) e RELP *Residual-Excited Linear Prediction* (Edgington et al., 1998). Em *Vocoders* LPC, modificações na fonte de excitação devem sempre vir acompanhadas de modificações no trato vocal. Se a interação entre a fonte de excitação e o trato vocal não for levada em consideração, então possíveis degradações de qualidade vocal poderão ser ouvidas no sinal sintetizado. O grande problema com *Vocoders* LPC é que esta interação entre fonte de excitação e trato vocal ainda não é um fenômeno bem explicado e modelado matematicamente. Esta interação, geralmente, se manifesta com maior intensidade entre locutores com altos valores de F_0, explicando, portanto, as dificuldades dos *Vocoders* LPC em produzirem boas vozes para locutores do sexo feminino e para crianças. Melhorias na qualidade das modificações prosódicas obtidas com a técnica RELP foram apresentadas por Edgington em (Edgington et al., 1998) (*British Telecom - Laurente Text-to-Speech system*). A técnica proposta por Edgington opera sincronamente com os pulsos glotais, "re-amostrando" o resíduo LPC em regiões próximas aos instantes de abertura da glote (a fase do ciclo glotal em que o sinal de fala é perceptivamente menos audível) e preservando o resíduo LPC em regiões próximas aos instantes de fechamento da glote.

A técnica HNM (Stylianou, 1996) assume que os segmentos sonoros do sinal de fala são sons mistos, compostos por uma componente harmônica somada a uma componente ruidosa. A componente harmônica representa a parte "quase-periódica" do sinal de fala e a componente ruidosa a parte não-periódica (por exemplo, os ruídos presentes nas fricativas sonoras). Estas duas componentes são separadas no domínio da freqüência por um parâmetro variável no tempo, denominado *maximum voiced frequency*, $F_m(n)$. Os componentes de freqüência do espectro abaixo da freqüência $F_m(n)$ são associados à componente harmônica, enquanto as componentes de freqüência do espectro acima de $F_m(n)$ são associados à componente ruidosa. Nos segmentos não-sonoros da fala, a variável $F_m(n)$ é forçadamente ajustada para 0, isto é, todos os componentes de freqüência dos segmentos não-sonoros são considerados não-harmônicos. Apesar de estas suposições não serem claramente válidas do ponto de vista da produção da fala, elas representam uma boa aproximação do ponto de vista perceptivo, garantindo à técnica HNM análise/ressíntese e modificações prosódicas de alta qualidade.

Com o objetivo de superar alguns problemas apresentados pelas técnicas TD-PSOLA, MBROLA, Vocoders LPC, RELP e HNM, este Capítulo apresenta um novo algoritmo para análise e síntese do sinal de fala, denominado OPWI (*Optimized Prototype Waveform Interpolation*). O algoritmo OPWI explora conceitos da técnica HNM e de duas outras técnicas já bem estabelecidas nas áreas de codificação de fala: TFI (*Time Frequency Interpolation*) (Shoham, 1993) e WI (*Waveform Interpolation*) (Kleijn and Paliwal, 1998). A técnica OPWI garante não só ressínteses com relações sinal/ruído acima de 30 dB (ligeiramente acima da técnica HNM), mas também permite modificações prosódicas (TSM e PSM) contínuas e de alta qualidade. Além disso, a estrutura paramétrica da técnica OPWI permite que diferentes técnicas de interpolação (simples e eficientes) possam ser utilizadas para minimizar descontinuidades na concatenação entre unidades de síntese.

Outra importante vantagem da estrutura paramétrica do algoritmo OPWI é que ela permite imple-

mentações elaboradas do processo de fusão de unidades de síntese proposto por Kagoshima (Mizutani and Kagoshima, 2005). Conforme mencionado no Capítulo 1, este método de fusão de unidades tem como objetivo aumentar o *nível de estabilidade vocal* dos sistemas CTF-SCAUS.

Este Capítulo limita-se a apresentar os aspectos descritivos da formulação teórica/matemática do algoritmo OPWI. A seção 7.2 apresenta os princípios básicos do algoritmo OPWI. A seção 7.3 descreve os processos de estimativa dos instantes de análise e de segmentação sonoro/não-sonoro do sinal de fala. A decomposição CEL/CER é descrita na seção 7.4. Os critérios para a estimativa do nível de estacionariedade do sinal de fala é apresentado na seção 7.5. A seção 7.6 introduz os conceitos de protótipos ótimos e de suas respectivas representação temporais. O primeiro e o segundo critério para a estimativa dos protótipos ótimos são descritos nas seções 7.7 e 7.8, respectivamente. A seção 7.9 apresenta as operações envolvidas no processo de análise da componente CER. O processo de síntese da componente CEL é descrito na seção 7.10. As operações de modificação prosódica (PSM e TSM) são apresentadas na seção 7.11. A seção 7.12 descreve a síntese da componente CER. Uma proposta para suavização espectral na junção entre unidades de síntese é apresentada na seção 7.13. A seção 7.14 discute alguns aspectos sobre o custo computacional do algoritmo OPWI (tanto nas etapas de análise quanto de síntese) e apresenta técnicas de otimização capazes de reduzir drasticamente estes custos. Por último, a seção 7.15 encerra este Capítulo com algumas considerações finais.

Resultados experimentais e análises de desempenho do algoritmo OPWI serão apresentados no Capítulo 8. Estes resultados e análises serão de fundamental importância para auxiliar na compreensão dos fundamentos teóricos/matemáticos a serem apresentados neste Capítulo.

7.2 Formulação do Problema

De forma semelhante à técnica HNM, o algoritmo OPWI também assume que o sinal de fala pode ser decomposto em duas componentes, representando as estruturas quase-periódica e aperiódica do sinal de fala. Entretanto, diferentemente da técnica HNM, a decomposição utilizada no algoritmo OPWI não utiliza o conceito de *maximum voiced frequency* e suas duas componentes cobrem todo o espectro de freqüência do sinal. As componentes "quase-periódica" e aperiódica da técnica OPWI são denominadas CEL - *Componente de Evolução Lenta* e CER - *Componente de Evolução Rápida*, respectivamente. A componente CER é modelada como um processo auto-regressivo, por meio de análise preditiva linear (Quatieri, 2002). A síntese da componente CER é realizada utilizando-se o algoritmo LP-PSOLA (*Linear Prediction Pitch Synchronous Overlap and Add*). A componente CEL é modelada como uma interpolação no domínio tempo-freqüência (Shoham, 1993) de uma seqüência de protótipos ótimos estimados a partir de segmentos da componente CEL extraídos sincronamente com os pulsos glotais (e em posições próximas aos IFGs).

A decomposição CEL/CER do algoritmo OPWI inspirou-se na decomposição SEW/REW (*Slowly Evolving Waveform/Rapid Evolving Waveform*) da técnica WI (Kleijn and Paliwal, 1998). Entretanto, diferentemente da decomposição SEW/REW, a decomposição CEL/CER opera diretamente sobre o sinal de fala e não sobre o resíduo de predição linear. A estimação dos protótipos ótimos utilizados

para análise e ressíntese da componente CEL foi inspirada, conjuntamente, no conceito de estrutura harmônica da técnica HNM (Stylianou, 1996) e na interpolação de protótipos no domínio tempo-freqüência da técnica TFI.

7.2.1 Etapas de Análise e Síntese

A Figura 7.1 apresenta um diagrama de blocos com as principais operações das etapas de análise e síntese do algoritmo OPWI. Em sistemas CTF-SCAUS as operações envolvidas na etapa de análise do algoritmo OPWI podem ser realizadas *off-line* e, portanto, o custo computacional envolvido em tais operações não é uma questão relevante. Por outro lado, em sistemas CTF-SCAUS, todas as operação da etapa de síntese do algoritmo OPWI, normalmente, devem ser executadas *on-fly* (em tempo de execução). Por conseguinte, é fundamental que estas operações apresentem um baixo custo computacional.

Operações envolvidas na etapa de análise

- *Estimativa dos Instantes de Análise.* Em trechos sonoros do sinal de fala todas as análises do algoritmo OPWI são realizadas sincronamente com os pulsos glotais. A cada pulso glotal, um instante de análise é posicionado junto ao ponto de máximo ou de mínimo do sinal que se encontra próximo ao instante de fechamento da glote (IFG). Nos trechos não-sonoros os instantes de análise são definidos através de um processo de interpolação dos períodos fundamentais dos trechos sonoros adjacentes (à esquerda e à direita).
- **Classificação sonoro/não-sonoro.** Todos os quadros de análise (segmentos do sinal de fala delimitados por dois instantes de análise consecutivos) são submetidos a uma classificação sonoro *versus* não-sonoro (S/NS).
- *Decomposição CEL/CER* - O sinal é decomposto nas componentes CEL e CER.
- *Modelagem da componente CEL.* O processo de modelagem da componente CEL consiste na determinação do nível de estacionariedade de cada quadro de análise e na estimativa dos protótipos ótimos.

  - *Determinação do nível de estacionariedade da componente CEL* - Um procedimento baseado em características temporais e espectrais é utilizado para a determinação do nível de estacionariedade de cada quadro de análise da componente CEL.
  - *Estimativa dos protótipos ótimos* - A cada instante de análise um protótipo ótimo é estimado. A estimativa destes protótipos ótimos pode ser realizada através de dois métodos distintos (com diferentes resoluções tempo/freqüência). A escolha de qual desses métodos deve ser utilizado a cada quadro de análise será função do nível de estacionariedade associado ao quadro de análise em questão.

- *Modelagem da componente CER* - Esta modelagem inicia com uma análise auto-regressiva, por meio de análise preditiva linear, seguida por uma filtragem inversa, para estimar o resíduo de predição.

Operações envolvidas na etapa de síntese

- *Síntese da componente CEL* - Realizada através de um procedimento de interpolação tempo-freqüência dos protótipos ótimos (de forma semelhante ao método proposto em (Shoham, 1993)).
- *Síntese da componente CER* - Realizada segundo o algoritmo LP-PSOLA. Este método emprega um procedimento de *Overlap and Add* para sintetizar o resíduo de predição da componente CER e em seguida este resíduo sintetizado é filtrado pelo filtro de predição linear, gerando a componente CER sintetizada.

7.3 Estimativa dos Instantes de Análise e Segmentação Sonoro/Não-Sonoro

O primeiro passo da etapa de análise do algoritmo OPWI consiste na estimativa dos instantes de análise (Tuan and d'Alessandro, 1999), (Huang and Acero, 1998), e na classificação sonoro *versus* não-sonoro dos quadros de análise. Uma maneira eficiente para a estimativa dos instantes de análise é a gravação simultânea do sinal de fala e da dinâmica dos pulsos glotais por meio de um laringógrafo digital (Morais et al., 2000).

Neste trabalho, a marcação dos os instantes de análise e a classificação sonoro *versus* não-sonoro foram realizadas diretamente a partir do sinal de fala empregando o *software* Praat (Boersma and Weenink, 2005), versão 4.4. A Figura 7.2 mostra os instantes de análise estimados pelo *software* Praat, para um segmento sonoro do sinal de fala.

Durante a etapa de análise do algoritmo OPWI, a distância entre os instantes de análise (imediatamente consecutivos) n_i^a e n_j^a será definida como N_{ij}^a. Nos quadros de análise classificados como sonoros, a distância N_{ij}^a será igual ao período fundamental T_{0i} em amostras, que por sua vez está relacionado à freqüência fundamental F_{0_i} em Hz através da equação $T_{0_i} = \frac{Fs}{F_{0_i}}$; sendo Fs a freqüência de amostragem empregada na aquisição do sinal. Entretanto, durante a etapa de síntese, esses instantes de análise passarão a ser denominados instantes de síntese, os quais poderão ser movidos de suas posições originais, devido a modificações prosódicas de TSM ou PSM, conforme será descrito na seção 7.11. A distância entre os instantes de síntese n_i^s e n_j^s será denominada N_{ij}^s.

O *software* Praat estima os instantes de análise apenas para os segmentos sonoros do sinal fala. Os instantes de análise dos segmentos não-sonoros são obtidos através de um processo de interpolação entre o período fundamental associado ao último instante de análise do segmento sonoro precedente, ${T_{0_i}}^{ant}$, e o período fundamental associado ao primeiro instante de análise do segmento sonoro seguinte, ${T_{0_i}}^{seg}$. Este procedimento garante uma transição suave dos períodos fundamentais localizados nas

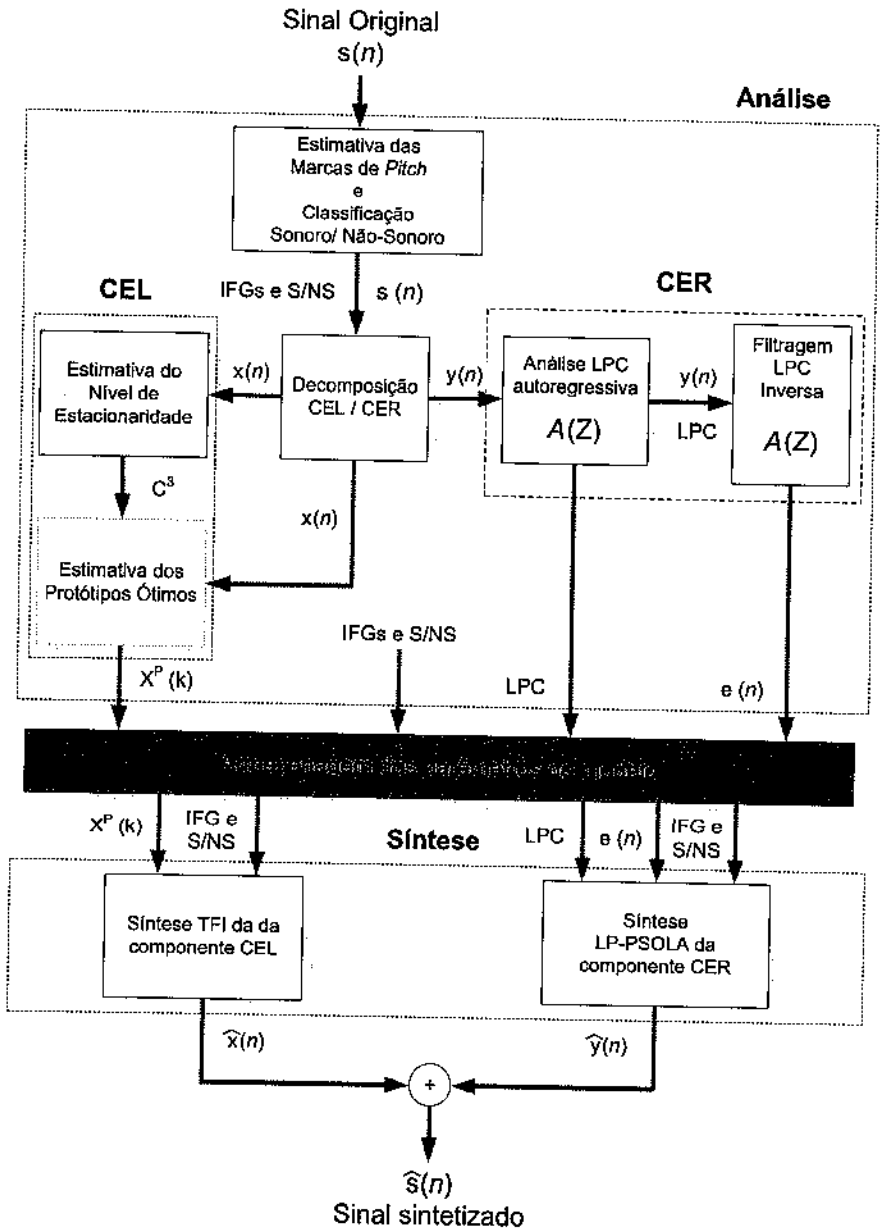


Figura. 7.1: Diagrama de blocos das etapas de análise e síntese do algoritmo OPWI.

transições entre os segmentos sonoros e não-sonoros e vice-versa. A Figura 7.3 ilustra os instantes de análise para um segmento não-sonoro do sinal de fala.

7.4 Decomposição CEL × CER

O processo de decomposição CEL/CER do algoritmo OPWI opera diretamente sobre o sinal de fala, $s(n)$, decompondo-o em uma Componente de Evolução Lenta (CEL) e em uma Componente de

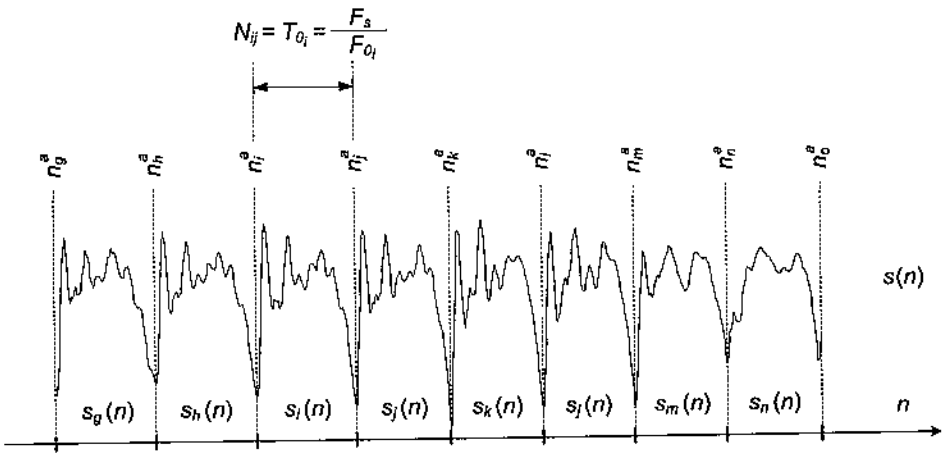


Figura. 7.2: Estimativa dos instantes de análise nos segmentos sonoros (localizados próximos aos IFGs).

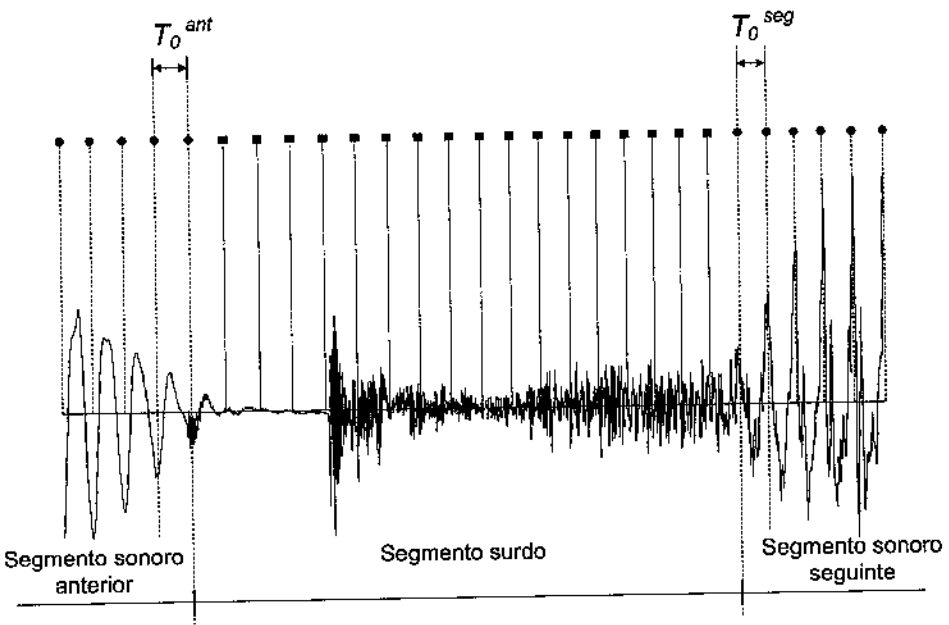


Figura. 7.3: Estimativa dos instantes de análise nos segmentos não-sonoros, por interpolação linear dos períodos fundamentais (T_0) dos segmentos sonoros adjacentes (anterior e seguinte).

Evolução Rápida (CER). Nos trechos sonoros de $s(n)$ a decomposição CEL/CER assume que todos os componentes espectrais (ao longo de toda a banda de freqüências) são representados pela combinação aditiva de uma componente harmônica e uma componente ruidosa. Por outro lado, nos segmentos não-sonoros a decomposição CEL/CER assume a não existência de componentes harmônicos. As principais operações envolvidas nas etapas de análise e síntese da decomposição CEL/CER são:

- Operações da etapa Análise

- As amplitudes dos trechos não-sonoros do sinal de fala $s(n)$ são reduzidas a zero, gerando o sinal $\bar{s}(n)$;
- O sinal $\bar{s}(n)$ é mapeado em uma seqüência de espectros $\widehat{S_W}(k,m)$, sendo m o eixo definido pelos instantes de análise, $m = \{\cdots n_h, n_i, n_j, n_k, n_l \cdots\}$;
- A seqüência de espectros $\widehat{S_W}(k,m)$ é filtrada ao longo do eixo m gerando uma seqüência de espectros suavizados, denominada $\widehat{X_W}(k,m)$.

• Síntese

- A partir da seqüência de espectros suavizados $\widehat{X_W}(k,m)$, retorna-se ao eixo dos tempos obtendo-se uma versão suavizada de $s(n)$, denominada $x(n)$;
- Esta seqüência $x(n)$ é denominada componente CEL de $s(n)$;
- A seqüência $y(n) = s(n) - x(n)$, é denominada componente CER de $s(n)$. É importante observar que $y(n)$ contém toda a energia dos trechos não-sonoros de $s(n)$.

Obviamente vários detalhes sobre o processo de decomposição CEL/CER descrito acima necessitam ser explicitados, como por exemplo: (1) Qual a melhor maneira para estimar os espectros $\widehat{S_W}(k,m)$? (2) Como garantir que todos os espectros $\widehat{S_W}(k,m)$ possuam a mesma dimensão ao longo do eixo k, para poderem ser filtrados ao longo do eixo m? (3) Como estimar as freqüências de corte apropriadas para os filtros de suavização (filtragem ao longo do eixo m) da seqüência de espectros $\widehat{S_W}(k,m)$? (4) Como mapear a seqüência de espectros suavizados $\widehat{X_W}(k,m)$ na seqüência $x(n)$, e garantir que $x(n)$ seja uma versão suavizada de $s(n)$? O restante desta seção é dedicado a responder, em detalhes, cada uma dessas perguntas.

7.4.1 Etapa de Análise da Decomposição CEL/CER

A Figura 7.4 ilustra as principais operações envolvidas na etapa de análise da decomposição CEL/CER. Inicialmente, as amplitudes dos segmentos não-sonoros do sinal de fala $s(n)$ são reduzidas a zero, gerando o sinal $\bar{s}(n)$. Em seguida o sinal $\bar{s}(n)$ é submetido a um processo de janelamento utilizando-se janelas de *Hanning* assimétricas centradas em cada um dos instantes de análise e estendendo-se do instante de análise anterior até o instante de análise seguinte, conforme Figura 7.4(a). Estes segmentos janelados definem uma seqüência de segmentos de forma de onda ao longo do eixo $m = \{\cdots n_h, n_i, n_j, n_k, n_l \cdots\}$ (eixo dos instantes de análise), o qual será denominada $s_w(n,m)$ e que se encontra ilustrada na Figura 7.4(b). Em seguida, as durações dos segmentos janelados de $s_w(n,m)$ devem ser devidamente normalizadas, ao longo do eixo n, para garantir que todos os segmentos de $s_w(n,m)$ possuam a mesma dimensão. Esta normalização dos segmentos de $s_w(n,m)$ é realizada através do acréscimo de zeros no início e no final (ao longo do eixo n) de cada um dos segmentos de $s_w(n,m)$. A quantidade de zeros a ser acrescentada no início e no final dos segmentos de $s_w(n,m)$, deverá garantir não somente que todos os segmentos de $s_w(n,m)$ possuam a mesma duração, mas também que eles sejam alinhados entre si, por seus respectivos instantes centrais de análise. Esta versão normalizada de $s_w(n,m)$ será denominada $\widehat{s_w}(n,m)$, e pode ser visualizada na Figura 7.4(c).

Em seguida, toma-se a transformada Discreta de Fourier, DFT (*Discrete Fourier Transform*) de cada segmento $\widehat{s_w}(n,m)$ (ao longo do eixo n), definindo-se a seqüência de espectros $\widehat{S_W}(k,m)$, cujo módulo encontra-se ilustrado na Figura 7.4(d). Num passo seguinte, esta seqüência de espectros $\widehat{S_W}(k,m)$ é filtrada ao longo do eixo m utilizando-se filtros passa-baixas com freqüência de corte $f_c(k)$ (são utilizadas diferentes freqüências de corte para cada componente k). Este processo de filtragem dá origem à seqüência de espectros suavizados $\widehat{X_W}(k,m)$, que se encontra ilustrada na Figura 7.4(e).

7.4.2 Etapa de Síntese da Decomposição CEL/CER

As principais operações da etapa de síntese da decomposição CEL/CER encontram-se descritas na Figura 7.5. Nesta etapa a seqüência suavizada de espectros $\widehat{X_W}(k,m)$, cuja magnitude se encontra ilustrada na Figura 7.5(a), é convertida na seqüência $\widehat{x_w}(n,m)$, Figura 7.5(b), tomando-se a transformada Discreta de Fourier Inversa, IDFT (*Inverse Discrete Fourier Transform*), de cada espectro da seqüência $\widehat{X_W}(k,m)$. Após isto, de posse das informações sobre os instantes de análise anterior, central e seguinte de cada segmento de $\widehat{x_w}(n,m)$ (que são os mesmos dos segmentos de $s_w(n,m)$), deriva-se a seqüência de segmentos $x_w(n,m)$, Figura 7.5(c), com dimensões (ao longo do eixo n) idênticas às dos segmentos de $s_w(n,m)$. Finalmente, empregando-se um procedimento de *Overlap and Add* sobre as seqüências $x_w(n,m)$, conforme descrito na Figura 7.5(d), sintetiza-se o sinal $x(n)$ ilustrado na Figura 7.5(e). Esse sinal $x(n)$, será a componente CEL do sinal $s(n)$.

Uma vez obtida a componente CEL (sinal $x(n)$), a componente CER (sinal $y(n)$) é obtida subtraindo-se o sinal $x(n)$ do sinal original $s(n)$, isto é: $y(n) = s(n) - x(n)$.

7.4.3 Freqüências de Corte do Filtro de Decomposição CEL/CER

O procedimento adotado para estimar as freqüências de corte $f_c(k)$ do filtros passa-baixas, utilizados para suavizar a seqüência de espectros $\widehat{S_W}(k,m)$, gerando a seqüência de espectros $\widehat{X_W}(k,m)$, segue os seguintes passos:

- Calcula-se o logaritmo da magnitude de cada um dos espectros de $\widehat{S_W}(k,m)$, definindo a seqüência $\log\left(|\widehat{S_W}(k,m)|\right)$.
- Interpreta-se a seqüência bidimensional $\log\left(|\widehat{S_W}(k,m)|\right)$, como N_k seqüências unidimensionais de números reais, ao longo do eixo m, sendo N_k a dimensão de $\log\left(|\widehat{S_W}(k,m)|\right)$ ao longo do eixo k, e toma-se a DFT de cada uma destas seqüências, ao longo do eixo m, definindo-se a seqüência bidimensional $Z(k,l)$ dada pela equação 7.1. As componentes ao longo do eixo l da seqüência $|Z(k,l)|$ (magnitude da seqüencia $Z(k,l)$), estão associadas à taxa de variação ao longo do eixo m da seqüência $\log\left(|\widehat{S_W}(k,m)|\right)$.

$$Z(k,l) = \sum_{m=0}^{N_m-1} \log\left(|\widehat{S_W}(k,m)|\right) \cdot e^{-j\cdot\frac{2\cdot\pi\cdot l\cdot m}{N_m}} \tag{7.1}$$

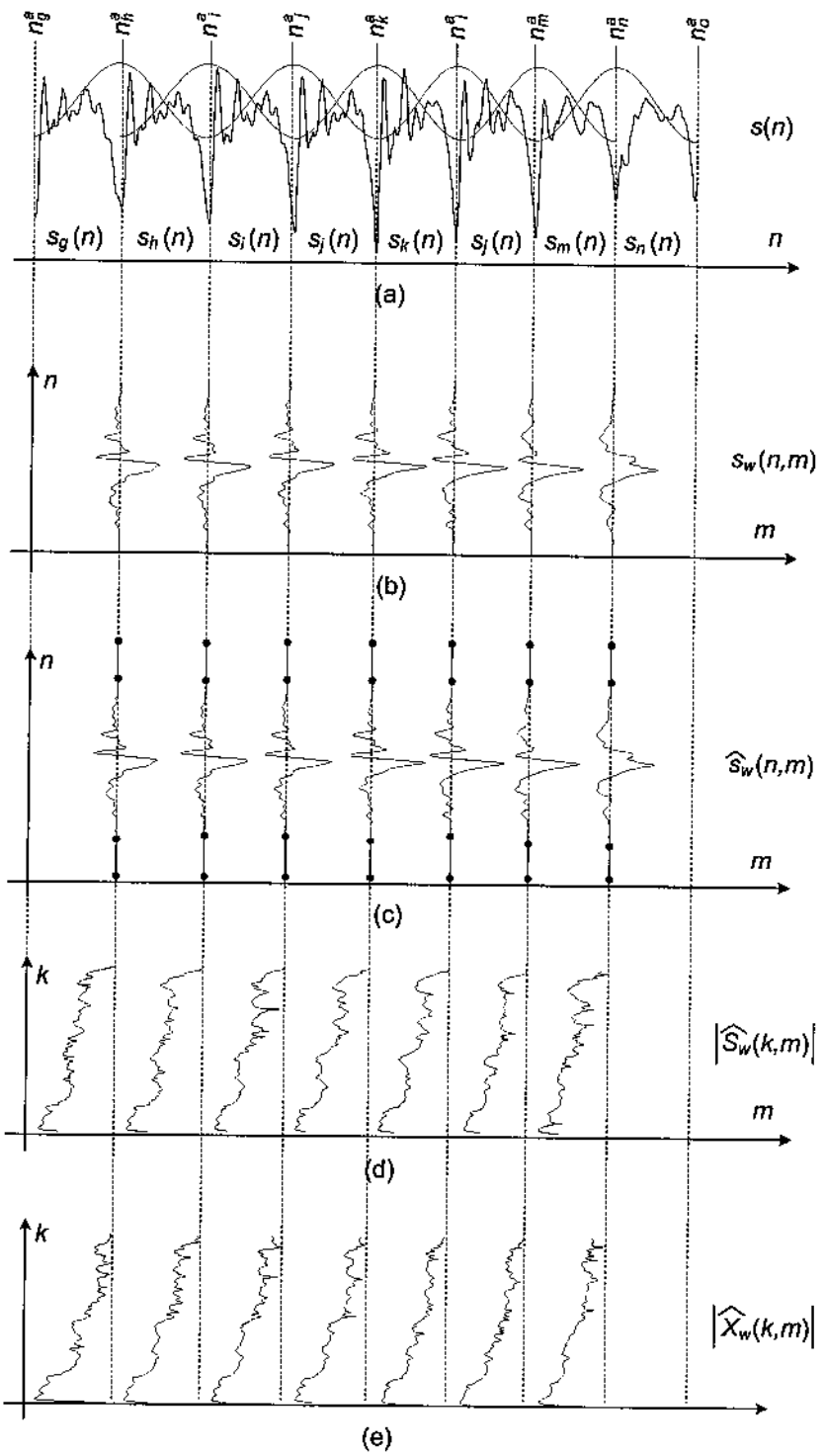


Figura. 7.4: Processo de decomposição CEL X CER: Etapa de análise.

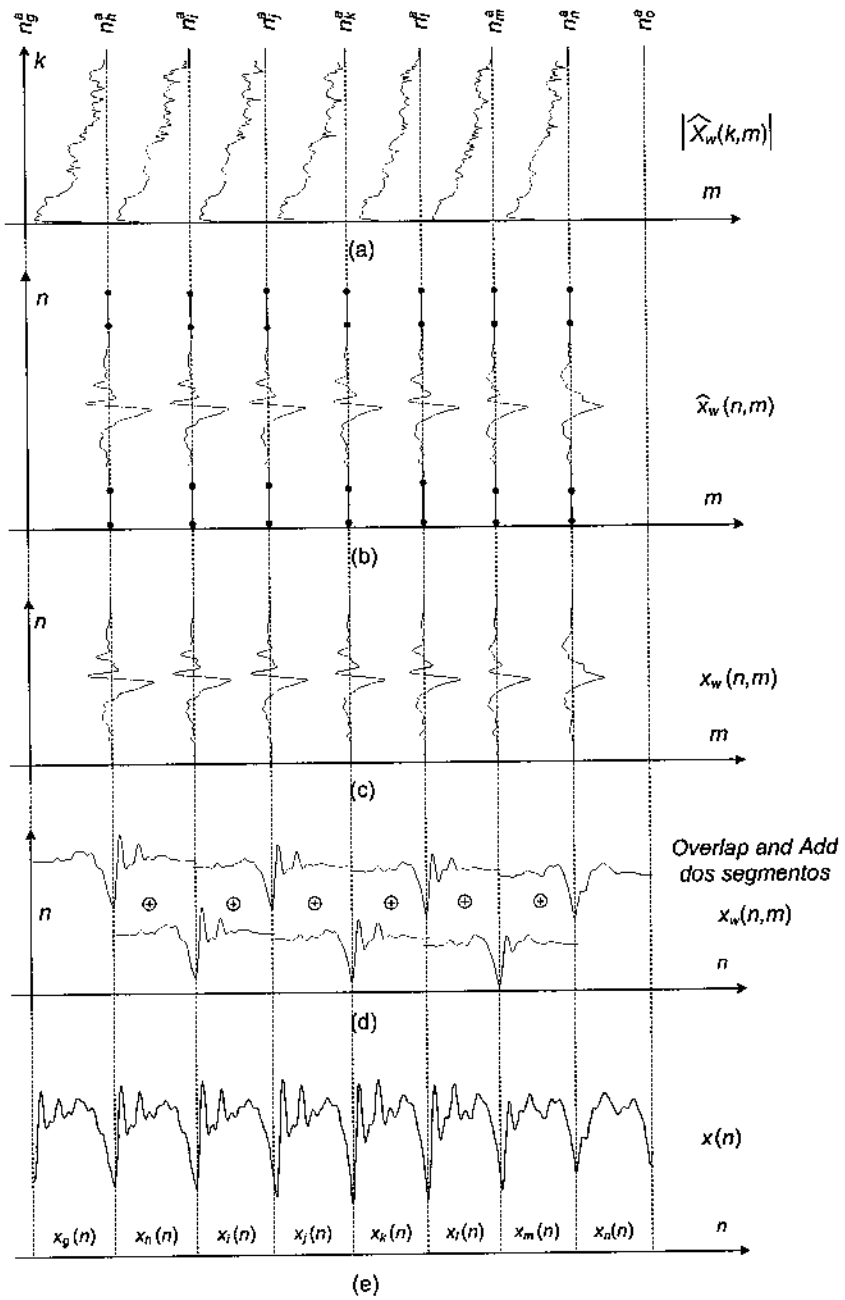


Figura. 7.5: Processo de decomposição CEL X CER: Etapa de síntese.

sendo N_m a dimensão da seqüência $\log\left(|\widehat{S_W}(k,m)|\right)$ ao longo do eixo m e $l = -\frac{\lfloor N_m \rfloor}{2}, -\frac{\lfloor N_m \rfloor}{2} + 1, \cdots, \frac{\lfloor N_m \rfloor}{2} - 2, \frac{\lfloor N_m \rfloor}{2} - 1$. O operador $\lfloor \cdot \rfloor$ representa o maior inteiro contido em.

- Calcula-se a magnitude da seqüência $Z(k,l)$, gerando $|Z(k,l)|$. Em seguida estima-se a seqüência de desvios-padrão $\sigma(k) = \{\sigma_0, \sigma_1, \cdots, \sigma_{N_k-2}, \sigma_{N_k-1}\}$ obtida a partir de cada componente k de

$|Z(k,l)|$ ao longo do eixo l. De posse da seqüência $\sigma(k)$, as freqüências de corte dos filtros passa-baixas $f_c(k)$ são definidas como:

$$f_c(k) = \frac{F_s}{2} \cdot \xi \cdot \left(1 - \frac{\sigma_k}{max\left(\sigma(k)\right)}\right) \tag{7.2}$$

sendo F_s a freqüência de amostragem do sinal em análise e $0 < f_c(k) < \frac{F_s}{2}$ e $0 < \xi \leq 1$. O valor de ξ deve ser obtido experimentalmente em função das características do locutor em análise.

7.5 Estimativa do Nível de Estacionariedade do Sinal de Fala

Conforme será descrito na seção 7.6, a estimativa dos protótipos ótimos poderá ser realizada de duas maneiras distintas. No caso de segmentos considerados pouco estacionários, será utilizada uma estimativa com maior resolução temporal (e, conseqüentemente com menor resolução espectral). Por outro lado, em segmentos com um alto nível de estacionariedade, será utilizada uma estimativa com maior resolução espectral (às custas de uma menor resolução temporal). Portanto, como a estimativa dos protótipos ótimos depende do nível de estacionariedade da componente CEL, essa seção será dedicada à apresentação de três critérios, originalmente propostos em (Kapilow et al., 1999), para estimar esse nível de estacionariedade.

7.5.1 Primeiro Critério, C^1

O primeiro critério baseia-se na taxa de transição do valor RMS (*Root Mean Square*) da amplitude dos segmentos da componente CEL, e é definido como:

$$C_j^1 = \frac{E_j - E_i}{E_j + E_i} \tag{7.3}$$

sendo E_j o valor RMS de amplitude associado ao instante de análise n_j^a, tomado ao longo de $2 \cdot T_{0_j}$ amostras centradas em n_j^a.

$$E_j = \sqrt{\frac{1}{T_{0_j}} \cdot \sum_{n=-T_{0_j}}^{T_{0_j}} w_j^2(n) \cdot x^2\left(n_j^a + n\right)} \tag{7.4}$$

sendo $w_j(n)$ uma janela de *Hanning* centrada no instante de análise n_j^a e estendendo-se de $n_j^a - T_{0_j}$ até $n_j^a + T_{0_j}$.

Analisando-se a equação 7.3, verifica-se que:

$$C_j^1 = \begin{cases} \approx 1 & \text{se } |E_j - E_i| \gg 0 \text{ - pouco estacionário} \\ \approx 0 & \text{se } |E_j - E_i| \approx 0 \text{ - bastante estacionário} \end{cases} \tag{7.5}$$

A definição em 7.3 estima apenas variações no envelope de energia da componente CEL ao longo do tempo, não apresentando qualquer informação sobre a variação de componentes espectrais, tais como freqüência central e largura de banda dos formantes.

7.5.2 Segundo Critério, C^2

Este critério emprega o gradiente de regressão linear dos coeficientes LSF (*Line Spectral Frequency*) (Quatieri, 2002) ao longo do tempo, conforme definido em (Kapilow et al., 1999). No instante de análise n_j^a, o gradiente de regressão linear de segunda ordem será dado por:

$$g_j^p = \frac{-2 \cdot LSF_h(p) - 1 \cdot LSF_i(p) + 0 \cdot LSF_j(p) + 1 \cdot LSF_k(p) + 2 \cdot LSF_l(p)}{(-2)^2 + (-1)^2 + (0)^2 + (1)^2 + (2)^2} \tag{7.6}$$

sendo $LSF_j(p)$, com $p = 1, 2, \cdots, P$, os coeficientes LSF de ordem P, estimados no instante de análise n_j^a.

Em função de g_j^p, define-se a taxa de transição de coeficientes LSF como sendo:

$$m_j = \sum_{p=1}^{P} \left(g_j^p\right)^2 \tag{7.7}$$

O segundo critério, C_j^2, é definido em função de m_j. Além disso, para garantir que $0 < C_j^2 < 1$, este critério é expresso como:

$$C_j^2 = \frac{2}{1 + e^{-\alpha_2 \cdot m_j}} - 1 \tag{7.8}$$

sendo α_2 uma constante estritamente positiva e obtida experimentalmente.

Diferentemente do critério C^1, o critério C^2 contempla, primordialmente, variação de componentes espectrais, tais como freqüência central e largura de banda dos formantes.

7.5.3 Terceiro Critério, C^3

O terceiro critério C^3 combina os critérios C^1 e C^2 na tentativa de obter medidas mais consistentes tanto em termos de características temporais como espectrais. No instante de análise n_j^a, esse critério é definido como:

$$C_j^3 = \frac{2}{1 + e^{-(\alpha_3 \cdot C_j^1 + \beta_3 \cdot C_j^2)}} - 1 \tag{7.9}$$

sendo α_3 e β_3 constantes estritamente positivas e obtidas experimentalmente.

7.6 Protótipo Ótimo e sua Representação Temporal

O protótipo ótimo associado ao instante de análise n_j^a será denominado $X_j^P(k)$, e definido como a DFT de um segmento de forma de onda, de extensão T_{0_j}, extraído de uma seqüência $x_j^P(n)$, ao longo do trecho $n = 0, 1, \cdots, T_{0_j} - 1$. Em outras palavras,

$$X_j^P(k) = \sum_{n=0}^{n=T_{0_j}-1} x_j^P(n) \cdot e^{-j \cdot \frac{2 \cdot \pi \cdot k \cdot n}{T_{0_j}}} \tag{7.10}$$

A seqüência $x_j^P(n)$ será denominada representação temporal do protótipo $X_j^P(k)$ e deverá satisfazer, entre outras, as seguintes propriedades:

P1. $x_j^P(n)$ deverá ser uma seqüência de números reais e de extensão infinita, $-\infty < n < \infty$;

P2. $x_j^P(n)$ deverá ser uma função periódica com período T_{0_j};

P3. $|x_j^P(0) - x_j^P(T_{0_j} - 1)| < \delta$, com $\delta \rightarrow 0$;

Como conseqüência da definição apresentada na equação 7.10, a seqüência periódica $x_j^P(n)$ poderá ser calculada a partir do seu protótipo ótimo correspondente, $X_j^P(k)$, através da equação 7.11 a seguir:

$$x_j^P(n) = \frac{1}{T_{0_j}} \cdot \sum_{k=0}^{T_{0_j}-1} X_j^P(k) \cdot e^{j \cdot \frac{2 \cdot \pi \cdot k \cdot n}{T_{0_j}}}, \quad -\infty < n < \infty \tag{7.11}$$

Apesar de estabelecida a relação entre os protótipos ótimos $X_j^P(k)$ e sua respectiva representação temporal $x_j^P(n)$, bem como definida algumas das propriedades que $x_j^P(n)$ deve satisfazer, ainda não foi apresentado nenhum procedimento para estimar os protótipos ótimos $X_j^P(k)$, e/ou sua representação temporal $x_j^P(n)$ a partir de parâmetros conhecidos, como por exemplo, o próprio sinal $x(n)$ (componente CEL do sinal de fala $s(n)$).

O restante desta subseção será dedicado justamente a apresentar dois métodos para a estimativa dos protótipos ótimos $X_j^P(k)$ a partir de $x(n)$. Os dois métodos a serem apresentados não apenas satisfazem as propriedades P1, P2 e P3 associadas a $x_j^P(n)$, como também garantem ressínteses e modificações prosódicas (TSM e PSM) de $x(n)$ de altíssima qualidade (assumindo-se, obviamente, que

sejam utilizados os procedimentos de síntese e modificações prosódicas definidos, respectivamente, nas seções 7.10 e 7.11 deste Capítulo).

7.7 Estimativa dos Protótipos Ótimos: Método I

Este método estima o protótipo $X_j^P(k)$ em função de um segmento do sinal $x(n)$ centrado em n_j^a e de extensão $2 \cdot T_{0_j}$, e dos protótipos $X_i^P(k)$ e $X_k^P(k)$, os quais são adjacentes a $X_j^P(k)$ e associados aos instantes de análise n_i^a e n_k^a, respectivamente. Essa dependência de $X_j^P(k)$ em relação aos protótipos anterior ($X_i^P(k)$) e seguinte ($X_k^P(k)$), condiciona esse método a uma baixa resolução temporal (segundo esse método, a estimativa de $X_j^P(k)$ depende de mais de quatro quadros de análise). Devido a esta baixa resolução temporal, esse método somente será aplicado aos quadros de análise da componente CEL que apresentarem um elevado grau de estacionariedade. Antes de apresentar o critério de otimização utilizado por esse método para estimar $X_j^P(k)$, é necessário introduzir o conceito de normalização de protótipos.

7.7.1 Normalização de Protótipos

Para que os protótipos $X_i^P(k)$ e $X_k^P(k)$ possam ser utilizados na estimativa do protótipo $X_j^P(k)$, suas extensões devem ser normalizadas (caso necessário) para se tornarem iguais a T_{0_j} (extensão de $X_j^P(k)$). O processo de normalização empregado pelo algoritmo OPWI consiste no acréscimo de zeros (*Zero Padding*) ou na eliminação de amostras (*Truncagem*) nas partes centrais de $X_i^P(k)$ e $X_k^P(k)$ (próximas à freqüência normalizada π). Esses processos de *Zero Padding* e *Truncagem* encontram-se descritos no Apêndice B. As versões normalizadas dos protótipos $X_i^P(k)$ e $X_k^P(k)$, com extensões iguais a T_{0_j}, serão denominadas $\widehat{X_i^P}(k)$ e $\widehat{X_k^P}(k)$, respectivamente.

Extensões periódicas das representações temporais dos protótipos normalizados $\widehat{X_i^P}(k)$ e $\widehat{X_k^P}(k)$) podem ser obtidas através do uso da IDFT (considerando-se suas extensões periódicas), conforme indicado nas equações 7.12 e 7.13:

$$\widehat{x_i^P}(n) = \frac{1}{T_{0_j}} \cdot \sum_{k=0}^{T_{0_j}} \widehat{X_i^P}(k) \cdot e^{j \cdot \frac{2 \cdot \pi \cdot k \cdot n}{T_{0_j}}}, \quad -\infty < n < +\infty \tag{7.12}$$

$$\widehat{x_k^P}(n) = \frac{1}{T_{0_j}} \cdot \sum_{k=0}^{T_{0_j}} \widehat{X_k^P}(k) \cdot e^{j \cdot \frac{2 \cdot \pi \cdot k \cdot n}{T_{0_j}}}, \quad -\infty < n < +\infty \tag{7.13}$$

7.7.2 Critério de Otimização no Domínio Temporal

A partir das representações temporais $\widehat{x_i^P}(n)$ e $\widehat{x_k^P}(n)$ (equações 7.12 e 7.13, respectivamente), pode-se definir a seqüência $\widehat{x_{ik}^P}$ a seguir:

$$\widehat{x_{ik}^P}(n) = \begin{cases} \widehat{x_i^P}(n) & \text{para} \quad -\infty < n < 0 \\ \\ \widehat{x_k^P}(n) & \text{para} \quad 0 \leq n < +\infty \end{cases} \tag{7.14}$$

De posse das seqüências $x(n)$ e $\widehat{x_{ik}^P}(n)$, define-se o protótipo $X_j^P(k)$ como sendo aquele cuja correspondente representação temporal $x_j^P(n)$ minimiza o erro quadrático dado pela equação 7.15,

$$\epsilon = \sum_{n=-T_{0_j}}^{T_{0_j}-1} w^2(n) \cdot \left(x(n_j^a + n) - \left(\gamma(n) \cdot x_j^P(n) + (1-\gamma(n)) \cdot \widehat{x_{ik}^P}(n)\right)\right)^2 \tag{7.15}$$

sendo $w(n)$ uma janela de ponderação que aplica uma maior penalização aos erros próximos ao instante de análise n_j^a. O uso desta janela de ponderação é extremamente importante para garantir que as propriedades P2 e P3 associadas a $x_j^P(n)$ sejam satisfeitas. A janela $w(n)$ utilizada nesta Tese foi a janela de *Hanning* definida pela equação 7.16.

$$w(n) = \begin{cases} \left[\frac{1}{2} \cdot \left(1 + cos\left(\frac{\pi \cdot n}{T_{0_j}}\right)\right)\right] & \text{para} \quad -T_{0_j} \leq n \leq 0 \\ \\ \left[\frac{1}{2} \cdot \left(1 + cos\left(\frac{\pi \cdot n}{T_{0_j}-1}\right)\right)\right] & \text{para} \quad 1 \leq n \leq T_{0_1} - 1 \end{cases} \tag{7.16}$$

As funções $\gamma(n)$ e $1-\gamma(n)$ são utilizadas para interpolar as seqüências x_j^P e $\widehat{x_{ik}^P}$ (ao longo de $2 \cdot T_{0_j}$ amostras). É importante ressaltar que a função $\gamma(n)$ deve, necessariamente, ser definida de forma consistente com as funções de interpolação $\alpha(n)$ e $\beta(n)$ a serem utilizadas na etapa de síntese do algoritmo OPWI, descrita na seção 7.10. Neste trabalho, a função $\gamma(n)$ foi definida como sendo a própria janela de *Hanning*, $w(n)$, descrita na equação 7.16.

Analisando o critério de otimização definido na equação 7.15, pode-se verificar que o erro ϵ a ser minimizado é igual à soma ponderada (pela função $w^2(n)$) e acumulada, do quadrado da diferença entre as seqüências $x(n)$ e a seqüência resultante da interpolação entre $x_j^P(n)$ e $\widehat{x_{ik}^P}(n)$ (pelas funções $\gamma(n)$ e $1-\gamma(n)$), ao longo de um intervalo de $2 \cdot T_{0_j}$ amostras centradas em n_j^a.

7.7.3 Critério de Otimização no Domínio Espectral

Com o objetivo de estabelecer uma relação direta entre o critério proposto na equação 7.15 e o conceito de componentes harmônicas utilizado por Stylianou, em seu modelo HNM (*Harmonic Plus Noise Model*) (Stylianou, 2001), a equação 7.15 será reescrita em termos dos protótipos $X_j^P(k)$, $\widehat{X_i^P}(k)$ e $\widehat{X_k^P}(k)$, e não de suas respectivas representações temporais $x_j^P(n)$, $\widehat{x_i^P}(n)$ e $\widehat{x_k^P}(n)$. Além disso, por

uma questão de conveniência matemática, a notação escalar da equação 7.15 será substituída por uma notação matricial. Para reescrever a equação 7.15 neste novo formato, torna-se necessária a definição de alguns vetores:

$$\mathbf{X_j^P} = \left[X_j^P(0), X_j^P(1), \cdots, X_j^P(T_{0_j}-2), X_j^P(T_{0_j}-1)\right]^T \tag{7.17}$$

$$\widehat{\mathbf{X_i^P}} = \left[\widehat{X_i^P}(0), \widehat{X_i^P}(1), \cdots, \widehat{X_i^P}(T_{0_j}-2), \widehat{X_i^P}(T_{0_j}-1)\right]^T \tag{7.18}$$

$$\widehat{\mathbf{X_k^P}} = \left[\widehat{X_k^P}(0), \widehat{X_k^P}(1), \cdots, \widehat{X_k^P}(T_{0_j}-2), \widehat{X_k^P}(T_{0_j}-1)\right]^T \tag{7.19}$$

$$\mathbf{x_{ij}} = \left[x(n_j^a - T_{0_j}), x(n_j^a - T_{0_j}+1), \cdots, x(n_j^a), \cdots, x(n_j^a + T_{0_j}-1)\right]^T \tag{7.20}$$

Dados os vetores definidos em 7.17, 7.18, 7.19 e 7.20, então a equação 7.15 pode ser reescrita na forma matricial e, explicitamente, em função do protótipo $X_j^P(k)$ e dos protótipos normalizados $\widehat{X_i^P}(k)$ e $\widehat{X_k^P}(k)$, como:

$$\begin{aligned}\epsilon \;=\; & \left(\mathbf{W}\cdot\left(\mathbf{x_{ij}} - \left(\mathbf{\Delta}\cdot\mathbf{B_{ij}}\cdot\mathbf{X_j^P} + \nabla\cdot\mathbf{C_{ij}}\right)\right)\right)^T \cdot \\ & \left(\mathbf{W}\cdot\left(\mathbf{x_{ij}} - \left(\mathbf{\Delta}\cdot\mathbf{B_{ij}}\cdot\mathbf{X_j^P} + \nabla\cdot\mathbf{C_{ij}}\right)\right)\right)\end{aligned} \tag{7.21}$$

sendo,

$$\mathbf{B_{ij}} = \begin{bmatrix}\mathbf{B_i} \\ \mathbf{B_j}\end{bmatrix} \tag{7.22}$$

com $\mathbf{B_i}$ e $\mathbf{B_j}$ dados por:

$$\mathbf{B_i} = \left[\mathbf{b_0^i} \vdots \mathbf{b_1^i} \vdots \mathbf{b_2^i} \vdots \cdots \vdots \mathbf{b_{T_{0_j}-2}^i} \vdots \mathbf{b_{T_{0_j}-1}^i}\right] \tag{7.23}$$

$$\mathbf{B_j} = \left[\mathbf{b_0^j} \vdots \mathbf{b_1^j} \vdots \mathbf{b_2^j} \vdots \cdots \vdots \mathbf{b_{T_{0_j}-2}^j} \vdots \mathbf{b_{T_{0_j}-1}^j}\right] \tag{7.24}$$

e sendo $\mathbf{b_k^i}$ e $\mathbf{b_k^j}$ os vetores de dimensão $2T_{0_j} - por - 1$ dados pelas equações 7.25 e 7.26 a seguir:

$$\mathbf{b_k^i} = \frac{1}{T_{0_j}} \cdot \left[e^{j\frac{2\pi k(-T_{0_j})}{T_{0_j}}} e^{j\frac{2\pi k(-T_{0_j}+1)}{T_{0_j}}} \cdots e^{j\frac{2\pi k(-2)}{T_{0_j}}} e^{j\frac{2\pi k(-1)}{T_{0_j}}} \right]^T \tag{7.25}$$

$$\mathbf{b_k^j} = \frac{1}{T_{0_j}} \cdot \left[e^{j\frac{2\pi k(0)}{T_{0_j}}} e^{j\frac{2\pi k(1)}{T_{0_j}}} \cdots e^{j\frac{2\pi k(T_{0_j}-2)}{T_{0_j}}} e^{j\frac{2\pi k(T_{0_j}-1)}{T_{0_j}}} \right]^T \tag{7.26}$$

O termo $\mathbf{C_{ij}}$ representa a contribuição dos protótipos adjacentes $X_i^P(k)$ e $X_k^P(k)$ (por meio de suas versões normalizadas $\widehat{X_i^P}(k)$ e $\widehat{X_k^P}(k)$) na estimação do protótipo $X_j^P(k)$.

$$\mathbf{C_{ij}} = \begin{bmatrix} \mathbf{B_i} \cdot \widehat{\mathbf{X_i}}^{\mathbf{P}} \\ \mathbf{B_j} \cdot \widehat{\mathbf{X_k}}^{\mathbf{P}} \end{bmatrix} \tag{7.27}$$

A matriz $\mathbf{W}$ é uma matriz diagonal de dimensão $2T_{0_j} - por - 2T_{0_j}$ e cujos elementos de sua diagonal são iguais aos valores da janela de *Hanning* definida em 7.16. A matriz $\mathbf{\Delta}$ também é uma matriz diagonal de dimensão $2T_{0_j} - por - 2T_{0_j}$ e contendo como elementos de sua diagonal os valores da função de interpolação $\gamma(n)$. A matriz ∇ é igual a $\mathbf{I}_{2T_{0_j},2T_{0_j}} - \mathbf{\Delta}$, sendo $\mathbf{I}_{2T_{0_j},2T_{0_j}}$ a matriz identidade de dimensão $2T_{0_j} - por - 2T_{0_j}$.

7.7.4 Solução do Critério de Otimização no Domínio Espectral

Desenvolvendo 7.21, tem-se:

$$\begin{aligned} \epsilon \;=\; & \left(\mathbf{W} \cdot \mathbf{x_{ij}} - \mathbf{W} \cdot \mathbf{\Delta} \cdot \mathbf{B_{ij}} \cdot \mathbf{X_j}^{\mathbf{P}} - \mathbf{W} \cdot \nabla \cdot \mathbf{C_{ij}}\right)^T \cdot \\ & \left(\mathbf{W} \cdot \mathbf{x_{ij}} - \mathbf{W} \cdot \mathbf{\Delta} \cdot \mathbf{B_{ij}} \cdot \mathbf{X_j}^{\mathbf{P}} - \mathbf{W} \cdot \nabla \cdot \mathbf{C_{ij}}\right) \end{aligned} \tag{7.28}$$

$$
\begin{aligned}
\epsilon \;=\; & (\mathbf{W}\cdot\mathbf{x_{ij}})^T\cdot(\mathbf{W}\cdot\mathbf{x_{ij}}) \\
& -(\mathbf{W}\cdot\mathbf{x_{ij}})^T\cdot\mathbf{W}\cdot\boldsymbol{\Delta}\cdot\mathbf{B_{ij}}\cdot\mathbf{X_j}^{\mathbf{P}} \\
& -(\mathbf{W}\cdot\mathbf{x_{ij}})^T\cdot\mathbf{W}\cdot\nabla\cdot\mathbf{C_{ij}} \\
& -\left(\mathbf{W}\cdot\boldsymbol{\Delta}\cdot\mathbf{B_{ij}}\cdot\mathbf{X_j}^{\mathbf{P}}\right)^T\cdot(\mathbf{W}\cdot\mathbf{x_{ij}}) \\
& +\left(\mathbf{W}\cdot\boldsymbol{\Delta}\cdot\mathbf{B_{ij}}\cdot\mathbf{X_j}^{\mathbf{P}}\right)^T\cdot\left(\mathbf{W}\cdot\boldsymbol{\Delta}\cdot\mathbf{B_{ij}}\cdot\mathbf{X_j}^{\mathbf{P}}\right) \\
& +\left(\mathbf{W}\cdot\boldsymbol{\Delta}\cdot\mathbf{B_{ij}}\cdot\mathbf{X_j}^{\mathbf{P}}\right)^T\cdot(\mathbf{W}\cdot\nabla\cdot\mathbf{C_{ij}}) \\
& -(\mathbf{W}\cdot\nabla\cdot\mathbf{C_{ij}})^T\cdot(\mathbf{W}\cdot\mathbf{x_{ij}}) \\
& +(\mathbf{W}\cdot\nabla\cdot\mathbf{C_{ij}})^T\cdot\left(\mathbf{W}\cdot\boldsymbol{\Delta}\cdot\mathbf{B_{ij}}\cdot\mathbf{X_j}^{\mathbf{P}}\right) \\
& +(\mathbf{W}\cdot\nabla\cdot\mathbf{C_{ij}})^T\cdot(\mathbf{W}\cdot\nabla\cdot\mathbf{C_{ij}})
\end{aligned}
\tag{7.29}
$$

Como os termos $(\mathbf{W}\cdot\mathbf{x_{ij}})^T\cdot\mathbf{W}\cdot\boldsymbol{\Delta}\cdot\mathbf{B_{ij}}\cdot\mathbf{X_j}^{\mathbf{P}}$ e $\left(\mathbf{W}\cdot\boldsymbol{\Delta}\cdot\mathbf{B_{ij}}\cdot\mathbf{X_j}^{\mathbf{P}}\right)^T\cdot(\mathbf{W}\cdot\mathbf{x_{ij}})$ são números escalares, então eles podem ser somados. O mesmo acontece com os termos $\left(\mathbf{W}\cdot\boldsymbol{\Delta}\cdot\mathbf{B_{ij}}\cdot\mathbf{X_j}^{\mathbf{P}}\right)^T\cdot(\mathbf{W}\cdot\nabla\cdot\mathbf{C_{ij}})$ e $(\mathbf{W}\cdot\nabla\cdot\mathbf{C_{ij}})^T\cdot\left(\mathbf{W}\cdot\boldsymbol{\Delta}\cdot\mathbf{B_{ij}}\cdot\mathbf{X_j}^{\mathbf{P}}\right)$. Logo a equação 7.29 pode ser reescrita como:

$$
\begin{aligned}
\epsilon \;=\; & (\mathbf{W}\cdot\mathbf{x_{ij}})^T\cdot(\mathbf{W}\cdot\mathbf{x_{ij}}) \\
& -(\mathbf{W}\cdot\mathbf{x_{ij}})^T\cdot\mathbf{W}\cdot\nabla\cdot\mathbf{C_{ij}} \\
& -2\cdot\left(\mathbf{W}\cdot\boldsymbol{\Delta}\cdot\mathbf{B_{ij}}\cdot\mathbf{X_j}^{\mathbf{P}}\right)^T\cdot(\mathbf{W}\cdot\mathbf{x_{ij}}) \\
& +\left(\mathbf{W}\cdot\boldsymbol{\Delta}\cdot\mathbf{B_{ij}}\cdot\mathbf{X_j}^{\mathbf{P}}\right)^T\cdot\left(\mathbf{W}\cdot\boldsymbol{\Delta}\cdot\mathbf{B_{ij}}\cdot\mathbf{X_j}^{\mathbf{P}}\right) \\
& +2\cdot\left(\mathbf{W}\cdot\boldsymbol{\Delta}\cdot\mathbf{B_{ij}}\cdot\mathbf{X_j}^{\mathbf{P}}\right)^T\cdot(\mathbf{W}\cdot\nabla\cdot\mathbf{C_{ij}}) \\
& -(\mathbf{W}\cdot\nabla\cdot\mathbf{C_{ij}})^T\cdot(\mathbf{W}\cdot\mathbf{x_{ij}}) \\
& +(\mathbf{W}\cdot\nabla\cdot\mathbf{C_{ij}})^T\cdot(\mathbf{W}\cdot\nabla\cdot\mathbf{C_{ij}})
\end{aligned}
\tag{7.30}
$$

O ponto de mínimo da superfície de erro quadrática dada pela equação 7.30, pode ser obtido tomando-se a derivada parcial de ϵ em relação a $\mathbf{X_j^P}$ e igualando-se o resultado a zero, $\frac{\partial\epsilon}{\partial\mathbf{X_j^P}}=\mathbf{0}$,

$$
\begin{aligned}
\frac{\partial\epsilon}{\partial\mathbf{X_j^P}} \;=\; & -\mathbf{B_{ij}}^T\cdot\boldsymbol{\Delta}^T\cdot\mathbf{W}^T\cdot\mathbf{W}\cdot\mathbf{x_{ij}} \\
& +\mathbf{B_{ij}}^T\cdot\boldsymbol{\Delta}^T\cdot\mathbf{W}^T\cdot\mathbf{W}\cdot\boldsymbol{\Delta}\cdot\mathbf{B_{ij}}\cdot\mathbf{X_j^P} \\
& +\mathbf{B_{ij}}^T\cdot\boldsymbol{\Delta}^T\cdot\mathbf{W}^T\cdot\mathbf{W}\cdot\nabla\cdot\mathbf{C_{ij}}=\mathbf{0}
\end{aligned}
\tag{7.31}
$$

o que implica em,

$$\mathbf{B_{ij}}^T \cdot \mathbf{\Delta}^T \cdot \mathbf{W}^T \cdot \mathbf{W} \cdot \mathbf{\Delta} \cdot \mathbf{B_{ij}} \cdot \mathbf{X_j^P} = \mathbf{B_{ij}}^T \cdot \mathbf{\Delta}^T \cdot \mathbf{W}^T \cdot \mathbf{W} \cdot \mathbf{x_{ij}} - \mathbf{B_{ij}}^T \cdot \mathbf{\Delta}^T \cdot \mathbf{W}^T \cdot \mathbf{W} \cdot \nabla \cdot \mathbf{C_{ij}} \quad (7.32)$$

Definindo **D**, **E** e **F** segundo as equações 7.33, 7.34 e 7.35,

$$\mathbf{D} = \mathbf{B_{ij}}^T \cdot \mathbf{\Delta}^T \cdot \mathbf{W}^T \cdot \mathbf{W} \cdot \mathbf{\Delta} \cdot \mathbf{B_{ij}} \quad (7.33)$$

$$\mathbf{E} = \mathbf{B_{ij}}^T \cdot \mathbf{\Delta}^T \cdot \mathbf{W}^T \cdot \mathbf{W} \quad (7.34)$$

$$\mathbf{F} = \mathbf{B_{ij}}^T \cdot \mathbf{\Delta}^T \cdot \mathbf{W}^T \cdot \mathbf{W} \cdot \nabla \quad (7.35)$$

então, o protótipo ${X_j}^P(k)$ (representado pelo vetor $\mathbf{X_j^P}$) pode ser escrito como:

$$\mathbf{X_j^P} = \mathbf{D}^{-1} \cdot (\mathbf{E} \cdot \mathbf{x_{ij}} - \mathbf{F} \cdot \mathbf{C_{ij}}) \quad (7.36)$$

7.7.5 Aproximação para a Estimativa do Protótipo Seguinte

Ao analisar as equações 7.36, 7.27, 7.17, 7.18 e 7.19, verifica-se que a estimação ótima de $X_j^P(k)$ assume o conhecimento dos protótipos normalizados $\widehat{X_i^P}(k)$ e $\widehat{X_k^P}(k)$. Assumindo que a estimação dos protótipos se faz a partir do início da componente CEL (da esquerda para a direita), então no momento da estimação do protótipo $X_j^P(k)$, o protótipo $\widehat{X_i^P}(k)$ já se encontrá disponível, entretanto o mesmo não será verdadeiro para o protótipo $\widehat{X_k^P}(k)$. A maneira encontrada para contornar este problema foi utilizar, como estimativa do protótipo ótimo $\widehat{X_k^P}(k)$, o protótipo $\widetilde{X_k^P}(k)$, cuja correspondente representação temporal $\widetilde{x_k^P}(n)$ minimiza o erro quadrático definido na equação 7.37,

$$\epsilon = \sum_{n=-T_{0_j}}^{T_{0_j}-1} w^2(n) \cdot \left(x(n_j^a + n) - \widetilde{x_k^P}(n)\right)^2 \quad (7.37)$$

sendo a relação entre $\widetilde{x_k^P}(n)$ e $\widetilde{X_k^P}(k)$ definida como:

$$\widetilde{x_k^P}(n) = \frac{1}{T_{0_j}} \cdot \sum_{k=0}^{T_{0_j}-1} \widetilde{X_k^P}(k) \cdot e^{j \cdot \frac{2\pi k(n)}{T_{0_j}}}, \quad -\infty < n < \infty \tag{7.38}$$

Solução Para $\widetilde{X_k^P}(k)$

Definindo-se os vetores,

$$\widetilde{\mathbf{X_k^P}} = \left[\widetilde{X_k^P}(0), \widetilde{X_k^P}(1), \cdots, \widetilde{X_k^P}(T_{0_j}-1)\right]^T \tag{7.39}$$

$$\mathbf{x_{jk}} = \left[x(n_k - T_{0_j}), x(n_k - T_{0_j}+1), \cdots, x(n_k), \cdots, x(n_k + T_{0_j}-1)\right]^T \tag{7.40}$$

então, a equação 7.37 pode ser reescrita de forma matricial e explicitamente em função de $\widetilde{X_k^P}(k)$,

$$\epsilon = \left(\mathbf{W} \cdot \left(\mathbf{x_{jk}} - \mathbf{B_{jk}} \cdot \widetilde{\mathbf{X_k^P}}\right)\right)^T \cdot \left(\mathbf{W} \cdot \left(\mathbf{x_{jk}} - \mathbf{B_{jk}} \cdot \widetilde{\mathbf{X_k^P}}\right)\right) \tag{7.41}$$

sendo $\mathbf{B_{jk}}$ definida como:

$$\mathbf{B_{jk}} = \left[\mathbf{b_0} \vdots \mathbf{b_1} \vdots \mathbf{b_2} \vdots \cdots \vdots \mathbf{b_{T_{0_j}-2}} \vdots \mathbf{b_{T_{0_j}-1}}\right] \tag{7.42}$$

e sendo $\mathbf{b_k}$ o vetor de dimensão $2T_{0_j} - por - 1$ definido a seguir.

$$\mathbf{b_k} = \frac{1}{T_{0_j}} \cdot \left[e^{j\frac{2\pi k(-T_{0_j})}{T_{0_j}}} \; e^{j\frac{2\pi k(-T_{0_j}+1)}{T_{0_j}}} \; \cdots \; e^{j\frac{2\pi k(0)}{T_{0_j}}} \; e^{j\frac{2\pi k(T_{0_j}-2)}{T_{0_j}}} \; e^{j\frac{2\pi k(T_{0_j}-1)}{T_{0_j}}}\right]^T \tag{7.43}$$

Como no caso anterior, $\mathbf{W}$ será uma matriz diagonal de dimensão $2T_{0_j} - por - 2T_{0_j}$, cujos elementos de sua diagonal são iguais aos valores da janela de *Hanning*, $w(n)$, definida em 7.16

Desenvolvendo 7.41, obtém-se:

$$\begin{aligned}
\epsilon &= \left(\mathbf{W} \cdot \mathbf{x_{jk}} - \mathbf{W} \cdot \mathbf{B_{jk}} \cdot \widetilde{\mathbf{X_k^P}}\right)^T \cdot \left(\mathbf{W} \cdot \mathbf{x_{jk}} - \mathbf{W} \cdot \mathbf{B_{jk}} \cdot \widetilde{\mathbf{X_k^P}}\right) \\
&= (\mathbf{W} \cdot \mathbf{x_{jk}})^T \cdot (\mathbf{W} \cdot \mathbf{x_{jk}}) - (\mathbf{W} \cdot \mathbf{x_{jk}})^T \cdot \left(\mathbf{W} \cdot \mathbf{B_{jk}} \cdot \widetilde{\mathbf{X_k^P}}\right) - \\
&\quad \left(\mathbf{W} \cdot \mathbf{B_{jk}} \cdot \widetilde{\mathbf{X_k^P}}\right)^T \cdot \mathbf{W} \cdot \mathbf{x_{jk}} + \left(\mathbf{W} \cdot \mathbf{B_{jk}} \cdot \widetilde{\mathbf{X_k^P}}\right)^T \cdot \left(\mathbf{W} \cdot \mathbf{B_{jk}} \cdot \widetilde{\mathbf{X_k^P}}\right)
\end{aligned} \tag{7.44}$$

Como os termos $(\mathbf{W}\cdot\mathbf{x_{jk}})^T\cdot\left(\mathbf{W}\cdot\mathbf{B_{jk}}\cdot\widetilde{\mathbf{X_k^P}}\right)$ e $\left(\mathbf{W}\cdot\mathbf{B_{jk}}\cdot\widetilde{\mathbf{X_k^P}}\right)^T\cdot\mathbf{W}\cdot\mathbf{x_{jk}}$ são escalares, então eles podem ser somados. Logo a equação 7.44 pode ser reescrita como:

$$\begin{aligned}\epsilon &= \mathbf{x_{jk}}^T\cdot\mathbf{W}^T\cdot\mathbf{W}\cdot\mathbf{x_{jk}}+\widetilde{\mathbf{X_k^P}}^T\cdot\mathbf{B_{jk}}^T\cdot\mathbf{W}^T\cdot\mathbf{W}\cdot\mathbf{B_{jk}}\cdot\widetilde{\mathbf{X_k^P}}-\\ &\quad 2\cdot\widetilde{\mathbf{X_k^P}}^T\cdot\mathbf{B_{jk}}^T\cdot\mathbf{W}^T\cdot\mathbf{W}\cdot\mathbf{x_{jk}}\end{aligned}\tag{7.45}$$

Tomando-se a derivada parcial do erro ϵ em relação a $\widetilde{\mathbf{X_k^P}}$ e igualando-se o resultado a zero, $\frac{\partial\epsilon}{\partial\widetilde{\mathbf{X_k^P}}}=\mathbf{0}$, tem-se:

$$\frac{\partial\epsilon}{\partial\widetilde{\mathbf{X_k^P}}} = 2\cdot\mathbf{B_{jk}}^T\cdot\mathbf{W}^T\cdot\mathbf{W}\cdot\mathbf{B_{jk}}\cdot\widetilde{\mathbf{X_k^P}}-2\cdot\mathbf{B_{jk}}^T\cdot\mathbf{W}^T\cdot\mathbf{W}\cdot\mathbf{x_{jk}}=\mathbf{0}\tag{7.46}$$

o que implica em:

$$\mathbf{B_{jk}}^T\cdot\mathbf{W}^T\cdot\mathbf{W}\cdot\mathbf{B_{jk}}\cdot\widetilde{\mathbf{X_k^P}}=\mathbf{B_{jk}}^T\cdot\mathbf{W}^T\cdot\mathbf{W}\cdot\mathbf{x_{jk}}\tag{7.47}$$

Definindo **G** e **H** segundo as equações 7.48 e 7.49,

$$\mathbf{G} = \mathbf{B_{jk}}^T\cdot\mathbf{W}^T\cdot\mathbf{W}\cdot\mathbf{B_{jk}}\tag{7.48}$$

$$\mathbf{H} = \mathbf{B_{jk}}^T\cdot\mathbf{W}^T\cdot\mathbf{W}\tag{7.49}$$

então $\widetilde{\mathbf{X_k^P}}$ pode ser estimado como:

$$\widetilde{\mathbf{X_k^P}}=\mathbf{G}^{-1}\cdot\mathbf{H}\cdot\mathbf{x_{jk}}\tag{7.50}$$

7.8 Estimativa dos Protótipos Ótimos: Método II

O segundo método foi desenvolvido para ser utilizado nos quadros de análise da componente CEL que não apresentam um elevado grau de estacionaridade. A estimativa do protótipo ótimo segundo este método, dependerá de apenas dois quadros de análise e, por isso, ele será considerado de alta resolução temporal. Segundo este critério, o protótipo ótimo $X_j^P(k)$ será aquele cuja correspondente representação temporal $x_j^P(n)$ minimiza o erro quadrático definido na equação 7.51.

$$\epsilon = \sum_{n=-T_{0_j}}^{T_{0_j}-1} w^2(n) \cdot \left(x(n_j^a + n) - x_j^P(n)\right)^2 \tag{7.51}$$

Desenvolvendo-se este critério de forma semelhante ao apresentado na equação 7.37, tem-se:

$$\mathbf{X_j^P} = \mathbf{R}^{-1} \cdot \mathbf{S} \cdot \mathbf{x_{ij}} \tag{7.52}$$

sendo

$$\mathbf{X_j^P} = \left[X_j^P(0), X_j^P(1), \cdots, X_j^P(T_{0_j} - 2), X_j^P(T_{0_j} - 1)\right]^T \tag{7.53}$$

$$\mathbf{R} = \mathbf{B_{ij}}^T \cdot \mathbf{W}^T \cdot \mathbf{W} \cdot \mathbf{B_{ij}} \tag{7.54}$$

$$\mathbf{S} = \mathbf{B_{ij}}^T \cdot \mathbf{W}^T \cdot \mathbf{W} \tag{7.55}$$

O vetor $\mathbf{x_{ij}}$ e a matriz $\mathbf{B_{ij}}$ são os mesmos definidos em 7.20 e 7.22, respectivamente. Novamente a matriz $\mathbf{W}$ será uma matriz diagonal de dimensão $2T_{0_j} - por - 2T_{0_j}$ e cujos elementos de sua diagonal são iguais aos valores da janela de *Hanning* definida em 7.16.

7.9 Análise da Componente CER: Modelo Auto-regressivo

A componente CER, representada pelo sinal $y(n)$, é modelada como um processo auto-regressivo e variante no tempo. Para isto, o sinal $y(n)$ foi submetido a uma modelagem preditiva linear segundo o método da autocorrelação (Quatieri, 2002). Os coeficientes dessa modelagem, coeficientes LPC (*Linear Predictive Coefficients*), são estimados a cada instante de análise utilizando-se uma janela de *Hanning* assimétrica (conforme descrito na Figura 7.6(b)). Por exemplo, para o instante de análise n_i^a, a janela de *Hanning* é centrada em n_i^a e se estende do instante de análise precedente n_h^a até o instante de análise seguinte n_j^a, como mostrado na Figura 7.6(b). A ordem da análise preditiva linear será denominada P (o qual será função da taxa amostragem do sinal) e o vetor de coeficientes LPC associado ao instante de análise n_i^a será denominado $LPC_i^0 = \{a_i^0(1), a_i^0(2), \cdots, a_i^0(P)\}$.

Com o objetivo de suavizar o processo de análise LPC ao longo do tempo, a cada dois instantes de análise consecutivos, são gerados 3 novos vetores de coeficientes LPC através de um processo de

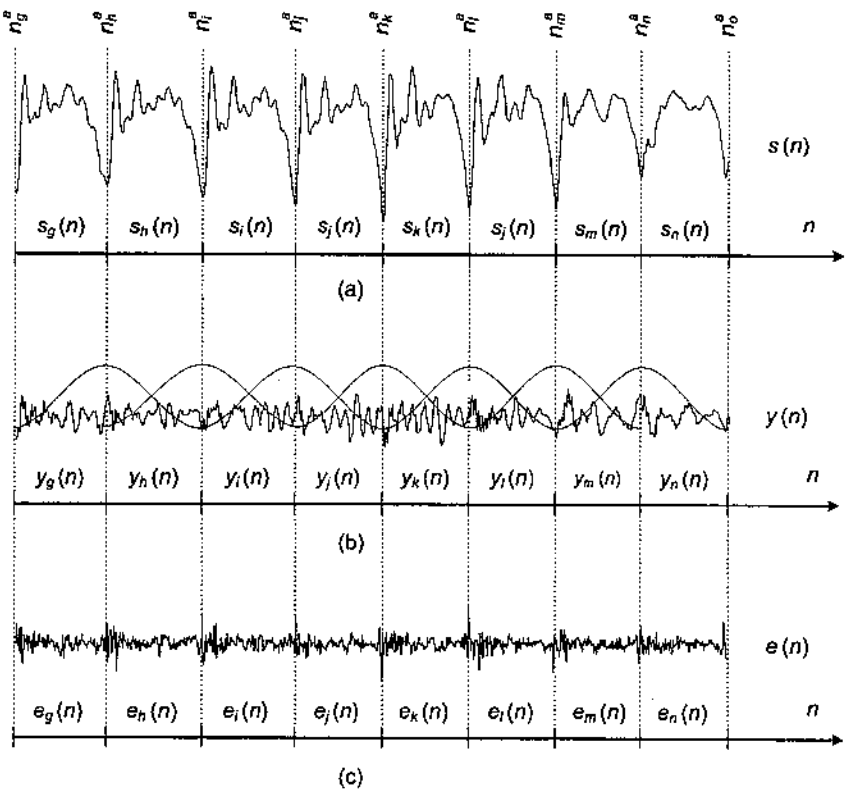


Figura. 7.6: (a) Segmento sonoro do sinal de fala. (b) Componente CER, correspondente ao sinal em (a). (c) Resíduo de predição correspondente a componente CER em (b).

interpolação dos vetores de coeficientes LPC já existentes. Entretanto, como a interpolação direta de coeficientes LPC não garante estabilidade do filtro de síntese, esse processo é realizado com o auxílio dos coeficientes LSF (*Line Spectral Frequency*). Esse procedimento de interpolação dos coeficientes LPC, através do uso de coeficientes LSF, segue os seguintes passos:

- Inicialmente os coeficientes LPC associados aos instantes de análise são transformados em *Line Spectral Frequency Coefficients* - LSF ($\cdots LPC_i^0 \longrightarrow LSF_i^0$, $LPC_j^0 \longrightarrow LSF_j^0 \cdots$).
- A cada dois intantes de análise consecutivos n_i^a e n_j^a, os correspondentes pares de vetores de LSF_i^0 e LSF_j^0 são interpolados gerando 3 novos vetores de coeficientes LSF_i^1, LSF_i^2 e LSF_i^3.
- Estes 3 novos vetores de coeficientes LSF, juntamente com o vetor de coeficientes LSF_i^0, são representados pelo conjunto $\mathcal{LSF}_i = \{LSF_i^0, LSF_i^1, LSF_i^2, LSF_i^3\}$.
- Em seguida o conjunto de vetores de coeficientes $\mathcal{LSF}_i$ é transformado no conjunto de vetores de coeficientes $\mathcal{LPC}_i = \{LPC_i^0, LPC_i^1, LPC_i^2, LPC_i^3\}$.

Portanto, após este processo de interpolação, cada instante de análise irá contar com 4 vetores de coeficientes LPC, de ordem P. Dado estes vetores de coeficientes LPC, a estimativa do resíduo de predição linear $e(n)$, do trecho do sinal $y(n)$ entre os instantes n_i^a e n_j^a, será dado por:

$$e(n) = \begin{cases} y(n) - \sum_{m=1}^{P} a_i^0(m) \cdot y(n-m) & \text{se } n_i^a \leq n < n_i^a + \lfloor \frac{N_{ij}^a}{4} \rfloor; \\ y(n) - \sum_{m=1}^{P} a_i^1(m) \cdot y(n-m) & \text{se } n_i^a + \lfloor \frac{N_{ij}^a}{4} \rfloor \leq n < n_i^a + 2 \cdot \lfloor \frac{N_{ij}^a}{4} \rfloor; \\ y(n) - \sum_{m=1}^{P} a_i^2(m) \cdot y(n-m) & \text{se } n_i^a + 2 \cdot \lfloor \frac{N_{ij}^a}{4} \rfloor \leq n < n_i^a + 3 \cdot \lfloor \frac{N_{ij}^a}{4} \rfloor; \\ y(n) - \sum_{m=1}^{P} a_i^3(m) \cdot y(n-m) & \text{se } n_i^a + 3 \cdot \lfloor \frac{N_{ij}^a}{4} \rfloor \leq n < n_i^a + N_{ij}^a - 1. \end{cases} \tag{7.56}$$

Em resumo, os parâmetros resultantes da modelagem da componente CER são: o resíduo de predição $e(n)$ e os vetores de coeficientes LPC (um conjunto de 4 vetores de coeficientes LPC a cada instante de análise).

7.10 Síntese da Componente CEL: Interpolação Tempo-Freqüência

A Figura 7.7 ilustra os protótipos ótimos extraídos para um segmento sonoro da componente CEL $x(n)$. A Figura 7.7(b) mostra as magnitudes dos espectros dos protótipos ótimos e a Figura 7.7(c) mostra as respectivas representações temporais (ao longo de dois períodos fundamentais) dos protótipos da Figura 7.7(b).

O processo de síntese da componente CEL consiste na interpolação tempo-freqüência dos protótipos estimados por 7.36 ou 7.52 (de acordo com o nível de estacionariedade do quadro de análise). Cada segmento de análise $x_i(n) = \{x(n_i^a), x(n_i^a + 1), \cdots, x(n_i^a + T_{0_i} - 1)\}$ é estimado pela seguinte transformação inversa:

$$x_i(n) \approx \hat{x}_i(n) = \mathbf{T}^{-1}\{\widehat{X}_i(k,n)\} \tag{7.57}$$

sendo $\widehat{X}_i(k,n)$ obtido pela interpolação dos protótipos $X_i^P(k)$ e $X_j^P(k)$ segundo o operador $\mathbf{I}_n$ da equação 7.58.

$$\widehat{X}_i(k,n) = \mathbf{I}_n\left(X_i^P(k), X_j^P(k)\right) \tag{7.58}$$

Entretanto, a interpolação dos protótipos $X_i^P(k)$ e $X_j^P(k)$ requer que ambos tenham a mesma dimensão. Esse problema é resolvido utilizando-se as operações de *Zero Padding* ou *Truncagem*, descritas no Apêndice B, de tal forma que o protótipo normalizado $\widehat{X_j^P}(k)$ passe a ter dimensão igual a $M_{ij} = T_{0i}$. Além disso, o operador de interpolação $\mathbf{I}_n$ deve, necessariamente, ser definido a partir de funções de interpolação $\alpha(n)$ e $\beta(n)$ que sejam consistentes com as funções de interpolação $\gamma(n)$ e $1 - \gamma(n)$ (equação 7.15), utilizadas no primeiro método para estimação dos protótipos ótimos. Por

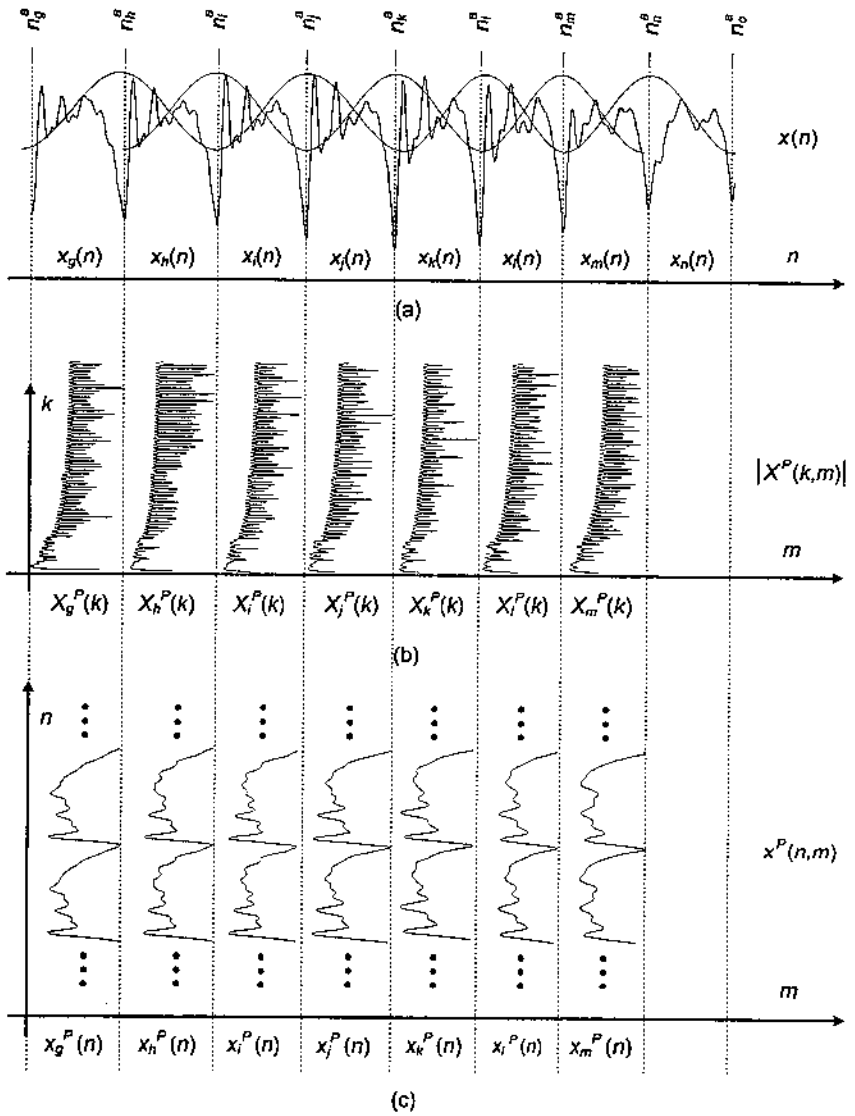


Figura. 7.7: (a) Segmento da componente CEL. (b) Protótipos ótimos. (c) Extensão periódica (dois períodos fundamentais) das respectivas representações temporais dos protótipos apresentados em (b).

conseguinte, o processo de interpolação dos protótipos da equação 7.58 foi definido como:

$$\widehat{X}_i(k,n) = \alpha(n) \cdot X_i^P(k) + \beta(n) \cdot \widehat{X_j^P}(k) \tag{7.59}$$

sendo as funções de interpolação $\alpha(n)$ e $\beta(n)$ definidas como:

$$\alpha(n) = \left[\frac{1}{2} \cdot \left(1 + cos\left(\frac{\pi \cdot n}{T_{0_i}}\right)\right)\right], \quad n = 0, 1, \cdots, T_{0_i} - 1 \tag{7.60}$$

$$\beta(n) = 1 - \alpha(n), \quad n = 0, 1, \cdots, T_{0_i} - 1 \tag{7.61}$$

A transformação da seqüência de espectros $\widehat{X_i}(k,n)$ na seqüência temporal $\widehat{x_i}(n)$, com $0 \leq n < N_{ij}^s$, é realizada pela transformada inversa $\mathbf{T}^{-1}$, definida na equação 7.62 a seguir:

$$\widehat{x_i}(n) = \begin{cases} \frac{\widehat{X_i}(0,n)}{M_{ij}} + \frac{\widehat{X_i}\left(\frac{M_{ij}}{2},n\right)}{M_{ij}} \cdot (-1)^n + \frac{1}{M_{ij}} \cdot \sum\limits_{k=1}^{\left(\frac{M_{ij}}{2}-1\right)} C_i(k,n) & \text{se } M_{ij} \text{ é par} \\ \frac{\widehat{X}(0,n)}{M_{ij}} + \frac{1}{M_{ij}} \cdot \sum\limits_{k=1}^{\left(\frac{M_{ij}-1}{2}\right)} C_i(k,n) & \text{se } M_{ij} \text{ é ímpar} \end{cases} \tag{7.62}$$

sendo,

$$C_i(k,n) = 2 \cdot \Re\left(\widehat{X_i}(k,n) \cdot e^{j \cdot \varphi_i(n) \cdot k}\right) \tag{7.63}$$

A função $\varphi_i(n)$ é definida como,

$$\varphi_i(n) = \left[\left(\frac{2 \cdot \pi}{T_{0i}}\right) - \Phi_i\right] \cdot n \tag{7.64}$$

e a função de fase Φ_i é definida como:

$$\Phi_i = \Phi_h + \frac{2 \cdot \pi \cdot N_{ij}^s}{T_{0i}} \tag{7.65}$$

sendo Φ_h a função de fase calculada no $h-$*ésimo* instante de análise (n_h corresponde ao instante de análise imediatamente anterior a n_i).

Da equação 7.65, pode ser observado que se o sinal for ressintetizado sem modificações prosódicas, então a fase Φ_i será desnecessária, uma vez que $N_{ij}^s = N_{ij}^a = T_{0_i}$.

7.11 Modificações Prosódicas na Componente CEL

7.11.1 Modificações na Taxa de Articulação: TSM

Modificações na taxa de articulação - TSM (*Time Scale Modification*) podem ser obtidas, simplesmente, movendo-se os protótipos de seus instantes de análise, para os novos instantes de síntese. Nenhuma modificação necessita ser feita nos componentes de freqüência dos protótipos. Por exemplo, se o sinal for submetido a uma TSM igual a ν então os protótipos $X_i^P(k)$ e $X_j^P(k)$, que se localizavam originalmente nos instantes n_i^a e n_j^a, devem ser movidos para os novos instantes de síntese n_i^s e n_j^s, cuja distância entre eles será igual a $N_{ij}^s = n_j^s - n_i^s = \langle \nu \left(n_j^a - n_i^a \right) \rangle = \langle \nu \cdot N_{ij}^a \rangle$ (sendo que $\langle \circ \rangle$ indica o inteiro mais próximo) . Se for utilizado um fator $\nu > 1$ isso resultará em uma redução na taxa de articulação do sinal e a interpolação tempo-freqüência de $X_i^P(k)$ e $X_j^P(k)$ (segundo a equação 7.62), gerará mais de um período fundamental de $\widehat{x}_i(n)$. Por outro lado, se for utilizado um fator $\nu < 1$, isto resultará em um aumento na taxa de articulação do sinal e a interpolação tempo-freqüência entre $X_i^P(k)$ e $X_j^P(k)$ (segundo a equação 7.62) não se estenderá ao longo de um período fundamental completo de $\widehat{x}_i(n)$. Dada a possibilidade de o sinal sintetizado $\widehat{x}_i(n)$ apresentar um número fracionário de períodos fundamentais, ao longo do intervalo N_{ij}^s, exige-se que a equação 7.65 seja alterada, conforme descrito na equação 7.66, para garantir a continuidade da função de fase Φ_i,

$$\begin{aligned} \Phi_i &= \Phi_h + \frac{2 \cdot \pi \cdot \nu \cdot N_{ij}^a}{T_{0i}} \\ &= \Phi_h + \frac{2 \cdot \pi \cdot N_{ij}^s}{T_{0i}} \end{aligned} \tag{7.66}$$

7.11.2 Modificações do Contorno da Freqüência Fundamental: PSM

Modificações no contorno da freqüência fundamental - PSM (*Pitch Scale Modifications*) podem ser obtidas simplesmente reamostrando-se os envelopes dos protótipos em novos valores de freqüência. Nenhuma alteração na posição dos instantes de análise será necessária e N_{ij}^s se manterá igual a N_{ij}^a.

Sendo a freqüência fundamental associada ao protótipo $X_i^P(k)$ igual a $F_{0_i} = \frac{1}{T_{0_i}}$ então, suas componentes de freqüência se encontraram localizadas em múltiplos de F_{0_i}. Entretanto, após uma operação de PSM por um fator de $\frac{1}{\rho}$, a nova freqüência fundamental de $X_i^P(k)$ será igual a $\frac{F_{0_i}}{\rho} = \frac{1}{\rho \cdot T_{0_i}}$, e, portanto, esse protótipo deverá ser re-amostrado em múltiplos dessa nova freqüência fundamental. Se for aplicado uma PSM por um fator $\frac{1}{\rho} > 1$, então o sinal terá a sua freqüência fundamental aumentada e a utilização da equação 7.62 ao longo do intervalo N_{ij}^a resultará em mais de um período fundamental do sinal $\widehat{x}_i(n)$. Por outro lado, se for utilizado um fator $\frac{1}{\rho} < 1$, então o sinal terá a sua freqüência fundamental reduzida, e a aplicação da equação 7.62 ao longo do intervalo N_{ij}^a resultará em menos de um período fundamental do sinal $\widehat{x}_i(n)$. Dada a possibilidade do sinal sintetizado $\widehat{x}_i(n)$ apresentar um número fracionário de períodos fundamentais, ao longo do intervalo N_{ij}^s, exige-se que a equação 7.65 seja modificada, conforme descrito na equação 7.67, para garantir a continuidade da função de fase Φ_i,

$$\Phi_i = \Phi_h + \frac{2 \cdot \pi N_{ij}^a}{\rho \cdot T_{0i}} \tag{7.67}$$

A re-amostragem dos componentes de freqüência dos protótipos, para posições múltiplas (inteiras ou fracionárias) de suas novas freqüências fundamentais $\frac{F_{0_i}}{\rho}$, foi realizada por meio de uma interpolação/dizimação linear das partes reais e imaginárias destes protótipos (Stylianou, 2001). Duas outras possibilidades para re-amostragem dos componentes de freqüência dos protótipos em múltiplos da freqüência fundamental são: (1)interpolação/dizimação da magnitude e da fase dos protótipos (após operação de *unwrapping* da fase), conforme proposto em (Quatieri, 2002); (2) Uso de Cepstrum Discreto para representar o envelope de amplitude dos protótipos conforme proposto em (Stylianou, 1996). De posse desse Cepstrum Discreto e da função de fase dos protótipos (após *unwrapping*), novos componentes de freqüência dos protótipos podem ser re-amostrados.

7.11.3 Modificações Conjuntas de PSM e TSM

Modificações conjuntas na taxa de articulação e no contorno de freqüência fundamental podem ser obtidas reamostrando-se, inicialmente, as componentes de freqüência dos protótipos em valores múltiplos da nova freqüência fundamental (modificações de PSM) e, em seguida, redefinindo-se as posições dos protótipos de acordo com os novos intantes de síntese (modificações de TSM). Neste caso, a equação 7.65 deve ser alterada para:

$$\begin{aligned} \Phi_i &= \Phi_h + \frac{2 \cdot \pi \cdot \nu \cdot N_{ij}^a}{\rho \cdot T_{0i}} \\ &= \Phi_h + \frac{2 \cdot \pi \cdot N_{ij}^s}{\rho \cdot T_{0i}} \end{aligned} \tag{7.68}$$

7.12 Síntese da Componente CER

O processo de síntese da componente CER é realizado utilizando-se uma versão modificada do algoritmo LP-PSOLA (*Linear Prediction Pitch Synchronous Overlap and Add*) (Dutoit, 1997). Neste processo, o resíduo de predição, $e(n)$, é sintetizado utilizando-se um procedimento de *Overlap and Add* síncrono com os instantes de análise. Após isso, a componente CER é gerada filtrando-se o resíduo sintetizado $\hat{e}(n)$ através dos filtros de predição linear.

Para síntese sem modificações prosódicas de TSM ou PSM, o algoritmo LP-PSOLA produz um sinal sintetizado $\hat{e}(n)$ com altíssima qualidade segmental e com elevados valores de SNR (*Signal to Noise Ratio*). Entretanto, para síntese sujeita a modificações prosódicas de TSM e/ou PSM, são necessários alguns cuidados com o algoritmo LP-PSOLA com o objetivo de lidar com algumas das particularidades apresentadas pelo resíduo de predição $e(n)$ e, também, garantir um perfeito sincronismo temporal e

espectral, entre as componentes CER e CEL sintetizadas. Os cuidados tomados com o algoritmo LP-PSOLA na síntese do resíduo de predição da componente CER foram:

- Apesar de o sinal $e(n)$ não apresentar qualquer característica sonora ao longo de toda a sua extensão, os quadros de análise do resíduo $e(n)$, localizados em regiões classificadas como sonoras (pelo classificador sonoro *versus* não-sonoro do algoritmo OPWI), foram submetidos a operações de PSM (conforme será apresentado na Figura 7.8). Os fatores de PSM aplicados aos quadros classificados como sonoros foram os mesmos aplicados aos correspondentes quadros de análise da componente CEL. Esta operação de PSM (nos quadros classificados como sonoros) é de extrema importância para garantir o sincronismo temporal e espectral entre as componentes CER e CEL, conforme será discutido no Capítulo 8.
- Os quadros de análise do sinal $e(n)$ localizados em regiões classificadas como não-sonoras não foram submetidos a modificações prosódicas de PSM.
- As modificações prosódicas de TSM foram aplicadas a todos os quadros de análise de $e(n)$ (classificados como sonoros ou não). Contudo, as operações de TSM utilizadas para os quadros sonoros foram distintas das operações de TSM para os quadros não-sonoros:
  - Nos quadros classificados como sonoros, a mera repetição dos *segmentos de pitch*, realizada pelo algoritmo LP-PSOLA original, foi substituída por uma interpolação linear entre *segmentos de pitch* adjacentes. Esse processo de interpolação dos *segmentos de pitch* possui dois objetivos principais: (1) minimizar o excesso de periodicidade causado pela mera repetição de *segmentos de pitch*; (2) garantir o sincronismo entre as componentes CER e CEL ao longo de segmentos mistos (compostos por uma componente harmônica e uma componente ruidosa), como por exemplo segmentos fricativos sonoros.
  - Nos quadros classificados como não-sonoros, a cada dois *segmentos de pitch* a serem repetidos, um deles deve ser espelhado no tempo (rotacionado em torno do seu eixo central), conforme proposto por (Moulines and Charpentier, 1990). O objetivo desta operação é minimizar o excesso de periodicidade causado pela mera repetição de *segmentos de pitch*. O processo de interpolação, aplicado aos segmentos sonoros não é aconselhável nesse caso, porque pode gerar uma suavização indesejada dos segmentos que forem preponderantemente fricativos.

A Figura 7.8 ilustra a aplicação do algoritmo LP-PSOLA na síntese do sinal $\hat{e}(n)$, para o caso de uma modificação prosódica de TSM= 3.3 e PSM= $\frac{1}{1.5}$, ao longo de um segmento do sinal $s(n)$ considerado sonoro. Antes de discutir as operações envolvidas na Figura 7.8, torna-se necessária a definição dos termos *segmento de pitch à direita* e *segmento de pitch à esquerda*.

O termo *segmento de pitch à direita* associado ao instante de análise n_j^a é definido como:

$$we_j^d(n) = w_j(n - T_{0_j}) \cdot e(n_j^a + n - T_{0_j}) \tag{7.69}$$

para $n = 0, 1, \cdots, (2 \cdot T_{0_j}) - 1$. O termo $w_j(n)$ é definido pela equação 7.70 a seguir:

$$w_j(n) = \begin{cases} \frac{1}{2} \cdot \left[1 + cos\left(\frac{\pi \cdot n}{T_{0_j}}\right)\right] & \text{se } -T_{0_j} \leq n \leq 0 \\ \frac{1}{2} \cdot \left[1 + cos\left(\frac{\pi \cdot n}{T_{0_j}-1}\right)\right] & \text{se } 1 \leq n \leq T_{0_j} - 1 \\ 0 & \text{caso contrário.} \end{cases} \tag{7.70}$$

Observe que a extensão de $we_j^d(n)$ foi definida em função de T_{0_j}, período de *pitch* à direita do instante de análise central n_j^a.

O termo *segmento de pitch à esquerda* associado ao instante de análise n_k^a é definido como:

$$we_k^e(n) = w_j(n) \cdot e(n_k^a + n - T_{0_j}) \tag{7.71}$$

para $n = 0, 1, \cdots, (2 \cdot T_{0_j}) - 1$.

Observe que, diferentemente da equação 7.69 a extensão de $we_k^e(n)$ é definida em função do período de *pitch* à esquerda do instante de análise central, n_j^a.

Definidos os termos $we_j^d(n)$ e $we_k^e(n)$, pode-se verificar da Figura 7.8 que a as operações envolvidas na síntese do quadro de análise $e_j(n)$ para uma TSM por um fator $\nu = 3.3$ e uma de PSM por um fator $\frac{1}{\rho} = \frac{1}{1.5}$, são:

1. Alteração no espaçamento entre os segmentos de *pitch* (operação de PSM). O novo espaçamento entre os *segmentos de pitch* será dado por:

$$T_{0_j}^s = \langle \rho \cdot T_{0_j}^a \rangle. \tag{7.72}$$

2. Repetição dos *segmentos de pitch* (somente para TSM por um fator $\nu > 1$). O número de repetições dos *segmentos de pitch* $we_j^e(n)$ e $we_k^d(n)$ é dado por:

$$M = \langle \frac{\nu}{\rho} \rangle + 1 \tag{7.73}$$

3. Interpolação entre $we_j^e(n)$ e $we_k^d(n)$. Os *segmentos de pitch* a serem repetidos (caso necessário) são interpolados segundo as funções lineares $\tau(m)$ e $1 - \tau(m)$. Sendo $\tau(m)$ definida como:

$$\tau(m) = 1 - \frac{m-1}{M}, \quad m = 1, 2, \cdots, M \tag{7.74}$$

4. Recorte do segmento sintetizado útil. Após a interpolação entre os *segmentos de pitch*, deve ser realizado o recorte do segmento estendendo-se do instante de análise n_j^a até $n_j^a + N_{jk}^s$, sendo $N_{jk}^s = (\nu \cdot N_{jk}^a)$.

5. Atraso para o próximo quadro de análise. Como o segmento sintetizado, em geral, não corresponde a um número inteiro de $T_{0_j}^s$ (equação 7.72), então um atraso em amostras deve ser calculado para ser aplicado ao próximo quadro de análise. O atraso em amostras associado ao instante de síntese n_j^s será denominado D_j e calculado segundo a equação 7.75 a seguir:

$$D_j = (T_{0_j}^s \cdot \frac{mod\,(\Phi_j,\, 2 \cdot \pi)}{2 \cdot \pi}) \tag{7.75}$$

sendo *mod* o operador resto da divisão inteira e Φ_j a função de fase calculada segundo a equação 7.65, durante a síntese da componente CEL. O uso desta função de fase proveniente da síntese da componente CEL é de fundamental importância para garantir o perfeito sincronismo temporal entre as componentes CEL e CER.

Uma vez sintetizado o resíduo de predição $\hat{e}(n)$, a síntese da componente CER, sinal $\hat{y}(n)$, entre os instantes de síntese n_i^s e n_j^s, será obtida utilizando-se a equação 7.76 a seguir:

$$\hat{y}(n) = \begin{cases} \sum_{m=1}^{P} a_j^0(m) \cdot \hat{y}\,(n-m) + \hat{e}(n) & \text{se } n_j^s \le n < n_j^s + \lfloor \frac{N_{jk}^s}{4} \rfloor; \\ \sum_{m=1}^{P} a_j^1(m) \cdot \hat{y}\,(n-m) + \hat{e}(n) & \text{se } n_j^s + \lfloor \frac{N_{jk}^s}{4} \rfloor \le n < n_j^s + 2 \cdot \lfloor \frac{N_{jk}^s}{4} \rfloor; \\ \sum_{m=1}^{P} a_j^2(m) \cdot \hat{y}\,(n-m) + \hat{e}(n) & \text{se } n_j^s + 2 \cdot \lfloor \frac{N_{jk}^s}{4} \rfloor \le n < n_j^s + 3 \cdot \lfloor \frac{N_{jk}^s}{4} \rfloor; \\ \sum_{m=1}^{P} a_j^3(m) \cdot \hat{y}\,(n-m) + \hat{e}(n) & \text{se } n_j^s + 3 \cdot \lfloor \frac{N_{jk}^s}{4} \rfloor \le n < n_j^s + N_{jk}^s - 1. \end{cases} \tag{7.76}$$

sendo $a_j^0(m)$, $a_j^1(m)$, $a_j^2(m)$ e $a_j^3(m)$, para $m = 1, 2, \cdots, P$, os coeficientes de predição linear pertencentes a $\mathcal{LPC}_j$ (os quais estão associados ao instante de síntese n_j^s).

7.13 Suavização Espectral

O algoritmo OPWI permite que várias técnicas de suavização espectral possam ser aplicadas na fronteira entre unidades de síntese. A seguir será apresentada uma dessas técnicas:

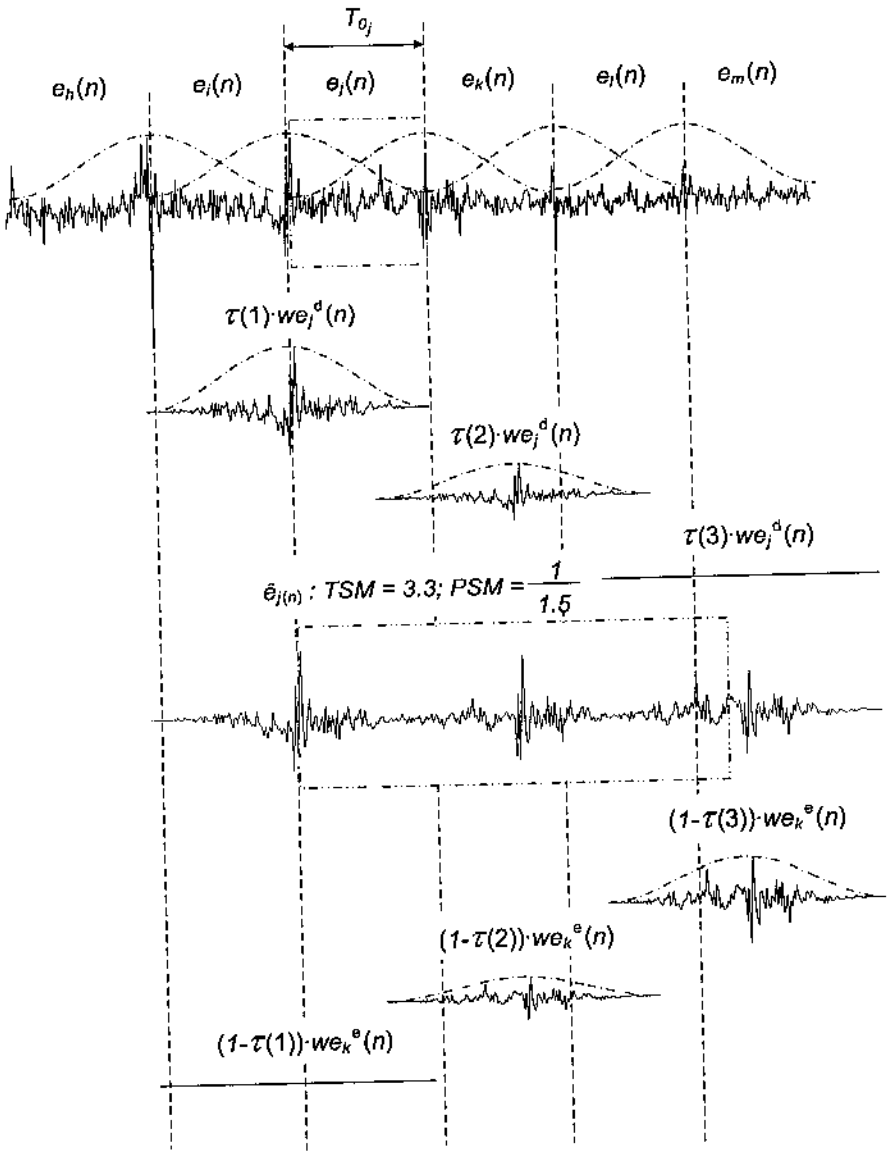


Figura. 7.8: (a) Processo de *Overlap and Add* utilizado para síntese do resíduo de predição da componente CER.

7.13.1 Suavização da Componente CEL

O método proposto para suavização da componente CEL será apresentado por meio de um exemplo. Suponha que se deseje suavizar os três últimos protótipos da unidade de síntese anterior, US_{ant}, com os três primeiros protótipos da unidade de síntese seguinte US_{seg}. Denominando os três últimos protótipos de US_{ant} como $\left[X_h^{P_{ant}}(k), X_i^{P_{ant}}(k), X_j^{P_{ant}}(k)\right]$ e os três primeiros protótipos de US_{seg} como $\left[X_j^{P_{seg}}(k), X_k^{P_{seg}}(k), X_l^{P_{seg}}(k)\right]$, então uma concatenação sem suavização espectral, entre estas

duas unidades de síntese definiria a seguinte seqüência de protótipos:

$$\left[X_h^{P_{ant}}(k)\ ,\ X_i^{P_{ant}}(k)\ ,\ X_j^{P_{ant}}(k)\ \ ,\ \ X_k^{P_{seg}}(k)\ ,\ X_l^{P_{seg}}(k)\ ,\ X_m^{P_{seg}}(k)\right] \tag{7.77}$$

A suavização espectral proposta consiste em definir os três últimos protótipos de US_{ant} como:

$$\begin{aligned} US_{ant} \ = \ & \eta_2 \cdot X_h^{P_{ant}}(k) + \kappa_2 \cdot \widehat{X_k^{P_{seg}}}(k)\ ,\ \eta_1 \cdot X_i^{P_{ant}}(k) + \kappa_1 \cdot \widehat{X_k^{P_{seg}}}(k)\ , \\ & \eta_0 \cdot X_j^{P_{ant}}(k) + \kappa_0 \cdot \widehat{X_k^{P_{seg}}}(k) \end{aligned} \tag{7.78}$$

e os três primeiros protótipos da unidade de síntese seguinte como:

$$\begin{aligned} US_{seg} \ = \ & \eta_1 \cdot X_k^{P_{seg}}(k) + \kappa_1 \cdot \widehat{X_j^{P_{ant}}}(k)\ ,\ \eta_2 \cdot X_l^{P_{seg}}(k) + \kappa_2 \cdot \widehat{X_j^{P_{ant}}}(k)\ , \\ & X_m^{P_{seg}}(k) \end{aligned} \tag{7.79}$$

sendo η_i e κ_i definidos pelas equações 7.80 e 7.81, respectivamente.

$$\eta_i = 0.5 + \frac{i}{2 \cdot N_{smooth}} \tag{7.80}$$

$$\kappa_i = 1 - \eta_i \tag{7.81}$$

N_{smooth} representa o número de protótipos a serem interpolados, nesse exemplo, $N_{smooth} = 3$. Os protótipos $\widehat{X_k^{P_{seg}}}(k)$ e $\widehat{X_j^{P_{ant}}}(k)$ são versões normalizadas de $X_k^{P_{seg}}(k)$ e $X_j^{P_{ant}}(k)$, respectivamente, obtidas segundo as operações de *Zero Padding* e *Truncagem* em freqüência descritas no Apêndice B.

7.13.2 Suavização da Componente CER

A suavização espectral para a componente CER pode ser realizada de maneira semelhante à suavização da componente CEL da subseção 7.13.1. Todavia, no caso da componente CER, a suavização deve ser aplicada tanto aos *segmentos de pitch* (no domínio do resíduo LPC da componente CER), quanto aos envelopes espectrais, representados pelos coeficientes LSFs.

7.14 Aspectos de Implementação

7.14.1 Etapa de Análise

A maior parcela do custo computacional associado ao algoritmo OPWI, durante a etapa de análise, deve-se à estimativa dos protótipos ótimos. Este custo computacional é conseqüência, principalmente, dos sistemas de equações definidos em 7.36, 7.50 e 7.52. Entretanto, analisando-se cuidadosamente as matrizes envolvidas nessas equações, verifica-se que todas elas dependem apenas do valor de T_0 associado ao instante de análise a ser processado, não dependendo de qualquer outra característica do sinal de fala. Portanto, como todo o processamento realizado neste Capítulo envolve apenas valores inteiros de T_0, e sabendo que a faixa de valores de T_0 de um locutor está limitada a algumas centenas de valores, então todas as matrizes envolvidas em 7.36, 7.50 e 7.52 (incluindo as matrizes inversas D^{-1}, G^{-1} e R^{-1}) podem ser pré-calculadas para a faixa prevista de valores inteiros de T_0. Uma vez estimadas essas matrizes, elas poderão ser utilizadas para estimar os protótipos ótimos do locutor em análise.

O armazenamento de todas essas matrizes requer obviamente uma elevada quantidade de memória. Logo, um bom compromisso entre custo computacional e espaço para armazenamento das matrizes consiste em se levantar um histograma dos valores inteiros de T_0 a serem processados e armazenar apenas as matrizes associadas aos $\mathcal{N}$ valores de T_0 mais freqüentes. As matrizes restantes (aquelas não associadas aos $\mathcal{N}$ valores de T_0 mais freqüentes) seriam estimadas *on-fly*.

7.14.2 Etapa de Síntese

Como já mencionado anteriormente, em sistema CTF-SCAUS, a etapa de síntese do módulo de *Back-End* deve operar *on-fly* (em tempo de execução). É justamente nesta etapa que reside um dos principais problemas associados ao algoritmo OPWI, uma vez que o custo computacional desta etapa de síntese é relativamente alto. Estima-se que a equação 7.62 seja responsável por 80% do custo computacional associado a etapa de síntese do algoritmo OPWI. Com o objetivo de reduzir a complexidade da etapa de síntese do algoritmo OPWI, esta subseção apresenta um algoritmo capaz de reduzir em até 90% o custo computacional da equação 7.62.

Um dos principais custos computacionais associados à equação 7.62 refere-se ao cálculo do termo $C_i(k,n)$, devido à presença da exponencial complexa $e^{j\cdot\phi_i(n)\cdot k}$. Apesar de as máquinas modernas possuírem algoritmos rápidos para o cálculo de funções trigonométricas, o cálculo do termo $C_i(k,n)$ ainda apresenta um custo computacional importante.

Utilizando-se o fato do número de possíveis valores inteiros de T_0 ser relativamente reduzido, é possível se obter excelentes aproximações para o termo $C_i(k,n)$, utilizando-se funções trigonométricas *seno* e *cosseno*, previamente calculadas. Reescrevendo a equação 7.63 em termos de $cos(\phi_i(n)\cdot k)$ e $sen(\phi_i(n)\cdot k)$, tem-se:

$$\begin{aligned} C_i(k,n) &= 2 \cdot \Re\left(\widehat{X_i}(k,n)\right) \cdot cos\left(\varphi_i(n) \cdot k\right) - \Im\left(\widehat{X_i}(k,n)\right) \cdot sen\left(\varphi_i(n) \cdot k\right) \\ &= 2 \cdot \Re\left(\widehat{X_i}(k,n)\right) \cdot cos\left(\left(\frac{2\pi}{T_{0_i}} \cdot n - \Phi_i\right) \cdot k\right) - \\ &\quad 2 \cdot \Im\left(\widehat{X_i}(k,n)\right) \cdot sen\left(\left(\frac{2\pi}{T_{0_i}} \cdot n - \Phi_i\right) \cdot k\right) \end{aligned} \tag{7.82}$$

Definindo-se os vetores **SEN** e **COS**,

$$\mathbf{SEN} = \left[sen\left(0 \cdot \frac{2\pi}{T_{0_i}}\right), sen\left(1 \cdot \frac{2\pi}{T_{0_i}}\right), \cdots, sen\left((T_{0_i} - 1) \cdot \frac{2\pi}{T_{0_i}}\right)\right] \tag{7.83}$$

$$\mathbf{COS} = \left[cos\left(0 \cdot \frac{2\pi}{T_{0_i}}\right), cos\left(1 \cdot \frac{2\pi}{T_{0_i}}\right), \cdots, cos\left((T_{0_i} - 1) \cdot \frac{2\pi}{T_{0_i}}\right)\right] \tag{7.84}$$

e estimando-se o atraso em amostras, d_i, correspondente à função de fase Φ_i da equação 7.68,

$$d_i = -\langle\left(\frac{T_{0_i} \cdot \Phi_i}{2 \cdot \pi}\right)\rangle \tag{7.85}$$

então, verifica-se que a equação 7.82 pode ser aproximada por:

$$\begin{aligned} C_i(k,n) \approx\ & 2 \cdot \Re\left(\widehat{X_i}(k,n)\right) \cdot \mathbf{COS}\left[mod\left(\left(n - d_i\right) \cdot k, T_{0_i}\right)\right] + \\ & 2 \cdot \Im\left(\widehat{X_i}(k,n)\right) \cdot \mathbf{SEN}\left[mod\left(\left(n \cdot k - k \cdot d_i\right), T_{0_i}\right)\right] \end{aligned} \tag{7.86}$$

sendo *mod* o operador "resto da divisão inteira".

7.15 Considerações Finais

Este Capítulo apresentou um novo algoritmo para operar como módulo de *Back-End* em sistemas CTF-SCAUS, o algoritmo OPWI (*Optimized Prototype Waveform Interpolation*). Algumas das principais características e diferenças do algoritmo OPWI em relação aos outros algoritmos de *Back-End* mencionados ao longo deste Capítulo são:

- O algoritmo OPWI não faz uso do conceito de *maximum voiced frequency* utilizado pelo método HNM (Stylianou, 1996) e suas duas componentes CEL e CER apresentam componentes de freqüência que se estendem ao longo de toda a largura de banda do espectro.

- A estimativa dos protótipos ótimos é realizada utilizando-se dois métodos com diferentes resoluções tempo-freqüência, de acordo com o nível de estacionariedade do sinal
  - No primeiro método, os protótipos são otimizados levando-se em consideração que eles serão interpolados segundo funções de interpolação específicas. Apesar de este método reduzir a resolução temporal dos protótipos, ele garante ressínteses e modificações prosódicas de excelente qualidade.
  - O segundo método é semelhante ao proposto por (Stylianou, 1996) para estimar os parâmetros da componente harmônica de seu modelo HNM. Apesar deste método não levar em consideração o processo de interpolação dos protótipos, ele garante ressínteses e modificações prosódicas de boa qualidade e, além disso, apresenta uma alta resolução temporal, correspondente a apenas dois períodos fundamentais.
- O algoritmo OPWI opera sincronamente com os pulsos glotais, com instantes de análise localizados próximos aos IFGs. Esta operação síncrona com os IFGs garante alta qualidade nas modificações prosódicas e facilita a suavização espectral entre unidades de síntese.
- Como o algoritmo OPWI opera com valores inteiros de T_0, então algoritmos rápidos que trabalham com funções trigonométricas previamente calculadas, podem ser utilizados na etapa de síntese do algoritmo.

Capítulo 8

Algoritmo OPWI: Resultados Experimentais e Análises

8.1 Introdução

Este Capítulo apresenta resultados e análises de vários experimentos realizados com o algoritmo OPWI, com o objetivo de avaliar o desempenho das principais operações envolvidas nos módulos de análise e síntese deste algoritmo. Para o módulo de análise, as operações avaliadas foram a decomposição CEL/CER (Componente de Evolução Lenta/Componente de Evolução Rápida), a determinação dos níveis de estacionariedade da componente CEL e a estimativa dos protótipos ótimos. Quanto ao módulo de síntese, os experimentos e análises se concentraram na síntese das componentes CEL e CER e na síntese do sinal de fala (soma das componentes CER e CEL) quando sujeito a uma larga variedade de modificações prosódicas de TSM (*Time Scale Modifications*) e PSM (*Pitch Scale Modifications*).

Os sinais de fala utilizados nos experimentos consistem em apenas 4 sentenças, pronunciadas por 4 locutores distintos. As duas primeiras sentenças são de locutores falantes do português brasileiro, um locutor masculino natural de Minas Gerais e uma locutora feminina natural do Mato Grosso do Sul. Estas sentenças vêm sendo utilizadas em experimentos sobre modelagem da estrutura rítmica da fala pelo Grupo de Estudos de Prosódia da Fala do IEL/Unicamp (Instituto de Estudos da Linguagem/Universidade Estadual de Campinas). A terceira sentença é de uma locutora feminina falante do inglês americano. Esta voz feminina faz parte da base de dados ARCTIC (Kominek and Black, 2004) do projeto FESTVOX (Black, 2006) da CMU (*Carnegie Mellon University*), sobre síntese de fala. A quarta sentença é de um locutor masculino falante do alemão. Esta voz masculina é a voz oficial dos sistemas de síntese de fala do projeto SmartKom (Schweitzer et al., 2004), conduzido pelo IMS (*Institute for Natural Language Processing*), da Universidade de Stuttgart, na Alemanha. Para uma melhor identificação destas vozes ao longo deste Capítulo, serão utilizadas as seguintes notações:

- SF_PB - Sinal de fala da locutora feminina falante do Português Brasileiro (PB). Amostrado a uma taxa de 22025 Hz e quantizado com 16 bits. Sentença pronunciada: "*O vento sul e o sol*

discutiam qual dos dois era o mais forte, quando passou um viajante envolto num casaco".

- SM_PB - Sinal de fala do locutor masculino falante do PB. Amostrado a uma taxa de 22025 Hz e quantizado com 16 bits. Sentença pronunciada: "*O vento sul e o sol discutiam qual dos dois era o mais forte, quando passou um viajante envolto num casaco*".
- SF_US - Sinal de fala da locutora feminina falante do Inglês Americano. Amostrado a uma taxa de 16000 Hz e quantizado com 16 bits. Sentença pronunciada: "*A combination of Canadian capital quickly organized in partition for the same privileges*".
- SM_GER - Sinal de fala do locutor masculino falante do Alemão. Amostrado a uma taxa de 16000 Hz e quantizado com 16bits. Sentença pronunciada: "*Herzlich willkommen bei SmartKom information System. Ich bin Aladin, wie kann ich Ihnen helfen?*".

O restante deste Capítulo está dividido em 6 outras seções. A seção 8.2 apresenta as formas de onda e os respectivos espectogramas dos sinais, SF_PB, SM_PB, SF_US e SM_GER. A seção 8.3 apresenta resultados e análises sobre o processo de decomposição CEL/CER, com destaque para as freqüências de corte dos filtros utilizados na decomposição CEL/CER e para os espectogramas e segmentos de forma de ondas resultantes do processo de decomposição CEL/CER. A seção 8.4 analisa o desempenho dos estimadores dos níveis de estacionariedade ao longo do tempo da componente CEL. A seção 8.5 avalia o processo de estimação dos protótipos ótimos. A seção 8.6 avalia o desempenho do algoritmo OPWI nas operações de análise e ressíntese do sinal de fala sem modificações prosódicas. A seção 8.7 avalia a síntese do sinal de fala quando sujeito a diversas modificações prosódicas de TSM e PSM. A seção 8.8 apresenta uma lista de vários sinais sintetizados (com e sem modificações prosódicas de PSM e TSM) a partir dos sinais SF_PB, SM_PB, SF_US e SM_GER, utilizando-se o algoritmo OPWI. Todos estes sinais sintetizados, bem como os sinais originais SF_PB, SM_PB, SM_US e SM_GER, encontram-se disponíveis no CD em anexo a esta Tese. Finalmente, a seção 8.9 apresenta algumas considerações finais sobre o algoritmo OPWI.

8.2 Sinais de Fala a Serem Utilizados

As Figuras 8.1, 8.2, 8.3 e 8.4 apresentam as formas de onda e espectogramas dos sinais SF_PB, SM_PB, SF_US e SM_GER, respectivamente. Conforme poderá ser verificado ao longo dos resultados, todos os espectogramas apresentados neste Capítulo são de banda estreita. Esta escolha tem como objetivo ressaltar a estrutura harmônica dos sinais.

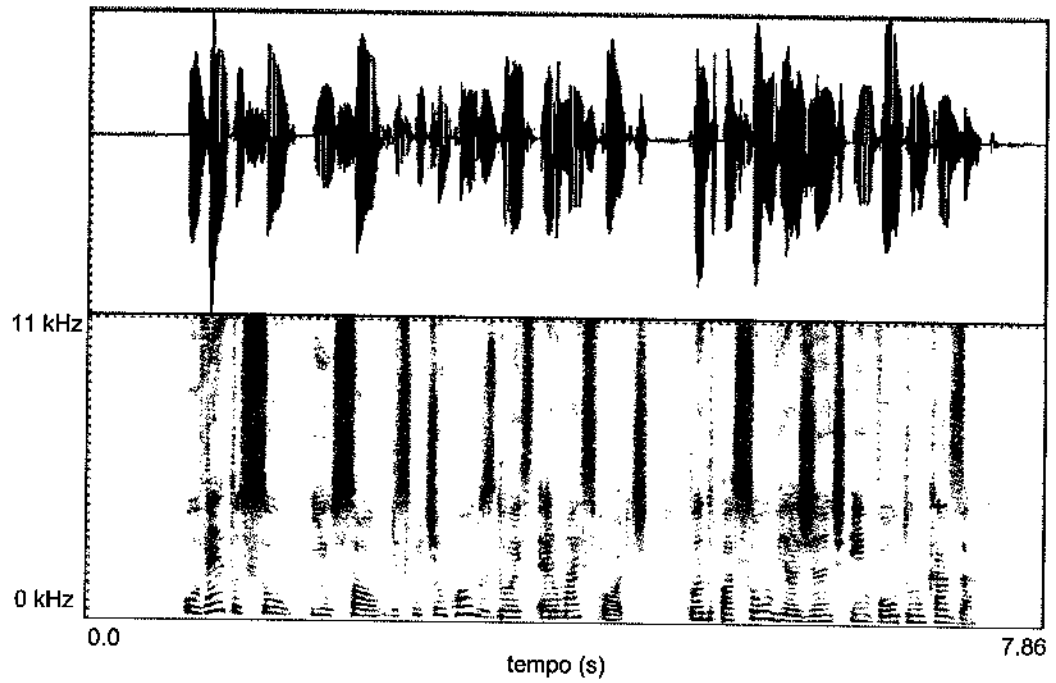


Figura. 8.1: Forma de onda e espectograma do sinal SF_PB. Taxa de amostragem 22050 Hz.

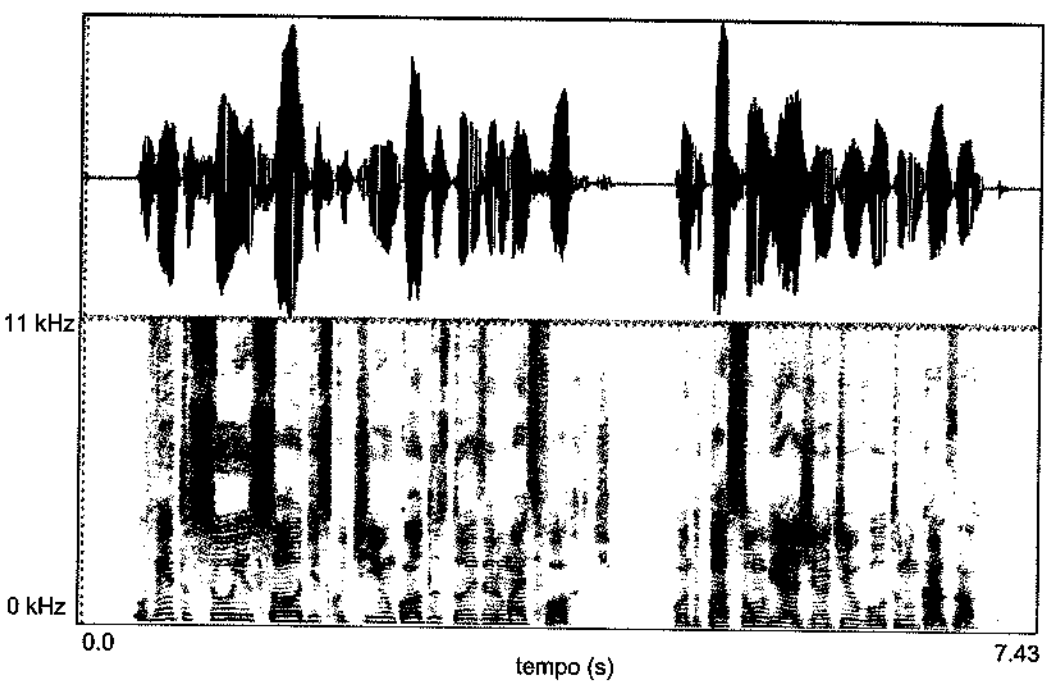


Figura. 8.2: Forma de onda e espectograma do sinal SM_PB. Taxa de amostragem 22050 Hz.

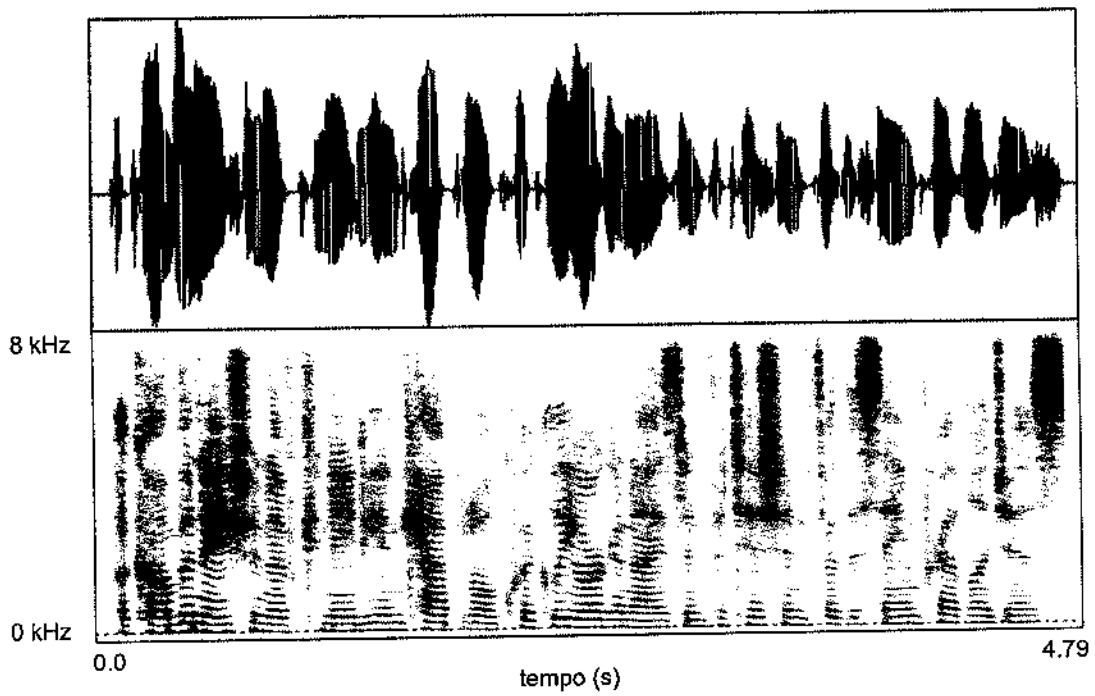


Figura. 8.3: Forma de onda e espectograma do sinal SF_US. Taxa de amostragem 16000 Hz

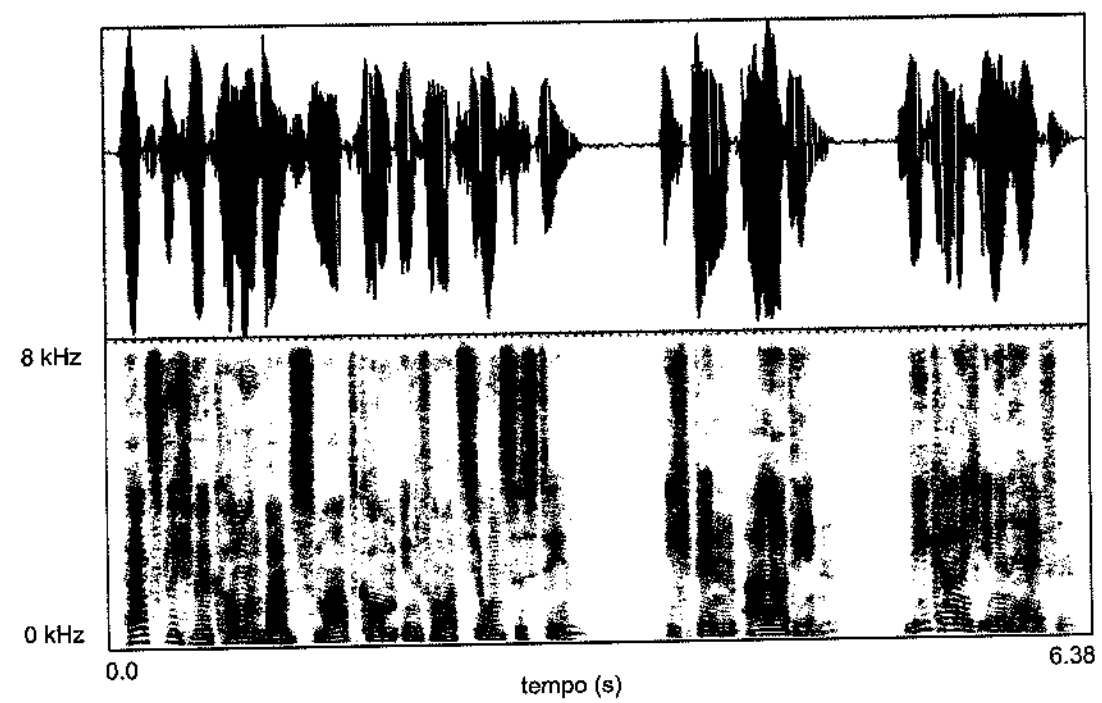


Figura. 8.4: Forma de onda e espectograma do sinal SM_GER. Taxa de amostragem 16000 Hz

Tabela. 8.1: Valores para o fator ξ.

	SF_PB	SM_PB	SF_US	SF_GER
ξ	0,75	0,5	1,0	0,7

8.3 Decomposição CEL/CER

8.3.1 Filtros para Decomposição CEL/CER

As Figuras 8.5, 8.6, 8.7 e 8.8, apresentam as freqüências de corte $f_c(k)$ dos filtros para decomposição CEL/CER a serem utilizados nos sinais SF_PB, SM_PB, SF_US e SM_GER, respectivamente. Estas freqüências de corte, $f_c(k)$, foram obtidas aplicando-se a equação 7.2 a cada um dos sinais SF_PB, SM_PB, SF_US e SM_GER e em seguida empregando-se um filtro de suavização para eliminar variações abruptas em $f_c(k)$ ao longo do eixo k. Análises experimentais (baseadas unicamente em conhecimentos de especialistas) com os sinais SF_PB, SM_PB, SF_US e SM_GER, indicaram os valores da Tabela 8.1 para o fator ξ da equação 7.2.

Nas Figuras 8.5, 8.6, 8.7 e 8.8, as freqüências de corte dos filtros $f_c(k)$, (eixo vertical) foram normalizadas para que o valor 1 correspondesse à máxima freqüência discreta π. O eixo de freqüências (eixo horizontal) também foi normalizado para que $k = 1$ correspondesse à máxima freqüência contida no sinal em análise. É importante relembrar que a máxima freqüência contida nos sinais SF_PB e SM_PB é 11025 Hz e a máxima freqüência contida no sinal SM_GER e SF_US é 8000 Hz.

Uma análise das Figuras 8.5, 8.6, 8.7 e 8.8 sugere as seguintes observações sobre as freqüências de corte dos filtros $f_c(k)$, e sobre o processo de decomposição CEL/CER:

- A freqüência de corte dos filtros para as componentes de freqüências próximas a 0 Hz é extremamente baixa. Esta freqüência de corte é responsável por minimizar (eliminar) flutuações no valor DC dos espectro $\widehat{S_W}(k, m)$ (Figura 7.4), ao longo do eixo m (eixo dos pulsos glotais).
- Apesar da existência de flutuações (mais ou menos abruptas), verifica-se um forte decaimento da freqüência de corte $f_c(k)$ na faixa de freqüências entre, aproximadamente, 220 Hz e 4000 Hz.
- Na faixa de freqüências entre, aproximadamente, 6000 Hz e 8000 Hz, verifica-se uma ligeira elevação das freqüências de corte $f_c(k)$, (apesar da existência de flutuações mais ou menos abruptas). Análises, ainda não muito criteriosas, indicam que este aumento das freqüências de corte deve-se à concentração de componentes harmônicas nesta faixa de freqüências.
- A função $f_c(k)$ tende a não se anular ao longo da faixa de freqüência $0 < k \leq 1$. Por conseguinte, conclui-se que a componente CEL tende a possuir componentes de freqüência com energia maior que zero, ao longo de toda a faixa de freqüências $0 < k \leq 1$.

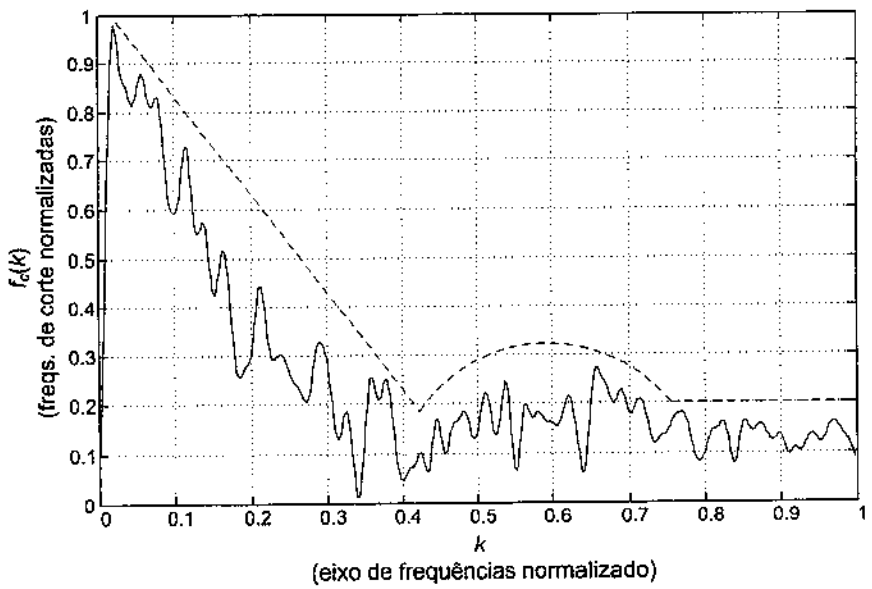


Figura. 8.5: Freqüências de corte dos filtros de decomposição CEL/CER, $f_c(k)$, para o sinal SF_PB (linha contínua). Contorno estilizado de $f_c(k)$ (linha pontilhada).

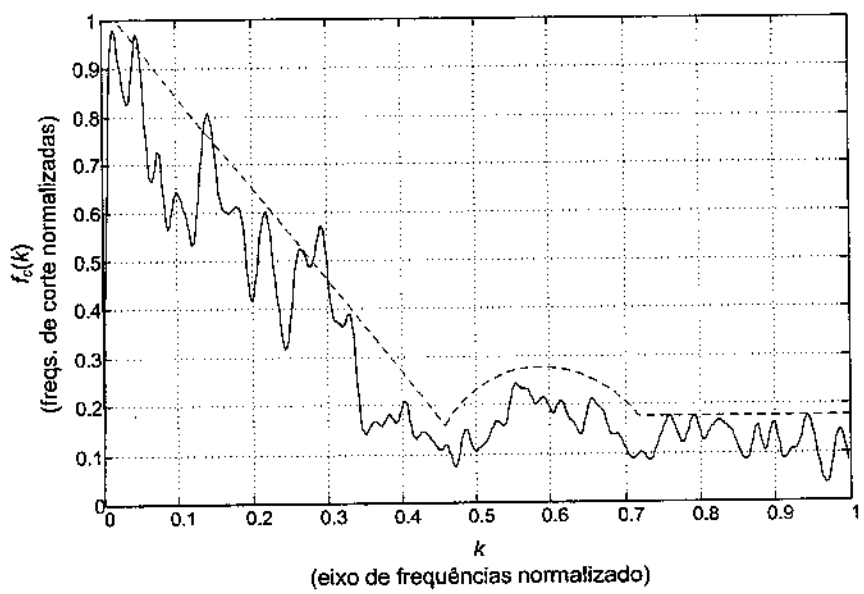


Figura. 8.6: Freqüências de corte dos filtros de decomposição CEL/CER, $f_c(k)$, para o sinal SM_PB (linha contínua). Contorno estilizado de $f_c(k)$ (linha pontilhada).

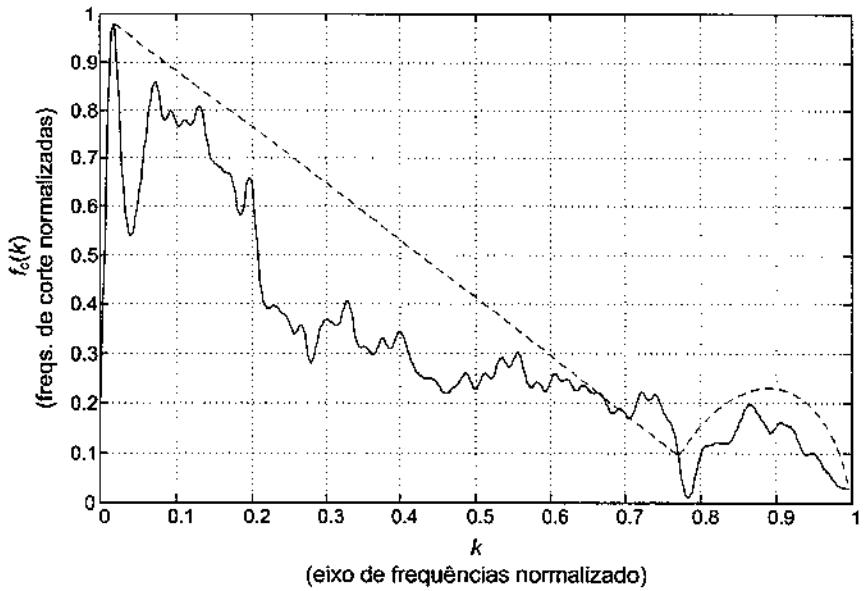


Figura. 8.7: Freqüências de corte dos filtros de decomposição CEL/CER, $f_c(k)$, para o sinal SF_US (linha contínua). Contorno estilizado de $f_c(k)$ (linha pontilhada).

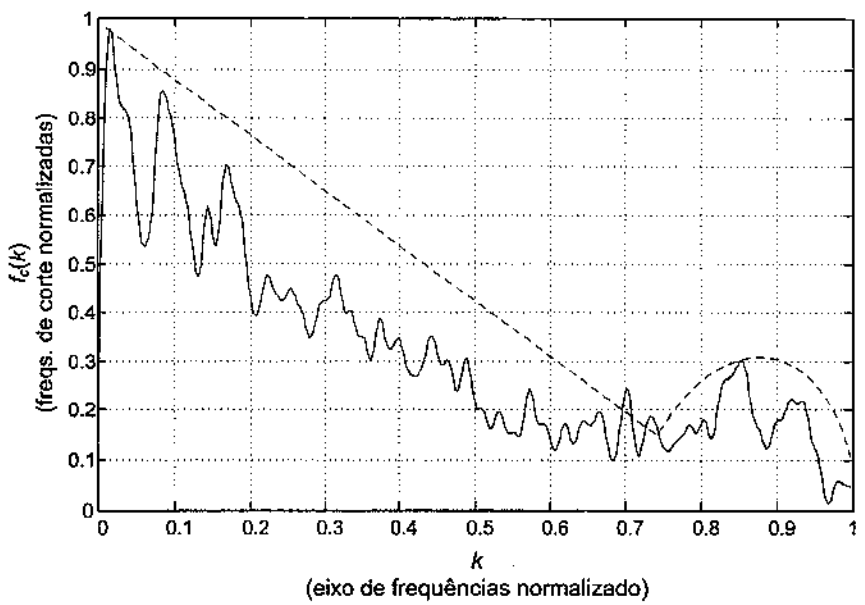


Figura. 8.8: Freqüências de corte dos filtros de decomposição CEL/CER, $f_c(k)$, para o sinal SM_GER (linha contínua). Contorno estilizado de $f_c(k)$ (linha pontilhada).

8.3.2 Avaliação Espectral

As Figuras 8.9 e 8.10 apresentam as formas de onda e os espectogramas resultantes da decomposição CEL/CER do sinal SF_PB. Uma análise da Figura 8.9 permite verificar que:

- A estrutura harmônica da componente CEL é mais proeminente (mais definida) do que a estrutura harmônica do sinal original (Figura 8.1).
- As amplitudes dos segmentos não-sonoros da componente CEL são reduzidas a zero.
- Há uma redução da componente de ruído que se encontra somada aos segmentos sonoros da componente CEL. Como será verificado posteriormente, nas Figuras 8.17, 8.18 e 8.19, esta redução se dá, principalmente, nos segmentos fricativos sonoros.
- O envelope de amplitude dos segmentos sonoros, na componente CEL, é praticamente idêntico ao envelope de amplitude dos segmentos sonoros no sinal original (Figura 8.1).

Uma análise da Figura 8.10 (Componente CER), permite verificar que:

- A estrutura harmônica ao longo dos segmentos sonoros do sinal SF_PB foi quase que totalmente removida.
- A componente CER contém todo o sinal correspondente aos segmentos não-sonoros, destacando-se:
  - Os segmentos ruidosos com energia aproximadamente constante ao longo do tempo, por exemplo, sons fricativos surdos.
  - Os segmentos transitórios tais como *bursts* de plosivas.
  - Segmentos sonoros resultantes de erros na classificação sonoro/não-sonoro dos quadros de análise do sinal SF_PB.
- Os segmentos sonoros da componente CER contêm:
  - Variações bruscas na estrutura espectral do sinal SF_PB ao longo do tempo. Estas variações incluem, principalmente, variações bruscas na evolução temporal da estrutura dos formantes (variações nas freqüências centrais e larguras de banda dos formantes).
  - Variações extremamente bruscas no contorno de energia do sinal.
  - Ruído ao longo de segmentos sonoros. Principalmente ao longo de sons fricativos sonoros.

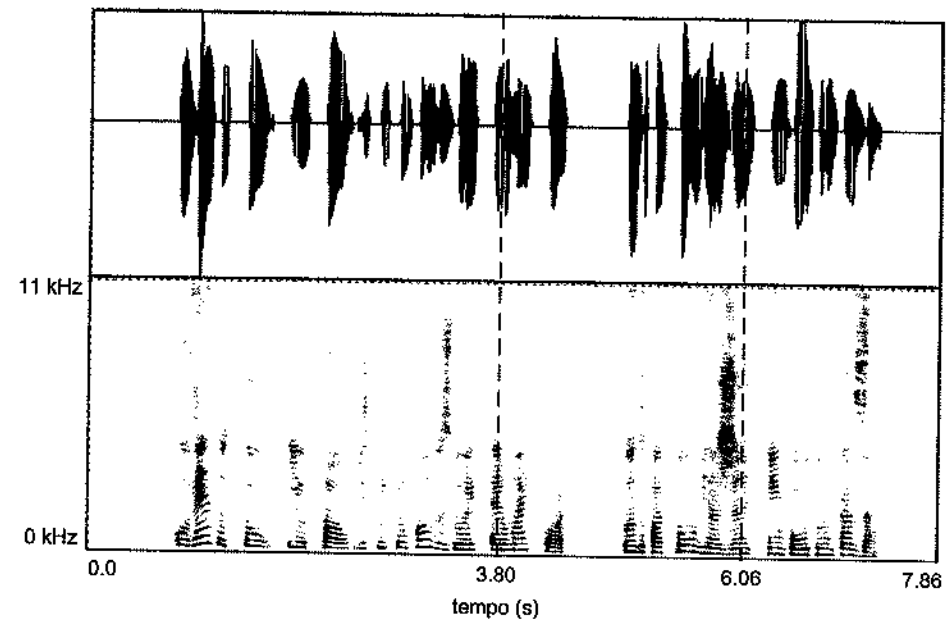


Figura. 8.9: Forma de onda e espectograma da componente CEL de SF_PB. Há uma ênfase na estrutura harmônica do sinal, os segmentos não-sonoros são eliminados e a componente de ruído dos segmentos sonoros é fortemente atenuada.

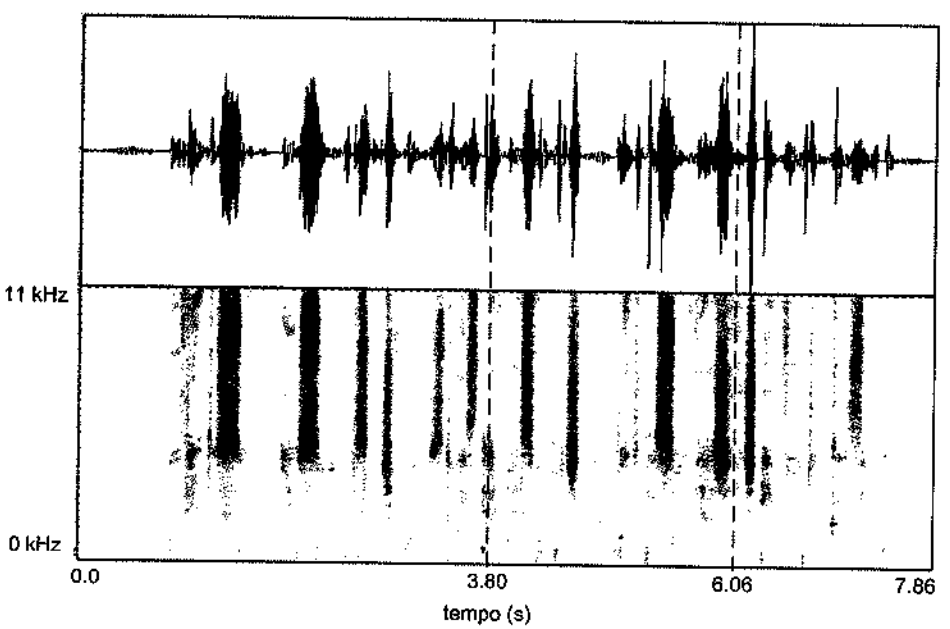


Figura. 8.10: Forma de onda e espectograma da componente CER de SF_PB. Esta componente contém toda a energia dos segmentos não-sonoros e também as componentes de ruído que se misturavam (ao longo de toda a banda de freqüências) à componente harmônica (CEL)

As Figuras 8.11, 8.12 e 8.13 apresentam espectros (apenas a magnitude) tomados ao longo de uma janela, associada a um segmento sonoro, centrada no instante de tempo igual a 3,8 segundos, do sinal, SF_PB e suas componentes CEL e CER, respectivamente. Destas Figuras, pode-se verificar que a magnitude do espectro da componente CEL se aproxima significativamente da componente harmônica do sinal SF_PB. Por outro lado, todos os ruídos presentes ao longo da magnitude do espectro do sinal SF_PB, sejam eles ruídos localizados entre os lóbulos da estrutura harmônica ou ruídos em regiões do espectro que aparentemente não apresentam estrutura harmônica, foram transferidos para o espectro da componente CER.

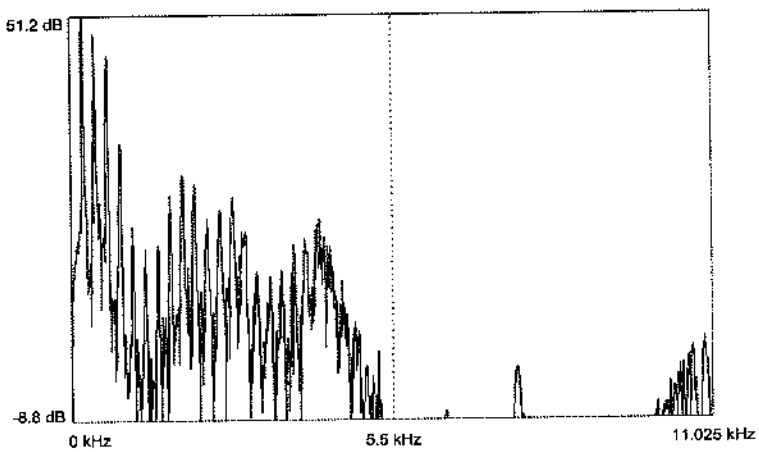


Figura. 8.11: Espectro de magnitude de um segmento janelado do sinal SF_PB, centrado no instante de tempo 3,8 segundos.

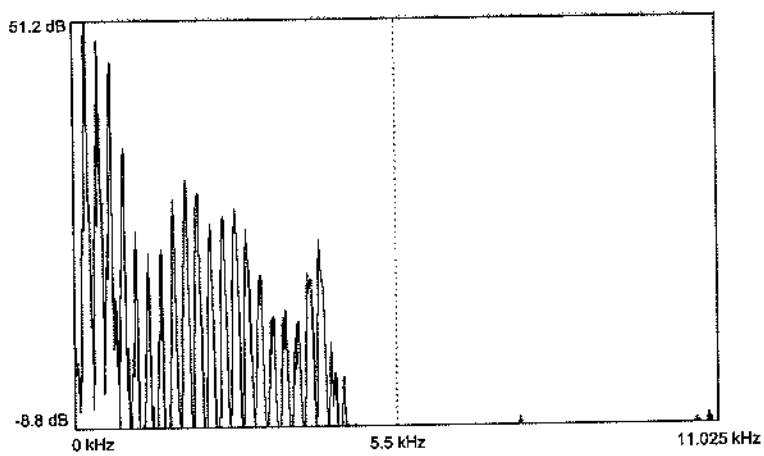


Figura. 8.12: Espectro de magnitude de um segmento janelado da componente CEL, do sinal SF_PB, centrado no instante de tempo 3,8 segundos.

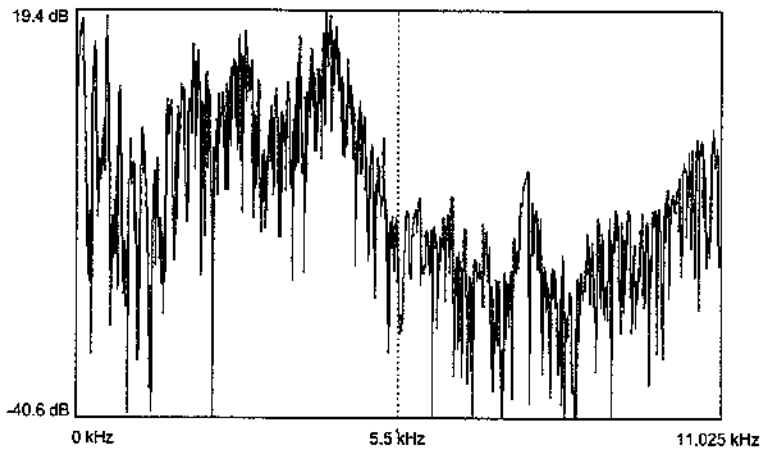


Figura. 8.13: Espectro de magnitude de um segmento janelado da componente CER, do sinal SF_PB, centrado no instante de tempo 3,8 segundos.

As Figuras 8.14, 8.15 e 8.16 apresentam espectros (apenas a magnitude) tomados ao longo de janelas, associadas a segmentos sonoros, centrados no instante 6,07 segundos, para os sinais SF_PB e suas componente CEL e CER, respectivamente. As mesmas análises feitas para as Figuras 8.11, 8.12 e 8.13, também são válidas para as Figuras 8.14, 8.15 e 8.16.

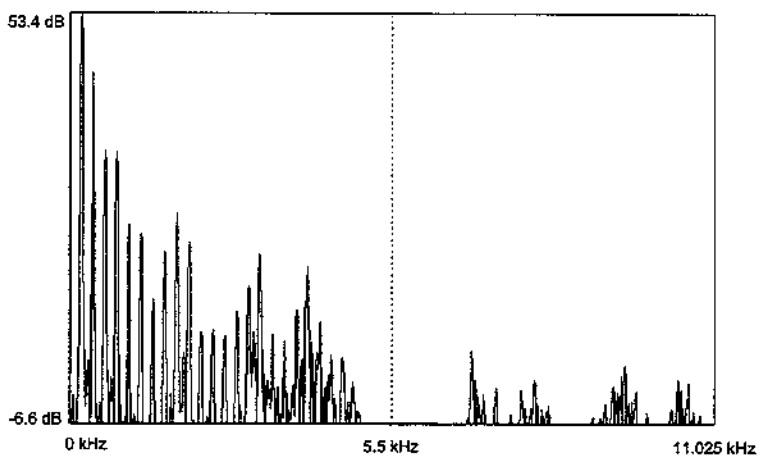


Figura. 8.14: Espectro de magnitude de um segmento janelado do sinal SF_PB, centrado no instante de tempo 6,06 segundos.

8.3.3 Avaliação Temporal

As Figuras 8.17, 8.18 e 8.19 apresentam formas de onda resultantes da decomposição CEL/CER de segmentos curtos do sinal SF_PB. Um análise destas Figuras permiter verificar, como já dito anteriormente, que a decomposição CEL/CER consegue remover praticamente todo o ruído misturado

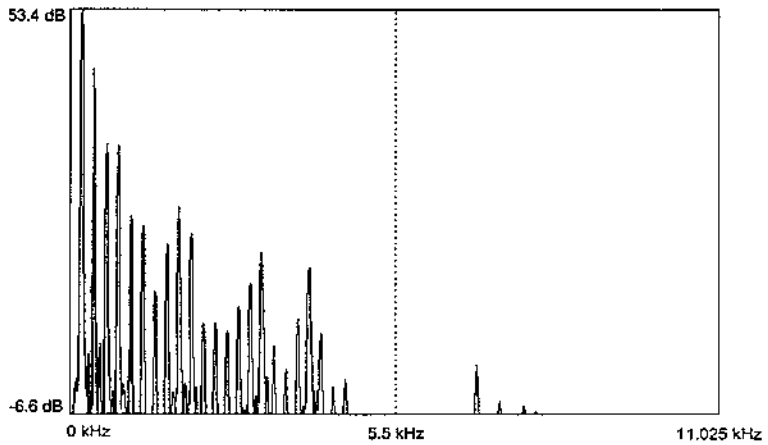


Figura. 8.15: Espectro de magnitude de um segmento janelado da componente CEL, do sinal SF_PB, centrado no instante de tempo 6,06 segundos.

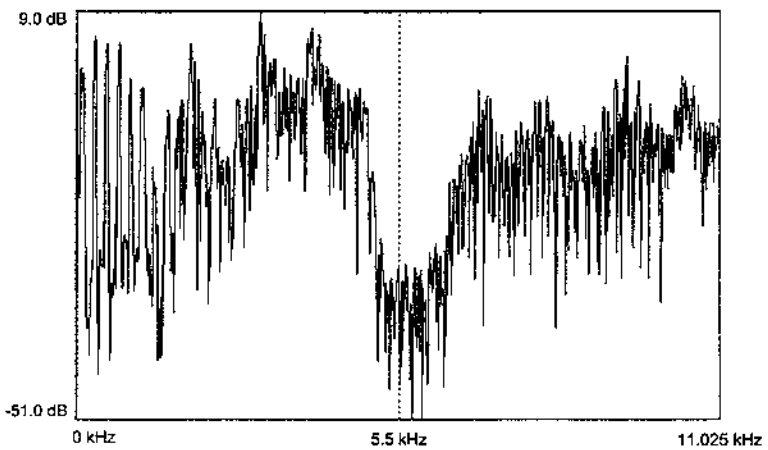


Figura. 8.16: Espectro de magnitude de um segmento janelado da componente CER, do sinal SF_PB, centrado no instante de tempo 6,06 segundos.

à componente sonora sem introduzir modificações significativas no contorno de amplitude do sinal. É importante ressaltar que apesar dos filtros de decomposição CEL/CER possuírem freqüências de corte $f_c(k)$ distintas, o processo de decomposição (filtragem) não causa qualquer alteração nos instantes de realização dos pulsos glotais (abertura e fechamento). A preservação dos instantes de realização dos pulsos glotais é de fundamental importância para os estágios seguintes de análise e síntese do algoritmo OPWI.

As Figuras 8.17 e 8.18, ilustram a capacidade da decomposição CEL/CER de, praticamente, eliminar os ruídos do sinal de fala ao longo dos segmentos fricativos sonoros, sem introduzir distorções significativas nas amplitudes da componente CEL. A Figura 8.19 mostra que a decomposiçao CEL/CER é capaz de eliminar da componente CEL, ruídos esporádicos localizados em intervalos de curtíssima

duração.

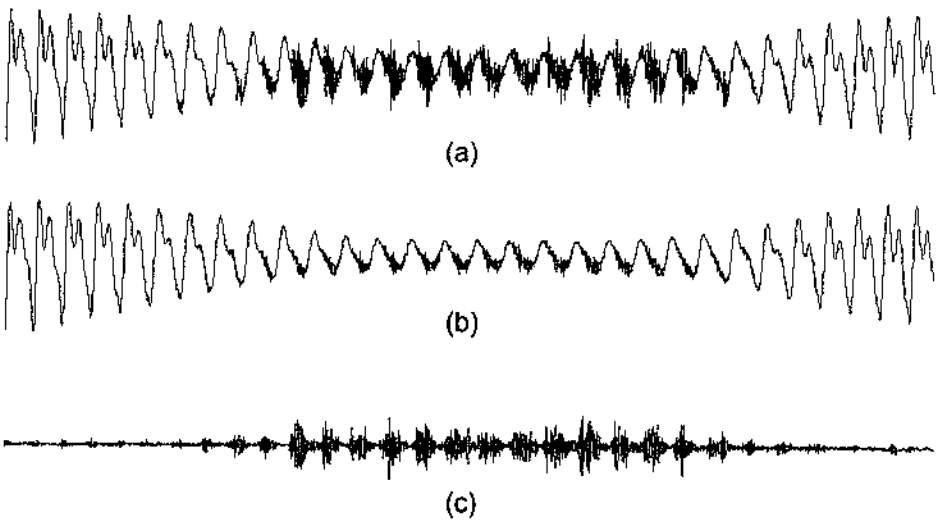


Figura. 8.17: Segmento do sinal SF_PB submetido ao processo de decomposição CEL/CER. (a) Sinal original, (b) componente CEL e (c) componente CER. (Destaque para o trecho fricativo sonoro)

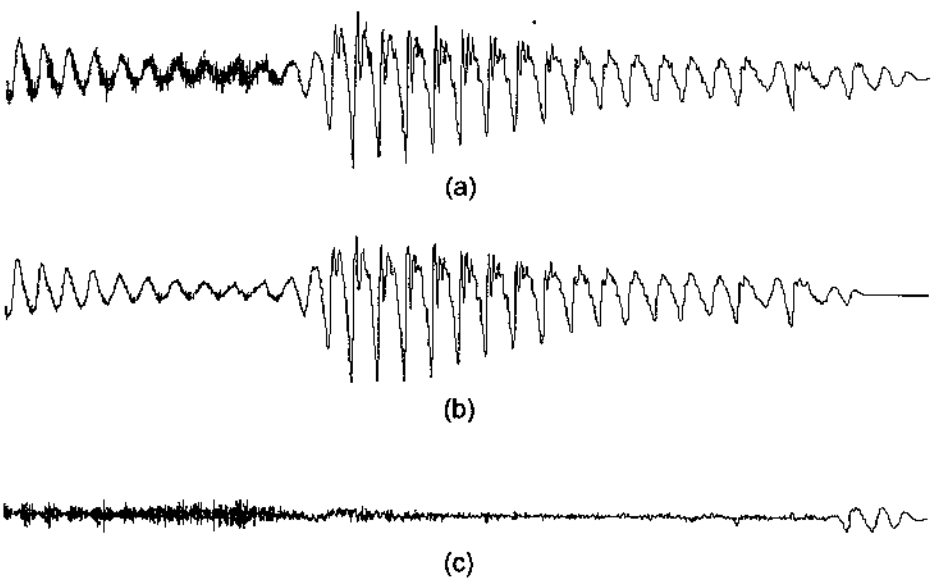


Figura. 8.18: Segmento do sinal SF_PB submetido ao processo de decomposição CEL/CER. (a) Sinal original, (b) componente CEL e (c) componente CER.

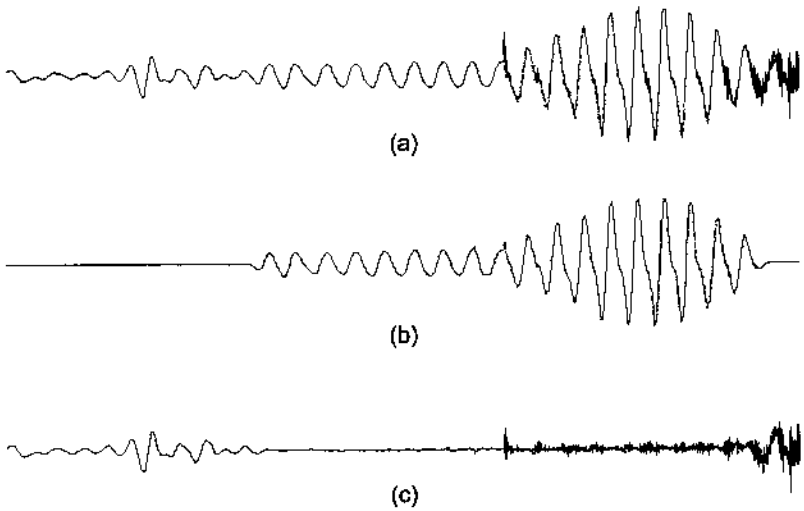


Figura. 8.19: Segmento do sinal SF_PB submetido ao processo de decomposição CEL/CER. (a) Sinal original, (b) componente CEL e (c) componente CER.

8.4 Estimativa do Nível de Estacionariedade

O nível de estacionariedade foi estimado empregando-se o critério C3 descrito na equação 7.9. A Figura 8.20 ilustra a estimativa do nível de estacionariedade da componente CEL do sinal SM_GER. Nesta Figura os segmentos delimitados pela faixas verticais (em cinza) correspondem aos segmentos classificados como estacionários. Todos os outros segmentos, incluindo os segmentos não-sonoros, são considerados não-estacionários. É importante ressaltar que os segmentos não-sonoros foram forçadamente classificados como não-estacionários.

Como descrito na equação 7.9, o critério C3 é uma combinação dos critérios C1 e C2, equações 7.3 e 7.8, respectivamente. Uma análise das Figuras 8.20(b), 8.20(c) e 8.20(d) confirma que a contribuição do critério C1 na formação do critério C3 foi a principal responsável pela detecção nas variações de energia, ao longo do tempo. Estas variações de energia se manifestam principalmente nas regiões de transição sonoro/não-sonoro e não-sonoro/sonoro. Por outro lado a contribuição do critério C2 na formação do critério C3 foi a principal responsável pela detecção das variações abruptas (ao longo do tempo) nas estruturas dos formantes.

As Figuras 8.21, 8.22 e 8.23 ilustram, em mais detalhes, que o critério C3 foi capaz não somente de detectar variação no contorno de energia do sinal ao longo do tempo, mas também de detectar transições abruptas na estrutura dos formantes da componente CEL do sinal SM_GER.

Através de análises experimentais (baseadas unicamente em conhecimentos de especialistas) com os sinais SM_GER, SF_PB, SM_PB e SF_US, as constantes α_2, α_3 e β_3 das equações 7.8 e 7.9 foram ajustadas para $\alpha_2 = 30$, $\alpha_3 = 2$ e $\beta_3 = 4$.

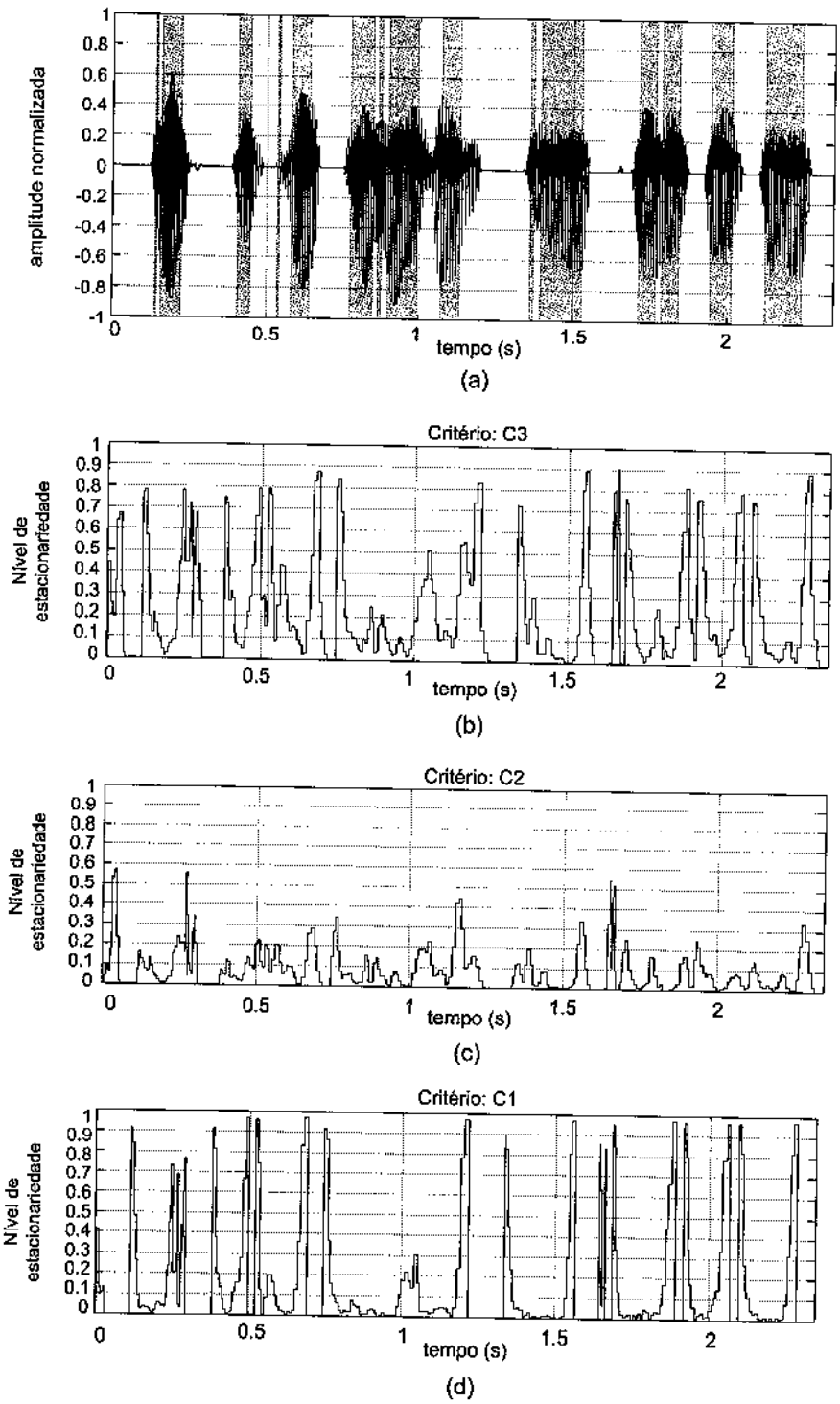


Figura. 8.20: Análise e detecção dos níveis de estacionariedade do sinal SM_GER. (a) As faixas verticais (cinza) indicam os segmentos estacionários. (b) Critério C3 que combina os critérios C1 e C2. (c) Critério C2, mede variações na estrutura dos formantes do sinal ao longo do tempo. (d) Critério C1, mede variações de energia ao longo do tempo.

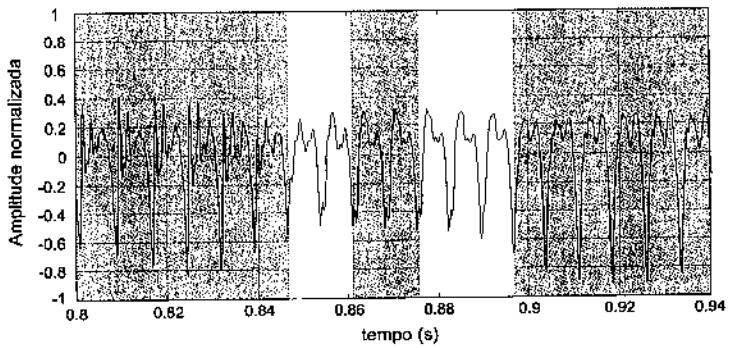


Figura. 8.21: Classificação estacionário × não-estacionário ao longo de um trecho da componente CEL do sinal SM_GER. (Segmentos estacionários são indicados pelas faixas cinzas).

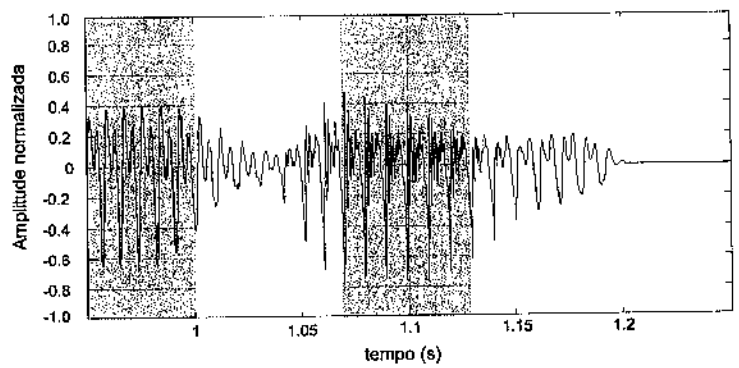


Figura. 8.22: Classificação estacionário × não-estacionário ao longo de um trecho da componente CEL do sinal SM_GER. (Segmentos estacionários são indicados pelas faixas cinzas).

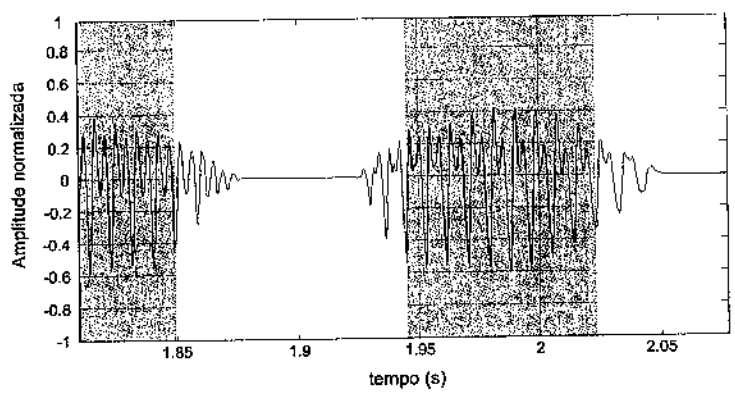


Figura. 8.23: Classificação estacionário × não-estacionário ao longo de um trecho da componente CEL do sinal SM_GER. (Segmentos estacionários são indicados pelas faixas cinzas).

8.5 Estimativa dos Protótipos Ótimos

As formas de onda representadas em linha contínua nas Figuras 8.24(a), 8.25(a) e 8.26(a) representam extensões periódicas, ao longo de dois períodos fundamentais, das representações temporais de três protótipos ótimos, estimados para diferentes instantes de análise ao longo da componente CEL do sinal SM_GER. Por outro lado, as formas de onda representadas em linha tracejada nas Figuras 8.24(a), 8.25(a) e 8.26(a) representam os segmentos originais da componente CEL (ao longo de dois períodos de fundamentais), utilizados para a estimativa dos respectivos protótipos ótimos.

Denominando estes protótipos ótimos de $X_j^P(n)$ e sua representação temporal de $x_j^P(n)$ (independente do instante de análise em questão), então pode-se verificar que as propriedades desejadas para as representações temporais destes protótipos, que haviam sido previamente estabelecidas na seção 7.6, foram de fato satisfeitas.

P1. $x_j^P(n)$ é um seqüência de números reais;

P2. $x_j^P(n)$ é uma função periódica com período T_{0_j};

P3. $|x_j^P(0) - x_j^P(T_{0_j} - 1)| < \delta$, com $\delta \to 0$;

sendo n_j^a o instante de análise associado ao protótipo X_j^P.

Em outras palavras, as representações temporais dos protótipos estimados, x_j^P, são de fato funções periódicas com perído T_{0_j}. Além disso a extensão periódica das representações temporais dos protótipos x_j^P não apresentam qualquer tipo de descontinuidade indesejada. Finalmente, nas proximidades dos instantes de análise (pulsos glotais), o erro entre a representação temporal dos protótipos e o sinal original é extremamente reduzido (próximo de 0).

As Figuras 8.24(b), 8.25(b) e 8.26(b) apresentam os espectros de magnitude das extensões periódicas das representações temporais dos protótipos x_j^P, apresentadas, respectivamente, nas Figuras 8.24(a), 8.25(a) e 8.26(a). As Figuras 8.24(c), 8.25(c) e 8.26(c) apresentam os espectros de magnitude dos segmentos originais da componente CEL apresentados, respectivamente, nas Figuras 8.24(a), 8.25(a) e 8.26(a).

É importante alertar que o eixo de freqüências das Figuras 8.24(b), 8.25(b), 8.26(b), 8.24(c), 8.25(c), e 8.26(c) foram normalizados para que o valor $\frac{1}{4}$ corresponda à freqüencia discreta $\pi/4$ (com o objetivo de ampliar os detalhes presentes na faixa de freqüência discreta entre 0 e $\pi/4$).

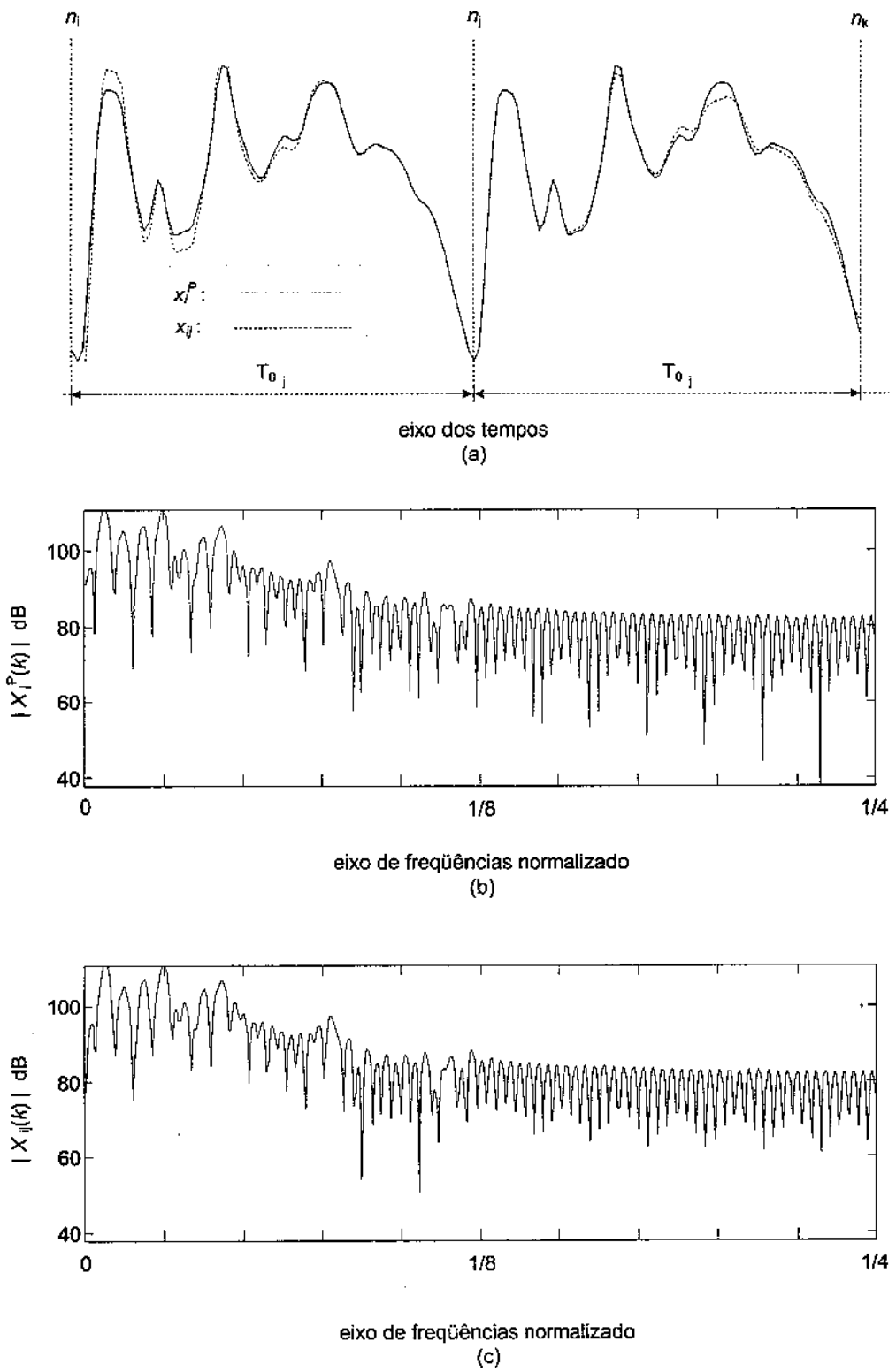


Figura. 8.24: Análise da representação temporal dos protótipos ótimos estimados. (a) sinal original (linha tracejada); extensão periódica da representação temporal do protótipo estimado ao longo de dois períodos fundamentais (linha contínua). (b) Espectro de magnitude correspondente à extensão periódica (2 períodos) da representação temporal do protótipo em (a). (c) Espectro de magnitude correspondente ao sinal original em (a) (ao longo dos 2 períodos).

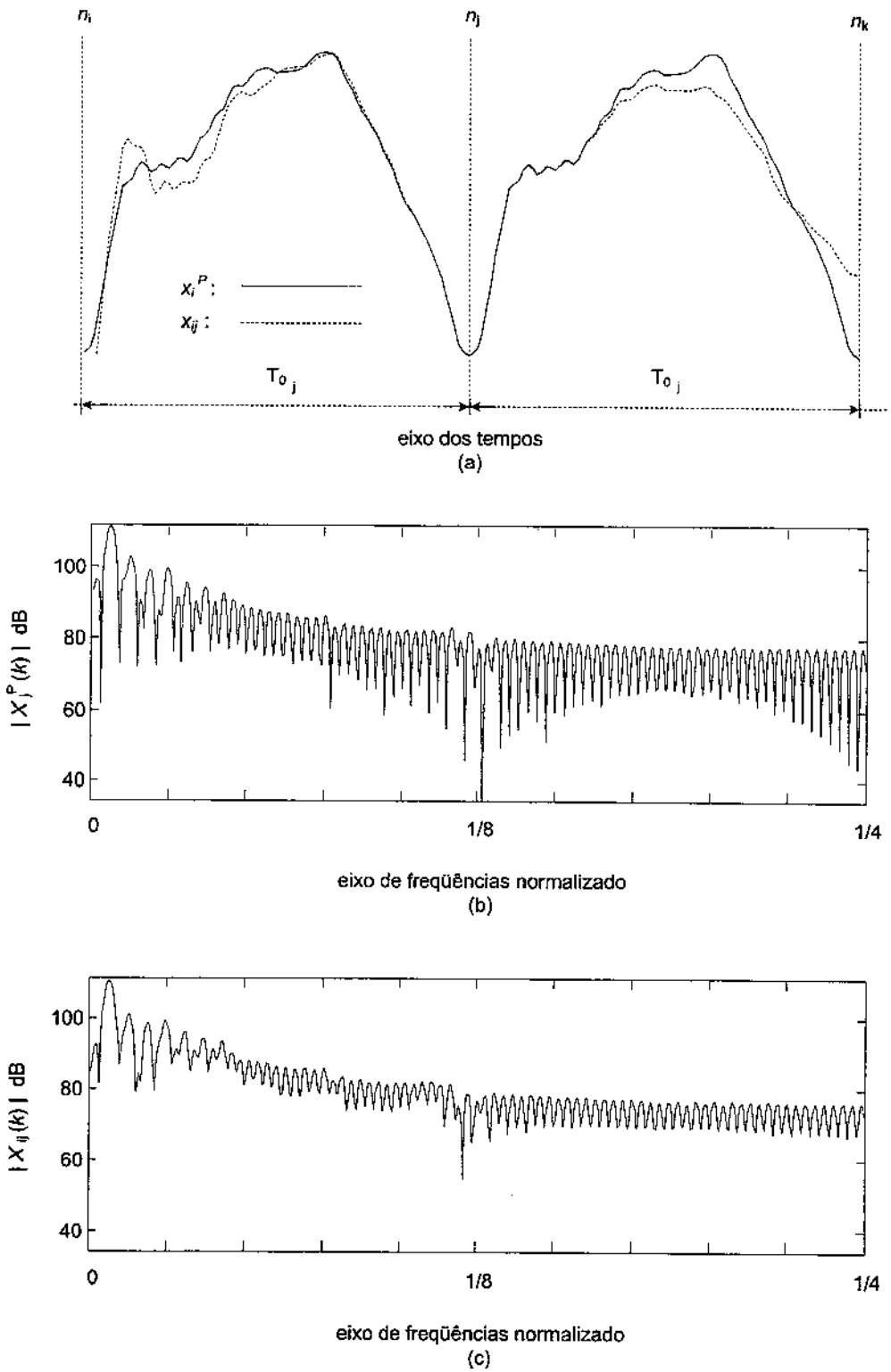


Figura. 8.25: Análise da representação temporal dos protótipos ótimos estimados. (a) sinal original (linha tracejada); extensão periódica da representação temporal do protótipo estimado ao longo de dois período fundamentais (linha contínua). (b) Espectro de magnitude correspondente à extensão periódica (2 períodos) da representação temporal do protótipo em (a). (c) Espectro de magnitude correspondente ao sinal original em (a) (ao longo dos 2 períodos).

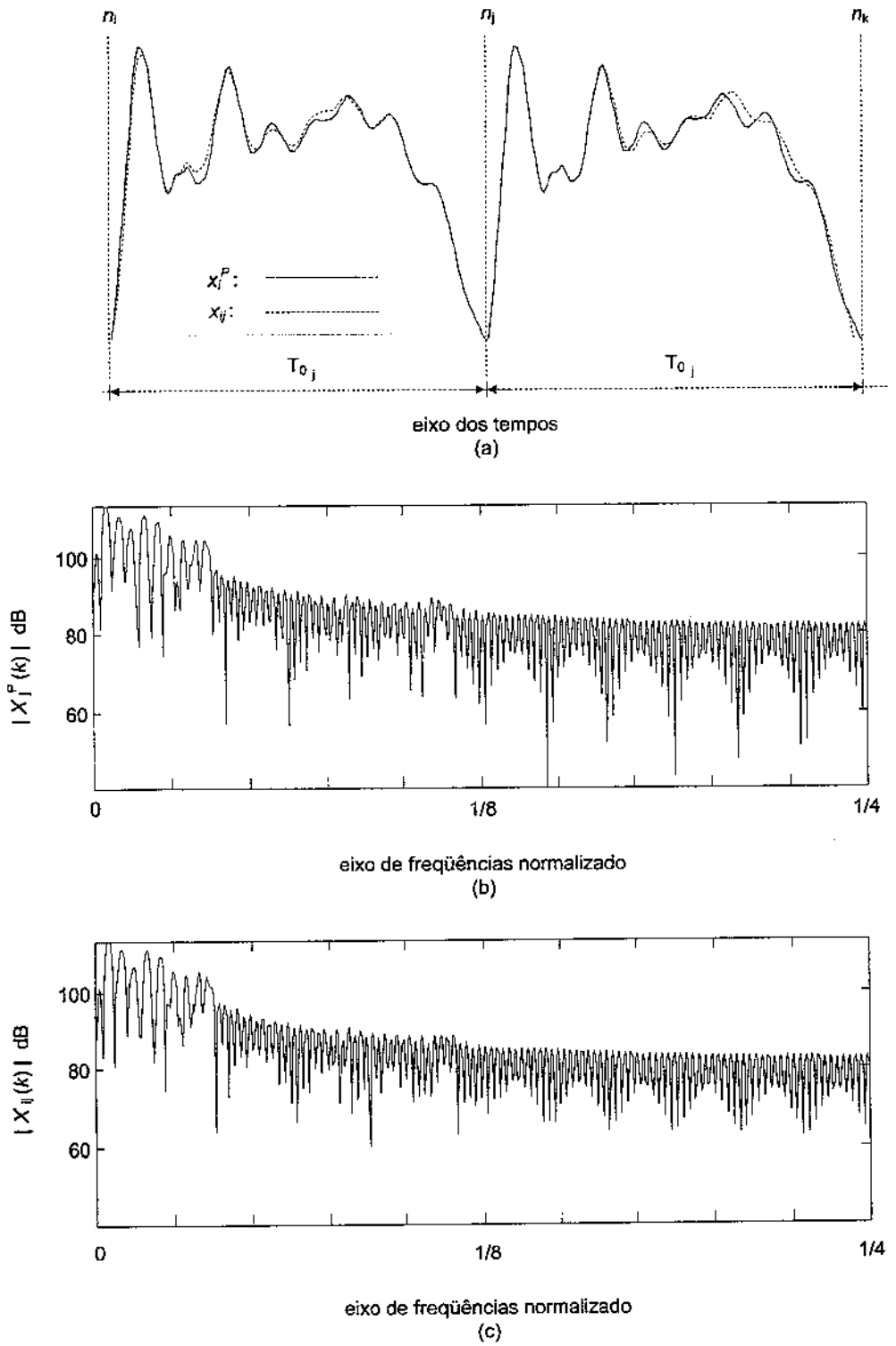


Figura. 8.26: Análise da representação temporal dos protótipos ótimos estimados. (a) sinal original (linha tracejada); extensão periódica da representação temporal do protótipo estimado ao longo de dois períodos fundamentais (linha continua). (b) Espectro de magnitude correspondente à extensão periódica (2 períodos) da representação temporal do protótipo em (a). (c) Espectro de magnitude correspondente ao sinal original em (a) (ao longo dos 2 períodos).

Tabela. 8.2: Valores de SNR obtidos pelo algoritmo OPWI no processo de análise/ressíntese.

Experimento	SF_PB	SM_PB	SF_US	SM_GER	Média
E_1 (Métodos I e II)	30,51 dB	30,03 dB	29,81 dB	31,82 dB	30,54 dB
E_2 (Métodos II)	28,74 dB	26,90 dB	27,30 dB	29,11 dB	28,01 dB
E_3 (Métodos I)	34,29 dB	33,19 dB	33,72 dB	34,82 dB	34,01 dB

8.6 Análise e Ressíntese

Para avaliar o desempenho do algoritmo OPWI nas operações de análise e ressíntese (sem qualquer tipo de modificação prosódica), foram realizados três experimentos. No primeiro experimento (E_1) os protótipos ótimos foram estimados empregando-se o Método I e o Método II (seções 7.7 e 7.8, respectivamente), conforme o nível de estacionariedade dos quadros de análise da componente CEL. No segundo experimento (E_2) os protótipos ótimos foram estimados utilizando-se apenas o Método II (seção 7.8). No terceiro experimento (E_3) os protótipos ótimos foram estimados empregando-se apenas o Método I (seção 7.7).

A Tabela 8.2 mostra os valores de relação sinal ruído (SNR - *Signal to Noise Ratio*) obtidos pelo algoritmo OPWI no processo de análise e ressíntese segundo estes três experimentos. Os valores de SNR foram estimados segundo a equação 8.1 a seguir:

$$SNR = \frac{\sum_{n=0}^{N_s-1} s(n)^2}{\sum_{n=0}^{N_s-1} s(n)^2 - \hat{s}(n)^2} \tag{8.1}$$

sendo $s(n)$ o sinal de fala original, $\hat{s}(n)$ o sinal de fala ressintetizado e N_s o número total de amostras de $s(n)$.

Analisando-se os resultados da Tabela 8.2 pode-se verificar que o experimento E_3 produziu um SNR médio de 34,01 dB, o que corresponde a 6 dB acima do SNR médio para o experimento E_2. Por outro lado, o experimento E_1 apresentou um SNR médio de 30,54 dB, o que corresponde a 2,54 dB acima do experimento E_2 e 3,47 dB abaixo do experimento E_3. É importante ressaltar que o desempenho do experimento E_1 se aproximará do desempenho do experimento E_3 nas seguintes condições:

- Se a componente CEL apresentar um elevado nível de estacionariedade. Este nível de estacionariedade pode ser conseqüência das características do locutor (por exemplo, se o locutor apresentar uma tendência por transições suaves entre segmentos não-sonoro e sonoros) ou do processo de decomposição CEL/CER (por exemplo, se o fator ξ da equação 7.2 for reduzido, então a componente CEL será mais suave).

- Se o critério C^3, utilizado para medir o nível de estacionariedade da componente CEL, favorecer a classificação de quadros de análise como estacionários.

8.7 Modificações Prosódicas

8.7.1 TSM por um Fator Constante

Análise Espectral

A Figura 8.27 apresenta um segmento do sinal original SM_PB no intervalo de 0,57 segundo a 4,88 segundos. Este segmento é submetido a modificações prosódicas de TSM iguais a 2,4 e 0,7 respectivamente, e o resultado é mostrado nas Figuras 8.28 e 8.29. Pode ser verificado da Figura 8.28 que uma operação de TSM por um fator de 2,4 preserva a dinâmica espectral (das estruturas harmônicas e ruidosas) do sinal sem qualquer efeito indesejável de compressão espectral. De forma semelhante a Figura 8.29 mostra que uma operação de TSM por um fator de 0,7 também preserva a dinâmica espectral do sinal sem introduzir qualquer efeito indesejável de expansão espectral. Além disso, pode ser verificado das Figuras 8.28 e 8.29, que as operações de TSM mantiveram os contornos de amplitude do sinal SM_PB praticamente intactos.

Análise Temporal

A Figura 8.30 mostra um segmento do sinal SF_PB submetido a uma TSM igual a 0,7. Dois aspectos importantes merecem destaque nesta Figura, a componente CEL sintetizada é capaz de preservar com elevada precisão os contornos de amplitude do sinal original e a componente CER é sintetizada em perfeito sincronismo temporal com a componente CEL. Este sincronismo entre as componentes CEL e CER garante que os ruídos dos segmentos fricativos sonoros sejam adequadamente sintetizados em regiões próximas aos instantes de abertura da glote.

A Figura 8.31 mostra um segmento do sinal SF_PB submetido a uma TSM igual a 2,4. O aspecto mais importante a ser observado nesta Figura é o sincronismo temporal entre as componentes CEL e CER sintetizadas.

As Figuras 8.32 e 8.33 mostram novos segmentos do sinal SF_PB submetidos a uma TSM igual a 2,4. Novamente, os aspectos mais importantes a serem observados são as preservações dos contornos de amplitude e o sincronismo entre as componentes CEL e CER sintetizadas. Da Figura 8.33(c) pode-se observar que a operação de TSM foi capaz de concentrar os ruídos da componente CER nas regiões de abertura da glote.

A Figura 8.34 mostra um segmento do sinal SF_GER submetido a uma TSM igual a 1,6. Esta Figura ressalta a capacidade do algoritmo OPWI em lidar com variações bruscas na estrutura dos formantes do sinal de fala.

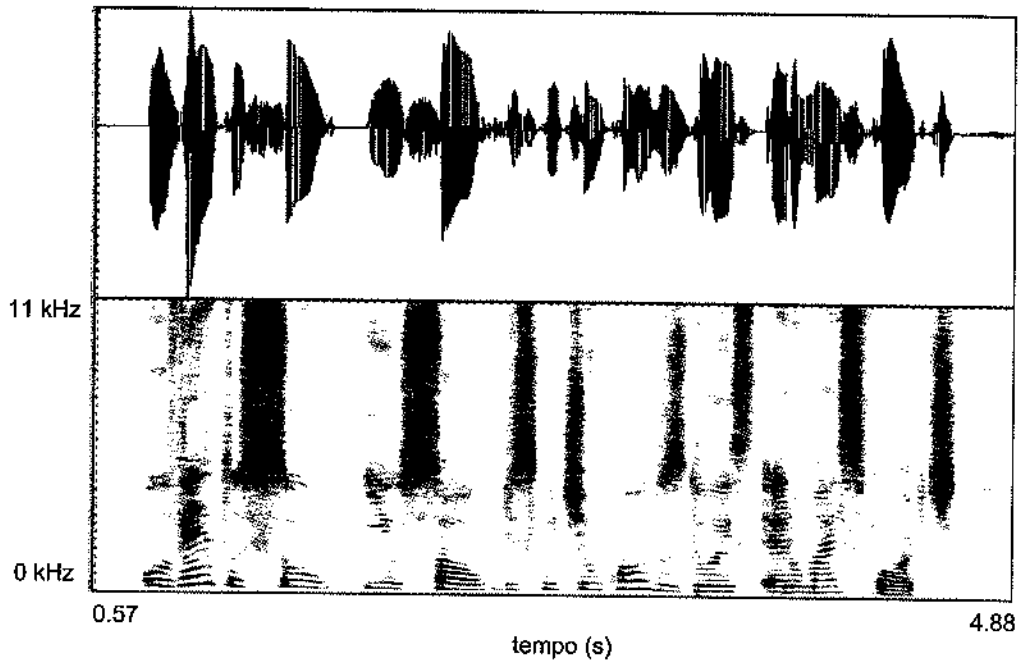


Figura. 8.27: Segmento original do sinal SF_PB no intervalo de 0,57 seg. a 4,88 seg.

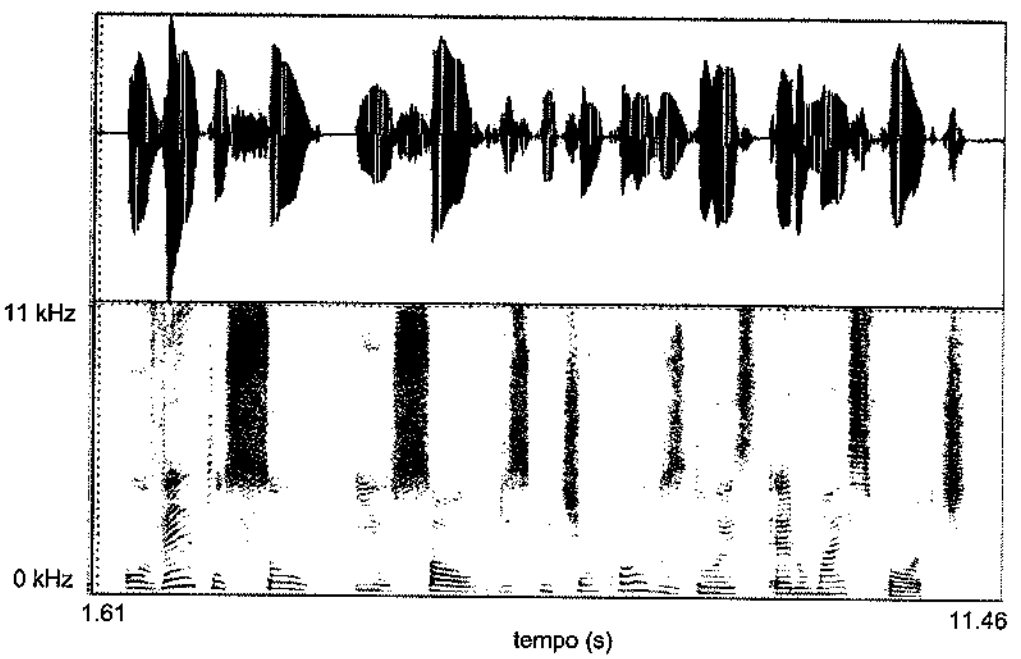


Figura. 8.28: Segmento da Figura 8.27 submetido a uma TSM igual a 2,4.

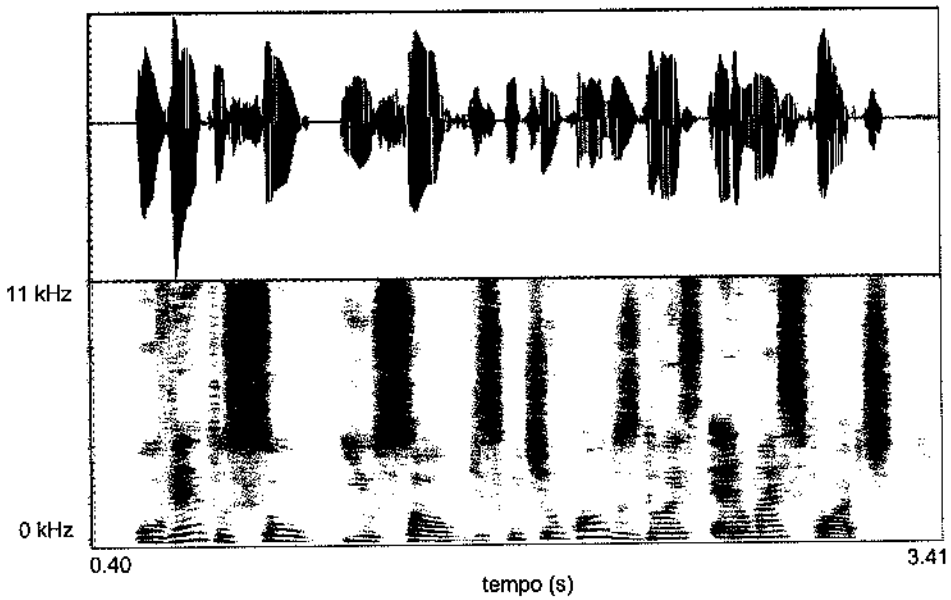


Figura. 8.29: Segmento da Figura 8.27 submetido a uma TSM igual a 0,7.

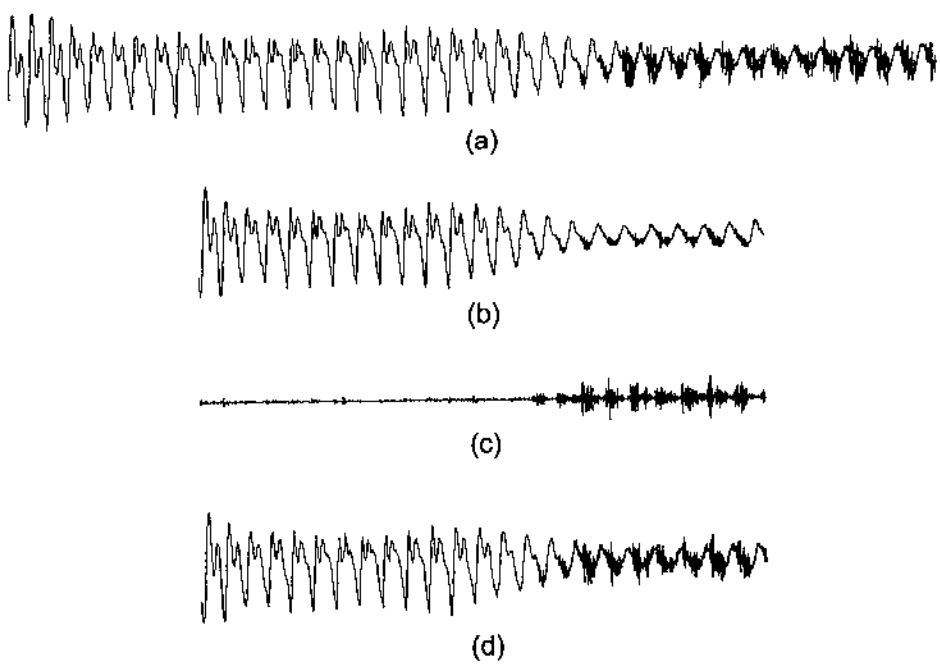


Figura. 8.30: Segmento do sinal SF_PB submetido a uma TSM igual a 0,7. (a) Sinal original; (b) Componente CEL sintetizada; (c) Componente CER sintetizada; (d) Sinal sintetizado (CEL + CER)

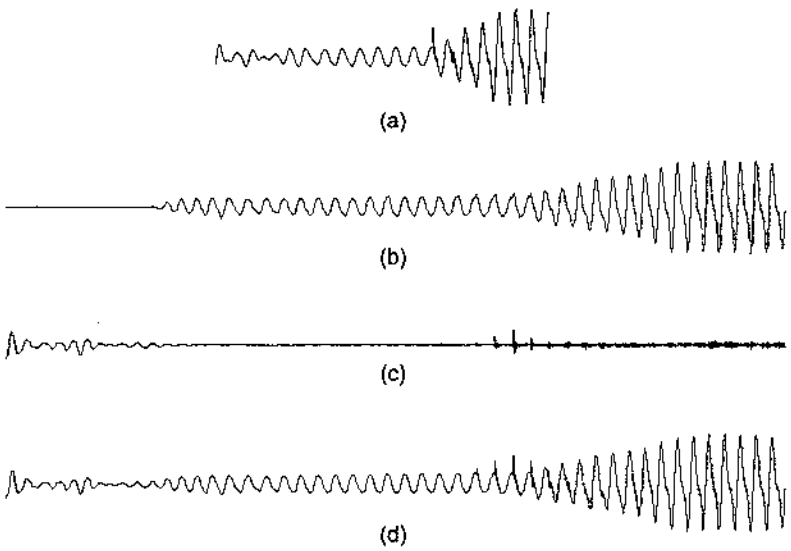


Figura. 8.31: Segmento do sinal SF_PB submetido a uma TSM igual a 2,4. (a) Sinal original; (b) Componente CEL sintetizada; (c) Componente CER sintetizada; (d) Sinal sintetizado (CEL + CER)

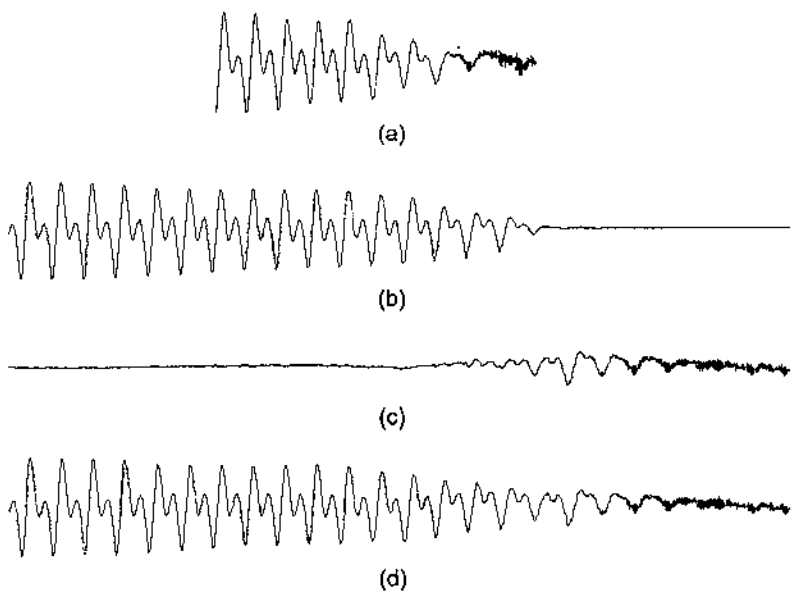


Figura. 8.32: Segmento do sinal SF_PB submetido a uma TSM igual a 2,4. (a) Sinal original; (b) Componente CEL sintetizada; (c) Componente CER sintentizada; (d) Sinal sintetizado (CEL + CER)

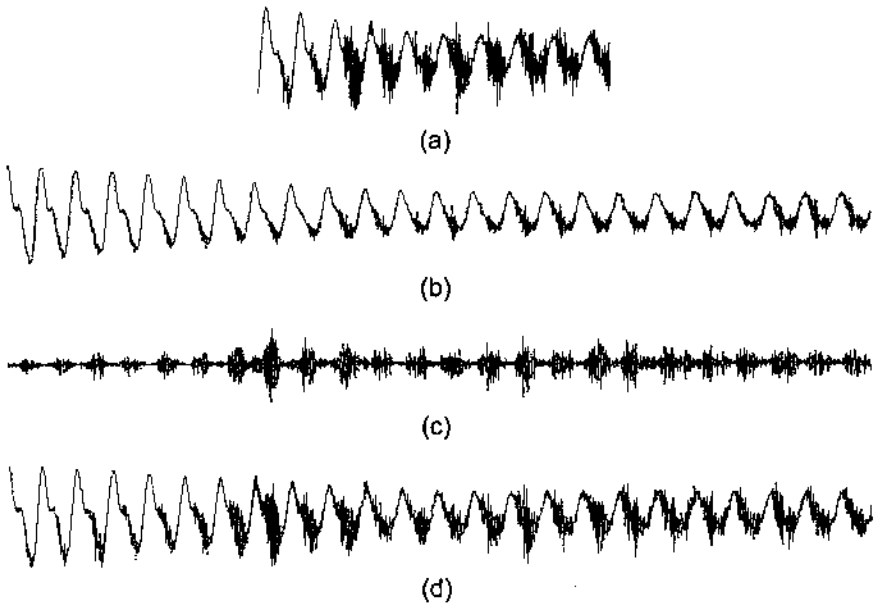


Figura. 8.33: Segmento do sinal SF_PB submetido a uma TSM igual a 2,4. (a) Sinal original; (b) Componente CEL sintetizada; (c) Componente CER sintentizada; (d) Sinal sintetizado (CEL + CER)

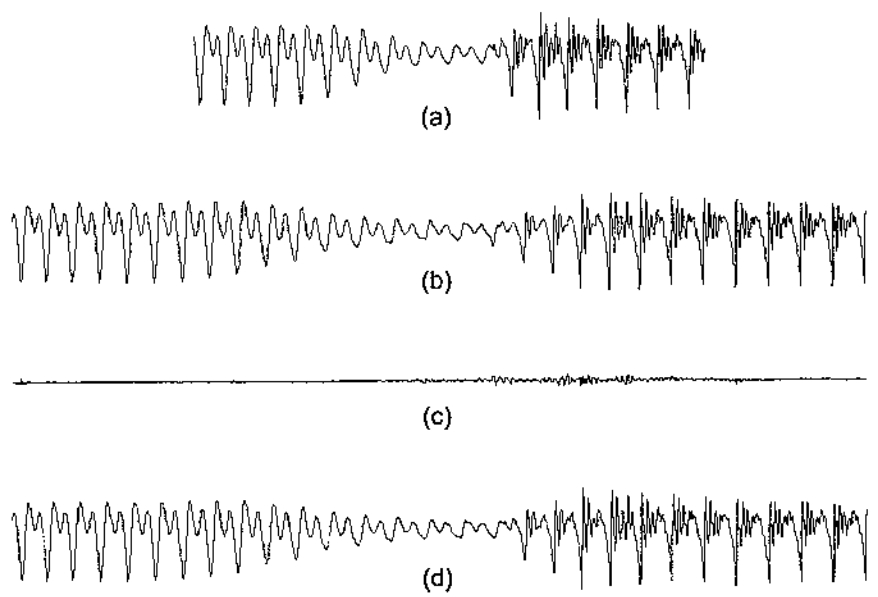


Figura. 8.34: Segmento do sinal SM_GER submetido a uma TSM = 1,6. (a) Sinal original; (b) Componente CEL sintetizada; (c) Componente CER sintentizada; (d) Sinal sintetizado (CEL + CER)

8.7.2 PSM por um Fator Constante

Análise Espectral

As Figuras 8.35 e 8.36 apresentam, respectivamente, uma redução na freqüência fundamental do sinal SF_PB por um fator multiplicativo igual a $\frac{1}{2,2}$ e um aumento na freqüência fundamental do sinal SF_PB por uma fator multiplicativo igual a $\frac{1}{0,7}$. Pode ser verificado das figuras que os processos de reamostragem dos protótipos, (interpolação por um fator de $2,2$ e dezimação por um fator de $0,7$), foram capazes de comprimir (Figura 8.37) e expandir (Figura 8.40) adequadamente a estrutura harmônica da componente CEL sem causar efeitos indesejáveis de compressão e/ou expansão em sua estrutura dos formantes. Além disso a interpolação destes protótipos reamostrados segundo a equação 7.62 foi capaz de gerar sinais sintetizados que preservam a estrutura dinâmica do sinal original sem a introdução de qualquer tipo de descontinuidade temporal indesejada.

Análise Temporal

A Figura 8.37 mostra um segmento do sinal SF_PB submetido a uma PSM igual a $\frac{1}{2,2}$. O processo de re-harmonização (reamostragem dos protótipos) é caracterizado ao longo do eixo dos tempos através de 4 fenômenos principais. Os dois primeiros fenômenos dizem respeito à preservação da estrutura formântica do sinal original e são caracterizados pela preservação dos valores de amplitude do sinal, próximos aos pulsos glotais (instantes de fechamento da glote) e pela não alteração das distâncias, ao longo do tempo, entre picos do sinal próximos aos pulsos glotais. Os dois últimos aspectos dizem respeito à redução da freqüência de pulsação das pregas vocais e são evidenciados através do aumento na distância entre os pulsos glotais e do aparecimento de uma região de baixa energia entre pulsos glotais consecutivos.

A Figura 8.38 ilustra dois aspectos importantes. Os segmentos não-sonoros da componente CER não são submetidos a qualquer tipo de PSM. Os segmentos ruidosos que ocorrem ao longo de segmentos sonoros da componente CER (por exemplo, segmentos fricativos sonoros) são adequadamente modificados para garantir o perfeito sincronismo entre as componentes CEL e CER.

A Figura 8.39 enfatiza a necessidade de sincronismo entre as componente CEL e CER. Pode ser verificado da Figura 8.39 que ruídos presentes na componente CER (provenientes de segmentos fricativos sonoros) são devidamente sintetizados para que estejam sincronizados com as regiões de abertura da glote, ao longo do sinal original.

A Figura 8.40 ilustra que as variações na componente DC dos espectros tomados a cada quadro de análise ao longo do sinal original se manifestam na componente CER e não na componente CEL (observar os instantes finais das Figuras 8.40(b) e 8.40(c)). Isto enfatiza, mais uma vez, a necessidade de perfeito sincronismo entre as componente CEL e CER.

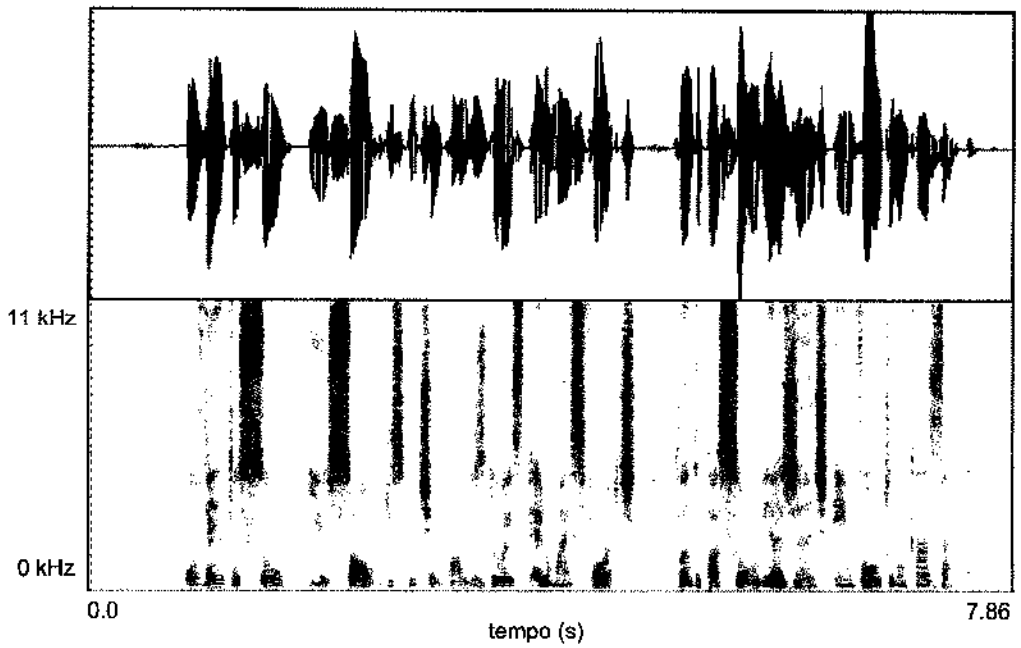


Figura. 8.35: Segmento do sinal SF_PB submetido a uma PSM igual a $\frac{1}{2,2}$. A estrutura harmônica se encontra altamente comprimida, porém sem nenhum prejuizo para a estrutura formântica do sinal.

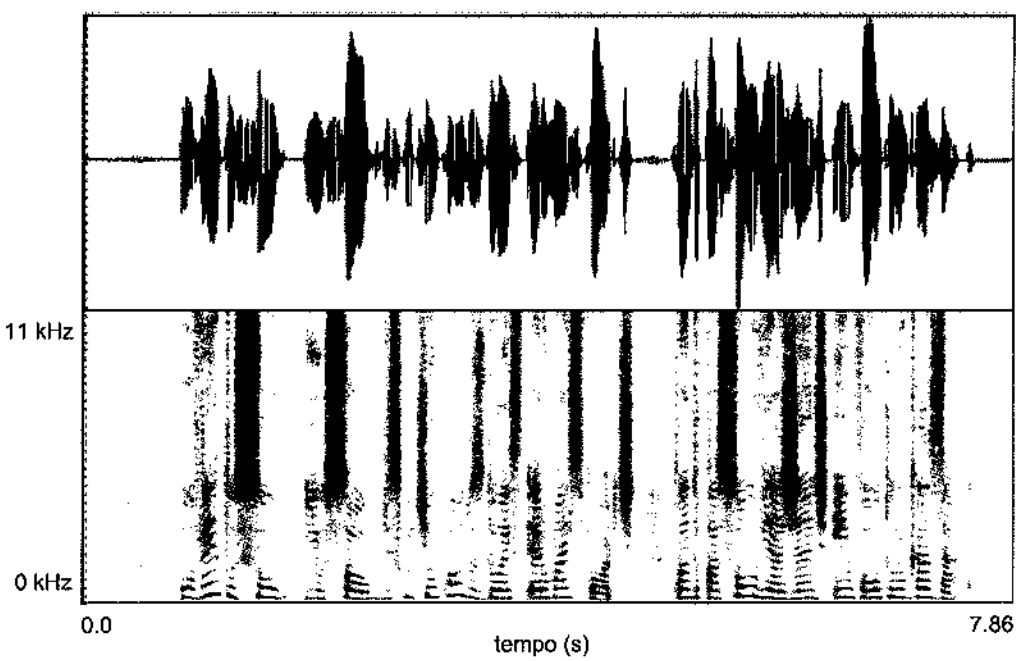


Figura. 8.36: Segmento do sinal SF_PB submetido a uma PSM igual a $\frac{1}{0,7}$. A estrutura harmônica se encontra expandida, porém a estrutura formântica do sinal encontra-se inalterada.

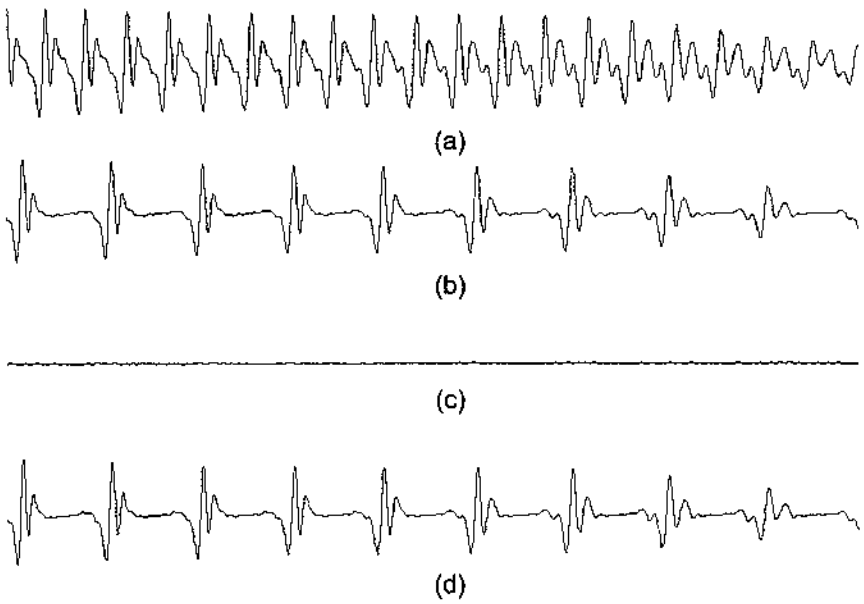


Figura. 8.37: Segmento do sinal SF_PB submetido a uma PSM igual a $\frac{1}{2,2}$. (a) sinal original, (b) componente CEL, (c) componente CER e (d) componentes CEL + CER.

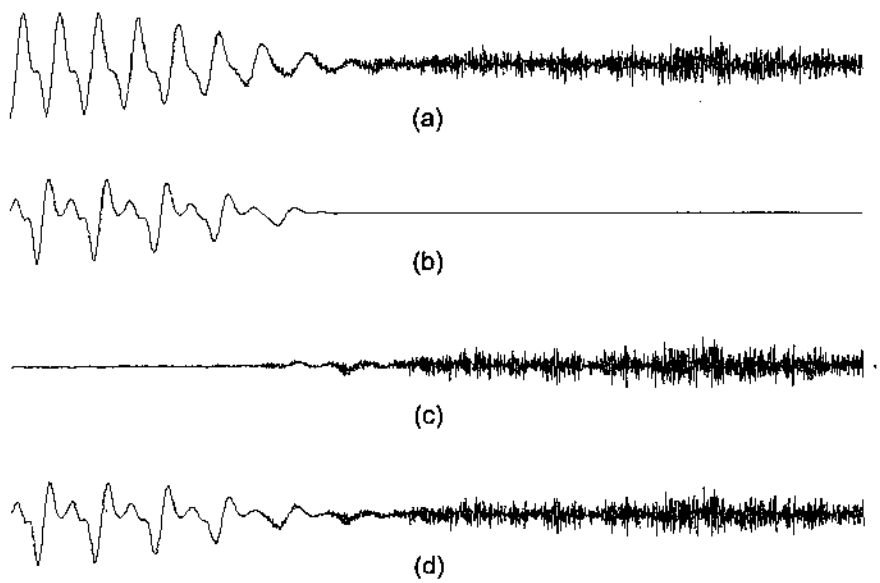


Figura. 8.38: Segmento do sinal SF_PB submetido a uma PSM igual a $\frac{1}{1,5}$. (a) sinal original, (b) componente CEL, (c) componente CER e (d) componentes CEL + CER.

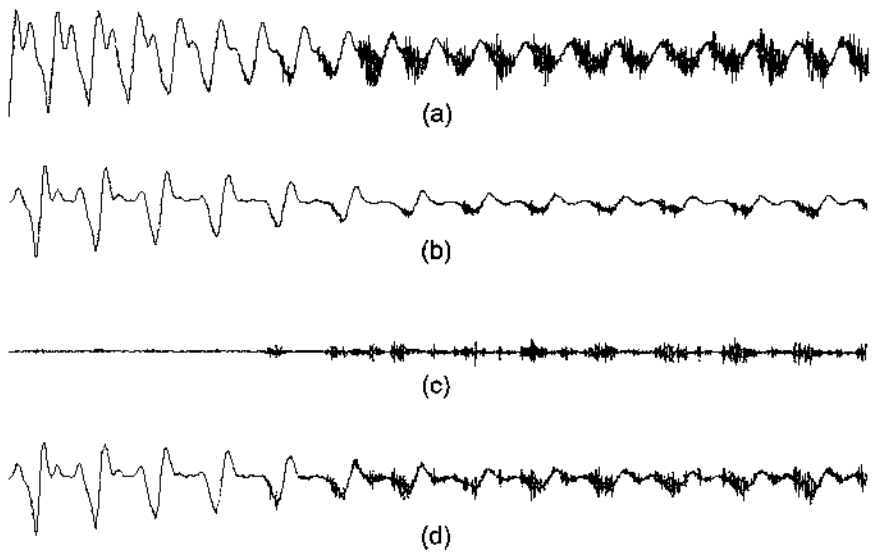


Figura. 8.39: Segmento do sinal SF_PB submetido a uma PSM igual a $\frac{1}{1,5}$. (a) sinal original, (b) componente CEL, (c) componente CER e (d) componentes CEL + CER. Destaque para o trecho fricativo sonoro.

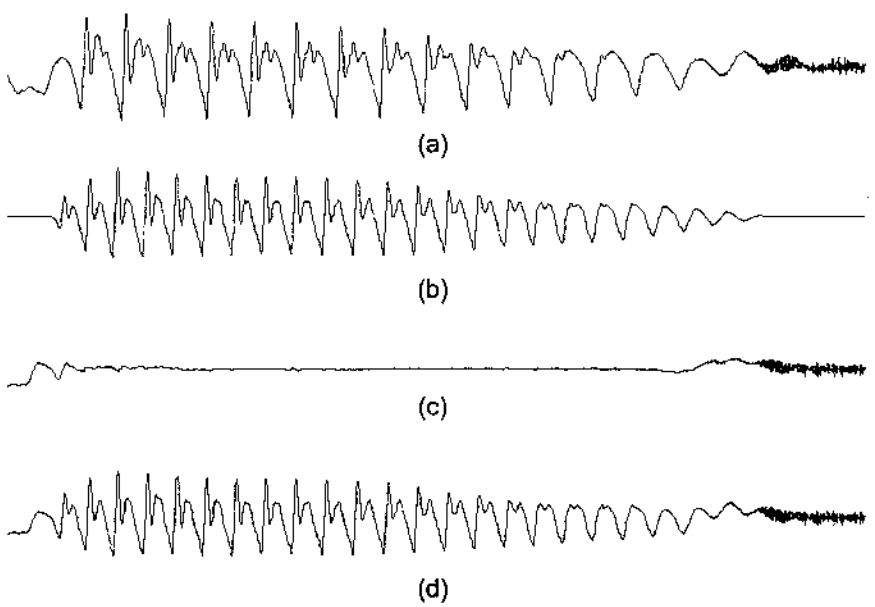


Figura. 8.40: Segmento do sinal SF_PB submetido a uma PSM $\frac{1}{0,7}$. (a) sinal original, (b) componente CEL, (c) componente CER e (d) componentes CEL + CER. Destaque para a presença de uma leve oscilação na componente DC, no final da componente CER.

8.7.3 PSM e TSM por Fatores Variáveis

As Figuras 8.41, 8.42 e 8.43 mostram o sinal SF_PB e suas componentes CEL e CER, respectivamente, quando submetidas a variações cossenoidais de freqüência fundamental dadas por:

$$F_0(m) = \frac{22050}{90 + 40 \cdot cos\left(10 \cdot \frac{2 \cdot \pi}{N_s} \cdot m\right)} \text{ Hertz} \tag{8.2}$$

para $m \in \{\cdots n_g, n_h, n_i, n_j, n_k \cdots\}$ (instantes de análise), e sendo N_s o número de total de amostras do sinal.

É importante observar que os espectogramas foram limitados à faixa de freqüências de 0 Hz a 4000 Hz para poder evidenciar as alterações na estrutura harmônica da componente CEL do sinal SF_PB ao longo do tempo.

As Figuras 8.44, 8.45 e 8.46 mostram os sinais SF_PB e suas componentes CEL e CER, respectivamente, quando submetidas a variações cossenoidais nas distâncias entre os instantes de análise. As novas distâncias, em amostras, entre os instantes de análise são dadas pela seguinte expressão:

$$D(m) = 110 + 40 \cdot cos\left(10 \cdot \frac{2 \cdot \pi}{N_s} \cdot m\right) \quad \text{amostras} \tag{8.3}$$

para $m \in \{\cdots n_g, n_h, n_i, n_j, n_k \cdots\}$ (instantes de análise), sendo N_s o número total de amostras do sinal. Por exemplo, para $m = n_i$, $D(n_i)$ corresponde a N^s_{ij}.

Pode ser observado das Figuras 8.44, 8.45 e 8.46 que a variação cossenoidal imposta à distância entre os instantes de análise ($D(m)$) não acarreta qualquer alteração no contorno de freqüência fundamental do sinal.

As Figuras 8.47, 8.48 e 8.49 mostram os sinais SF_PB e SM_GER quando submetidos simultaneamente a variações cossenoidas tanto nas distâncias entre os instantes de análise quanto no contorno de freqüência fundamental. As variações de duração e freqüência fundamental são dadas pelas equações 8.3 e 8.2, respectivamente.

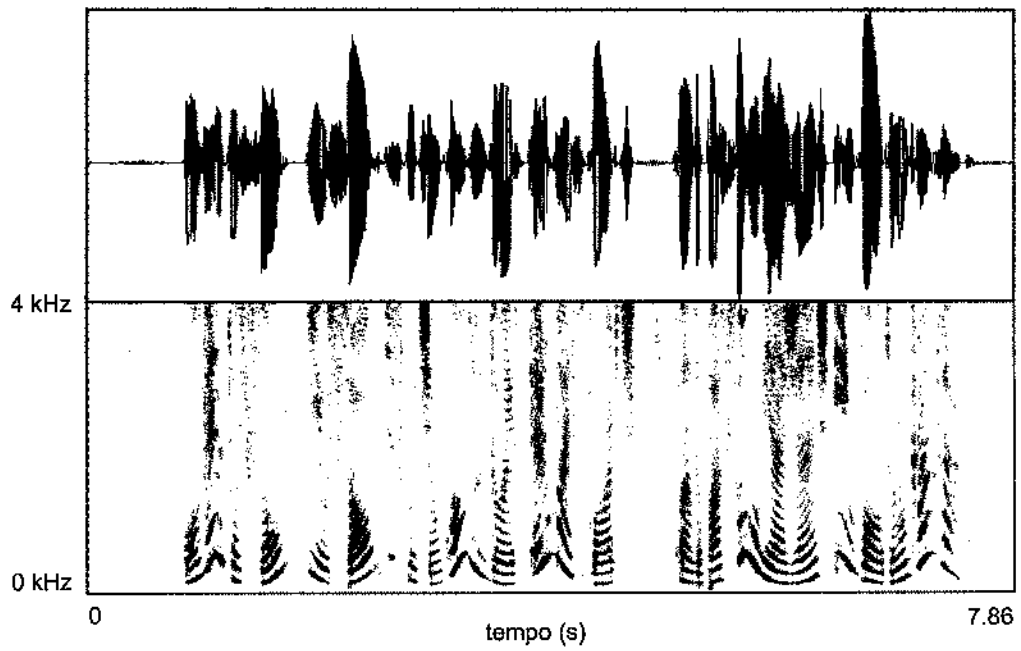


Figura. 8.41: Segmento do sinal SF_PB (CEL + CER) submetido a uma PSM com variação cossenoidal.

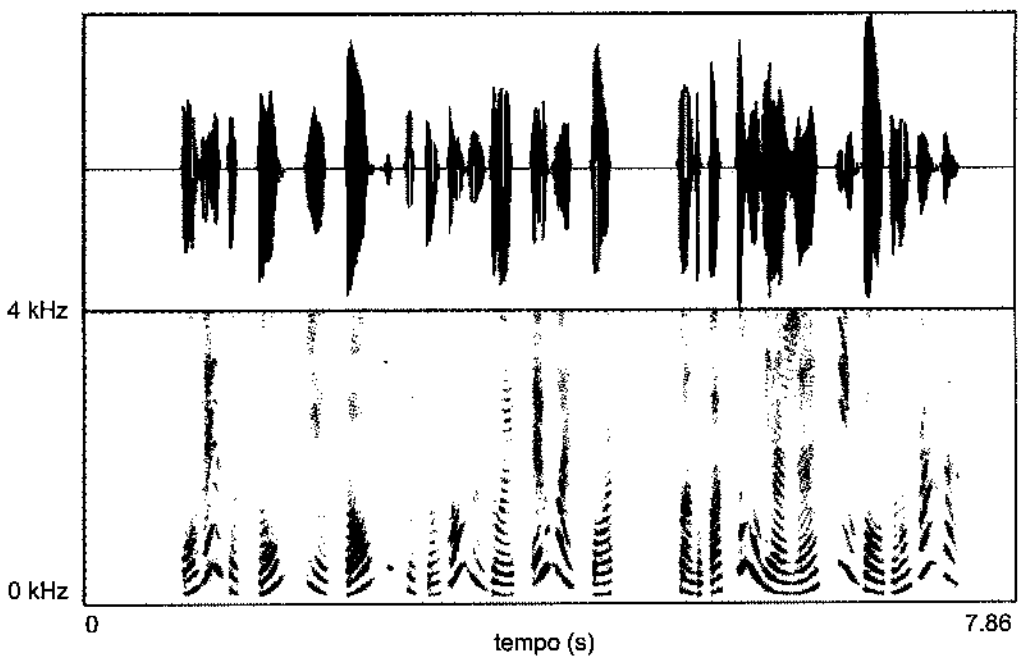


Figura. 8.42: Segmento da componente CEL do sinal SF_PB submetido a uma PSM com variação cossenoidal.

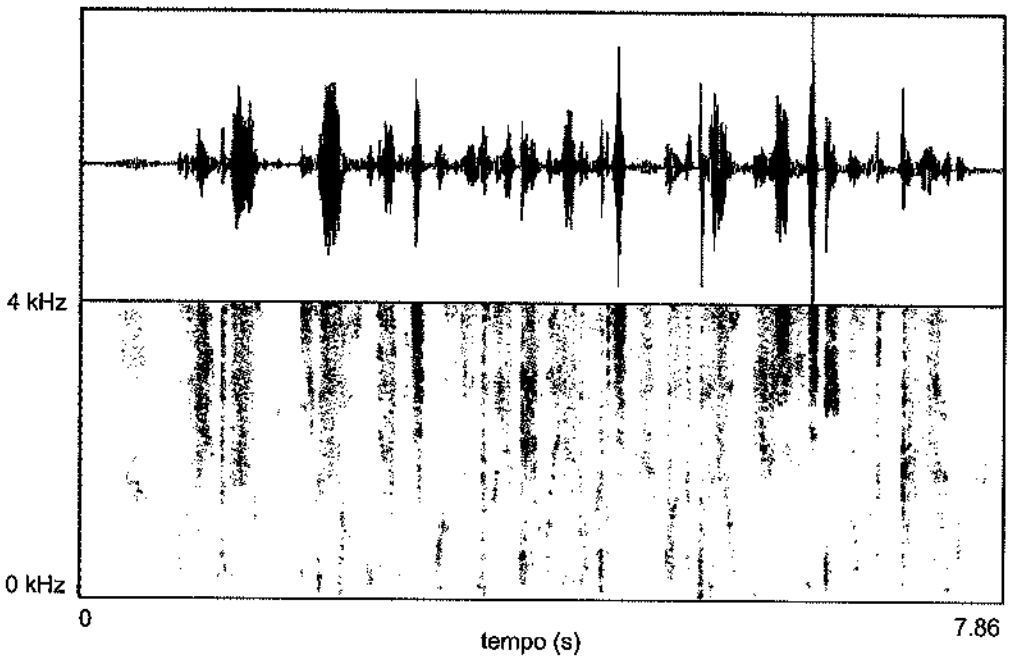


Figura. 8.43: Segmento da componente CER do sinal SF_PB submetido a uma PSM com variação cossenoidal.

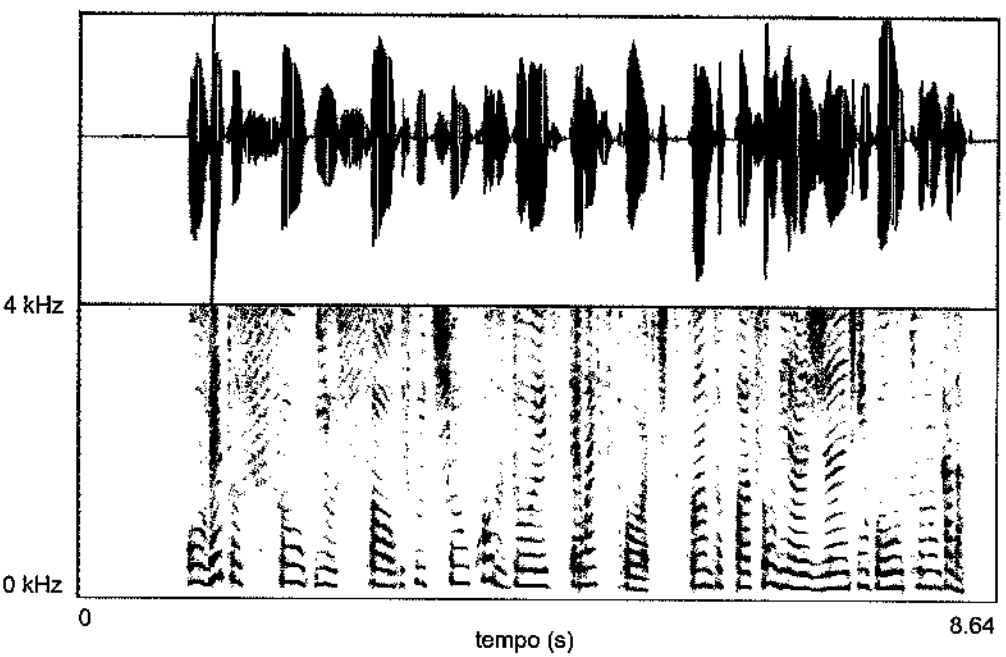


Figura. 8.44: Segmento do sinal SF_PB (CEL + CER) submetido a uma TSM com variação cossenoidal.

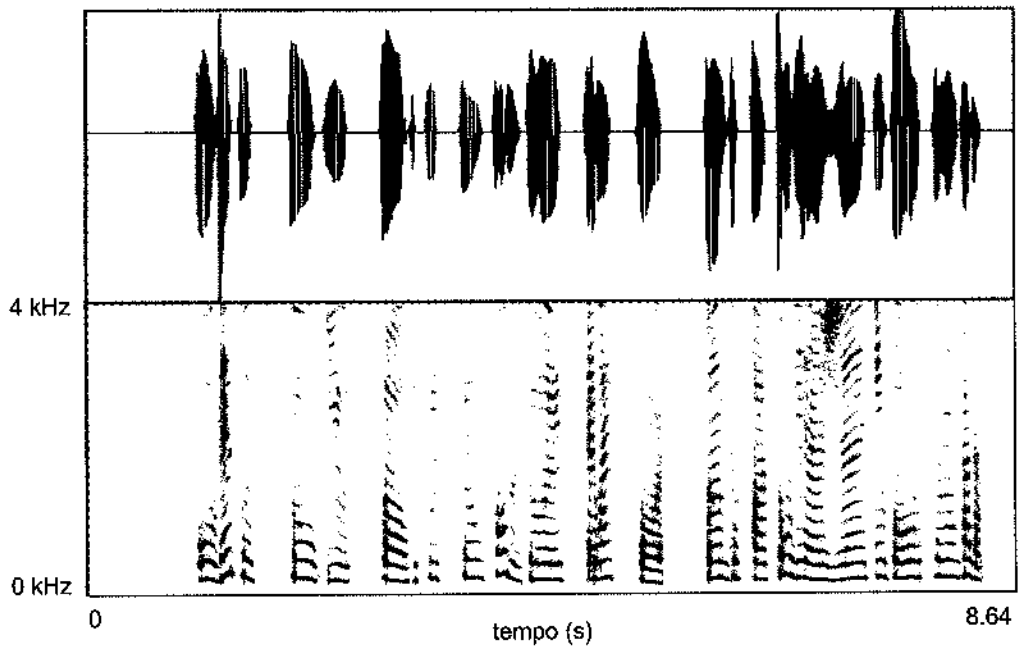


Figura. 8.45: Segmento da componente CEL do sinal SF_PB submetido a TSM com variação cossenoidal.

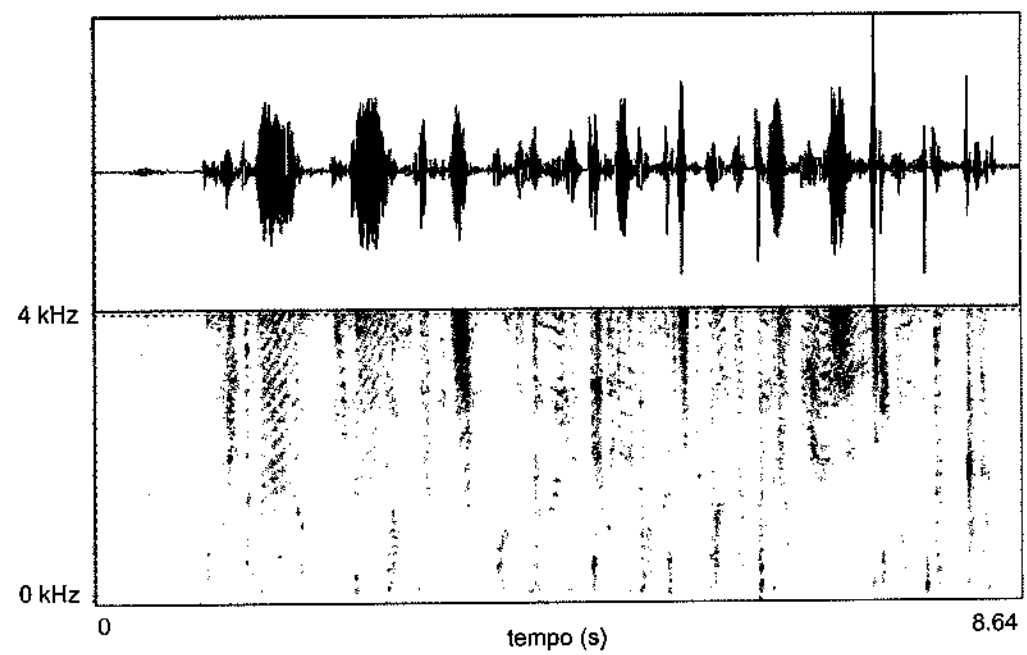


Figura. 8.46: Segmento da componente CER do sinal SF_PB submetido a TSM com variação cossenoidal.

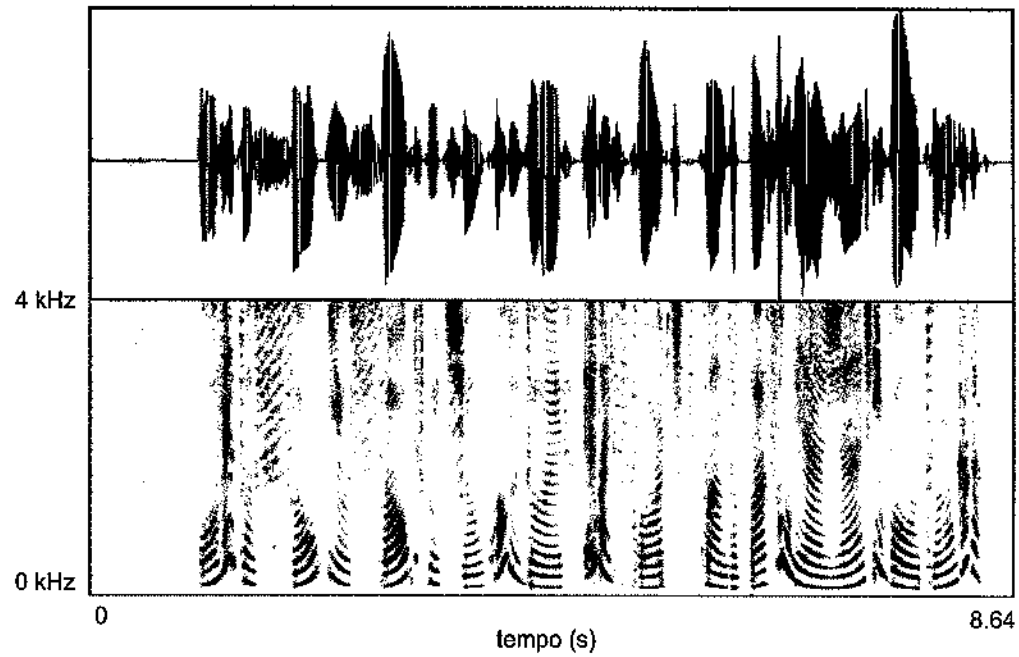


Figura. 8.47: Segmento do sinal SF_PB (CEL + CER) submetido a TSM e PSM com variação cossenoidal.

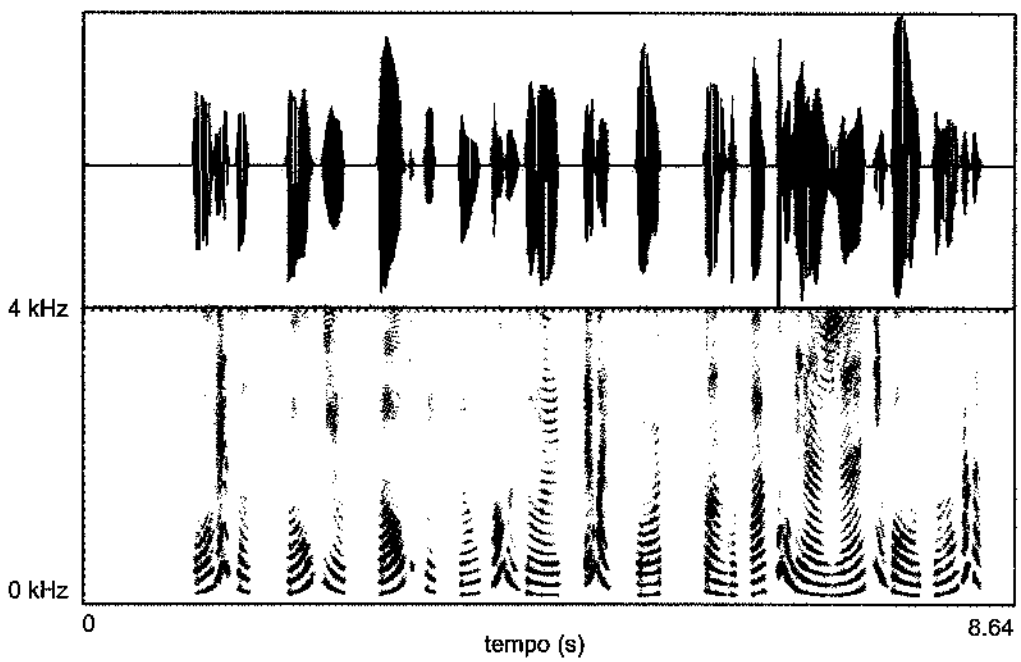


Figura. 8.48: Segmento da componente CEL do sinal SF_PB, submetido a TSM e PSM com variação cossenoidal.

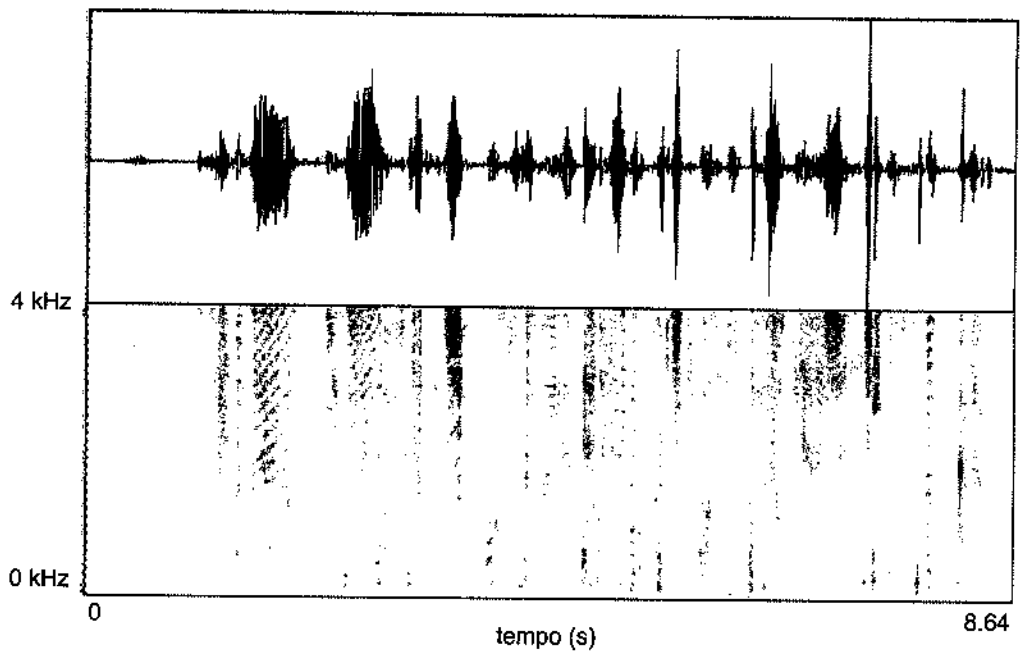


Figura. 8.49: Segmento da componente CER do sinal SF_PB, submetido a TSM e PSM com variação cossenoidal.

8.8 Exemplos para Avaliação Audível

O CD em anexo a esta Tese contém vários exemplos audíveis, resultantes de análises e sínteses dos sinais SF_PB, SM_PB, SF_US e SM_GER, utilizando-se o algoritmo OPWI. Estes exemplos podem ser utilizados para avaliar o desempenho do algoritmo OPWI quanto a aspectos auditivo-perceptivos. As subseções a seguir apresentam a descrição dos vários exemplos presentes no CD. O termo SINAL é utilizado para representar um dos 4 sinais, SF_PB, SM_PB, SF_US ou SM_GER. O termo COMPONENTE indica se o sinal corresponde a componente CEL, CER ou FULL (CER + CER).

8.8.1 Sinais Originais

- SINAL.wav

8.8.2 Sinais Sintetizados Sem Modificações Prosódicas

- SINAL_TSM_1-0_PSM_1-0_COMPONENTE.wav

8.8.3 Sinais Sintetizados Com Modificações Prosódicas de TSM

TSM igual a $2,4$

- SINAL_TSM_2-4_PSM_1-0_COMPONENTE.wav

TSM igual a $1,5$

- SINAL_TSM_1-5_PSM_1-0_COMPONENTE.wav

TSM igual a $0,7$

- SINAL_TSM_0-7_PSM_1-0_COMPONENTE.wav

Alteração de TSM por um fator cossenoidal

- SINAL_TSM_COS_PSM_1-0_COMPONENTE.wav

8.8.4 Sinais Sintetizados Com Modificações Prosódicas de PSM

PSM igual a $\frac{1}{2,2}$

- SINAL_TSM_1-0_PSM_0-454_COMPONENTE.wav

PSM igual a $\frac{1}{1,5}$

- SINAL_TSM_1-0_PSM_0-667_COMPONENTE.wav

PSM igual a $\frac{1}{0,7}$

- SINAL_TSM_1-0_PSM_1-428_COMPONENTE.wav

Alteração de PSM por um fator cossenoidal

- SINAL_TSM_1-0_PSM_COS_COMPONENTE.wav

8.8.5 Sinais Sintetizados Com TSM e PSM Cossenoidais

- SINAL_TSM_COS_PSM_COS_COMPONENTE.wav

8.9 Considerações Finais

Este Capítulo apresentou vários resultados experimentais e análises de desempenho dos processos de decomposição CEL/CER e de modificações prosódicas de PSM e TSM do algoritmo OPWI. Análises auditivo-perceptivas (ainda não rigorosas) realizadas por especialistas apontaram a alta qualidade das

operações de modificações prosódicas do algoritmo OPWI. Entretanto, testes mais rigorosos e melhorias no algoritmo OPWI ainda são necessárias. Como trabalhos futuros sugere-se:

- Realização de testes auditivo-perceptivos formais (MOS - *Mean Opinion Score*), para uma avaliação criteriosa do desempenho do algoritmo OPWI na realização de modificações prosódicas.
- Modelagem mais simples (porém eficiente) da componente CER.
- Avaliação de outros métodos para decomposição do sinal de fala em componente harmônica e componente ruidosa.
- Ajuste automático das constantes α_2, α_3 e β_3 das equações 7.8 e 7.9, para a determinação dos níveis de estacionariedade.
- Desenvolver o primeiro e o segundo método para estimativa dos protótipos ótimos diretamente no domínio do tempo, empregando as representações temporais dos protótipos ótimos.
- Realizar experimentos aplicando o algoritmo OPWI a sistemas CTF-SCAUS, com o objetivo de avaliar os processos de concantenação e suavização de unidades de síntese.

Capítulo 9

Conclusões

Este Capítulo conclui este trabalho apresentando inicialmente um sumário sobre as três principais contribuições desta Tese: (1) uma ampla revisão bibliográfica sobre sistemas CTF-SCAUS, (2) um novo algoritmo para análise exploratória de dados lingüísticos, algoritmo LDM-GA (*Linguistic Data Mining Using Genetic Algorithm*) e (3) um novo algoritmo de *Back-End*, algoritmo OPWI (*Optimized Prototype Waveform Interpolation*). Em seguida, são apresentadas algumas sugestões para melhorias de desempenho e avaliações mais criteriosas dos algoritmos LDM-GA e OPWI. Para encerrar, são traçadas algumas considerações finais sobre a área de conversão texto-fala.

9.1 Principais Contribuições

9.1.1 Ampla Revisão sobre Sistemas CTF-SCAUS

Os Capítulos 1, 2, 3 e 4 e as primeiras seções dos Capítulos 5 e 7 desta Tese apresentaram uma ampla revisão bibliográfica sobre os quatros principais módulos de um sistema CTF-SCAUS, módulo de *Front-End* (módulo lingüístico), módulo prosódico, módulo de seleção de unidades de síntese e módulo de *Back-End*.

O Capítulo 1 apresentou uma introdução geral sobre um sistema CTF-SCAUS, iniciando com uma descrição sobre a evolução destes sistemas na última década, passando por sua arquitetura e pelas principais técnicas estatísticas empregadas no seu treinamento, e finalizado com a descrição de alguns dos principais problemas enfrentados por estes sistemas.

O Capítulo 2 descreveu detalhes sobre o módulo de *Front-End*. Foram discutidos aspectos como estrutura de texto, normalização de texto, análise lingüística (análise sintática e semântica), análise fonética, análise morfológica, conversão grafema-fonema e projeto e compressão de léxico. Um destaque especial foi dado à construção de etiquetadores morfossintáticos empregando VMM (*Visible Markov Models*) e HMM (*Hidden Markov Models*).

O Capítulo 3 versou sobre o módulo prosódico abordando aspectos paralingüísticos (estilo de elocução), prosódicos (unidades entoacionais/segmentação prosódica) e fonéticos (modelagem de duração,

freqüência fundamental e contorno de intensidade). Discutiu-se brevemente as quatro principais classes de modelos entoacionais abordados na literatura, modelos baseados em tons de Pierrehumbert [pierrehumbert], modelos perceptuais [IPO], modelos superposicionais [Fujisaki] e modelos de estilização acústica [Fujisaki]. Destaques foram dados aos modelos entoacionais de Paul Taylor (*Tilt model of intonation*) [Taylor01] e Takehiko Kagoshima [Kagoshima01].

O Capítulo 4 apresentou o módulo de seleção automática de unidades de síntese. Fez uma discussão sobre as principais unidades de síntese utilizadas pela tecnologia SCAUS. Descreveu o processo de seleção de unidades em detalhes, dando ênfase à estimativa das funções de custo fonético-prosódico e de concatenação e também ao processo de busca empregando programação dinâmica, DTW (*Dynamic Time Warping*), ou algoritmo de Viterbi. Discutiu técnicas para clusterização e poda de unidades de síntese dando ênfase à definição de métricas para medir as distâncias acústico-prosódicas entre unidades de síntese. Detalhou técnicas para projeto, segmentação, etiquetagem e compressão de corpora de fala.

A seção 5.1 do Capítulo 5 discutiu os problemas associados a distribuições LNRE (*Large Number of Rare Events*) e a seção 5.2 formulou em detalhes o problema de predição da duração segmental da fala a partir de modelos de predição linear.

A seção 7.1 do Capítulo 7 apresentou uma breve revisão sobre os principais algoritmos empregados no módulo de *Back-End* de sistemas CTF-SCAUS: TD-PSOLA (*Time-Domain Pitch Synchronous Overlap and Add*), Modelos senoidais, Modelos baseados em análise preditiva linear (LPC) e HNM (Harmonic + Noise Models).

9.1.2 Algoritmo LDM-GA

Além de ter apresentado em detalhes a formulação teórica do algortimo LDM-GA (Capítulo 5), este trabalho também realizou vários experimentos (Capítulo 6) para avaliar o desempenho deste algoritmo na otimização de modelos de regressão linear a partir de variáveis nominais (modelos QMTI) aplicados à modelagem da duração segmental da fala. Estes resultados mostraram a eficiência do algoritmo LDM-GA no processo de determinação do número "ótimo" de modelos QMTI a serem utilizados (processo de clusterização de fones) e na seleção de quais fatores lingüísticos devem fazer parte dos modelos QMTI para os fones clusterizados (processo de determinação das topologias ótimas dos modelos).

Para avaliar o desempenho do algoritmo LDM-GA na determinação de topologias ótimas dos modelos QMTI/Ph (modelos QMTI por fone), vários experimentos foram realizados comparando os modelos QMTI/Ph + LDM-GA (modelos QMTI/Ph com topologias selecionadas pelo algoritmo LDM-GA) com modelos QMTI/Ph Cheios (modelos QMTI/Ph utilizando todos os fatores lingüísticos), com modelos QMTI/Ph + ANOVA (modelos QMTI/Ph com topologias selecionadas pelo método ANOVA) e com modelos RT/Ph (modelos estimados a partir de árvores de regressão - RT). Os resultados obtidos mostraram que:

Os modelos QMTI/Ph + LDM-GA apresentam uma capacidade de generalização significativamente superior aos modelos QMTI/Ph Cheios.

Os modelos QMTI/Ph + LDM-GA apresentam uma capacidade de generalização ligeiramente superior aos modelos QMTI/Ph + ANOVA.
Os modelos QMTI/Ph + LDM-GA apresentam uma melhor capacidade de generalização que as árvores de regressão. Este resultado mostra que, apesar das árvores de regressão serem métodos não-lineares (porém, sendo linear por partes), elas não apresentam um boa capacidade de generalização, mesmo quando submetidas a técnicas de *pruning*.

Avaliações do processo de clusterização do algoritmo LDM-GA mostraram que o método de clusterização hierárquica binária do algoritmo LDM-GA se mostrou eficiente na seleção de classes de fones que maximizam o desempenho dos modelos QMTI.

Apesar de os modelos QMTI terem se mostrado úteis na análise e validação do algoritmo LDM-GA, eles se mostraram demasiadamente simples para modelar adequadamente a duração segmental da fala. Muito provavelmente, esta limitação dos modelos QMTI deve-se ao fato de eles não contemplarem interações entre fatores lingüísticos. O desempenho dos modelos QMTI/Ph + LDM-GA para alguns fones pode ser observado nas Figuras 6.28, 6.29, 6.30, 6.31, 6.32, 6.33, 6.34, 6.35, 6.36.

9.1.3 Algoritmo OPWI

Além de ter apresentado uma formulação teórica detalhada para o algoritmo OPWI, vários experimentos foram realizados para analisar a eficiência deste algoritmo na análise e ressíntese do sinal de fala com modificações prosódicas de TSM (*Time Scale Modifications*) e PSM (*Pitch Scale Modifications*). Estes experimentos comprovaram o bom desempenho do algoritmo OPWI na realização de modificações prosódicas. Entre as principais características do algoritmo OPWI, destacam-se:

- O algoritmo OPWI não faz uso do conceito de *maximum voiced frequency* utilizado pelo método HNM (Stylianou, 1996). Além disso, diferentemente do método HNM, as componentes "harmônica" (componente CEL) e ruidosa (componente CER) do algoritmo OPWI se estendem ao longo de toda a banda de freqüencia do espectro.
- A estimativa dos protótipos ótimos é realizada utilizando-se dois métodos com diferentes resoluções tempo/frequência, de acordo com o nível de estacionaridade do sinal.

  - No primeiro método, os protótipos são otimizados levando-se em consideração, explicitamente, que eles serão interpolados segundo funções de interpolação específicas. Apesar de este método reduzir a resolução temporal dos protótipos, ele garante ressínteses e modificações prosódicas de excelente qualidade.
  - O segundo método é semelhante ao proposto por (Stylianou, 1996) para estimar os parâmetros da componente harmônica de seu modelo HNM. Apesar deste método não levar em consideração o processo de interpolação dos protótipos, ele garante ressínteses e modificações prosódicas de boa qualidade e, além disso, apresenta uma alta resolução temporal, correspondente a apenas dois períodos de *pitch*.

- O algoritmo OPWI opera sincronamente com os pulsos glotais (mais especificamente com os Instantes de Fechamento da Glote - IFG) . Esta operação síncrona com os IFGs garante alta qualidade nas modificações prosódicas e facilita a suavização espectral entre unidades de síntese.
- Como o algoritmo OPWI opera com períodos fundamentais (T_0) inteiros, então algoritmos rápidos que trabalham com funções trigonométricas previamente calculadas (conforme proposto na seção 7.14.2), podem ser utilizados na etapa de síntese do algoritmo.

9.2 Sugestões para Trabalhos Futuros

9.2.1 Algoritmo LDM-GA

- Repetir os experimentos do Capítulo 6 para o português brasileiro.
- Incluir interações entre os fatores lingüísticos.
- Adaptar o algoritmo LDM-GA para operar com outros modelos de regressão como SoP, ANN.
- Adaptar o algoritmo LDM-GA para auxiliar na estimativa das funções de custo do processo de seleção de unidades de síntese de sistemas CTF-SCAUS.
- Adaptar o algoritmo LDM-GA para auxiliar na estimativa de modelos para estilização do contorno de F_0, como, por exemplo o modelo de Tilt de Paul Taylor (Taylor, 2000).

9.2.2 Algoritmo OPWI

- Integrá-lo a um sistema CTF-SCAUS para avaliá-lo quanto sua capacidade de suavização espectral na fronteira entre unidades de síntese.
- Submetê-lo a avaliações subjetivas formais, através de MOS (*Mean Opinion Score*).
- Utilizá-lo no processo de seleção e fusão proposto por Kagoshima, (Mizutani and Kagoshima, 2005).
- Analisar métodos alternativos para realizar a decomposição do sinal de fala em componente harmônica e componente ruidosa, como por exemplo o método proposto em (Yegnanarayana et al., 1998).
- Desenvolver procedimentos automáticos para ajustar a constante ξ da equação 7.2.
- Desenvolver procedimentos automáticos para ajustar as constantes α_2 da equação 7.8 e α_3 e β_3 da equação 7.9.
- Analisar a possibilidade de desenvolver os Métodos I e II para estimativa dos protótipos ótimos, diretamente no domínio do tempo.
- Modelar/sintetizar a componente CER empregando um método com menor custo computacional, como por exemplo, o algoritmo TD-PSOLA.
- Utilizar sua capacidade para realizar TSM e PSM para melhorar o desempenho do processo de clusterização de unidades dos sistemas CTF-SCAUS (várias das métricas utilizadas no processo

de clusterização de unidades de síntese requerem que as unidades sejam normalizadas em duração e em frequência fundamental).

9.3 Considerações Finais

Se por um lado a tecnologia SCAUS permitiu grandes avanços na qualidade da fala produzida por sistemas CTF, ela representou uma certa estagnação nas pesquisas fundamentais sobre os processos de produção e percepção da fala. A tecnologia SCAUS é uma tecnologia não-paramétrica e sintetiza fala através de um processo de busca ao longo de extensas bases de unidades de síntese previamente gravadas e etiquetadas lingüisticamente. Nenhum dos processos biomecânicos envolvidos na produção da fala são modelados explicitamente pela tecnologia SCAUS. Além disso, vários dos métodos estatísticos empregados na modelagem dos sistemas CTF-SCAUS não permitem interpretações lingüísticas detalhadas dos processos envolvidos. Isto não significa que a tecnologia SCAUS não venha contribuído para estudos lingüísticos e de produção e percepção da fala. Vários trabalhos importantes foram realizados nas áreas de modelagem prosódica, avaliações perceptuais e processamento digital de sinais no contexto de sistemas CTF-SCAUS. Entretanto, é importante ter consciência de que esta tecnologia ainda não resolveu por completo o problema de conversão-texto-fala e muito provavelmente não será capaz de resolver. Portanto, pesquisas básicas sobre os processo de produção e percepção da fala e da linguagem ainda continuam, e provavelmente continuarão por muito tempo a serem de fundamental importância para a área de conversão texto-fala.

A convergência entre as áreas de conversão texto-fala e reconhecimento de fala é um outro aspecto muito importante do atual cenário do estado-da-arte da área de ciência e tecnologia da fala e da linguagem. A nova tecnologia proposta por Keichi Tokuda (Tokuda et al., 2002), (Yoshimura et al., 1999), (Shichiri et al., 2002), (Tokuda et al., 1999) é um bom exemplo disto. Esta tecnologia produz fala sintetizada utilizando um princípio muito similar ao utilizado nos modernos sitemas de reconhecimento de fala. A tecnologia de Tokuda é muito versátil, modela os processos acústicos e prosódicos simultaneamente, possui um reduzido *footprint*, permite a alteração de vozes sem a necessidade da gravação de novos corpora e a cada ano tem melhorado significativamente a qualidade vocal de sua fala sintetizada.

Apesar de as Universidades brasileiras terem iniciado seus trabalhos nas áreas de CTF e ASR no inicio na década de 90, pode-se considerar que o conhecimento/domínio tecnológico brasileiro nestas áreas ainda é relativamente incipiente. Poucos são os grupos de pesquisa que dispõem de corpora de fala para o português brasileiro devidamente projetados, gravados, segmentados e etiquetados (tanto para sistemas CTF quanto ASR). Além disso a área de tecnologia da fala é altamente interdisciplinar envolvendo aspectos de engenharia, ciência da computação e lingüística, e poucos são os grupos brasileiros atuando nesta área que apresentam tal característica. Espero que esta Tese, bem como os frutos que dela poderão surgir, possam de alguma forma contribuir para o avanço da área de ciência e tecnologia da fala no Brasil.

Referências Bibliográficas

J. Adell and A. Bonafonte. Towards phone segmentation for concatenative speech synthesis. In *5th ISCA Speech Synthesis Workshop*, Pittsburgh, USA, 2004.

M. Akamine and T. Kagoshima. Analytic generation of synthesis units by closed loop training for totally speaker driven text to speech (TOS driven TTS). In *ICSLP*, pages 1927–1930, Sidney, Australia, 1998.

M. Atterer. Assigning prosodic structure for speech synthesis: A rule-based approach. In *First International Conference on Speech Prosody*, Aix-en-Provence, Laboratoire Parole et Language, France, 2002.

M. Atterer and E. Klein. Integrating linguistic and performance-based constraints for assigning phrase breaks. In *COLING*, Taipei, Tawan, 2002.

G. Bailly, N. Campbell, and B. Möbius. ISCA special session: Hot topics in speech synthesis. In *Eurospeech'2003*, pages 37–40, Geneva, Switzerland, 2003.

P. A. Barbosa. Explaining Brazilian Portuguese resistance to stress shift with a coupled-oscillator model of speech rhythm production. *Cadernos de Estudos Lingüísticos, Unicamp*, 43:71–92, July/Dec 2002.

P. A. Barbosa. Caractérisation et génération automatique de la structuration rythmique du français. Thése de doctorat, INPG/ICP, Université Stendhal, Grenoble, France, 1994.

P. A. Barbosa. A dynamical model for generating prosodic structure. In *3rd International Conference on Speech Prosody*, Dresden, Germany, May 2006.

J. Bellegarda and K. Silverman. Statistical prosodic modeling: From corpus design to parameter estimation. *IEEE Transaction on Speech and Audio Processing*, 9(1):52–66, 2001.

M. Beutnagel, A. Conke, J. Schroeter, Y. Stylianou, and A. Syrdal. The AT&T next-generation TTS system. In *Joint Meeting of ASA, EAA and DAGA*, Berlin, Germany, 1999a.

M. Beutnagel, M. Mohri, and M. Riley. Rapid unit selection from a large speech corpus for concatenative speech synthesis. In *Eurospeech*, pages 607–610. ISCA, 1999b.

C. M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, New York, USA, 1995.

A. Black. Festivox. Página na internet, Carnigie Melon University, http//www.festivox.org, Junho 2006.

A. Black. Chart, version 0.8, a generic speech synthesizer. Technical report, ATR-Interpreting Telecommunications Laboratories, Japan, March 1996.

A. Black and P. Taylor. Automatic clustering similar units for unit selection in speech synthesis. In *Eurospeech*, Rhodes, Greece, September 1997.

P. Boersma and D. Weenink. *Praat: Doing Phonetics by Computer.* http://www.praat.org, 2005.

A. Botinis, B. Granstrom, and B. Möbius. Developments and paradigms in intonation research. *Speech Communication*, 33:263–296, 2001.

L. Breiman. Bagging predictors. Technical report number 421, Department of Statistics, University of California, Berkeley, California, USA, September 1994.

L. Breiman, J. H. Friedman, and C. J. Stone. *Classification and Regression Trees.* Chapman & Hall, New York, 1993.

I. Bulyko. Flexible speech synthesis using weighted finite state transducers. Phd thesis, University of Washington, Seatle, USA, March 2002.

T. Bäck, D. B. Fogel, and Z. Michalewicz. *Evolutionary Computation 1: Basic Algorithms and Operators.* Institute of Physics Publishing, Philadelphia, PA, USA, 2000a.

T. Bäck, D. B. Fogel, and Z. Michalewicz. *Evolutionary Computation 2: Advanced Algorithms and Operators.* Institute of Physics Publishing, Philadelphia, PA, USA, 2000b.

D. Chappell and J. H. L. Hansen. Spectral smoothing for concatenative speech synthesis. In *ICSLP*, pages 1935–1938, Sydney, Australia, December 1998.

R. J. Cirigliano, C. Monteiro, F. Leandro, F. Barbosa, F. G. V. Resende, L. R. Couto, and J. A. Moraes. Um conjunto de 1000 frases foneticamente balanceadas para o português brasileiro obtido utilizando a abordagem de algoritmos genéticos. In *Simpósio Brasileiro de Telecomunicações*, Campinas, SP, September 2005. SBrT.

G. Coorman, J. Fackrell, P. Rutten, and B. Van Coile. Segment selection in the L&H realspeak laboratory TTS system. In *ICSLP*, pages 395–398, 2000.

T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience, New York, USA, 1991.

R. Donovan. Trainable speech synthesis. Phd thesis, Cambridge University, Cambridge, UK, 1996.

R. Donovan. A new distance measure for costing spectral discontinuities in concatenative speech synthesizers. In *Proc. 4th ISCA Tutorial and Research Workshop on Speech Synthesis*, Scotland, 2001.

R. E. Donovan. Topics in decision tree based speech synthesis. *Computer Speech and Language*, 17: 43–67, 2003.

R. E. Donovan. Segment pre-selection in decision-tree based speech synthesis. In *ICASSP*, Istanbul, Turkey, 2000.

N. Duffy and D. P. Helmbold. Leveraging for regression. In *13th COLT - Conference on Computational Learning Learning Theory*, pages 208–219, San Francisco, USA, 2000.

K. Dusterhoff and A. Black. Generating f0 contours for speech synthesis using the tilt intonation theory. In *Workshop on Intonation*, pages 867–870, Atenas, Grecia, 1997.

T. Dutoit. *An Introduction To Text-To-Speech Synthesis.* Kluwer Academic Publishers, Dordrecht, The Netherlands, 1997.

T. Dutoit and H. Leich. MBR-PSOLA: Text-to-speech synthesis based on a mbe re-synthesis of segments database. *Speech Communication*, 13:435–440, 1993.

M. Edgington, A. Lowry, P. Jackson, A. P. Breen, and S. Minnis. *Speech Technology for Telecommunications*, chapter Overview of Current Text-To-Speech Techniques - Part II: Prosody and Speech Generation, pages 181–210. J. F. A. Westall and A. Lewis (Eds.). Chapman & Hall, London, UK, 1998.

E. Eide. Improvements to the IBM trainable speech synthesis system. In *ICSLP*, Hong Kong, China, 2003.

H. Fujisaki. *Vocal Fold Physiology: Voice Production, Mechanisms and Functions*, chapter A Note on the Physiological and Physical Basis for the Phrase and Accent Components in the Voice Fundamental Frequency Contour, pages 347–355. In Osamu Fujimura (Ed.). Raven, New York, USA, 1988.

H. Fujisaki. The role of quantitative modeling in the study of intonation. In *Proceedings of the International Symposium on Japanese Prosody*, pages 163–174, 1992.

J. D. Gibbons. *Nonparametric Statistical Inference.* M. Dekker, 2nd edition, 1985.

D. Griffin and J. Lim. Multiband-excitation vocoder. *Transactions on Acoustic, Speech and Signal Processing*, 36:236–243, February 1988.

C. Hansa and Y. Sagisaka. Analysis of segmental duration for thai speech synthesis. In *Speech Prosody*, Nara, Japan, March 2004. ISCA.

C. Hayashi. On the quantification of qualitative data from the mathematico-statistical point of view. In *Annals of the Institute of Statistical Mathematics*, 1950.

S. Haykin. *Neural Networks: A comprehensive Foundation.* Prentice Hall, New Jersey, 1994.

T. Hirai, N. Iwahashi, N. Higuchi, and Y. Sagisaka. *Progress In Speech Synthesis*, chapter Automatic Extraction of F0 Control Rules Using Statistical Analysis, pages 333–346. J. V. Santen, R. W. Sproat, J. P. Olive and J. Hirschberg (Eds.). Springer-Verlag, 1996.

H. Hon, A. Acero, X. Huang, J. Liu, and M. Plumpe. Automatic generation of synthesis units for trainable text-to-speech systems. In *ICASSP*, pages 293–296, 1998.

X. Huang and A. Acero. Recent improvements on Microsoft's trainable text-to-speech system - Whistler. In *ICASSP'98*, Seattle, Washington, USA, May 1998.

X. Huang, A. Acero, and H. W. Hon. *Spoken Language Processing.* Prentice Hall - PTR, Upper Saddle River, New Jersey, USA, 2001.

A. J. Hunt and A. Black. Unit selection in a concatenative speech synthesis system using a large speech database. In *ICASSP*, pages 373–376, Atlanta, USA, 1996.

F. Jelinek. *Statistical Methods for Speech Recognition.* MIT Press, Cambridge, MA, 1997.

J. D. Jobson. *Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design.* Springer-Verlag, New York, USA, 1991.

T. Kagoshima, M. Morita, and S. Seto. An f0 contour model for totally speaker driven text to speech system. In *ICSLP*, Sidney, Australia, 1998.

D. Kapilow, Y. Stylianou, and J. Schroeter. Detection of non-stationarity in speech signal and its application to time-scaling. In *6th European Conference on Speech Communication and Technology*, pages 2307–2310, Budapest, Hungary, September 1999.

B. Z. Keller. *Volume on Speech Synthesis*, chapter Prediction of Temporal Structures for Various Speech Rates. N. Campbell (Ed.). Springer-Verlag, Forthcoming.

F. N. Kepler. Um etiquetador morfo-sintático baseado em cadeias de markov de tamanho variável (tese de mestrado). Tese de mestrado, IME-USP - Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo, Brasil, 2005.

E. Klabbers and R. Veldhuis. Reducing audible spectral discontinuities. *IEEE Transaction on Speech and Audio Processing*, 9:39–51, 2001.

D. Klatt. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 82(3):737–793, 1987.

B. Kleijn and K. Paliwal. *Speech Coding and Synthesis*. Elsevier, Amesterdam, The Netherlands, 1998.

K. Knill, E. Morais, P. Jackson, and Tina Burrows. Toshiba internal report, year 2002. Internal report, Speech Technology Group - Toshiba Cambridge Laboratory, Cambridge, UK, 2002.

K. Knill, E. Morais, P. Jackson, and Tina Burrows. Toshiba internal report, year 2003. Internal report, Speech Technology Group - Toshiba Cambridge Laboratory, Cambridge, UK, 2003.

J. Kominek and A. W. Black. The cmu arctic databases. In *5th Speech Synthesis Workshop*, pages 223–224, Pittsburgh, PA, USA, 2004.

W. Leben. The tones in English intonation. *Linguistic Analysis*, 2:69–107, 1976.

C. J. Leggetter and P. Woodland. Speaker adaptation using maximum likelihood linear regression. *Computer Speech and Language*, 9(2):171–185, 1995.

M. Liberman and A. Prince. On stress and linguistic rhythm. *Linguistic Inquiry*, 8(2):249–336, 1977.

E. López. Estudio de técnicas de procesado lingüístico y acústico para sistemas de conversión texto voz en español basados en concatenación de unidades. Tesis doctoral, E.T.S.I., de Telecomunicaciones, Universidad Politécnica de Madrid, Madrid, España, 1993.

D. E. Mancebo. Modelado estadístico de entonación con funciones de Bézier: Aplicaciones a la conversión texto-voz en español. Tesis doctoral, Universidad de Valladolid, Valladolid, España, 2002.

C. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

R. Meir and G. Rätsch. *Lecture Notes In Machine Intelligence*, chapter An Introduction to Boosting and Leaveraging. Advanced Lectures on Machine Learning. Springer-Verlag, New York, USA, 2003.

P. Mertens, F. Beaugendre, and C. R. d'Alexandro. *Progress In Speech Synthesis*, chapter Comparing Approaches to Pitch Contour Sylization for Speech Synthesis, pages 347–363. J. V. Santen, R. W. Sproat, J. P. Olive and J. Hirschberg (Eds.). Springer-Verlag, 1996.

M. Mitchell. *An Introduction to Genetic Algorithms*. The MIT Press, Cambridge, Massachusetts, USA, 1998.

T. Mizutani and T. Kagoshima. Concatenative speech synthesis based on the plural unit selection and fusion method. *IEICE Transaction on information and systems*, E88-D(11):2565–2572, 2005.

R. C. Monteiro, F. Barbosa, R. Resende, L. Couto, and J. Moraes. Um conjunto de 1000 frases foneticamente balanceadas para o português brasileiro obtido utilizando a abordagem de algoritmos genéticos. In *Simpósio Brasileiro de Telecomunicações*, Campinas, SP, Brasil, 2005.

E. Morais and F. Violaro. Análise exploratória de dados linguísticos para uma modelagem linear robusta da duração segmental da fala. In *TIL - Workshop em Tecnologia da Informação e da Linguagem*, São Leopoldo, RS, Brasil, Julho 2005a.

E. Morais and F. Violaro. Exploratory analysis of linguistic data based on genetic algorithm for robust modeling of the speech segmental duration. In *Interspeech*, Lisboa, Portugal, Setembro 2005b.

E. Morais and F. Violaro. Tutorial extendido: Data-driven text-to-speech synthesis. In *XXII Simpósio Brasileiro de Telecomunicações (SBrT'05)*, pages 1256–1271, Campinas, SP, Brasil, September 2005c.

E. Morais, P. Taylor, and F. Violaro. Concatenative text-to-speech synthesis based on prototype waveform interpolation (a time frequency approach). In *6th ICSLP*, Beijin, China, October 2000.

E. Morais, F. Violaro, and F. Meireles. Exploratory analysis of linguistic data based on genetic algorithm and its application to robust modeling of speech segmental duration. In *Simpósio Brasileiro de Telecomunicações*, Campinas, SP, Brasil, 2005.

E. Moulines and F. Charpentier. Pitch synchronous waveform processin techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:435–467, 1990.

B. Möbius. Rare events and closed domains: Two delicate concepts in speech synthesis. In *4th Tutorial and Research Workshop on Speech Synthesis*, pages 41–46, Blair Atholl, UK, 2001.

B. Möbius. Corpus-based speech synthesis: Methods and challenges. (report aims 6(4)), IMS, University of Stuttgart, Stuttargt, Germany, 2000.

B. Möbius and J. V. Santen. Modeling segmental duration in german text-to-speech synthesis. In *ICSLP*, pages 2395–2398, Philadelphia, USA, 1996.

B. Möebius. *Progress In Speech Synthesis*, chapter Synthesizing German Intonation Contours, pages 401–415. J. V. Santen and R. W. Sproat and J. P. Olive and J. Hirschberg (Eds.). Springer-Verlag, 1996.

M. Ostendorf and I. Bulyko. The impact of speech recognition on speech synthesis. In *Workshop on Speech Synthesis*, Santa Monica, CA, USA, September 2002.

F. Pacheco and R. Seara. Prosodic speech modification using relp. In *International Telecommunication Symposium - ITS*, pages 1–6, Natal, RN, 2002.

J. B. Pierrehumbert. The phonology and phonetics of English intonation. Phd thesis, MIT, MA, USA, 1980.

T. Quatieri. *Discrete-Time Speech Signal Processing - Principles and Practice*. Prentice Hall PTR, Upper Saddle River, New Jersey, USA, 2002.

S. Quazza, L. Donetti, L. Moisa, and P. Salza. Actor: A multilingual unit-selection speech synthesis system. In *4th ISCA Tutorial and Research Workship on Speech Synthesis*, 2001.

J. Ross Quinlan. *C4.5 Programs for Machine Leraning*. Morgan Kaufmann, California, USA, 1993.

L. Rabiner. A tutorial on hidden markov models and selected applications on speech recognition. *Transactions on Speech and Signal Processing*, 77(2), February 1989.

K. Ross. Modeling of intonation for speech synthesis. Phd thesis, College of Engineering, Boston University, Boston, USA, 1994.

K. Ross and M. Ostendorf. Predicting abstract prosodic labels for speech synthesis. *Computer Speech and Language*, 10:155–185, 1996.

J. V. Santen. Assignment of segmental duration in text-to-speech synthesis. *Computer Speech and Language*, 8, 1994.

J. V. Santen. Quantitative modeling of pitch accent alignment. In *Speech Prosody*, Aix-en-Provence, Laboratoire Parole et Language, 2002.

J. V. Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg. *Progress in Speech Synthesis*. Springer-Verlag, New York, USA, 1997.

H. Schmid. Probabilistic part-of-speech tagging using decision trees. In *International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

A. Schweitzer, N. Braunschweiler, E. Morais, B. Möebius, and G. Dogil. *SmartKom - Foundations of Multimodal Dialogue Systems*, chapter Multimodal Speech Synthesis. In Wolfgang Wahlster (Ed.). Springer-Verlag, Germany, 2004.

K. Shichiri, A. Sawabe, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Eigenvoices for HMM-based speech synthesis. In *Eurospeech*, 2002.

Y. Shoham. High-quality speech coding at 2.4 to 4.0 kbps based on time-frequency interpolation. In *ICASSP'93*, pages 167–179, Minneapolis, MN, USA, April 1993.

K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg. Tobi: A standard for labeling English prosody. In *ICSLP*, pages 867–870, Banff, Canada, 1992.

R. Sproat. *Multilingual Text-to-Speech Synthesis, The Bell Labs Approach*. Kluwer Academic Publishers, Dordrecht, 1998.

Y. Stylianou. Applying the harmonic plus noise model in concatenative speech synthesis. *Transactions on Speech and Audio Processing*, 9(1), January 2001.

Y. Stylianou. Harmonic plus noise models for speech combined with statistical methods for speech and speaker modification. Phd thesis, Ecole Nationale Supérieure des Télécommunications, Paris, France, January 1996.

P. Taylor. Analysis and synthesis of intonation using the tilt model. *Journal of Acoustical Society of America*, 107:1697–1714, 2000.

P. Taylor. The rise/fall/conection model of intonation. *Speech Communication*, 15:169–186, 1995.

P. Taylor. A phonetic model of English intonation. Phd thesis, University of Edinburgh, Edinburgh, UK, 1992.

K. Tokuda, T. Masuko, T. Kobayashi, N. Miyazaki, and T. Kitamura. Hidden markov models based on multi-space probability distribution for pitch pattern modeling. In *ICASSP*, 1999.

K. Tokuda, H. Zen, and A. Black. An HMM-based speech synthesis system applied to English. In *Workshop on Speech Synthesis*, Santa Mônica, CA, USA, September 2002.

V. N. Tuan and C. d'Alessandro. Robust glottal closure detection using the wavelet transform. In *Eurospeech'99*, pages 2805–2808, Budapest, Hungary, September 1999.

G. Valentini and F. Masulli. *Ensembles of Learning Machines*. Neural Nets WIRN Vietri-02, Series Lecture Notes in Computer Sciences. Springer-Verlag, Heidelberg, Germany, 2002.

E. V. Vlist. *XML Schema: The W3C's Object-Oriented Descriptions for XML*. O'Reilly, Sebastopol, CA, USA, 2002.

C. W. Wightman and M. Ostendorf. Automatic labeling of prosodic patterns. *IEEE Transaction on Speech and Audio Processing*, pages 469–481, October 1994.

J. Wouters. Analysis and synthesis of degree of articulation. Phd thesis, Oregon Graduate Institute, Oregon, USA, 2001.

J. Wouters and M. W. Macon. A perceptual evaluation of distance measures for concatenative speech synthesis. In *ICSLP*, Sydnei, Australia, October 1998.

H. Ye and S. Young. Perceptually wighted linear transformation for voice conversion. In *Eurospeech*, Geneva, Switzerland, 2003.

H. Ye and S. Young. High quality voice morphing. In *ICASSP*, Montreal, Quebec, Canada, May 2004.

B. Yegnanarayana, C. d'Alexandro, and V. Darsinos. An interactive algorithm for decomposition of speech signals into periodic and aperiodic components. *IEEE Transaction on Speech and Audio Processing*, 6:1–11, 1998.

T. Yoshimura, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. In *Eurospeech*, pages 2347–2350, ISCA, 1999.

W. Zhu and W. Zhang. Corpus building for data-driven TTS systems. In *Workshop on Speech Synthesis*, Santa Monica, CA, USA, 2002.

S. Öhman. Word and sentence intonation: A quantitative model. Technical report, KTH, Sweden, 1967.

Apêndice A

Histogramas das Durações dos Fones para Avaliação do Algoritmo LDM-GA

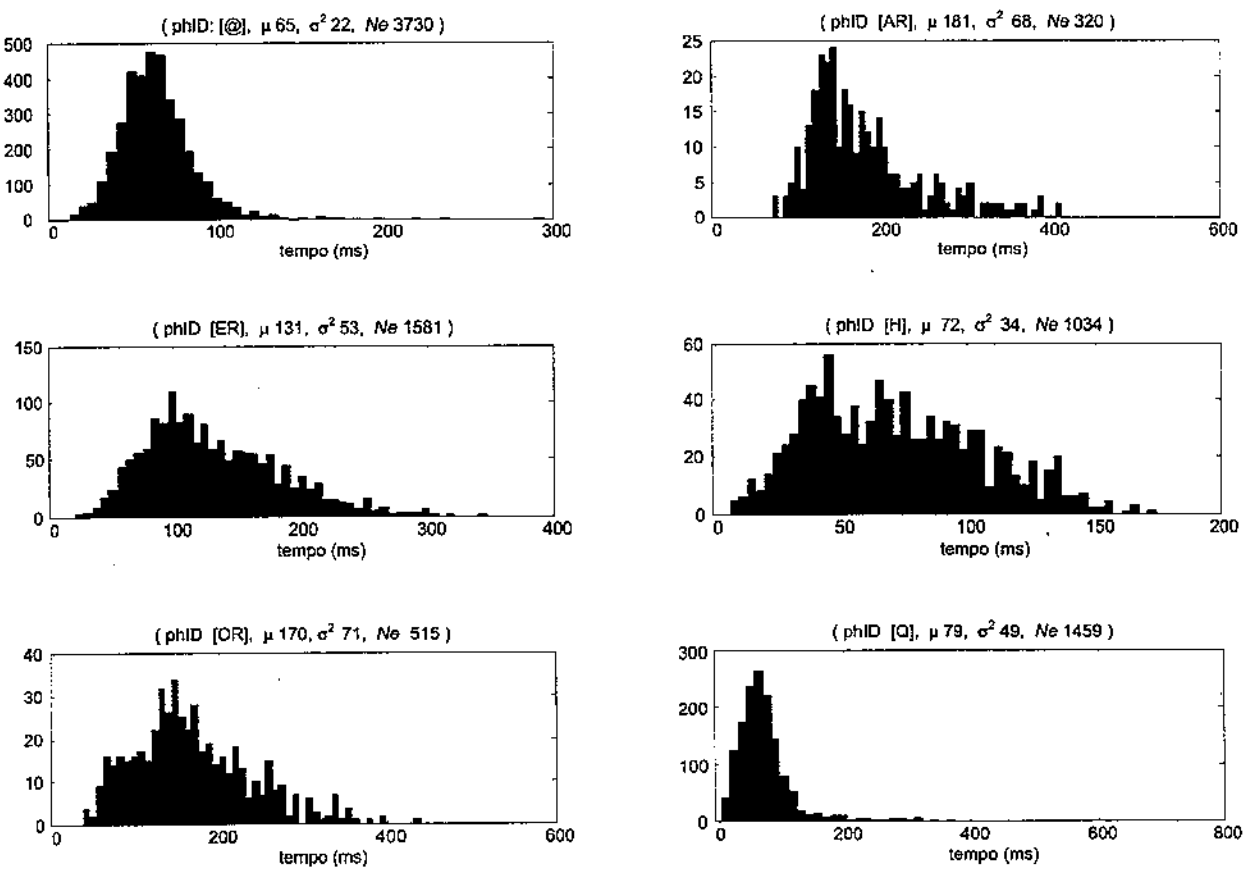


Figura. A.1: Histogramas dos fones [@], [AR], [ER], [H], [OR], [Q].

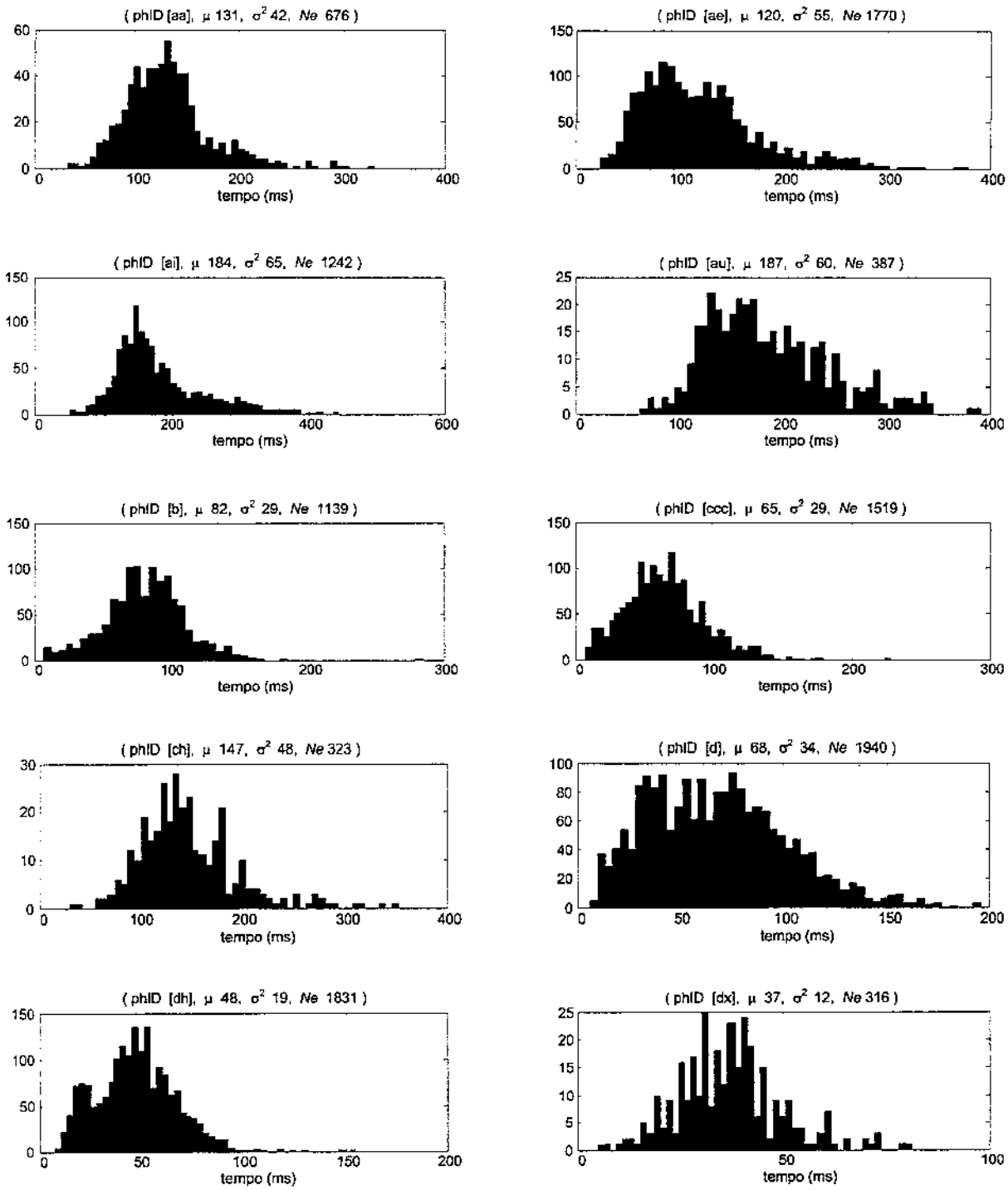


Figura. A.2: Histogramas dos fones [aa], [ae], [ai], [au], [b], [ccc], [ch], [d], [dh], [dx].

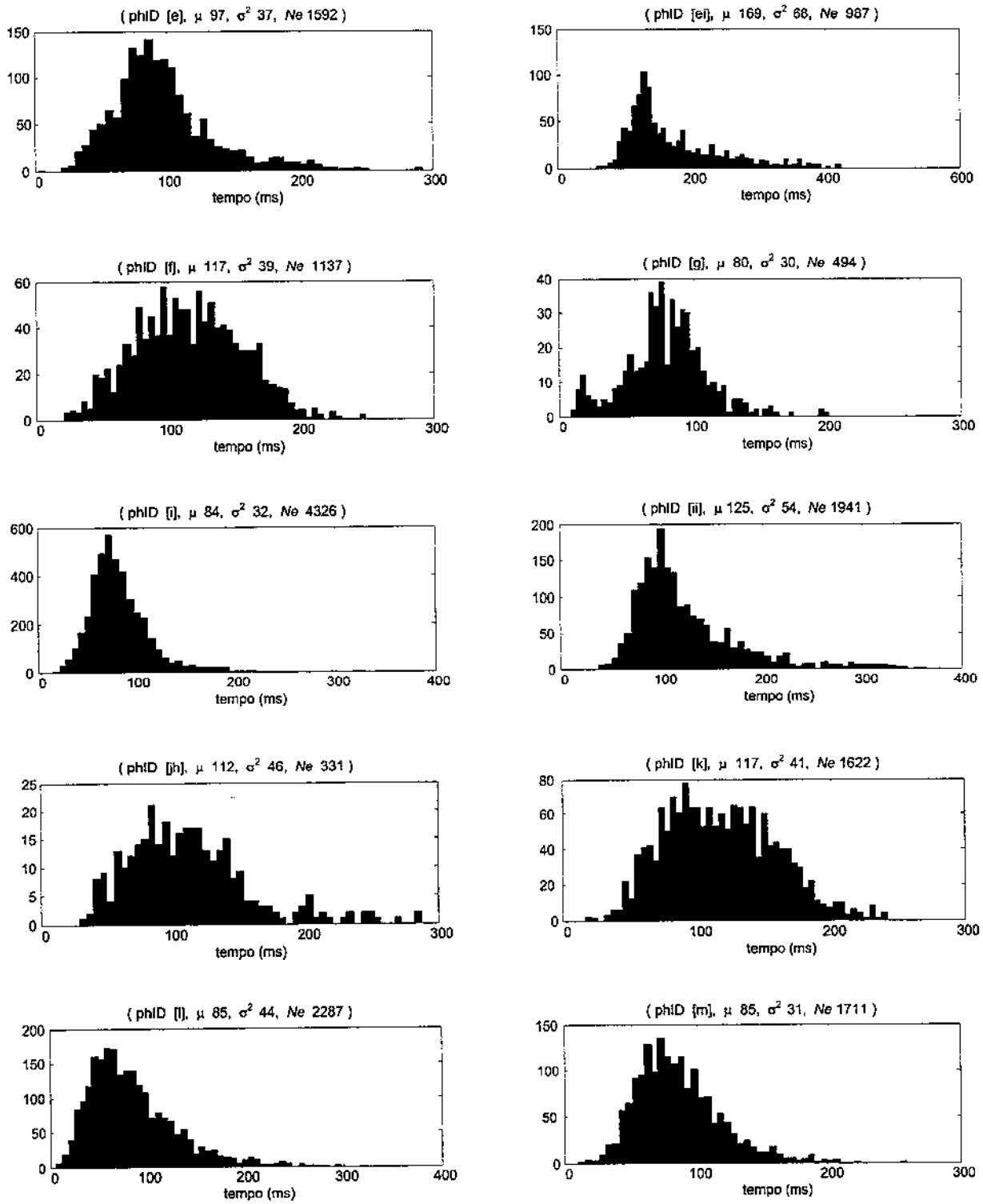


Figura. A.3: Histogramas dos fones [e], [ei], [f], [g], [i],[ii], [jh], [k], [l], [m].

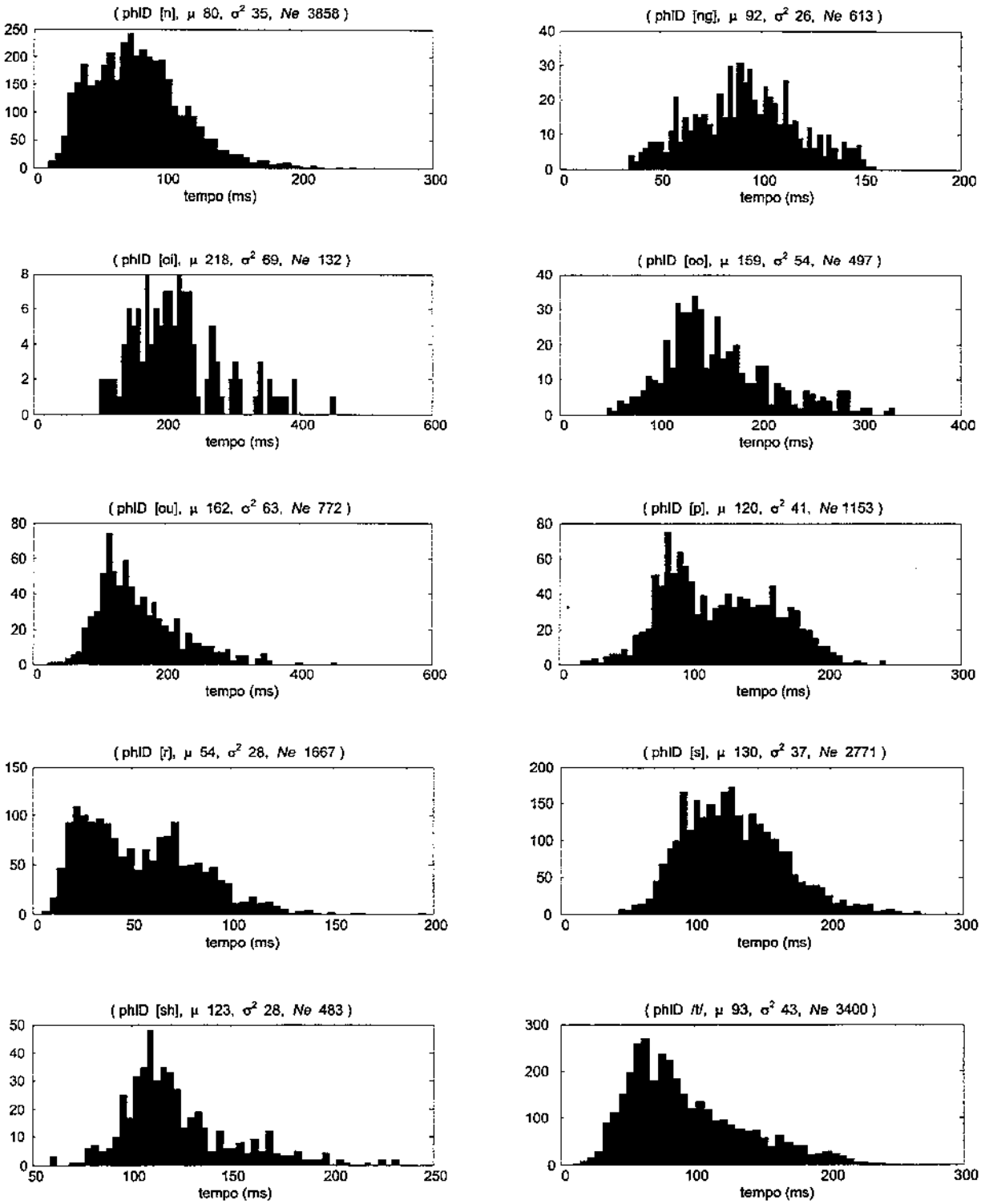


Figura. A.4: Histogramas dos fones [n], [ng], [oi], [oo], [ou], [p], [r], [s], [sh], [t].

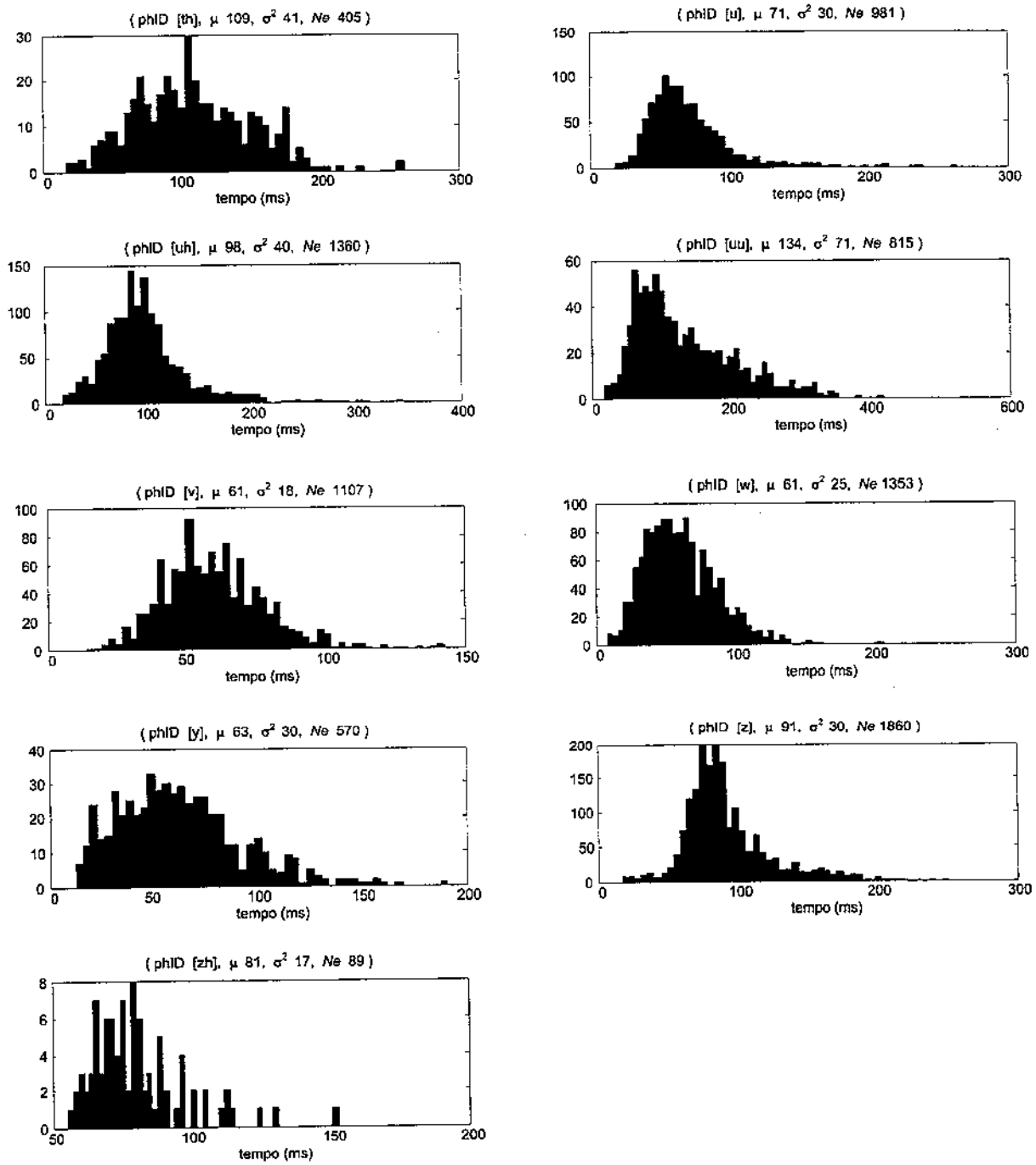


Figura. A.5: Histogramas dos fones [th], [u], [uh], [uu], [v], [w], [y], [z], [zh].

Apêndice B

Normalização dos Protótipos

O processo de normalização do protótipo $X^P(k)$ é realizado segundo duas operações denominadas *Zero Padding* e *Truncagem* em Freqüência. A operação de *Zero Padding* aumenta o número de componentes de $X^P(k)$ inserindo zeros em $X^P(k)$ em torno da freqüência normalizada π (sendo que $X^P(k)$ estende-se da freqüência normalizada 0 até $2 \cdot \pi$). O procedimento de *Truncagem* reduz o número de componentes de freqüência de $X^P(k)$ eliminando/modificando componentes de $X^P(k)$ localizados em torno da frequência normalizada π. Estes procedimentos de *Zero Padding* ou *Truncagem* em freqüência podem ser interpretados no domínio do tempo como um aproximação às operações, respectivamente, de interpolação ou de dizimação da seqüência temporal $x^P(n)$ (representação temporal do protótipo $X^P(k)$).

A normalização de um protótipo $X^P(k) = \{X^P(0), X^P(1), \cdot, X^P(K-2), X^P(K-1)$ de extensão K em um protótipo $\widehat{X^P}(k)$ de extensão L é realizada de acordo com a seguintes operações:

***Zero Padding* em Freqüência ($L > K$): K ímpar**

$$\widehat{X^P}(k) = \{X^P(0), \cdots, X^P(\frac{K-1}{2}), ZOdd, X^P(\frac{K+1}{2}), \cdots, X^P(K-1)\} \tag{B.1}$$

sendo ZOdd um vetor de dimensão $L-K$, e com todos seus elementos iguais a zeros.

***Zero Padding* em Freqüência ($L > K$): K par**

$$\widehat{X^P}(k) = \{X^P(0), \cdots, \frac{X^P(\frac{K}{2})}{2}, ZEven, \frac{X^P(\frac{K}{2})}{2}, \cdots, X^P(K-1)\} \tag{B.2}$$

sendo ZEven um vetor de dimensão $L-(K+1)$, e com todos seus elementos iguais a zeros.

Truncagem em Freqüência ($L < K$): K e L ímpares

$$\begin{aligned}\widehat{X^P}(k) &= \{X^P(0),\cdots,X^P(\frac{L-1}{2}),X^P(K-\frac{L-1}{2}-1),\\ &\quad \cdots,X^P(K-1)\}\end{aligned} \tag{B.3}$$

Truncagem em Freqüência ($L < K$): K ímpar e L par

$$\begin{aligned}\widehat{X^P}(k) &= \{X^P(0),\cdots,X^P(\frac{L}{2}-1),\frac{X^P(\frac{L}{2})+X^P(K-\frac{L}{2})}{2},\\ &\quad X^P(K-(\frac{L}{2}-1)),\cdots,X^P(K-1)\}\end{aligned} \tag{B.4}$$

Truncagem em Freqüência ($L < K$): K par e L ímpar

$$\widehat{X^P}(k) = \{X^P(0),\cdots,X^P(\frac{L-1}{2}),X^P(K-\frac{L-1}{2}),\cdots,X^P(K-1)\} \tag{B.5}$$

Truncagem em Freqüência ($L < K$): K e L pares

$$\begin{aligned}\widehat{X^P}(k) &= \{X^P(0),\cdots,X^P(\frac{L}{2}-1),\frac{X^P(\frac{L}{2})+X^P(K-\frac{L}{2})}{2},\\ &\quad X^P(K-(\frac{L}{2}-1)),\cdots,X^P(K-1)\}\end{aligned} \tag{B.6}$$